Validation of Synthea

A Synthetic Patient Population Simulator

A Major Qualifying Project for the degree of:

Bachelor of Science in Electrical and Computer Engineering

Cara Seely

Worcester Polytechnic Institute

Advised by:

Dr. Shamsnaz V. Bhada

Worcester Polytechnic Institute

Jason Walonoski & Dylan Hall

MITRE

2017

Table of Contents

Table of Figures

Table of Tables

Introduction

Synthetic data is used in a variety of industries in order to develop software and perform testing. Accurate synthetic data allows for all of the functionality of real data, but without privacy concerns. Synthetic health records can be used for research and development in the healthcare industry, without the risk of re-identification of real personal health records. Synthea is a synthetic patient generator that creates synthetic patient health records which can be used for the creation of new and improved electronic health record systems.

This Major Qualifying Project (MQP) is concerned with validating and verifying that Synthea is creating realistic synthetic patient data at the patient level, disease module level, and population level. Through interviewing medical doctors, it was found that the individual synthetic health records were somewhat realistic representation of actual health records. Additionally, it was found that the pregnancy disease module in Synthea was missing information contained in published standards of care. The pregnancy module was updated to be more aligned with the standards of care. It was also found that Synthea is producing realistic and consistent disease prevalence levels.

Background

The United States is a global leader in science and technology, home of forty of the world’s top universities and accounting for about forty percent of the world’s total spending on research and development (“U.S. Still Leads the World in Science and

Technology; Nation Benefits From Foreign Scientists, Engineers", 2008). Technology is advancing by the day, solving problems, bettering lives, and connecting people. With all of the advancements being made, it may be hard to believe that the United States healthcare industry still lags behind in technological advancements. In fact, many healthcare professionals do not advocate for the use of software in hospitals. Robert M. Wachter's 2015 New York Time's article *Why Health Care Tech Is Still So Bad* describes a hospital job posting in the Phoenix, Arizona area, boasting that they do not use an electronic medical record system (EMR) and using that as a "selling point".

Many sources including Athena Health, The New York Times, George Palma M.D., and Afia Health agree that while technology in the healthcare industry does promise to be transformative in the quality of care given to patients, it proves to be difficult to implement. EMRs can help to reduce human error in drug prescriptions, assist in real-time decision making, and make health records available to patients and outside hospitals in necessary situations including emergencies. However, with so many EMR solutions available, it is difficult to standardize the way patients are cared for, how payments and insurance are processed, how health records are organized, and how that data is shared. As of 2014, 76% of hospitals in the United States exchanged health information electronically (*Swain et al., 2015)*, however the format of that data differs widely among EMRs ("The Challenges of Sharing Data Between Separate EHRs", 2017). Non-uniform data can be difficult to parse in real-time situations that require immediate attention. Health interoperability must contain a standardized way of communicating information and being able to use that information effectively and timeefficiently. Software developers and researchers need access to many health records in order to test and implement new health interoperability solutions. However, obtaining health records can be expensive and poses privacy risks, even if the records are de-identified or anonymized. The use of synthetic data has been successful in the development of software in the financial industry. This synthetic data can model payments, withdrawals, and deposits, making a complete customer profile. A similar approach can be taken to model a synthetic patient health record in the healthcare industry.

Synthetic health records can be used for research and development in the healthcare field, without the risk of re-identification of real personal health records. Synthea is a Synthetic Patient Population Simulator which aims to provide the realistic health data needed to improve healthcare technology solutions. It is open-source, free of cost, and has no restrictions. The goal of Synthea is to produce realistic, yet synthetic, patient data. With any synthetically generated data, the use of that data is only as reliable as the data itself. Synthea must be both verified and validated on the patient, module, and population levels in order to be accredited for specific use cases such as health data interoperability solutions.

Health Data Interoperability

Health data interoperability is the ability of two or more systems to exchange and make use of health information (“Obstacles to Interoperability within Closed Systems”, 2017). Electronic Health Records (EHRs) contain information of a patient’s medical history including demographics, medications, vital signs, immunizations, laboratory results and more (“Electronic Health Records”, 2012). EHRs theoretically allow for the sharing of health data between hospitals, doctors offices, surgery centers etc. and can have digital processes which allow for decision support in important situations as well as medical coding and billing. There are many benefits to EHRs including improved patient care, diagnostics, and patient participation. The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 is meant to promote the meaningful use of health information technology in order to better healthcare experiences for all (“HITECH Act Enforcement Interim Final Rule”, 2017).

While sharing EHRs aims to be beneficial for patients and healthcare providers alike, the data shared can be useless if one EHR cannot effectively communicate with another EHR. When EHRs were first coming into existence, the companies building them did not take into account the importance of communication across EHRs, rather they focused on building the system primarily for the purpose of medical billing. The data within EHRs must move towards a standardized format for communication between systems, and recorded in such a way that allows for complete and correct medical billing. For patients, it is also important that the data be thoroughly clinically detailed and be readily available to them through an online patient portal (Bresnick, 2015).

Data exchange across EHRs encompasses interactions with users, communication between systems, and how the data communicated is processed (“What Is EHR Interoperability and Why Is It Important?” 2013). Afia Health Inc. describes three shortcomings of data exchange between EHR systems. The first shortcoming they describe is the difficulty of finding shared patients between systems, because EHRs identify patients in different ways and store them in different formats. The second shortcoming is the translation of data between EHRs and their coding standards; there are multiple medical billing and coding practices that exist and not all systems use the same set of codes. The final shortcoming Afia discusses is data storage; as more and more healthcare practices move to EHRs, it must be guaranteed that the system has the capacity to hold all required data and process necessary exchanges (“The Challenges of Sharing Data Between Separate EHRs”, 2017). George Palma, MD also discusses shortcomings including the difficulty of sharing health records between hospitals, and delays in documentation in his 2013 article “Electronic Health Records: The Good, the Bad and the Ugly”. All of these shortcoming point to the need for a more standardized way of recording, sharing, and storing EHRs. New solutions with improved data storage, communication abilities, patient portal access, and medical billing, must be tested with realistic and scalable health data.

Synthetic Data Generation

Simulation models aim to be accurate representations of a real world system. Simulation models can be valuable in the healthcare field for the creation of new health data interoperability solutions and for research. These models must be realistic at the patient and population levels, in order to properly design solutions that fit the needs of both the patients and the health care practices using them. Invalid models result in invalid conclusions, making the simulation model useless or possibly harmful in the worst-case scenario. Validation and verification of a model helps to ensure it is an accurate and valuable representation. Validation is the process of determining if the model is an accurate representation of the system; the associated question for validation is “Did we build the right model?” Verification is the process of determining if the simulation model works as it is intended to; the associated question for verification is “Did we build the model right?”. There are several techniques in practice for performing validation and verification (Systems Engineering Guide, 2014).

The MITRE *Systems Engineering Guide (SEG)* describes their verification, validation, and accreditation (VV&A) process as not only desirable for outcomes, but also essential. The SEG describes the verification phases as an iterative process which determines if each phase is complete, consistent, and correct, in order to move on to the next phase. Alternatively, the validation phases focus on comparing the system to the simulated model, and determining if the differences between them are acceptable or if they need to be adjusted. The final stage in their process is accreditation, which will state that the simulation model and the associated data are able to be used for a specific purpose.

Verification Techniques

Verification of a simulation model is concerned with the correct implementation and usage of the model. In Robert G. Sargent’s 2016 paper *Verification and Validation of Simulation Models* he explains that when a higher-level programming language is used to create the model, the model should be designed, developed, and implemented using software engineering techniques (188-89). When using a higher-level programming language, verification is most often concerned with concluding that the simulator functions properly and that it was both programmed and implemented correctly. Testing simulation software is most often done through static and dynamic testing. Static testing uses techniques including structured analyses and examining the properties of the software. Dynamic testing takes advantage of techniques including traces, input-output investigations, internal checks, and the reprogramming of essential components (189).

Antibugging is another approach to verifying simulations. Antibugging involves putting in additional checks such as counters, to ensure that the output of the program is what it was intended to be at each moment in time. Another approach to verifying simulations is a one-step analysis, in which the developer will explain the simulator step-by-step to either another person or to themselves to see if they have missed anything within the model or if their logic was flawed. Another verification technique is deterministic modeling. Deterministic modeling is when random variables are replaced with static variables in order to see if the model is behaving properly. Once it is determined that the model is correct, the values can then be changed back to random variables (Hillston, 2013).

Validation Techniques

The MITRE SEG describes five commonly recommended simulation model validation techniques (462). The first technique is comparison to other models. This method involves comparing the model being created to other simulation models that have been previously validated for a similar purpose. Another technique described is face validity; this method involves consulting field experts regarding their opinion of the accuracy of the model's behavior, logic, relationships etc. The next validation technique described is historical data validation. This method can be used if historical data exists for the simulation model being created. Some of the historical data can be used to create and build the model, while the remaining data can be used to test if the model behaves appropriately. Another technique explained is parameter variability - sensitivity analysis. In this technique, one would alter the inputs and internal parameters of the model to see if the output reflects that of the actual system output under those conditions. The last technique described is predictive validation which is when the model is used to forecast the actual system's behavior, then the forecast and actual outcome are compared.

Additional techniques, not described in the MITRE SEG but still worth considering for validation, include internal validity, traces, and Turing tests. Internal validity involves running multiple iterations of the model, and looking for variability among results as this may indicate that the model needs to be altered. Traces are when specific modules within the simulation are followed throughout the simulation; this helps to determine if the logic of the model is correct. Turing tests involve consulting a field expert to see if they can determine differences between real world results and results from the simulation (Sargent, 2016).

Synthea

Synthea is a Synthetic Patient Population Simulator research project created by MITRE. MITRE is a not-for-profit company that operates Federally Funded Research and Development Centers (FFRDCs). Synthea was developed to address the need for health record datasets for use in the creation of innovative software development and other non-clinical use, where realistic data is necessary. Synthea can be used in academic settings for research and can be used by software developers to test the features of new health data interoperability solutions (Synthea Wiki, 2017).

De-identified and anonymized health records can be used for the purposes of testing and software development in the healthcare field, however it is costly to acquire and poses risks of re-identification of the real people in the dataset. The data produced in Synthea is completely synthetic, open-source, and can be used in academics, research, and development. Unlike other synthetic EHR generators that exist, Synthea does not use de-identified data in the creation of their patients but rather it takes advantage of regional datasets, clinical practice guidelines (CPGs), and input from healthcare professionals. It supports and models the ten most common reasons why patients visit their primary care physicians, as well at the ten most common causes of death (Synthea Wiki, 2017).

A Generic Module Framework is used for creating the state machines of the diseases supported within Synthea. Each module is based on publically available data regarding disease incidence, prevalence, and progression as well as CPGs. Synthea uses an agent-based approach to generate the patients, one patient at a time. Each patient generated will go through each disease module within the system, simulating the progression and treatment of disease (Synthea Wiki, 2017).

The modules are defined in JavaScript Object Notation (JSON) and describe the progression of states as well as the transitions between them. Each generated patient begins in an *initial state* when being processed through the disease modules, and will end in a *terminal state* meaning that no further steps will be taken within the module, or in a continuing loop of treatment. There are multiple possible states that can occur between the initial and terminal state, one of which being a simple state. A *simple state* will not add information or actions, it will simply progress to the next defined state in order to chain together a series of complicated or branching transitions. A *guard state* only allows the generated patient to progress beyond it if specific conditions are met, for example, a specified age or age range must be met. A *delay state* will not allow the patient to progress onto the following state until a certain amount of time has passed within the simulator. An *encounter state* will specify that a specific healthcare encounter has occurred and will add it to the patient record. There is also an *encounter end* state which specifies the end date of that encounter and will update the record (Synthea Wiki, 2017).

There are similar states for *conditions, allergies*, *medication orders*, and *care plans* which have an onset and an end state. A *procedure state* indicates that a procedure has occurred during an encounter. The procedure also has an associated reason for the act. The *vital sign state* will indicate the physical state of the person at a certain time, for example during a doctor's office visit. The *observation state* is processed during an encounter and includes observations such as laboratory tests and findings. Similarly, the *multi-observation state* specifies the multiple observations that should be taken at that time. A *diagnostic report state* will group multiple observations into a single report. A *symptom state* will record and update the severity of a patient’s symptoms on a defined scale of 1-100, and the symptoms drive the care seeking behavior. The *counter state* will update the number of times something occurs in the patient’s life such as a disease occurrence. The last state of *death*, indicates that the patient has died or that they are within a terminal state that will end after the date of the generation (Synthea Wiki, 2017).

Certain states have associated clinical codes for further information recording. Synthea supports Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) to describe clinical findings, diagnoses, symptoms, and more in the encounter, procedure, condition onset, and care plan start states. RxNorm codes are used within the medication order state to describe prescriptions and medications. Lastly, LOINC codes are used in the observation state for tests, measurements, and observations (Synthea Wiki, 2017).

There are a variety of transitions that can occur between states. A *direct transition* will simply transition to the next specified state. A *distributed transition* has a certain likelihood of the patient progressing through it; there are more than one distributed states, all of which will sum to 100% from a certain state. A *conditional tradition* will only occur is certain defined conditions are met. A *complex transition* combines direct, distributed, and conditional transitions (Synthea Wiki, 2017).

The generated health records can be formatted in a number of ways depending on the needs of the user. Patients can be exported as Fast Healthcare Interoperability Resources (FHIR) which is a standard created by HL7 for the purpose of exchanging healthcare data electronically. Additionally Synthea can export the generated patients as Consolidated Clinical Document Architecture (C-CDA) which is also defined by HL7. Synthea can also export in HTML or simple text, for a more human-readable format. Lastly the patients can be exported in CSV format which can be useful for relational databases and Microsoft Excel. Synthea exports a total of nine CSV files including patients, encounters, allergies, medications, conditions, care plans, observations, procedures, and immunizations (Synthea Wiki, 2017).

Over one million synthetic EHRs, generated by Synthea, contribute to another MITRE project, SyntheticMass. SyntheticMass provides realistic population and demographic data at the state, county, and town / city levels for Massachusetts at a 1/7th scale. Additionally it has the added ability to filter and view male and female population levels, as well as diabetes, heart disease, and opioid addiction prevalence.

Alternatively, one can choose to view SyntheticMass with the census dataset instead of the Synthea-generated dataset (“About Synthetic Mass”, 2016).

Care Maps / Clinical Pathways

A clinical pathway or care map, as defined by BioMed Central (BMC) Medicine, is a tool used to guide evidence-based healthcare (*Kinsman et al.*, 2010). Children’s Hospital of Philadelphia states that clinical pathways aim to standardize care for clinical problems, processes, and procedures. Clinical pathways serve as a way to perform best practices, based on existing evidence, and avoid unnecessary variations in treatments. As research furthers and better practices emerge, clinical pathways are updated to encourage best practices (“About the Clinical Pathways Program”, 2017).

Synthea models the disease modules based on these clinical pathways, referring to them as “care maps”. In the future, MITRE hopes to have a web interface which would allow healthcare professionals to edit or modify the modules with no programming experience required.

Synthea Review

In 2017 Scott McLachlan published a thesis titled *Realism in Synthetic Data Generation*. Part of this thesis included a review of the Synthea SDG method. In his review he focused on the clinical pathways and data surrounding type-2 diabetes in Massachusetts. He found major disagreements between the prevalence of type-2 diabetes among demographic populations in Synthea compared to the national CDC statistics. Additionally he found the clinical pathways implemented in the disease module for type-2 diabetes within Synthea resulted in amputations for 100% of those diagnosed, which is drastically different than the reported 0.65% rate of diabetic-related amputations in Massachusetts.

He also found that the diagnosis of kidney failure in Synthea among diabetics was not close to the actual average, with 87.06% of type-2 diabetic Synthea patients having kidney failure, in contrast to the 0.17% of diabetics actually affected by kidney failure. In addition to this inaccuracy, he also found that the Synthea patients would not undergo dialysis but continue to live for ten or more years; this is improbable as well because kidney failure requires ongoing dialysis or a kidney transplant in order to not result in a quick death, he explained.

McLachlan lastly focused on the age of diagnosis of type-2 diabetics. Synthea diagnosed most patients around the age of 46, which is the mean age of diagnosis in the United States. However, Synthea failed to diagnose anyone over the age of 52, which is not in alignment with the actual average ages of diagnosis. Synthea also did not meet statistics in diagnosing many patients in the 11 years - 28 years age group. Lastly, Synthea diagnosed too many patients in the less than 10 years of age group, as type-2 diabetes diagnosis typically do not occur in young children.

It is important to note that since that review was published, MITRE has actively worked to correct those abnormalities within Synthea. Synthea now produces much more realistic data regarding type-2 diabetes prevalence, treatment, and diagnosis.

McLachlan provided an in-depth analysis of one disease module that can be useful for future analysis of Synthea and its other existing modules.

Synthea can be a great tool for developers looking to improve EHRs. Synthetic patient data allows developers to reap all the benefits of using real health data in their work, but without the risk of data breaches and at no cost. In order to be considered reliable for creating these new solutions, Synthea must be realistic on the patient, disease module, and population levels.

Research Proposal

This MQP will accomplish three different validations and verifications of Synthea. Firstly, the synthetic patient records will be verified to be realistic representations of health records. Second, the pregnancy disease module will be edited to contain necessary care and procedures, and then verified to be a realistic representation of a standard of care. Third, Synthea will be validated at the population level, specifically validating that the disease prevalence levels are realistic compared to published data regarding disease prevalence. This project is concerned with the validating the realism of Synthea for the purposes of software development of EHR solutions and other non-clinical secondary use. It will also help the Synthea team to assess what changes they may need to make to the simulator to improve its suitability.

Synthetic Patient Health Record

The hypothesis for the synthetic patient health records is that they are realistic representations of actual health records, and that the missing information is not crucial to the completeness of the health records. The records will be verified through interviews with six medical doctors, employed by MITRE. These interviews will be semi-structured in order to allow for elaboration on answers and follow up questions. In these interviews the questions regarding how realistic they perceive the records to be, where information in the record appears to be missing, what type of information is missing, and how crucial it is to the health record that the information be there. The interview data will be recorded in a Qualtrics survey. This allows for quick access the data for analysis of the results. The results of the interviews, specifically where information is missing or incorrect, will help with correcting and providing suggestions for specific disease modules.

Disease Module

The hypothesis of the pregnancy disease module is that the lack of information in the module results in incorrect patient record detail. The disease modules in Synthea are based on standards of care and are detailed, but are also an abstraction of true care maps. This will require an initial analysis of the pregnancy disease module to identify where information is missing or incorrect. The module care plan will be compared against published care plans in order to find the necessary information to add.

After comparing the Synthea modules with actual care plans, the disease module, defined in JSON, will be edited to present realistic data in terms of prevalence and progression. This requires using parameter variability - sensitivity analysis as the validation technique, adjusting values within the module to see how they affect the outcomes and if those outcomes align with real world data. For verification, a one-step analysis will be conducted, in which the edited disease module is explained step by step to a field expert, to check that the logic of the module is clear and correct.

Population Report

The hypothesis for the population report is that the levels of disease prevalence across various demographics will be similar to real world data, but will vary slightly from actual statistics regarding prevalence, when those statistics are available, by a standard deviation of +/- 10%. A detailed report of disease prevalence across age, gender, and race will be created. This report will be created using the synthetic patients created in Synthea, and compared to United States and Massachusetts published statistics of prevalence. This report will be created by populating a database with the data, querying the results in Structured Query Language (SQL), and creating the report using Java.

Internal validity will be used as the validation technique, populating many synthetic records and looking for noticeable variations between the outcomes will help to assess how accurate the synthetic data is and how often it is accurate. A one-step analysis for verification will be conducted in which it is explained step-by-step how the report was created, how it was programmed, and how it can be edited to house additional information in the future to a MITRE colleague.

Methodology

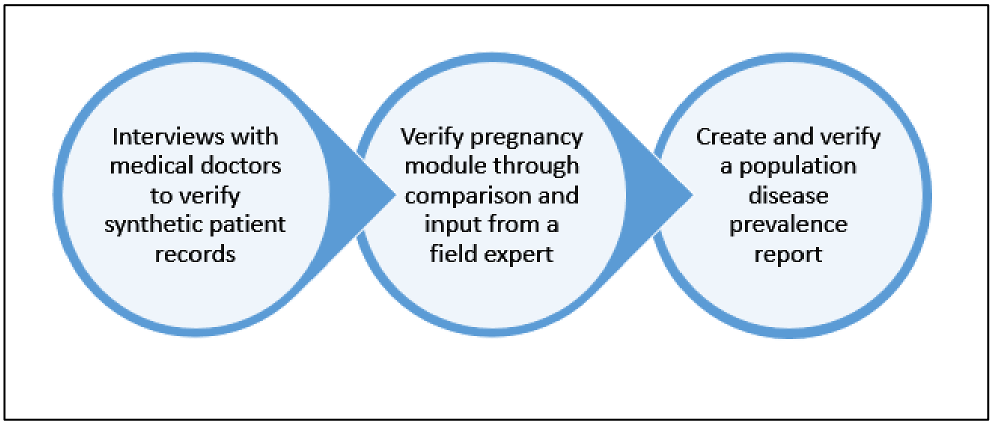

Interviews with medical doctors to verify synthetic patient records
Verify pregnancy module through comparison and input from a field expert
Create and verify a population disease prevalence report

Figure 1: Methodology

This methodology is divided into three main objectives which accomplish the goals stated in the research proposal. The first objective is to interview medical doctors to verify the synthetic patient health records are realistic representations of actual health records. This method takes advantage of the WPI Qualtrics tool, used for interviews and surveys, in order to successfully organize and evaluate information. The second objective is to validate and verify the pregnancy module in Synthea is a realistic representation of published standards of care. This requires editing the module, and researching to find the necessary information to include within the module. The final objective is to create, and validate a population disease prevalence report has realistic prevalence levels compared to published data regarding disease prevalence. This uses

Java and SQL to query an H2 database containing information from individual profiles and organize the results into a report.

Objective I: Interview Medical Doctors to Validate Synthetic Patient Health Records

This objective began with collaboration with MITRE and WPI advisors to compose semi-structured interview questions, which covered the major questions, but also allowed for elaboration and follow up questions. The questions were formatted into a WPI Qualtrics survey in order to keep track of the data in real time and for efficient access. This also minimized the amount of decoding to be done following the interviews. Additionally, the interview questions were reviewed and approved by WPI's Internal Review Board (IRB). Next, interviews were requested with six medical doctors, employed by MITRE.

Selecting which patient profiles to use in the interviews proved to be a challenge. When the records are formatted as simple text, they are easy for humans to read but contain less detailed information compared to when the records are formatted in FHIR; however the FHIR format is more difficult for humans without a computer science background to understand. The FHIR profiles contain personal information such as maiden name, address, socioeconomic status, education level, and primary language spoken. They also contain additional information for procedures and encounters that the text files do not have including a start and end time. The FHIR files contain patient goals that are not defined in the text files as well, such as keeping a blood pressure below a certain threshold. The FHIR profiles also contain additional observations including height, weight, BMI, blood pressure, and blood glucose levels.

Although the FHIR records are more detailed, they are also much longer. For example, one of the patient profiles used during the interviews was 329 lines in the text format, but was 50,273 lines in the FHIR format. Ultimately, five patient records, generated by the ruby version of Synthea and formatted as text files, were selected for use during the interviews with the M.D.s. Records with patients of various ages, race, gender, and diseases were selected.

The interviews were conducted over phone calls and the patient profiles were sent over to the M.D.s via email in a PDF format. At the beginning of each call, the format of the health record and what they could expect to see within the record was explained. The same set of questions were asked for each patient to ensure that similar data was collected for each record. The responses were recorded in the Qualtrics. The open-ended questions allowed for follow-up questions as well.

Objective II: Validate and Verify Pregnancy Module

This objective started with researching pregnancy, miscarriage, and abortion rates among various ages in the United States and Massachusetts, and prenatal clinical care plans to compare them to the Synthea care plan. The clinical care plans chosen to compare Synthea to were the University of Michigan Medicine: Prenatal Care Guidelines for Clinical Care Ambulatory and National Guideline Clearinghouse: Routine Prenatal Care.

Dr. Susan Haas gave suggestions for edits of this module as well, going through three rounds of review and feedback. After making edits to the module, the generated flow-chart image of the module were reviewed with Dr. Haas. She was able to provide insight into which practices are no longer standard, the timeline of procedures and appointments, follow-up care, and appropriate SNOMED-CT terminology.

Objective III: Create, Validate, and Verify Population Disease Prevalence Report

First, a prevalence CSV template was made, to be populated after the statistics have been calculated. Next, a database was populated with generated living patients, from the java version of Synthea. The first SQL queries were to find total counts for diabetes prevalence among different races and genders for adults (18 years and above), then this query was used this as a template to expand on to Hypertension, Coronary Heart Disease, and Asthma. All of these queries were written in SQL.

The queries were executed in Java by creating a string builder that adds additional lines to an initial query depending on the disease, gender, age group, and race that was on a specific line in the template. The additional strings, containing queries for these descriptors, would be executed if an “*” was not found on that line (which indicates all) and would insert the description (i.e. “native”) into the query. This was an efficient way to query based on the contents of the report template.

The total count of these queries is divided by the total count of living patients within that demographic in order to get the Synthea prevalence rate of that specific disease among the defined demographic. A similar string building technique was used to run the queries for population. The Synthea prevalence rate was also converted to a percent, and compared against the actual prevalence rate percentages. Finally a difference between the two was calculated, if an actual prevalence rate was found for that specific line in the report. It proved to be difficult to find prevalence rates for each demographic for every disease, as some diseases were not researched in such detail. However, as more data is published, the values can be inserted in the report and compared against the values generated in Synthea.

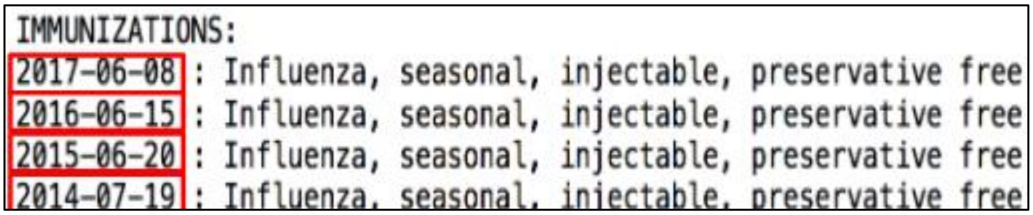
Another piece of this report includes prevalence rates of one disease given that a person has one or two defined pre-existing conditions. The same techniques were used to find the total count of people and the prevalence, however this required joining multiple attribute tables instead of querying on just one. After completing the template and ensuring the queries and calculations worked, ten reports each of which contained 1,000 living patients were generated. The results of the ten reports were then compared, looking for any prevalence percentages that were drastically different among the reports.

Implementation & Results

Synthetic Patient Health Record Interviews

On average, the doctors found the synthetic health records to be fairly realistic, with an average realism rating of 2.9375 on a scale of 1-5, 5 being very realistic. Additionally the doctors found that the importance of the missing information was around 3.875, 5 being very important. The interviewed doctors found the records to be realistic in how much information was missing at times, as it is common for records to be lacking in information as patients move from office to office. Another aspect found to be realistic is the smaller dosing of medications for children than adult dosages.

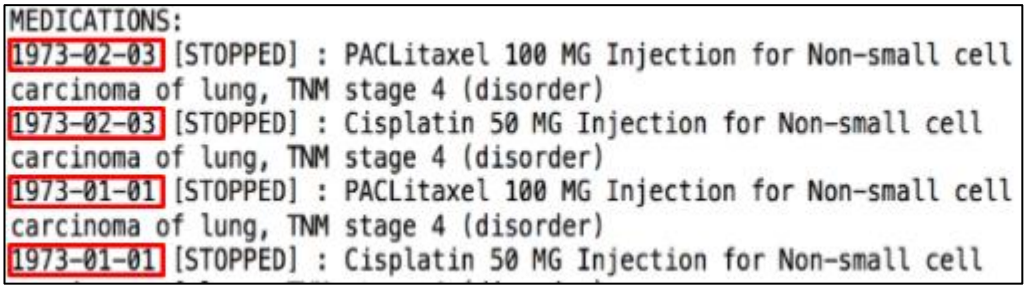
Some aspects of the records were unrealistic as well. For example, patients were very disciplined about getting their flu shot each year, which is ideal but not realistic. Additionally, most patients received flu shots in the spring or summer months, when the flu shot is typically administered in fall and winter.



```
IMMUNIZATIONS:
2017-06-08 : Influenza, seasonal, injectable, preservative free
2016-06-15 : Influenza, seasonal, injectable, preservative free
2015-06-20 : Influenza, seasonal, injectable, preservative free
2014-07-19 : Influenza, seasonal, injectable, preservative free
```

Figure 2: Off-season influenza immunization

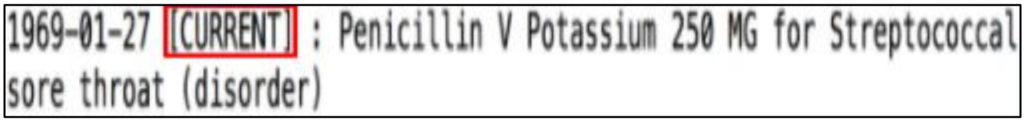
Another unrealistic aspect of the records was that patients would receive drugs at a time before they was created. For example, synthetic patient Blair872 Doyle966 received PAClitaxel and Cisplatin for non-small cell lung cancer from 1972-1973, when those drugs were not approved by the FDA until 1993.



```
MEDICATIONS:
1973-02-03 [STOPPED] : PACLitaxel 100 MG Injection for Non-small cell
carcinoma of lung, TNM stage 4 (disorder)
1973-02-03 [STOPPED] : Cisplatin 50 MG Injection for Non-small cell
carcinoma of lung, TNM stage 4 (disorder)
1973-01-01 [STOPPED] : PACLitaxel 100 MG Injection for Non-small cell
carcinoma of lung, TNM stage 4 (disorder)
1973-01-01 [STOPPED] : Cisplatin 50 MG Injection for Non-small cell
```

Figure 3: PACLitaxel and Cisplatin medications before FDA approval

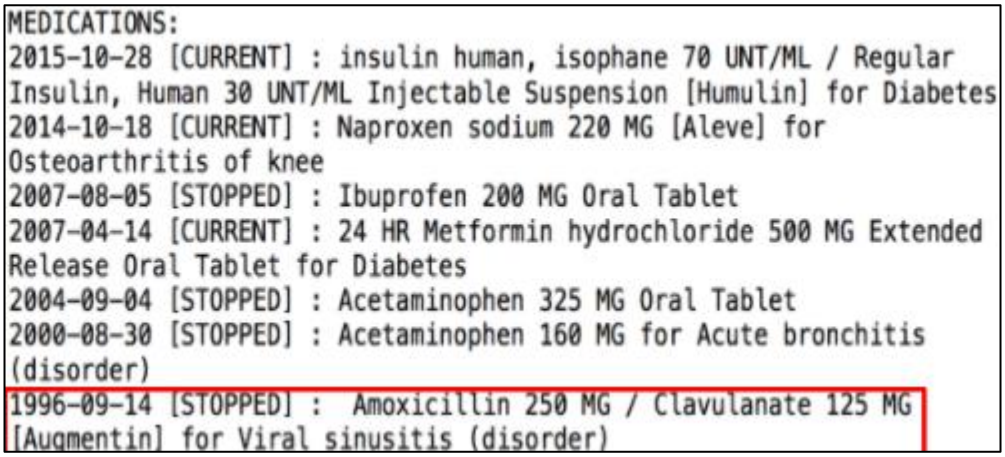
Another example of an inaccuracy was that all prescription medications for penicillin were “current” medications in the record, meaning that once a patient was prescribed penicillin they were prescribed it for life.



```
1969-01-27 [CURRENT] : Penicillin V Potassium 250 MG for Streptococcal
sore throat (disorder)
```

Figure 4: Penicillin prescription with no end date

Some prescriptions were also unnecessary including an antibiotic, amoxicillin, for viral sinusitis on synthetic patient record Manuel709 Schmitt79.

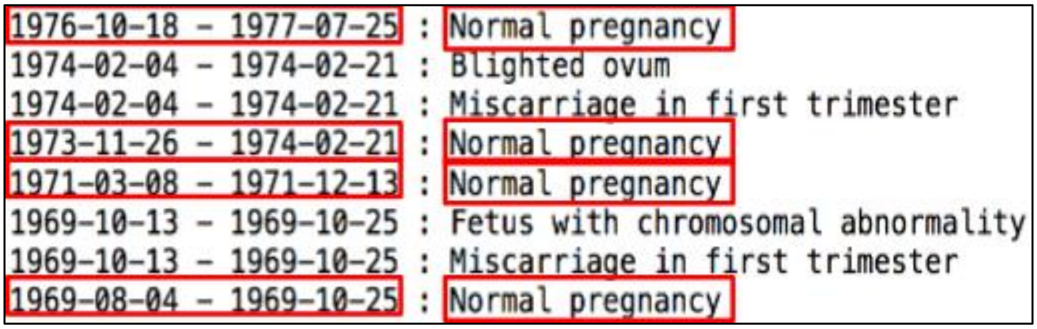


```
MEDICATIONS:
2015-10-28 [CURRENT] : insulin human, isophane 70 UNT/ML / Regular
Insulin, Human 30 UNT/ML Injectable Suspension [Humulin] for Diabetes
2014-10-18 [CURRENT] : Naproxen sodium 220 MG [Aleve] for
Osteoarthritis of knee
2007-08-05 [STOPPED] : Ibuprofen 200 MG Oral Tablet
2007-04-14 [CURRENT] : 24 HR Metformin hydrochloride 500 MG Extended
Release Oral Tablet for Diabetes
2004-09-04 [STOPPED] : Acetaminophen 325 MG Oral Tablet
2000-08-30 [STOPPED] : Acetaminophen 160 MG for Acute bronchitis
(disorder)
1996-09-14 [STOPPED] :  Amoxicillin 250 MG / Clavulanate 125 MG
[Augmentin] for Viral sinusitis (disorder)
```

Figure 5: Antibiotic for viral condition

Another anomaly is that the patient records did not specify if the patient had no known allergies, it was only specified if the patient had allergies. This was a point of concern from the doctors who believed that it should state that there are no known allergies in the record.

Additionally, the age of some pregnant mothers was unrealistic. For example, synthetic patient Christeen210 Bahringer247 was pregnant at ages 51, 47, 45, and 43, of which two of these pregnancies resulted in childbirth. Which is possible in the current day, likely through in vitro fertilization (IVF), but likely not in 1976 or earlier as the first successful birth from IVF occurred in 1978. Although she could have been pregnant, this would likely not be a normal pregnancy but rather a high risk pregnancy. It is also unlikely that she was able to have this many children at this age. Additionally this patient also had a history of heart disease and diabetes making these pregnancies even more unlikely at her age.



```
1976-10-18 - 1977-07-25 : Normal pregnancy
1974-02-04 - 1974-02-21 : Blighted ovum
1974-02-04 - 1974-02-21 : Miscarriage in first trimester
1973-11-26 - 1974-02-21 : Normal pregnancy
1971-03-08 - 1971-12-13 : Normal pregnancy
1969-10-13 - 1969-10-25 : Fetus with chromosomal abnormality
1969-10-13 - 1969-10-25 : Miscarriage in first trimester
1969-08-04 - 1969-10-25 : Normal pregnancy
```

Figure 6: Unlikely normal pregnancy

It is important to note that the data collected is from five patient records, with opinions from six medical professionals. More information on realism and missing or incorrect information in Synthea could be found by using additional profiles and interviewing additional medical professionals.

Pregnancy Module

The first shortcoming found is that the first prenatal appointment within Synthea is at 10 weeks, when this should actually be when the second prenatal appointment is occurring. This was fixed this to have an initial appointment at 6 weeks, with a follow up appointment at 10 weeks. There was a large gap between prenatal appointments two and three of 11 weeks, which is almost an entire trimester, so it is unlikely a pregnant women would not have an appointment for that long. To address this and be in accordance with clinical care plans, follow up appointments at weeks 16, 22, 28, 32, 36, 38, 39, 40, 41, and 42 were also implemented.

Another shortcoming was the lack of information recorded at each prenatal appointment. The types of information missing included blood tests, urine tests, ultrasounds, vaccinations, screenings, and care plans. The appropriate information, procedures, and observations were added to the appropriate prenatal appointments, according to the University of Michigan Medicine, National Guideline Clearinghouse, and expert suggestions. Additionally, an attribute for RH negative blood women was added, as they require additional screenings and procedures throughout the pregnancy module.

The next shortcoming was that Synthea does not have postpartum care plans within the pregnancy module. The module ended the moment the mother gave birth and contained no follow up appointments or care plans following the birth. Additionally, there were no follow up appointments or care plans assigned following an induced termination of pregnancy. However there were follow up appointments following miscarriages, but no assigned care plans. Follow up appointments were implemented and the appropriate care plans and necessary procedures performed at each of them. An option for a medically-induced termination of pregnancy following an ectopic pregnancy was implemented, as previously only surgically-induced termination of pregnancy was in the module. Additionally, the timing of birth was edited to be weighted more towards weeks 37, 38, and 39, as it was previously weighed a lot heavier to the later weeks, and categorized births before 40 weeks as premature.

After speaking with Dr. Susan Haas regarding the pregnancy module and asking for her insight on the edits and additions, a non-low-risk pregnancy pathway was implemented that can be expanded in the future. This non-low risk pathway would provide additional specialized care for expecting mothers including those with preeclampsia, diabetes, heart disease etc. The health records for pregnant women in Synthea are now much more complete with information about laboratory tests, ultrasounds, education, and follow-up appointments for birth, induced abortion, and miscarriages. It is important to note that the pregnancy module in Synthea is an abstraction of the standard of care, and does not account for additional factors such as insurance coverage in the care that is given, thus all patients receive the same level of care.

Disease Prevalence Report

Compiling the data from the ten reports, I found that Synthea is creating disease prevalence percentages that were on average +5.675% total difference from the actual prevalence percentages, for all living adults. Going into each disease individually, diabetes prevalence differed by +26.9%, hypertension differed by 1.1%, coronary heart disease differed by -1.0%, and asthma differed by -4.3%.

Table 1: Synthea disease prevalence compared to actual disease prevalence

	Actual Prevalence Percent	Synthea Prevalence Percent	Difference
Diabetes	8.0	39.4	26.9
Hypertension	29.1	30.2	1.1
Coronary Heart Disease	6.0	5.0	-1.0
Asthma	9.6	5.3	-4.3

Additionally, the prevalence levels for all living adults in each of the ten reports were extremely similar, differing on average by less than 1% from report to report. Error in this prevalence report could also come from the published data, as the prevalence rates may be different now than when the reports were last published. Additionally, comparing the total counts for unique conditions, and the prevalence percentages, Synthea was generating consistent counts in each report of 1,000 patients.

Conclusion

In conclusion, Synthea has proven to be a realistic yet synthetic patient generator. The majority of inaccuracies in the patient records are quick fixes, such as adding an end date to a penicillin prescription, putting a guard on certain drugs so that they cannot be administered before a certain year, and adding a restriction on the flu vaccine so that it is only available in the fall and winter months. The disease modules themselves will continue to get more sophisticated and realistic as more people contribute to them. As these modules become more realistic, realistic disease prevalence levels among various demographics will follow. The suggested course of action is to begin by fixing the little things at the patient level like the drugs and vaccines. Next focusing on the modules, modeling them off published standards of care, consulting field experts on the changes being made, and putting in detailed care plans. Once the small details are fixed and the modules are more detailed, the prevalence data will follow.

This MQP serves as a repeatable process for identifying areas of improvement of Synthea in the future. Interviewing medical doctors regarding the realism of patient records highlights the areas in Synthea that may need improvement. Comparing the modules to published care plans and consulting a field expert for multiple rounds of review is helpful in creating the most detailed and accurate abstractions of care possible. Finally, after editing the modules, and creating the disease prevalence report, it can be seen which diseases are occurring too often or not often enough. From there the disease modules can be altered again to produce more realistic values.

Works Cited

“About Synthetic Mass.” Synthetic Mass, The MITRE Corporation, 2016, syntheticmass.mitre.org/about.html.

“About the Clinical Pathways Program.” Children's Hospital of Philadelphia, 2017, www.chop.edu/pathways/about.

Bresnick, Jennifer. “What You Need to Know about Health Data Interoperability.” Health IT Analytics, 24 Apr. 2015, healthitanalytics.com/news/what-you-need-to-know-about-health-data-interoperability.

“Electronic Health Records.” CMS.gov, Centers for Medicare & Medicaid Services, 26 Mar. 2012, www.cms.gov/Medicare/E-Health/EHealthRecords/index.html.

Hillston, Jane. “Model Validation and Verification.” 19 Sept. 2003.

“HITECH Act Enforcement Interim Final Rule.” HHS.gov, US Department of Health and Human Services, 16 June 2017, www.hhs.gov/hipaa/for-professionals/special-topics/HITECH-act-enforcement-interim-final-rule/index.html.

McLachlan S. “Realism in Synthetic Data Generation.” MPhil Thesis. Computer Science and Information Technology, Massey University: New Zealand, 2016.

“Obstacles to Interoperability within Closed Systems.” Athena Health, 17 June 2017, www.athenahealth.com/knowledge-hub/clinical-integration/healthcare-interoperability.

Palma, George. “Electronic Health Records: The Good, the Bad and the Ugly.” Becker's Hospital Review, 14 Oct. 2013, www.beckershospitalreview.com/healthcare-information-technology/electronic-health-records-the-good-the-bad-and-the-ugly.html.

Sargent, Robert G. “Verification and Validation of Simulation Models .” L. C. Smith College of Engineering and Computer Science, Syracuse University, 2016.

Swain M, Charles D, Patel V, & Searcy T. Health Information Exchange among U.S. Non-federal Acute Care Hospitals: 2008-2014. ONC Data Brief, no.24. Office of the National Coordinator for Health Information Technology: Washington DC, April 2015.

Synthea Wiki. 2017, https://github.com/synthetichealth/synthea/wiki/.

“Systems Engineering Guide.” The MITRE Corporation, 2014.

“The Challenges of Sharing Data Between Separate EHRs.” Afia, 17 Mar. 2017, afiahealth.com/challenges-sharing-data-separate-ehrs/.

“U.S. Still Leads the World in Science and Technology; Nation Benefits From Foreign Scientists, Engineers.” RAND Corporation, 12 June 2008, www.rand.org/news/press/2008/06/12.html.

Wachter, Robert M. “Opinion | Why Health Care Tech Is Still So Bad.” The New York Times, The New York Times, 21 Mar. 2015, www.nytimes.com/2015/03/22/opinion/sunday/why-health-care-tech-is-still-so-bad.html.

“What Is EHR Interoperability and Why Is It Important?” HealthIT.gov, 15 Jan. 2013, www.healthit.gov/providers-professionals/faqs/what-ehr-interoperability-and-why-it-important.