HHS Synthetic Health Data Challenge—Challenge Category I: Enhancements to Synthea

On Improving Realism of Disease Modules in Synthea: Social Determinant-Based Enhancements to Conditional Transition Logic

Organization Name: LMI	**Contact**: Brant Horio; bhorio@lmi.org; 571-633-7838
Video: https://youtu.be/udG26HYjTK0	**Source Code:** https://github.com/LMISyntheticHealthTeam/synthea.git
Authors: Brant Horio, Gregory Pekar, Simon Whittle, Maureen Merkl, and Linna Qiao	

Abstract

The opioid crisis in the U.S. is an ongoing battle, with the COVID-19 pandemic exacerbating the problem. May 2019 to May 2020 had the highest number of overdose deaths ever recorded in a single year. To understand, and respond to, this crisis, many refer to social determinants of health (SDOH), which emphasize patient-centered outcomes research (PCOR) focused on the individual. However, clinical data necessary for PCOR is generally inaccessible to the research community, limiting its usability. Synthea™, a synthetic patient generator, bridges this clinical data gap effectively until real data become available. In response to the HHS Synthetic Health Data Challenge, Team LMI executed an innovative approach to Challenge Category I: Enhancements to Synthea. Our software development effort enhanced the Prescribing Opioids for Chronic Pain and Treatment of Opioid Use Disorder (OUD) module with greater realism. We included occupation as an SDOH to represent the diversity of chronic pain patients better—enabling more representative population simulation through SDOH-based branching logic. Our effort addresses occupation as a key opioid abuse–related SDOH domain for the geographic area of Bangor, Maine, selected for its higher than average national rates of opioid overdoses and accessible data for validation efforts. For the occupation SDOH, we specify open-source secondary data sets for creating new SDOH-related attributes in the appropriate agent classes. Open-source secondary data attributes support compatibility with Synthea functionality when developing or updating disease modules.

Currently, Synthea users can include some SDOH-related logic in the disease modules but the limited functionality requires complex chains of conditional logic that are time-consuming to arrange, difficult to update, and impossible to reuse. By drawing from hyperlocal (census tract) demographic data, we enable users to account for more nuanced, reusable, and succinct SDOHs conveniently through the Synthea Module Builder.

Our efforts produced greater realism in the OUD module, and Synthea, through better characterization of key SDOH-related features at the level of individual person agents. We expect the creation of new software functionality to enable Synthea users to access these SDOH features easily when developing conditional transition logic in disease modules. These outcomes will improve the realism of simulated populations and supply greater flexibility for research questions.

1 Introduction

The national opioid crisis is an ongoing battle, with more than 81,000 Americans dying from drug overdoses between May 2019 and May 2020, the highest number of overdose deaths ever recorded in a single year. Contributing to the crisis, the street drug supply has become more contaminated and dangerous—the Centers for Disease Control and Prevention (CDC) cites fentanyl, a common additive, as the primary driver of the increases in overdose deaths. Research supports medication-assisted treatment (MAT) as the gold standard for opioid survival and recovery. However, the resources for this standard are not available readily (particularly in rural communities) for patients with opioid abuse disorders along with the concurrent psychosocial support that could end repeated use of emergency departments (EDs) to handle overdoses. SDOH significantly influence opioid abuse and treatment outcomes and enhance understanding of increases in abuse. In the context of the COVID-19 pandemic, SDOH, such as social isolation, loss of employment, and homelessness, have influenced opioid abuse outcomes, stressing community EDs beyond their limits.

The opioid crisis, and its correlation with the situations of victims related to health conditions and SDOH, emphasizes the importance of PCOR focused on the individual and the study of medical prevention and treatment option effectiveness. Researchers and developers often cannot access or acquire the patient data for PCOR due to cost, patient privacy concerns, or other legal restrictions. To address this clinical data gap, MITRE developed Synthea, an open-source software application simulating the lifespans and medical history of patients. Though its use has increased, Synthea has key limitations affecting the realism of simulated outcomes and reducing trust in the tool by policymakers. Greater accuracy of representation for real-world populations (e.g., key SDOH related to opioids abuse and resources in the community) and accurate details on their clinical pathways (e.g., protocols and treatments in the ED for an opioid overdose encounter) will increase Synthea's viability as tool until real clinical data become more readily available.

2 Research Question and Key Technical Idea

Opioids are prescribed to treat a number of ailments, including chronic back pain, surgery, and broken bones. Occupations can cause chronic back pain, with some, such as fishing and forestry, having a higher risk (Yang et al. 2016). These occupations are common in more rural communities. Our research investigates occupation as an SDOH of interest related to OUD. By studying the fishing and forestry industries and their links to higher probabilities of chronic pain, we sought validation of Synthea's representation of the relationship between an increased rate of chronic pain and a proportional rise in opioid prescriptions. We examined the *Prescribing Opioids for Chronic Pain and Treatment of OUD* module and the transition probability for adults to *Condition_Chronic_Low_Back_Pain*. An increased rate of prescriptions can lead to higher levels of opioid misuse and addiction for those in the forestry and fishing industries. We hypothesize that representing chronic pain more precisely leads to increased accuracy in prescribed directed use of opioids, and, finally, better calculations in Synthea for representing population health outcomes for opioid addiction and misuse.

The key technical idea does not account for all SDOH contributing factors that lead to OUD but demonstrates

- a software workflow that mines open-source secondary data sources for SDOH,
- modifications to the Synthea codebase to operationalize the SDOH, and
- increased realism in Synthea outputs with respect to OUD.

We scoped our research to the geographic area of rural communities surrounding Bangor due to the availability of population health data for validation (including participation in the National Syndromic Surveillance Program [NSSP]). The NSSP collects chief complaint data in near real time on people who seek care in EDs. It includes work-related injuries and illnesses, such as chemical exposures and tree-related injuries (of interest to our research) (CDC, 2021). In addition, Bangor is economically depressed with a significant opioid addiction problem above the national average.

We validated our results by comparing the performance of the existing software (legacy_Synthea) and our enhanced software (LMI_Synthea) against ground truth validation patterns.

3 Methods

Our research focused on enhancing Synthea for the use case of opioids with the *Prescribing Opioids for Chronic Pain and Treatment of OUD* disease module.

In legacy_Synthea, including social-determinant–related logic in disease modules is limited and requires complex chains of time-consuming to arrange, difficult to update, and impossible to reuse conditional logic. We seek to make such logic succinct and reusable by drawing from new sources and Synthea data to create new attributes for the *person* class. Figure 1 shows our approach, followed by a description of each step.

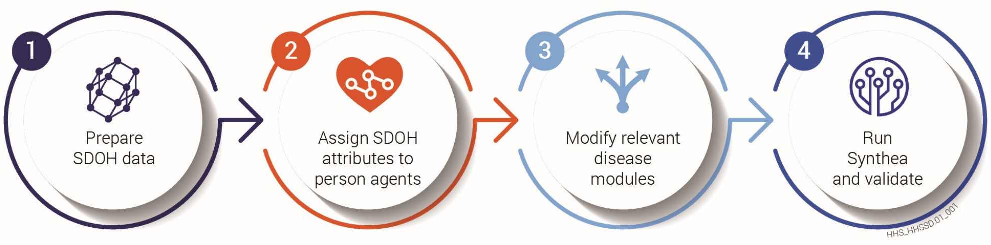
1
Prepare SDOH data
2
Assign SDOH attributes to person agents
3
Modify relevant disease modules
4
Run Synthea and validate
HHS_HHSSD.01_001

Figure 1. Proposed technical approach to move from data to codebase changes to validation tests.

Step 1—Prepare SDOH data: We use American Community Survey data from the United States Census Bureau to quantify the percentages of people in the fishing and forestry industries by census tract. The tracts.csv file includes fields such as the percentage of the population in the fishing and forestry industries, percentage of adults, percentage of unemployment, percentage of people uninsured, and other relevant demographic fields. All the fields are defined at the census tract level. This file augments the work in the demographics.csv file by providing more granular detail.

Step 2—Assign SDOH attributes to person agents: As part of the initialization step for a generator, we load the census tract data and then the block data. For tracts, loading reads the input.csv and creates a set of CensusTract objects. These objects store their tract ID code and some demographic information relevant to our study, specifically the percentage of people in the fishing and forestry industries. For blocks, loading reads the input.csv and creates a set of CensusBlock objects. These objects store their geoid code, the CensusTract object they belong to, and the geocoordinates of the census block's centroid point. During person object generation, these CensusBlock and CensusTract objects add granular demographic data to person agents by enabling the person to find the CensusBlock closest to the geolocation generated for it. Then, that CensusBlock object can access the CensusTract object containing demographic information, from which the person can randomly sample realistic information about itself to store in its attributes map. Once the information is stored in the person's attributes map, it can be accessed in Synthea modules with a new *Occupation* transition logic type that works much the same as existing Synthea *Age* or *Gender* logic types.

Step 3—Modify relevant disease modules: We modified the *Prescribing Opioids for Chronic Pain and Treatment of OUD* module to enable the use of our new SDOH features in the ConditionalTransition class. This modification inspects a person and, based on the attributes from Step 2, changes the transition probability into the *Condition_Chronic_Low_Back_Pain* state. Since this transition state leads to prescriptions of opioids, we expect Synthea to generate more opioid prescriptions for people in the fishing and forestry industries. Women in the fishing and forestry industries are 2.66 times more likely to have low back pain injuries and men are 0.86 times as likely to have low back pain injuries (Yang et al. 2016). We multiplied the base rate transition probability of 9% into *Condition_Chronic_Low_Back_Pain* by these two factors. Figure 2 shows our modifications.

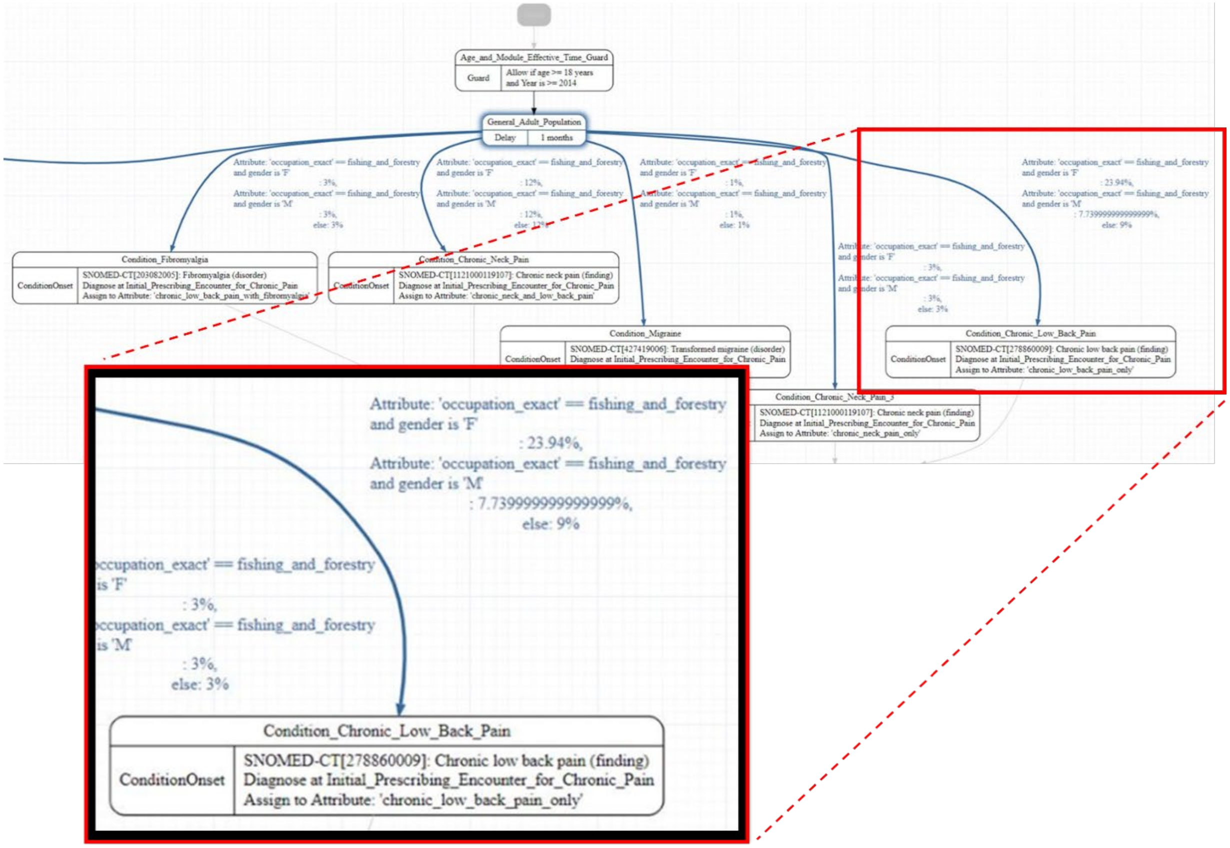
Age_and_Module_Effective_Time_Guard
Guard
Allow if age >= 18 years and Year is >= 2014
General_Adult_Population
Delay
1 months
Attribute: 'occupation_exact' == fishing_and_forestry and gender is 'F' : 3%,
Attribute: 'occupation_exact' == fishing_and_forestry and gender is 'M' : 3%,
else: 3%
Attribute: 'occupation_exact' == fishing_and_forestry and gender is 'F' : 12%,
Attribute: 'occupation_exact' == fishing_and_forestry and gender is 'M' : 12%,
else: 12%
Attribute: 'occupation_exact' == fishing_and_forestry and gender is 'F' : 1%,
Attribute: 'occupation_exact' == fishing_and_forestry and gender is 'M' : 1%,
else: 1%
Attribute: 'occupation_exact' == fishing_and_forestry and gender is 'F' : 23.94%,
Attribute: 'occupation_exact' == fishing_and_forestry and gender is 'M' : 7.739999999999999%,
else: 9%
Condition_Fibromyalgia
ConditionOnset
SNOMED-CT[203082005]: Fibromyalgia (disorder)
Diagnose at Initial_Prescribing_Encounter_for_Chronic_Pain
Assign to Attribute: 'chronic_low_back_pain_with_fibromyalgia'
Condition_Chronic_Neck_Pain
ConditionOnset
SNOMED-CT[1121000119107]: Chronic neck pain (finding)
Diagnose at Initial_Prescribing_Encounter_for_Chronic_Pain
Assign to Attribute: 'chronic_neck_and_low_back_pain'
Condition_Migraine
ConditionOnset
SNOMED-CT[427419006]: Transformed migraine (disorder)
Diagnose at Initial_Prescribing_Encounter_for_Chronic_Pain
Condition_Chronic_Low_Back_Pain
ConditionOnset
SNOMED-CT[278860009]: Chronic low back pain (finding)
Diagnose at Initial_Prescribing_Encounter_for_Chronic_Pain
Assign to Attribute: 'chronic_low_back_pain_only'
Condition_Chronic_Neck_Pain_3
SNOMED-CT[1121000119107]: Chronic neck pain (finding)
Diagnose at Initial_Prescribing_Encounter_for_Chronic_Pain
Assign to Attribute: 'chronic_neck_pain_only'

Figure 2. Transition logic changes in the Prescribing Opioids for Chronic Pain and Treatment of OUD module leveraging the new occupation SDOH attribute for fishing_and_forestry and the synthetic person's gender to influence the probability for chronic low back pain.

Step 4—Run Synthea and validate: To validate our modifications to Synthea, we ran 10 trials, each with different random seeds, generating 32,000 patients in Bangor. We used the same 10 seeds[1] for legacy_Synthea as well as LMI Synthea. We chose 32,000 patients to match the approximate population of Bangor, effectively running 10 trials on the entire City of Bangor. We used the Prescription Monitoring Program Annual Report 2020 (Maine HSS 2021) and data from the CDC (CDC 2020) as our ground truth data source for the number of opioid prescriptions in Maine. See the following section for more details.

4 Results and Validation

Our validation used the following command to generate patients.

```
run_synthea -p 32000 -s 10000 -cs 12345 Maine Bangor
```

To assist with replicability, we used the same 10 random seeds for legacy_Synthea as well as LMI_Synthea. We chose 32,000 as the number of patients as this is approximately the population of Bangor.

Since we modified the probability of chronic low back pain, simply examining whether the output reflected this increased probability would have been trivial. Instead, we focused on how increased chronic low back pain changed the number of opioid prescriptions for patients in the fishing and forestry industries. We picked out the prescriptions from the medications.csv file for the year 2019, as this is the year of our ground truth statistics for comparison, with 2019 being the latest year of data without any effects from COVID-19, which exacerbated the opioid crisis. We wanted controlled validation without the unprecedented complications associated with the COVID-19 pandemic. Figure 3 shows our results for opioid prescriptions per capita. Ground truth data for prescriptions in Maine comes from the Prescription Monitoring Program Annual Report 2020 (Maine HSS 2021). Ground truth data for prescriptions in Penobscot County come from the CDC (CDC 2020).

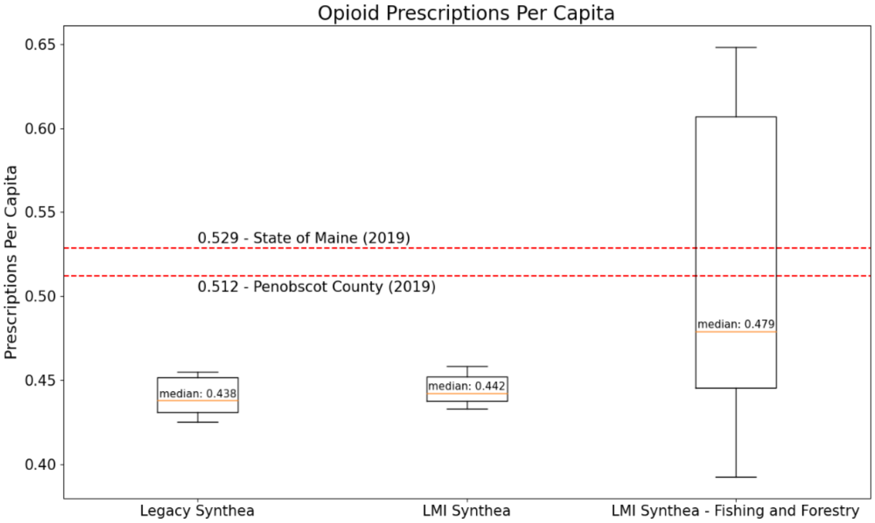

Figure 3. Synthea results for opioid prescriptions per capita, comparing legacy_Synthea to LMI_Synthea, with respect to the isolated subpopulation of fishing and forestry occupations from LMI_Synthea.

Our changes to the odds of chronic low back pain for loggers and fishers move the prescriptions per capita for LMI_Synthea closer to the ground truth, but only slightly, due to the low proportion of loggers and fishers to the entire workforce of Bangor. However, looking at people in the fishing and forestry industries, the box plot includes the ground truth in the inter-quartile range. We changed only one subset of the workforce. In future iterations, LMI can include construction workers, veterans, home contractors, and others predisposed to chronic pain, thus moving the opioid prescriptions per capita closer to the ground truth.

[1] The random seeds: 10000, 13370, 22222, 23123, 33555, 39093, 45000, 51327, 65888, and 74982.

5 Conclusion

While we integrated only a single new SDOH factor relevant to OUD into Synthea for this research effort, it demonstrated how additional SDOH features, based on census tracts and blocks, might be incorporated into future Synthea-based research. Researchers can use SDOH attributes in the person agents as levers to pose what-if questions. For example, experiment scenarios might assign representative subpopulations as unemployed artificially (to reflect COVID-19 economic effects), enabling an analysis of population health futures accounting for unemployment's SDOH influence on opioid abuse in the OUD module.

5.1 Impact and Innovation

Our solution was innovative with respect to the technical approach of assigning realistic census tracts and blocks to synthetic persons generated by Synthea. Our modifications to the Synthea code base and inclusion of new reference data files streamline future customization by researchers to assign SDOH-based attributes more easily to generated persons. These modifications enable greater realism of disease modules and more accuracy in the representation of real-world population health. The simple logic change of modifying a transition probability to the *Condition_Chronic_Low_Back_Pain* state can be followed for other use cases. For example, it can simulate a large-scale intervention, such as doctors prescribing twice as many non-opioids. By doubling the transition probability for directed use for non-opioids and generating patients using Synthea, a researcher can examine the population-level effects of this policy. This innovation can be a powerful tool for organizations such as the Centers for Medicare & Medicaid Services Innovation Center.

Our inclusion of occupation as a consideration for predicting long-term chronic pain and subsequent opioid abuse adds to the body of knowledge for safety in the workplace. Employers have not been included adequately in state and local efforts to combat the opioid crisis (Shaw, Roelofs, and Punnett 2020). Evidence regarding the predisposition to chronic pain in some occupations highlights the need for employer involvement.

5.2 Functionality and Implementation

Synthea treats most person agents the same with respect to transition probabilities (albeit with instances of socio-demographics, like income, forking state changes in the disease modules). Our effort added Synthea software infrastructure for individual-specific SDOH factors based on geolocation for more nuanced branching in the disease module state machines.

5.3 Validation

Close examination of a narrow geographic region exposed weaknesses in Synthea's ability to represent small populations. Our census tract demographic file will enable researchers more familiar with their communities to replicate them more accurately. These small, rural communities, especially those like Bangor, show the flaws of using national statistics most visibly because of their idiosyncrasies.

5.4 Above and Beyond the Challenge Requirements

We have been motivated to further democratize access to Synthea. We augmented the code base and included a new data file for the Synthea reference folder to increase ease of access to SDOH details in the Module Builder for practitioners and researchers who are not software engineers.

6 Future Work

Our improvements to the fishing and forestry industries move Synthea output closer to the ground truth but fail to consider other portions of the workforce. In future iterations, LMI can include construction workers and others predisposed to chronic pain. Support for including these occupations as a variable in Synthea models is evident in a report by the National Institute for Occupational Safety & Health-funded Massachusetts Department of Public Health state surveillance program. Massachusetts workers died of opioid overdoses at different rates depending on their jobs, with the highest rates in construction and extraction, including quarrying, mining, and oil and gas removal. These jobs were followed by agriculture, forestry, and fishing industries (Health 2018). The Department of Veteran Affairs reports that the number one disability claimed by veterans is for musculoskeletal conditions (Haskell 2012). We recommend including veterans and construction workers in future modeling.

7 References

CDC. June 16, 2021. "New Jersey Leverages Syndromic Surveillance to Combine Multiple Data Sources for Detecting Occupational Injuries and Exposures." Retrieved from National Surveillance Syndromic Program: https://www.cdc.gov/nssp/success-stories/NJ-Occupational-Injuries.html.

CDC. December 7, 2020. "U.S. County Opioid Dispensing Rates," 2019. Retrieved from CDC, National Center for Injury Prevention and Control: https://www.cdc.gov/drugoverdose/rxrate-maps/county2019.html.

Maine Department of Health & Human Services, Office of Behavioral Health. January 2021. *Prescription Monitoring Program Annual Report 2020*, page 8.

Haskell, Sally G. August 2012. "Post-Deployment Pain: Musculoskeletal Conditions in Male and Female OEF/OIF Veterans." *FORUM*. https://www.hsrd.research.va.gov/publications/forum/aug12/aug12-4.cfm.

Massachusetts Department of Public Health. 2018. "Opioid-Related Overdose Deaths in Massachusetts by Industry and Occupation 2011-2015." State Report.

Shaw, William S., Cora Roelofs, and Laura Punnett. 2020. "Work Environment Factors and Prevention of Opioid-Related Deaths." *Am J Public Health*, 1235-1241.

Yang, H., Haldeman, S., Lu, M.-L., & Baker, D. 2016. "Low Back Pain and Related Workplace Psychosocial Risk Factors: A Study Using Data From the 2010 National Health Interview Survey." *Journal of Manipulative and Physiological Therapeutics*, 459-472.