Health Insurance and Its Impact on the Survival Rates of Breast Cancer Patients in Synthea

Robert Francis Scalfani

Worcester Polytechnic Institute & MITRE

October 10, 2019

WORCESTER POLYTECHNIC INSTITUTE

COMPUTER SCIENCE PROGRAM

Health Insurance in Synthea Final Report

A Major Qualifying Project

Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the
Degree of Bachelor of Science by:

Robert Scalfani

Project Advisors:

Professor Shamsnaz V. Bhada

Professor Lane T. Harrison

Worcester Polytechnic Institute

Jason Walonoski

MITRE

Sponsored By:

The MITRE Corporation, Open Health Services

October 10, 2019

Acknowledgements

Thank you to Professor Shamsnaz V. Bhada, Jason Walonoski, Professor Lane T. Harrison, and Dr. Robert Lieberthal for advising this project and ensuring its success. Thank you for all of your time, teaching, and expertise without which this project would not have been possible. Thank you for the opportunity to work on and learn from this project.

Table of Contents

Acknowledgements 3

Table of Contents 4

Table of Figures 6

Table of Tables 6

Abstract 8

1 Introduction 9

2 Background 11

2.1 Health Insurance in the United States 12

2.2 Survival Rates as a Metric for Measuring Policy Impacts 16

2.3 Health Data Barriers and Concerns 16

2.4 Synthea 18

2.5 Breast Cancer 19

2.6 Data Sources 21

3 Research Proposal 24

3.1 Hypothesis 1: Determining the closeness of Synthea's breast cancer and health insurance outputs with real-world data 25

3.2 Hypothesis 2: Survival rates for breast cancer patients remain unchanged with or without insurance 27

3.3 Breast Cancer Focus 29

4 Methodology 31

4.1 Implementing Health Insurance in Synthea 32

4.2 Implementing Loss-Of-Care and its Impact on Survival 37

4.3 Implementing Real-World Health Insurance Levels 39

5 Results 42

5.1 Result 1: Verifying Synthea's Outputs 42

Real-World Data Used 43

Compare Synthea's Breast Cancer Outputs 46

Compare Synthea's Loss-Of-Care Impacts 54

Compare Synthea's Tuned Health Insurance Levels 56

5.2 Result 2: The Survival Rate of Uninsured Patients is Lower than Insured Patients. 58

6 Conclusion 62

Bibliography 64

Table of Figures

Figure 1: Uninsured Patient Loss-Of-Care 14

Figure 2: Implementation and Hypothesis 1 Breakdown 27

Figure 3: Hypothesis 2 Survival Rate Testing Breakdown 29

Figure 4: Sequence Diagram of Receiving Different Insurance 33

Figure 5: Synthea Class Diagram of Relevant New Insurance Features Added 35

Figure 6: Synthea Class Diagram of Relevant Components Prior to New Features 36

Figure 7: Comparison of Staging Incidences in Synthea and Real Data 51

Figure 8: Yearly Survival Rates in Synthea Compared to SEER Data 54

Figure 9: Comparison of Survival Rates By Stage of Insured and Uninsured Patients in Synthea 59

Table of Tables

Table 1: Underlying Population of Each Used Dataset 23

Table 2: Breakdown of Insurance Distributions of Real-World Breast Cancer Patients. 39

Table 3: Synthea's Tuned Insurance Incidence Distribution 40

Table 4: Comparison of Synthea's Insurance Distributions with Real-World Data 41

Table 5: SEERS Dataset Underlying Population 43

Table 6: Insurance Status Dataset Underlying Population 44

Table 7: Breastcancer.org Dataset Underlying Population 45

Table 8: Synthea's Underlying Population 46

Table 9: Conversion from Synthea Staging to Location and Macro Staging 47

Table 10: Synthea's Output of Staging Incidence and Survival Rates of Fully Insured Breast Cancer Patients 48

Table 11: Percentage of Incidences of Real-World Location Stages 49

Table 12: Comparison of Synthea Output to Real Data of Staging Incidences 50

Table 13: Comparison of Survival Rates by Stage between Synthea and Real Data 51

Table 14: Comparison of Uninsured Patient Survival Rates Between Synthea and Real Data 55

Table 15: Number of Years Each Insurance Category Was Utilized in Synthea 57

Table 16: Comparison of Percent of Utilization of Insurance Categories Between Synthea and Real Data 57

Table 17: Comparison of Overall Survival Rates of Uninsured and Insured Breast Cancer Populations in Synthea 59

Abstract

To do any inference regarding healthcare policy, researchers need secure and protected health data which is restricted by privacy laws and interoperability issues. Synthetic health data provides a way to generate and investigate data without concerns of violating legal restrictions (HIPAA). In this research, we built health insurance and loss-of-care modules into a synthetic health data simulator (Synthea) to simulate and analyze the impact of health insurance on breast cancer survival rates. We successfully reflected real world insurance and loss-of-care impact statistics in Synthea.

1 Introduction

Health Insurance is a key part of the American healthcare system [1]. Because health insurance increases a people's access to care, it leads to better health outcomes for a population [1]. Health insurance is also subject to government policy over its implementation [2]. Like all policy, its implementation can be decided through the use of data which provide insight on the public health implications and statistics of policy. However, health data is subject to barriers to its access, including interoperability issues [3] and legal restrictions [4] [5]. Without access to the complete health data that policy analysis is based on, researchers and policy makers face difficulties in making fully informed conclusions and inferences about health policy, including health insurance.

One way to bypass the issues that accessing health data presents is through the use of Synthea. Synthea is an open-source health data simulator developed by the MITRE Corporation that creates synthetic, yet realistic, health data. It generates complete health record histories for each individual patient, allowing for the creation of realistic data that can be analyzed and utilized. However, Synthea assumes that patients receive all of the care that they need, which is unrealistic. Synthea also does not currently feature health insurance or loss-of-care impacts as realsitically modeled features. When policy makers and researchers cannot analyze health insurance and policy impacts, they cannot make the best decisions for the healthcare industry and the Americans who need it. In this MQP, we implemented health insurance and loss-of-care modules into Synthea to analyze the impacts that health insurance has on the survival rates of breast cancer patients. Breast cancer patients were the population focus used for

this MQP as a way to limit the scope for our initial research. We chose survival rates as the metric to measure because it is used in real-world policy analysis [6]. The survival rates of breast cancer patients are well documented which allowed us to simulate and analyze it in Synthea.

Our implementation of health insurance in Synthea could be used for future health policy and research analysis. We established our newly developed health insurance and loss-of-care modules to be reflective of real life and that health insurance has a positive impact on survival rates. Health insurance, as an important aspect of health policy, is regulated based on health data. Because of the difficulties in accessing health data, both for analysis and to measure the impact of different policies, we modelled and simulated health insurance in Synthea to find its impact on the survival rates of breast cancer patients.

2 Background

Health insurance in the United States has undergone years of policy change and debate. But, like any public policy, the decision over how health insurance should be implemented requires a complete understanding of the situation. Understanding the complete picture requires using data and analysis. One way that health data determines health policy, including with regard to health insurance, is through the use of population survival rates. However, researchers struggle to evaluate survival rate data due to the barriers of access to health data. Health data barriers include interoperability and privacy restrictions which have resulted in incomplete and inaccessible data. One way that a researcher could get a complete and realistic picture of health and survival rate data is through the use of Synthea, a synthetic health record and patient generation software. By simulating a population of health records and health insurance companies, realistic, yet synthetic, data could be produced that allows for the analysis of survival rates based on different policies. It even allows for the ability to easily change variables to see the different impacts that they could have. For the scope of this MQP, we tested the impact that health insurance can have on the survival rate of breast cancer patients, with the expectation that this research and implementation could expand to include more, and eventually all, diseases. To provide a background for the topics of the MQP, this chapter describes:

- Health insurance in the United States and its impacts
- How researchers use survival rates

- The difficulties in accessing real health data
- Synthea, a health data simulator
- Breast cancer information
- The data sources utilized for this research

2.1 Health Insurance in the United States

Health insurance exists as a way for patients to overcome the high costs of healthcare expenses [7]. These expenses add up and include everything from treatment for serious disorders and prescription drugs to routine yearly checkups. For many patients, health insurance allows for and encourages medical care. When only a small payment is required as opposed to the full fee for care, a patient is far more likely to seek and utilize it. Otherwise, health conditions can easily go unchecked for fear of an expensive bill [7]. By acting as a way for patients to overcome the high monetary barrier-for-entry of healthcare, health insurance has vastly expanded the breadth of care that patients can receive [8].

Origins of Health Insurance in the United States

Healthcare costs have risen exponentially over the last century. In 1930, 3.5% American GDP was spent on health care, a number which has risen to over 15% today [9]. This is logical: considering that healthcare is expensive, labor-intensive, and constantly developing new treatments. These costs, coupled with the countless other

expenses that healthcare incurs, meant that a health insurance system would be the only way for medical care to continue sustainably in the United States.

Health insurance is a relatively modern phenomenon – only emerging as recently as the 1920's in the United States and, even then, on a very small scale [9]. The first health insurance system, known at the time as 'Hospital Insurance', was a trial carried out at a single hospital in Texas and charged its subscribers a $6 yearly premium [9]. Today, these premiums are charged monthly and cost hundreds to thousands of dollars per month [10]. The experiment was a success: the hospital's primary goal, increase cash flow, was accomplished. This 'scheme' as it was thought of at the time spread quickly to other medical institutions. Before long, groups of hospitals created insurance that was honored at all participating institutions [9], creating the first health insurance networks as we know them today. From the early notion of hospital insurance came the first modern health insurance company, Blue Cross, in 1932 [9].

Impacts of Health Insurance

In 2017, 91.2% of Americans had health insurance while 8.8% did not [11]. This uninsured percentage has widely been considered a metric of the health and success of the healthcare system because of its indication of a population's access to healthcare [12]. Health insurance has also been credited with improved healthcare outcomes, more regular care, and early detection and management of health conditions [12]. Uninsured breast cancer patients have even been shown to be 60% more likely to die from the disease than insured ones [13] [14]. A breakdown of the percentage of uninsured patients losing care is featured in Figure 1 from the Kaiser Family Foundation [15].

Overall as a key component of the healthcare system, health insurance has a positive impact on the health of populations.

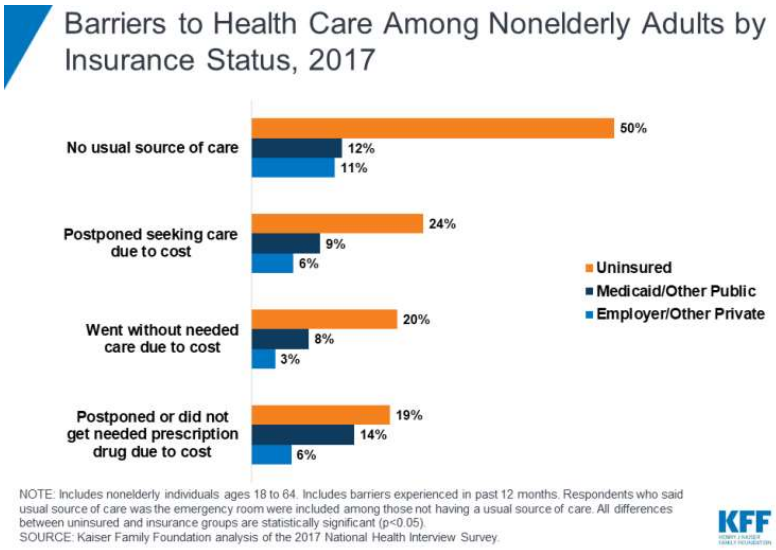

Figure 1: Uninsured Patient Loss-Of-Care

Private Health Insurance

Private health insurance is the most common type of insurance that patients have. 56% of Americans are covered by private insurance, 49% are covered by an employer and 7% is purchased directly from an insurance company [16]. With such an impact on health outcomes, health insurance's main source, private insurance, is a key component of the United States health system. However, it is also the most expensive for patients, leading to the creation of public government insurance.

Government Health Insurance

The two primary government insurances are Medicare and Medicaid. These health insurance programs were signed into law by President Lyndon Johnson in 1965 [17]. Medicare was created in order to provide insurance for those over the age of 65, for

whom it is near impossible to get private insurance due to affordability and companies rejecting this age group [18]. In 2019, 60.6 million Americans received health coverage through Medicare [19]. Medicaid was created as a way for Americans who cannot afford private insurance to have access to health care and health coverage [20]. Medicaid covers over 70 million Americans [20].

In order to qualify for Medicaid, a patient must meet the following requirement:

- Have a yearly income less than the federal poverty level multiplied by 1.33. The federal poverty level is currently $12490 [21]

In order to qualify for Medicare, a patient must meet one of the following requirements [22]:

- Be over the age of 65
- Suffer from Amyotrophic Lateral Sclerosis
- Have Kidney Failure

Health insurance has had a significant positive impact on the health and survival rates of the American population [12]. There have been a multitude of health insurance companies in the country, including private insurance and government insurance. With the continued expansion of government funded health insurance, both from Medicare, Medicaid, and Obamacare, policymakers and researchers have needed to evaluate the impacts that health insurance has to institute public health policy. Survival rates are one way to measure policy, and health insurance's, impacts.

2.2 Survival Rates as a Metric for Measuring Policy Impacts

Survival rates have been used to measure the burden and impact of disease for centuries [6]. These include analyses like determining the leading causes of death in the United States as well as age-specific mortality rates that can influence policy decisions [6]. Survival rates are simply calculated as the number of deaths divided by the size of the sample [23]. Survival rates are based on health data, however analyzing this data can be incredibly difficult because of the barriers to its access.

2.3 Health Data Barriers and Concerns

Despite the benefits that sharing health data would provide, both to researchers and policy makers for data analysis, health data is subject to privacy, legal, and interoperability issues that prevent its accessibility [4]. These barriers to health data are characterized by HIPAA, the *Health Insurance Portability and Accountability Act*, which enforces patient protections on health data [4]. Health record formats differ between each hospital, resulting in a lack of interoperability and inability to gather and integrate large amounts of data for analysis [24].

Legal and Privacy Regulations

Health data is subject to legal restrictions to protect patient the privacy. HIPAA, "The *Standards for Privacy of Individually Identifiable Health Information* ("Privacy Rule") establishes, for the first time, a set of national standards for the protection of certain health information." [25]. HIPAA makes it very difficult to share data and do research. Within individual hospitals, a hospital has control and access over their small

sample of data. However, an outsider academic researcher is barred from accessing the data, even for a rigorous, controlled study because of HIPAA [5]. However, the data that exists holds huge potential for research and doctors to determine policy and treatment impacts. Because of privacy and legal restrictions preventing the use of health data in research studies, it can be difficult for policy makers and researchers to analyze health data for information about survival rates.

Interoperability

Health data is also difficult to collect in a large-scale and complete format because of interoperability issues. Issues include patient identification (for cross-hospital health records) with studies showing that "up to one in five patient records are not matched even within the same health care system." [3] and "as many as half of patient records are mismatched when data is transferred between health systems" [3]. Without the ability to access patients' complete health records, especially when amplified when trying to analyze an entire population, it can be near-impossible to have full and accurate datasets. A federal patient identification system, while proposed, has faced backlash because of HIPAA [3]. Another interoperability issue is the lack of technological and record format standards across healthcare facilities. Without consistent record formats, both digital and paper, data exchange between different healthcare facilities can be near-impossible [3]. Complicated data transactions tend to fail, posing additional barriers to the flow of information [3]. Analyzing data relies on a complete and accurate dataset, a requirement which is difficult to create and access with real-world health data.

While the healthcare industry continues to struggle with data issues, many organizations are making an effort to ease interoperability and privacy issues: to improve the efficiency of care and allow for the analysis for complete data. One such solution for providing complete datasets is the open-source health care modeling and simulation software, Synthea.

2.4 Synthea

Synthea is an open-source health data modeling simulator developed by the MITRE Corporation. It simulates and outputs synthetic but realistic health records at an individual level for each patient in a population, modeling everything from diseases and procedures to wellness encounters. It uses real-world incidence rates, conditions, and treatment plans based on real diseases and illnesses. Medications, procedures, or simply time, result in each disease's conclusion or, potentially, the person's death, reflecting the real-world. All of the demographic data that is used to generate synthetic people also directly reflects real life, with real incomes and job levels.

Synthea as a Solution to Health Data Barriers

Synthea's synthetic data is free from all of the issues associated with real health data. There are no privacy or legal concerns because Synthea creates synthetic data. Because it simulates a synthetic person's life from birth to death, Synthea provides complete health records with no interoperability issues. As a realistic, yet synthetic, alternative to real health data, Synthea is able to overcome the barriers to real data, allowing for the analysis complete health records. The data is highly useful to analyzing

treatments, medications, and, of course, survival rates. However, Synthea does not currently feature one critical component of the real-world healthcare system: health insurance and the loss-of-care impacts associated with it.

Health Insurance and Loss-of-Care in Synthea Prior to the MQP

In the iteration of Synthea prior to this MQP, every person receives all of the healthcare that they need for free. There are not realistic health insurance or loss-of-care modules programmed in Synthea. Of course, this is unrealistic because a person's health insurance and income has a significant impact on a person's access to healthcare [26].

In the real world, a person must be able to: afford private insurance, have employer-sponsored insurance, or qualify for Medicare/Medicaid to have insurance. If a person does not meet any of these thresholds, then they will not receive health insurance. In this case, all medical bills will be out-of-pocket, and, in many cases, the person will not be able to afford or receive the treatment and care that they need [26]. When a person does not receive the requisite healthcare, then it is more likely that they will continue to suffer from any conditions and, eventually die [26]. One goal of this MQP, in order to analyze Synthea's health insurance and survival rate outputs, is to implement a realistic health insurance system and loss-of-care impacts on survival rates.

2.5 Breast Cancer

This MQP focuses on breast cancer patients as the scope of the population we analyzed and tested. Further research would be to expand our implementation of health

insurance impacts to all diseases in Synthea. In order to do research with a pilot program and have a specific dataset to work with, a single disease, breast cancer, was chosen.

We chose breast cancer because it has a high incidence rate at 12% of all females in their lifetimes [27], is an important social and policy topic, and there is survival rate and insurance information available to analyze Synthea's outputs as accurate.

An important aspect of our results is understanding the staging and characteristics of breast cancer. Breast cancer diagnosis includes two staging techniques: TNM staging and Location-based staging. Synthea outputs its staging diagnoses as TNM staging, however the most reliable data sources online presented its information in the location-based staging method.

- The TNM Staging method is defined as the following [28]:
    - Stage I
    - Stage II
    - Stage III
    - Stage IV
- The Location-based staging is based on where in the body the cancer has metastasized :
    - Localized
    - Regional
    - Distant

Health insurance impacts survival rates, and specifically breast cancer survival rates. For instance, one study showed that uninsured breast cancer patients were 60%

more likely not to survive the disease than insured ones. We used these data sources to tune Synthea's health insurance, breast cancer, and loss-of-care impacts as reflective of real life.

2.6 Data Sources

To verify and tune Synthea's breast cancer, health insurance, and loss-of-care impact outputs, we needed the real-world data. "The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is an authoritative source of information on cancer incidence and survival in the United States" [29]. SEERS datasets were the primary source of comparing real-world data to Synthea outputs for tuning and seeing how well it reflected the real world. The Center for Disease Control (CDC) even references SEER data as one of the primary sources of cancer statistics [30]. Our other two data sources used were studies conducted on SEER data, so indirectly all of our sources depended on SEER data. The three datasets and their reliability and underlying datasets are described as follows.

The real data that was gathered for verifying and comparing Synthea's outputs was received from:

- SEERS, the National Cancer Institute.
    - The first Source of Data is from SEERS which provided information on Location-based staging distributions and the survival rate by stage-at-diagnosis. It also detailed the overall survival rate of breast cancer patients.

- The Study: *Breast Cancer Stage Variation and Survival in association with insurance status and socioeconomic factors in US women aged 18-64 years Old*
    - This study was based off of the *Surveillance, Epidemiology, and End Results 18 Registries Database* (SEER 18). The information used from this study was the distribution of insurance statuses among breast cancer patients. It also provided information on the increased likelihood of death if a patient does not have insurance. Insurance data became available in SEERS in 2007, which the study utilized and aggregated.
    - The study found that uninsured breast cancer patients were 60% more likely to die than insured ones with a survival rate of 80.4% for uninsured patients. With a sample of this size, and 97,055 patients diagnosed with breast cancer in 2007 and 2008, it is an encompassing dataset. Because it is based on an analysis of SEERs data, which is highly regarded, it can be considered reliable.
- Breastcancer.org, an organization that aggregates SEERS data.
    - As a source that aggregates data in an easy to understand format, breastcancer.org used SEERs datasets in order to generate information about the survival rate by stage-at-diagnosis.

Each of the used datasets aggregates information about the stage-at-diagnosis and survival rates of breast cancer patients. In these aggregate datasets, staging distributions and survival rates do not change over time and are instead based solely on the initial stage of diagnosis, just as Synthea is. Each dataset has a similar, but slightly differing underlying population on which it is based. We also verified that each dataset

is reliable and realistic because the data that our hypothesis was tested on from Synthea needed to accurately reflect the real world. Table 1 features a breakdown of each source's underlying population.

Table 1: Underlying Population of Each Used Dataset

Metric	SEER Dataset	Breast Cancer Insurance Study	Breastcancer.org Information
Patient Type	Breast Cancer Only	Breast Cancer Only	Breast Cancer Only
Region	United States	United States	United States
Population Size	516,079	52,048	497,931
Gender	Female	Female	Female
Races	All	All	All
Year Range	2009 - 2015	2007 - 2008	2007 – 2013
Age Range	All	18 - 64	All

As evidenced by this chapter, because health insurance has had such a significant impact on the healthcare industry, it has become a key component of American health policy. For policy makers and researchers to make informed healthcare decisions, they need health data – which is subject to privacy, legal, and interoperability barriers that prevent its access. To overcome these barriers, a person could use Synthea to generate complete and realistic data that is free from the issues of real data. For this MQP, we implemented health insurance and loss-of-care in Synthea so that we can generate populations for analysis on how health insurance impacts survival rates. For now, we focused on breast cancer with the expectation that this research could expand to all other diseases in Synthea.

3 Research Proposal

Health insurance has been a central part of the American health policy and healthcare industry for over 50 years. Policy makers make informed decisions about money and resource allocation for healthcare and health insurance. Resource allocation is based on analyzing the efficacy of current policies using metrics such as survival rates [31], which is positively impacted by health insurance [32]. Recently, policy makers in the United States have been focused on health insurance when crafting health policy [33]. Policy makers need to have access to all of the data available in order to assess the actual survival rates and the reasons for it. Health data becomes especially crucial when it comes to targeting a specific cause of mortality, such as infant mortality, maternal mortality, or breast cancer [34] [35]. Health data provides insight to the causes and history of survival rates [36] but is difficult to access, both for researchers and even policy makers. Health data, that survival rate analysis is based on, is subject to legality and privacy regulations imposed by HIPAA [37]. Another health data barrier is interoperability issues, in which healthcare facilities use incompatible record formats, leading to incomplete health records [38]. Accurate and complete health data is critical for policy makers to analyze the impacts of health policy on metrics such as survival rates [39].

We simulated health insurance and its impact on the survival rates of a population. Due to the difficulties in accessing the health data that survival rate analysis is based on, we used Synthea, a health data simulator created by the MITRE Corporation. Synthea generates realistic, complete, yet synthetic health data that is free

from the concerns of real health data. However, Synthea does not currently feature or model realistic health insurance or loss-of-care survival impacts. In this MQP, we added health insurance and loss-of-care to Synthea to generate data about its impact on survival. To limit the scope of our population, we chose breast cancer as our focus. We chose breast cancer because of its high prevalence, frequent treatment needs, and the information available online. Through these resources, we were able to obtain the data necessary to make our conclusions. Overall, our research question is: **How does Health Insurance coverage impact the Survival Rates of Patients with Breast Cancer?**

We answered this question by:

- Implementing health insurance and loss-of-care impacts in Synthea. (Figure 2)
- Testing that Synthea's breast cancer, insurance coverage, and loss-of-care impact outputs are reflective of real-world SEERS data. This is Hypothesis 1, serving the purpose of establishing the validity of Synthea's data for inference. (Figure 2)
- Comparing Synthea's survival rate outputs for patients with insurance and patients without insurance. This is Hypothesis 2 and established the significance of health insurance on the survival rates of breast cancer patients. (Figure 3)

3.1 Hypothesis 1: Determining the closeness of Synthea's breast cancer and health insurance outputs with real-world data

To make any conclusions about Synthea's outputs, we must know that closely reflects real-world data. We ascertained closeness by comparing the synthetic data with

real world data. Based on our comparisons, we determined whether or not Synthea's outputs align or if they have a significant difference. Without confirming that Synthea's health insurance outputs reflect the real world, we would not have been able to make any reliable inferences.

We compared Synthea's outputs with real world data metrics. These metrics can be separated into three main categories:

1. Correlate Synthea's breast cancer outputs with real-world data.
    a. Correlate rates of stage-at-diagnosis.
    b. Correlate survival rates by-stage.
    c. Correlate the overall survival rate.
2. Correlate the overall survival rates of a population with no insurance in Synthea with real-world data.
3. Correlate the distribution of health insurance types in Synthea with real- world data.

We used the available real data and studies based on real data in our analysis of Synthea's outputs. We compared the outputs of the distribution of breast cancer stages, The survival rate by-stage, the overall survival rates, the impact of loss-of-care on a person's prognosis, and if their insurance status reflects the real world.

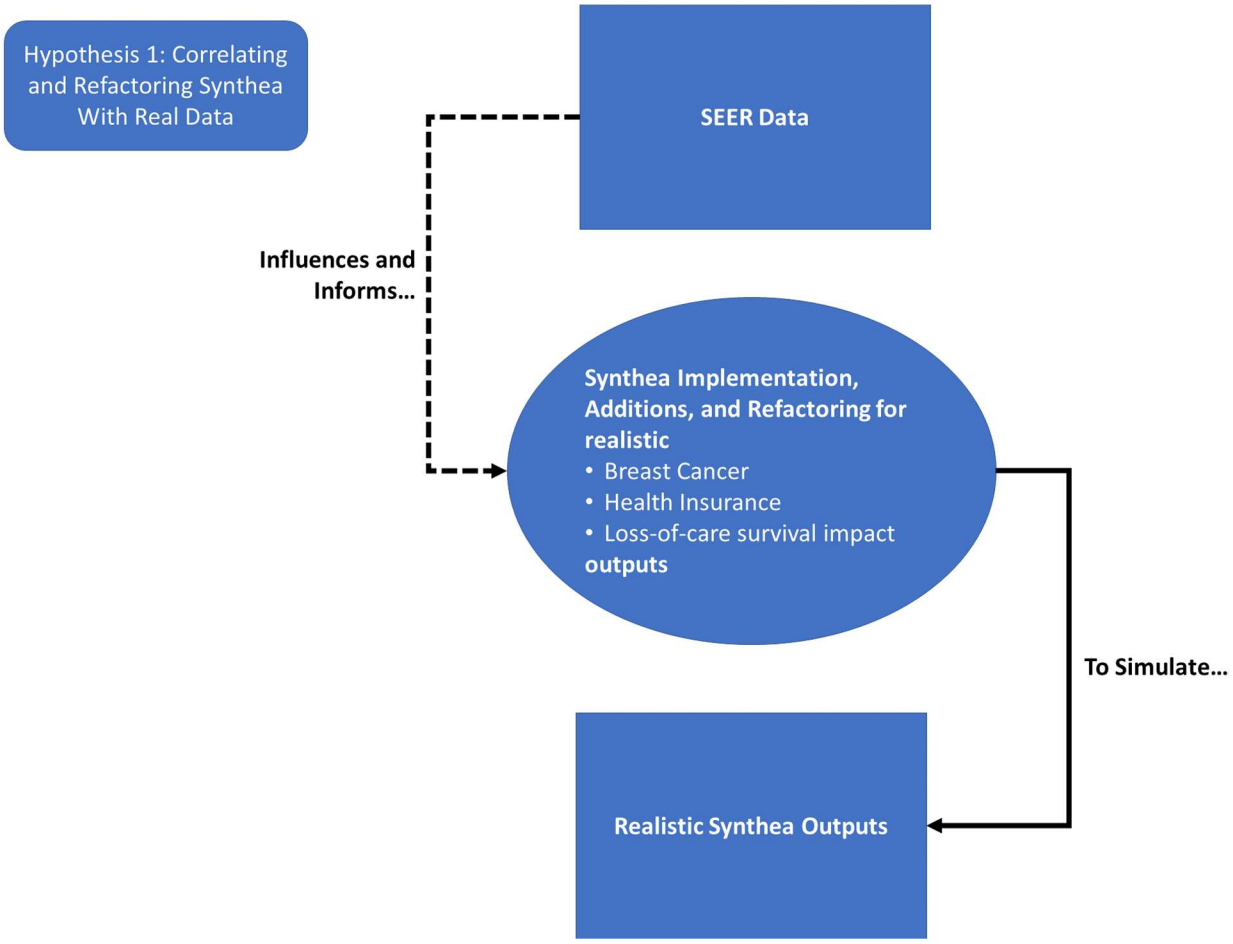

Figure 2: Implementation and Hypothesis 1 Breakdown

3.2 Hypothesis 2: Survival rates for breast cancer patients remain unchanged with or without insurance

We tested the hypothesis that survival rates of patients will not differ depending on their health insurance status. The survival rate was determined by dividing the number of breast cancer survivors of a Synthea population output by the number of

breast cancer deaths. The statistical formatting we used to analyze this hypothesis is as follows:

- $H_0 : \mu_I <= \mu_U$
- $H_1 : \mu_I > \mu_U$

Where μ_I is the average survival rate of insured breast cancer patients and μ_U is the average survival rate of uninsured breast cancer Patients.

The data was gathered through two simulations of Synthea in which, using real demographic data, patients were generated with individual incomes for each year. Based on this income, a person may be able to afford private insurance or qualify for Medicaid. In the event that a person falls in the range between the cost of private insurance and Medicaid qualification, they will have no insurance and will thus not receive healthcare. Throughout a given year, a person incurs costs, including healthcare costs (if they have no insurance), monthly premiums, and copays until, at one point, they will not be able to afford health insurance or healthcare. After not receiving requisite healthcare, such as chemotherapy or a lumpectomy, they will have a higher likelihood of death based on mortality rates for not receiving certain care. These survival rates were collected from SEERS, the National Cancer Institute and Cancer.net.

For this hypothesis, we compared the survival rates of those patients who had health insurance against the survival rates of patients who had no insurance. We used the survival rates of each population as our test statistic. To test this hypothesis, we generated two sample populations of size 1000 in Synthea. One population received universal, free health insurance and the other had no insurance available. Based on a two-sample pooled proportion p-test, we determined if there is a significant difference

between the two populations' survival rates, allowing us to potentially reject Hypothesis 1 and conclude that health insurance does have a positive effect on the survival rates of breast cancer patients in Synthea.

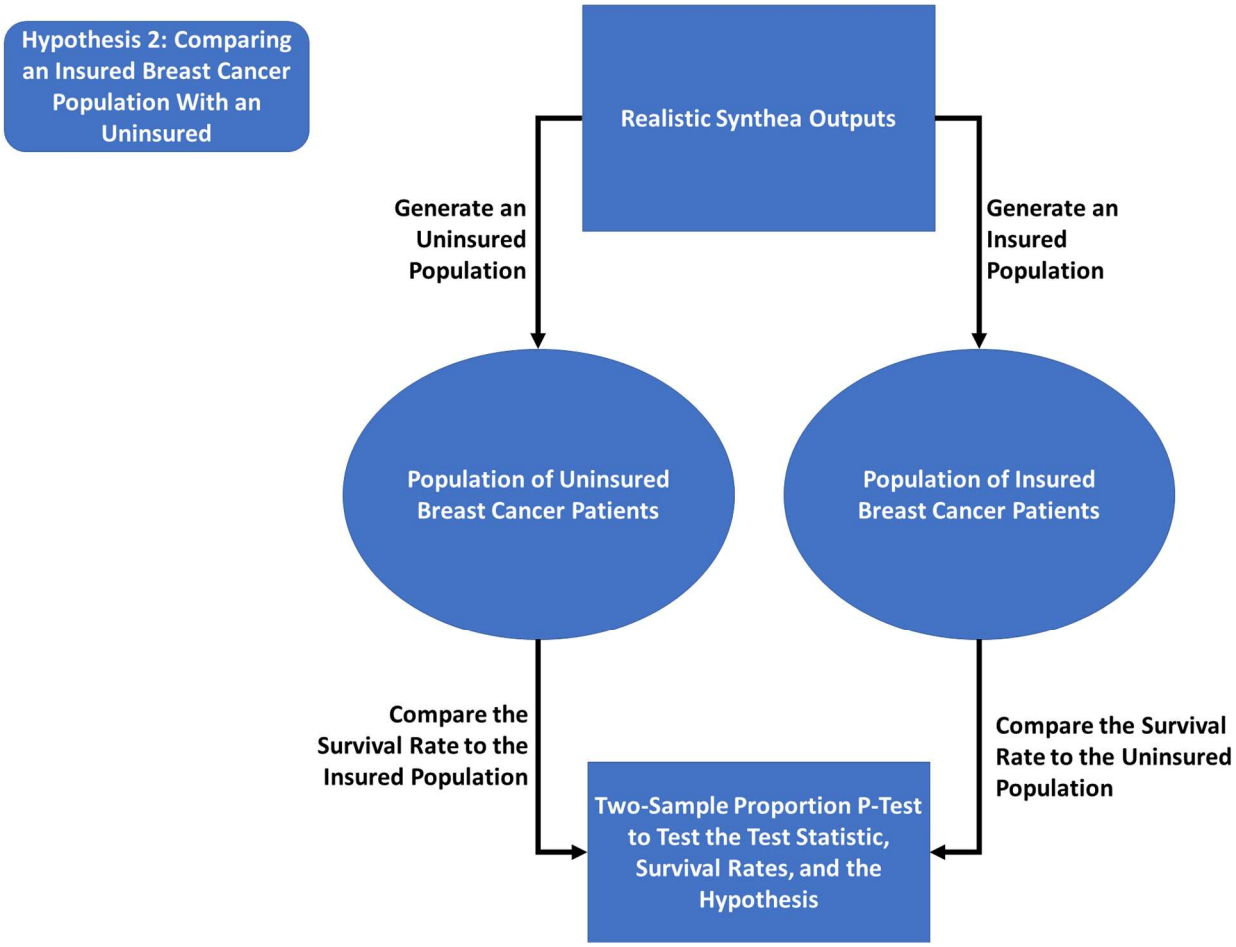

Figure 3: Hypothesis 2 Survival Rate Testing Breakdown

3.3 Breast Cancer Focus

In this MQP, we focused on breast cancer patients as our scope of population. Limiting our population to breast cancer patients allow us to focus on a well-documented disease [40]. Breast cancer is a significant disease in the modern world with incidence rates of about 12% of all women throughout their lifetimes [41]. The data

that provides information on breast cancer's staging, survival rates, and insurance statistic are in the range from 2007 – 2015. For these reasons, we used breast cancer patients from 2007 - 2015 as the population scope for this MQP.

The next chapter will describe our methodology and step-by-step process in implementing health insurance and loss-of-care survival impacts in Synthea.

4 Methodology

To address our research question, we needed to add a key missing component to Synthea: Health insurance. As it was prior to the MQP, Synthea held a very basic health insurance system that was based on enumerations of either Medicare, Medicaid, private insurance, or no insurance. It did not take into consideration any costs, copays, or deductibles of healthcare. Through this MQP, we implemented a more robust health insurance system, which includes real companies and allows for copays, monthly premiums, and other real-world insurance costs. Simply adding health insurance to the current implementation of Synthea would not be enough: because patients would still receive all the healthcare they need no matter their insurance status. To account for uncovered healthcare, we also implemented a loss-of-care feature that tracks the care that a patient should have received and how it impacted the patient's prognosis. Finally, we tuned Synthea so that its outputs would reflect real-world data, until we could assess it as accurate. The data we used to tune Synthea includes statistics about the insurance distributions of breast cancer patients as well as the survival rates of patients without insurance. Overall, our methodology in this MQP consisted of the following steps:

1. Implement health insurance in Synthea.
2. Implement loss-of-care survival impacts in Synthea.
3. Tune Synthea's health insurance and loss-of-care modules to reflect the real world.

4.1 Implementing Health Insurance in Synthea

To prepare Synthea to produce the requisite data, we removed the insurance enumerations and redefined the way that a person interacts with and receives health insurance. Throughout a person's simulated life in Synthea, they now enroll yearly in health insurance based on factors such as qualifying for Medicare or Medicaid and the affordability of private insurance. A person qualifies for Medicare if they are over the age of 65 and qualify for Medicaid if they are pregnant, have end-stage-renal-disease, or their income is less than the federal poverty level multiplied by 1.33; If the person qualifies for either of these programs, they automatically receive it. Otherwise, they will randomly choose any available private insurance that they can afford. Affordability is based on the cost of the payer's monthly premiums multiplied by 12 added to the deductible. If this value is greater than the person's income, which is generated based on real socioeconomic and demographic data, then they cannot afford the payer. We are making the assumption that every person is willing to spend the entirety of the yearly income on healthcare. We also assume that all private payers unconditionally accept any customer that can afford them. In the event that the person cannot afford any of the available insurance, then they receive no insurance. Figure 4 displays a sequence diagram describing the different ways that a person receives certain insurance in Synthea.

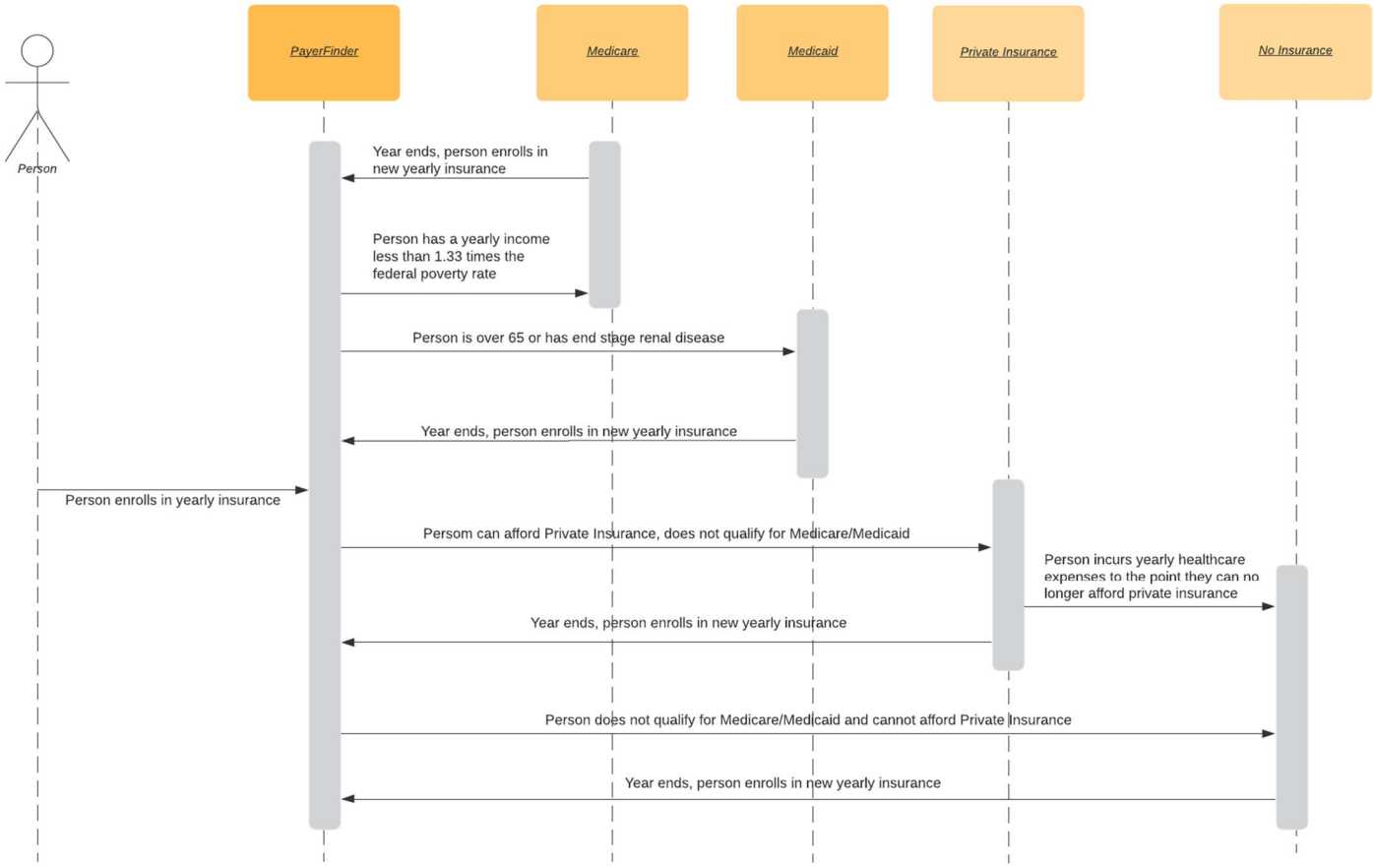

Figure 4: Sequence Diagram of Receiving Different Insurance

Choosing a person's insurance for the year is based on the HealthInsuranceModule class which features two methods: process() and determineInsurance(). The flow of time in Synthea is based on a time-step, which is a pre-defined period of time in which Synthea periodically executes. For every time step of Synthea, the HealthInsuranceModule uses process() for each person. Process() checks to see if the person needs to enroll in a new year's insurance and, if so, the person will enroll. In this case, the strategy design pattern was used in order to call a PayerFinder, which allows for a predefined insurance selection algorithm. In this iteration, we implemented an algorithm in which the person chooses random, affordable, insurance.

We implemented a new Payer class that sets, tracks, and processes all payer-related activities. The Payer class generates all payers that will exist in the simulation based on an input file. This input file sets the important attributes of each payer, including their name, the states they operate in, their deductible, their copay amount, their monthly premium, and whether they are private or government-owned. An additional Payer is generated from outside of the input file which acts as a null placeholder for no insurance and has all values set to zero. All statistics for each payer are tracked throughout the simulation, including the amount of expenses they covered, the amount of expenses they did not cover, the number of covered and uncovered encounters/medications/procedures/immunizations, the number of unique customers, the QOLS average of all of their customers by year, and the number of member months covered.

Whenever an encounter occurs, a Claim is created in which costs are calculated and assigned to the participating parties. When the encounter ends, the total cost of the encounter, including any associated procedures, medications, and immunizations is determined based on realistic costs. The person is then assigned the cost of the copay, or the full cost of the encounter if their payer does not cover it. The expenses that the person must pay are added to their annual incurred health costs. This metric is useful when determining if a person can still afford healthcare throughout the year. The payer is then assigned the rest of the cost of the claim that the person did not cover.

Throughout the year, a person will incur healthcare expenses including copays, full encounter costs, and monthly premiums. As expenses are incurred, they are added

to the person's yearly healthcare expenses which may, at one point, exceed their income. In the event that this occurs, they will immediately switch to no insurance.

The relevant features of the implementation of the payer class can be seen in the Class Diagram displayed in Figure 5.

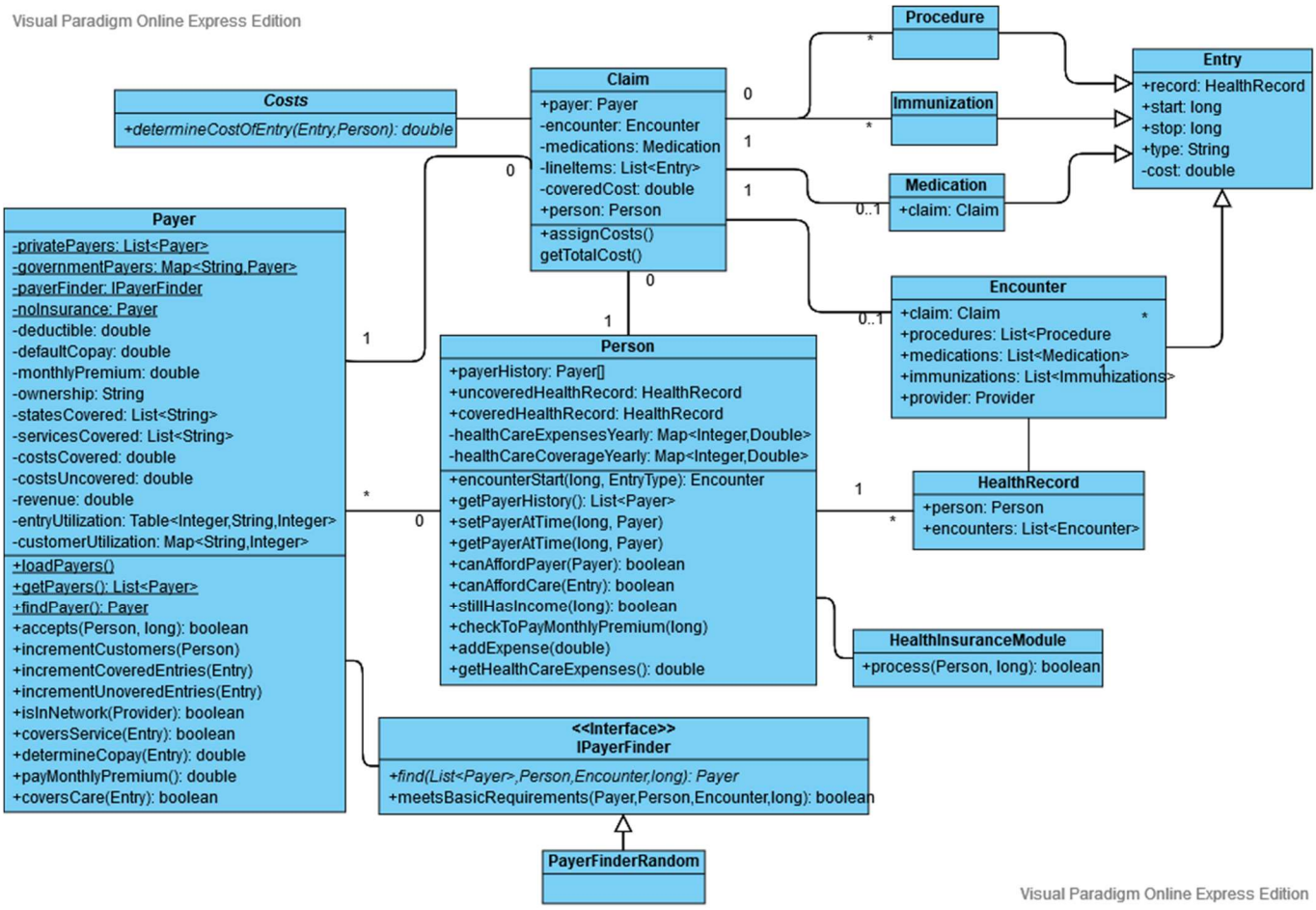

Figure 5: Synthea Class Diagram of Relevant New Insurance Features Added

Figure 5 shows the relevant classes, functions, and relationships that were added to Synthea. The following Class Diagram displayed in Figure 6 shows its structure prior to implementation.

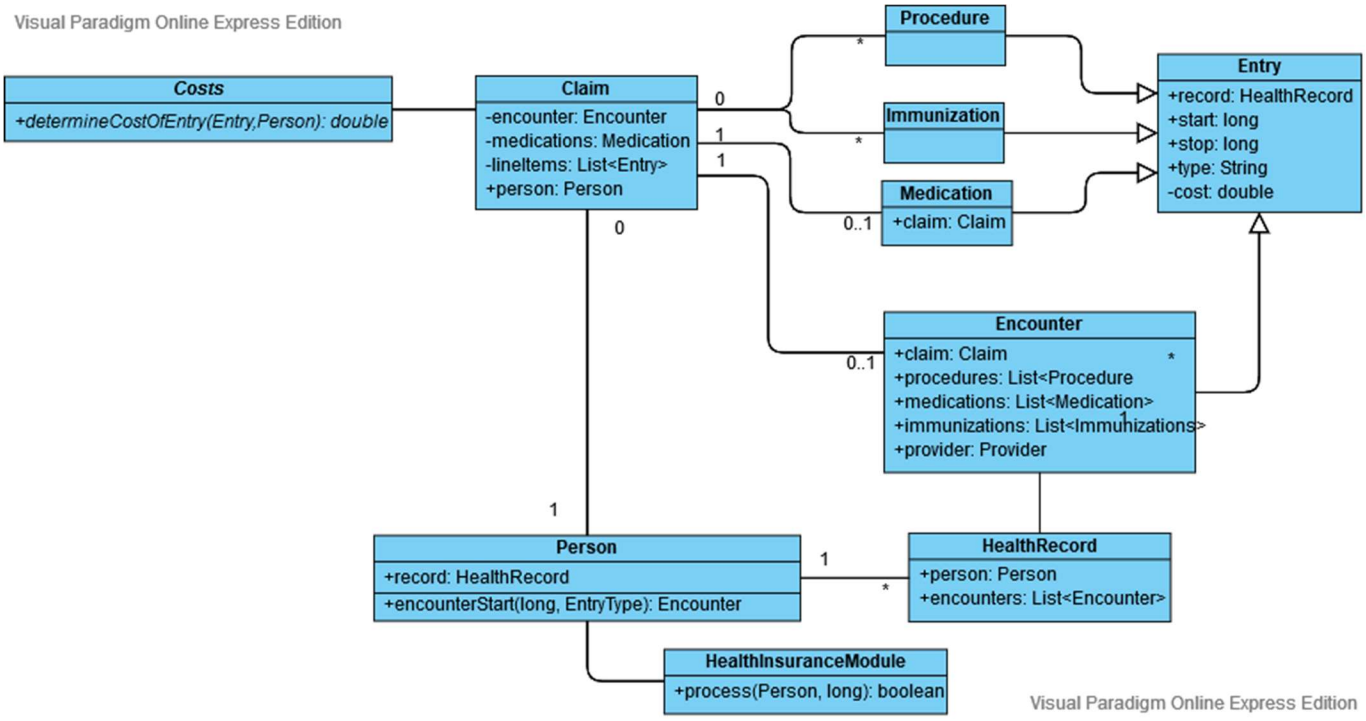

Figure 6: Synthea Class Diagram of Relevant Components Prior to New Features

To analyze health insurance in each simulation, we added an output file, payer_transitions.csv, to Synthea. Payer_transitions.csv details the health insurance held at every year of every patient's life. With a clear-cut described health insurance status for every given year, we gained the ability to determine exactly how many years each type of insurance was utilized for, giving us an overall percentage of the distributions of insurance types. Payer_transitions.csv would become useful for associating and tuning Synthea's outputs and analyzing the MQP's results.

To view the entire implementation, including code and a detailed explanation, visit https://github.com/synthetichealth/synthea/pull/527.

At this stage of Synthea's implementation, having no insurance and not receiving healthcare does not actually affect a person's survival rates. This was the next step to implement.

4.2 Implementing Loss-Of-Care and its Impact on Survival

In order for a person's insurance to have a real impact on their simulated lives, healthcare that is not covered must effect their survival probability. In order to implement this, a new health record for each person was created that acts as a tracker of what care the person should have received but did not. This health record is known as the *Uncovered Health Record*. In this case, if someone needed chemotherapy, but they did not have insurance and could not afford the procedure, then that treatment would go into the *Uncovered Health Record*.

Throughout a person's simulated life, Synthea will check their *Uncovered Health Record* for treatments that the person should have, but did not, receive. Based on real prognosis rates, each uncovered treatment will result in a higher likelihood of a person's death. Through research, we found that, according to the National Cancer Institute, based on SEERs information, uninsured breast cancer patients have a 60% higher likelihood of death than insured patients [13]. According to his dataset, the overall uninsured population survival rate should be around 80.4% [13]. We started by instituting a baseline 60% likelihood of death if a treatment was in the *Uncovered*

Health Record, but found that this resulted in far too low survival rates of around 56%. We then hit an issue where a person would constantly have a 60% likelihood of death for every Synthea timestep. So we needed to fix the issue by having Synthea calculate an increased likelihood of death only once for each missed treatment. We also needed to fine tune the percentage of death so that it would reach approximately 80.4%. With multiple simulations, each tweaking the likelihood of death, we eventually settled on probabilities of death that resulted near target 80.4% survival rate. The overall survival rate that Synthea's uninsured breast cancer patients experience after tuning was 81.0%.

With loss-of-care implemented, we needed an understandable data format that Synthea could output to describe its impacts in the simulation. We added a deathStatistics.csv output file that describes the following attributes of each patient:

- Their birthdate
- The date that they got breast cancer
- The Stage-at-diagnosis of their cancer
- Whether they survived the cancer
- Their date of death
- Their cause of death
    - Could only be: Natural Causes, Uncovered Treatment, or Inevitable Breast Cancer Death
- Whether they died due to uncovered treatment (loss-of-care)
- How many years after getting breast cancer they survived for

The deathStatistics.csv output file allowed us to analyze each of these attributes, gaining an understanding of how many people died due to loss-of-care impacts, how long people survived for, and what stages they were diagnosed at. These attributes became useful for tuning and testing its association with real data and analyzing the results of the MQP. Now that we implemented loss-of-care so that it affects a person's likelihood of death, we needed to implement real-world health insurance levels.

4.3 Implementing Real-World Health Insurance Levels

With payers and loss-of-care implemented and instituted in Synthea, the next step was ensuring that a population displays a distribution of health insurance statuses that is reflective of the real world. The implementation started with finding the real-world distributions of health insurance among breast cancer patients. The breast cancer health insurance study provided the following real-world data in Table 2 based on an aggregate of the SEERS 18 dataset.

Table 2: Breakdown of Insurance Distributions of Real-World Breast Cancer Patients

Real Data	
Insurance	**Percentage breakdown of each insurance status**
Private Insurance	86.00%
Medicaid	11.60%
No Insurance	2.40%

For Synthea to reflect these percentage breakdowns, we needed to fine-tune the cost of its private insurance, which acts as the only barrier for patients to receive it.

Based on patients' individual, and realistic, incomes, the may or may not be able to afford a payer, as previously described. We started with a baseline cost of a $1000 deductible, $55 copay, and $1000 monthly premium, but found that this resulted in far too high of a private insurance incidence rate. However, we did find that, regardless of any costs we input, the percentage of patients with Medicaid never changed and was highly consistent with the real-world Medicaid rate. This makes senses because the qualification for Medicaid is predefined at 1.33 multiplied by the federal poverty rate, $11000, and, since the patients have realistic income distributions, the amount qualifying for Medicaid naturally follows. As we gradually tweaked and adjusted insurance costs in order to tune Synthea to produce outputs reflective of the real world, we eventually settled on a $1000 deductible, $55 copay, and $1400 monthly premium. The breakdown of Synthea's insurance outputs and its comparison to real data is described in Table 3 and Table 4.

Table 3: Synthea's Tuned Insurance Incidence Distribution

Synthea Output	
Insurance	**Number of Years of Insurance Use**
Private Insurance	72850
Medicaid	9758
No Insurance	2017
Total Years	84625

Table 4: Comparison of Synthea's Insurance Distributions with Real-World Data

	Real Data	Synthea Data
Insurance	**Percentage breakdown of each insurance status**	**Percentage breakdown of each insurance status**
Private Insurance	86.00%	86.09%
Medicaid	11.60%	11.53%
No Insurance	2.40%	2.38%

With health insurance, loss-of-care, and data tuning implemented in Synthea, we were able to associate Synthea's outputs with real data and make conclusions on it so as to answer our research question. To answer the question, *How Does Health Insurance Impact Survival Rates of Breast Cancer Patients*? we needed to closely associate Synthea's outputs to be reflective of real-world data. We achieved this goal by obtaining the real-world data necessary to compare with and tune Synthea's outputs. Once Synthea was verified as reflective of the real world, we made conclusions about how different health insurance levels and policies affected the overall survival rates of breast cancer patients. To achieve this, we generated two different health insurance status populations from Synthea: a population with full insurance coverage and one with no insurance. We then compared the two samples with a statistical significance test, using two-sample proportion p tests. Next, we generated a Synthea population with realsitically tuned insurance distributions (private insurance, government insurance, no insurance) and compared them to the real data and Synthea's outputs.

5 Results

We used Synthea to generate two populations of 1000 patients. One population had full insurance and one population had no insurance. We started by comparing Synthea's breast cancer and insurance output data with real world data to assess how closely it reflected real life. When we were confident in its similarities, we ran Synthea to generate datasets detailing each patient's survival and health insurance status. We aggregated these outputs into percentages for comparison and statistical analysis in our hypotheses.

5.1 Result 1: Verifying Synthea's Outputs

To ensure that our hypotheses are valid, Synthea's outputs must be established to closely reflect real-world data. In comparing Synthea's outputs, we needed to analyze the following metrics:

1. Synthea's breast cancer outputs regarding cancer stage diagnoses and survival rates.
    a. Compare the percentage of incidence of stages in Synthea and real data.
    b. Compare the survival rates by-stage in Synthea with real data.
    c. Compare the overall survival rates of breast cancer patients in Synthea with real data.
2. Compare Synthea's overall survival rate of a population without insurance with real data.

3. Compare Synthea's levels of health insurance with real data.

Real-World Data Used

Each of the datasets used in verifying Synthea's outputs aggregates information about the cancer stage-at-diagnosis and survival rates of breast cancer patients. In these aggregate datasets, survival rates per-stage do not change over time and are instead based solely on the initial stage of diagnosis, just as Synthea is. Each dataset has a similar, but slightly differing underlying population on which it is based, the differences of which are described later in this section. We also needed to verify that each dataset is reliable and realistic because the data that our hypothesis were tested on from Synthea must accurately reflect the real world. What follows is a breakdown of each source's underlying population and its reliability as a valid data source.

Dataset: SEERS

The first Source of Data is from SEERS which provided information on the percentage of stages-at-diagnosis for Location-based the survival rate by stage-at-diagnosis. It also detailed the overall survival rate of breast cancer patients. It is based on the following underlying population in Table 5.

Table 5: SEERS Dataset Underlying Population

Patient Type	Breast Cancer Only
Region	Throughout United States
# Of Patients in Population	516,079
Gender	Female
Races	All
Years Range	2009 – 2015
Age Range	All

"The Surveillance, Epidemiology, and End Results (SEER) Programs provides information on cancer statistics in an effort to reduce the cancer burden among the US population" [42]. It is supported by both the Surveillance Research Program and the National Cancer Institute and is one of the CDC's official federal cancer registries [43].

Dataset: *Breast Cancer Stage Variation and Survival in association with insurance status and socioeconomic factors in US women aged 18-64 years Old* study

This study was based off of the *Surveillance, Epidemiology, and End Results 18 Registries Database* (SEER 18). The information used from this study was the distribution of insurance statuses among breast cancer patients. It also provided information on the increased likelihood of death if a patient does not have insurance. Insurance data became available in SEERS in 2007, which the study utilized and aggregated.

The study found that uninsured breast cancer patients were 60% more likely to die than insured ones with a survival rate of 80.4% for uninsured patients. With a sample of this size, and 97,055 patients diagnosed with breast cancer in 2007 and 2008, it is an encompassing dataset. Because it is based on an analysis of SEERs data, which is highly regarded, it can be considered reliable. It is based on the following underlying population in Table 6.

Table 6: Insurance Status Dataset Underlying Population

Patient Type	Breast Cancer Only
Region	United States
# Of Patients in Population	52,048
Gender	Female
Years Range	1/1/2007 – 12/31/2008
Age Range	18 - 64

Dataset: breastcancer.org

As a source that aggregates data in an easy to understand format, breastcancer.org used SEERs datasets in to generate information about the survival rate by stage-at-diagnosis. It is based on the following underlying population in Table 7.

Table 7: Breastcancer.org Dataset Underlying Population

Patient Type	Breast Cancer Only
Region	United States
# Of Patients in Population	497,931
Gender	Female
Range of Time	2007 - 2013
Age Range	All Ages

Synthea Output Data

In obtaining data from Synthea, we used the following run command for every dataset simulated:

- `./run_synthea -p 1000 -g F -m breast_cancer`

This command generates a population of size 1000 patients, all of which are female, with breast cancer as the only disease in the simulation. Its underlying

population, which is mostly similar to the populations of the sources used, is described in Table 8.

Table 8: Synthea's Underlying Population

Patient Type	Breast Cancer Only
Breast Cancer Patient Region	United States
# Of Patients in Population	1,000
Gender	Female
Range of Time	2007 – 2015, reflects the time ranges of utilized datasets.
Age Range	All Ages

Compare Synthea's Breast Cancer Outputs

To compare Synthea's breast cancer outputs, we used the SEERS dataset as our real-world source. Because a full statistical analysis to validate all of the components of Synthea's realism would constitute a full project in itself, we instead compared our outputs to show association. To comprehensively associate it, we analyzed the following three components of Synthea's breast cancer outputs with the SEERS data.

1. Show a close association between Synthea's breast cancer stage-at-diagnosis incidence with real data.
2. Compare the survival rate by-stage of diagnosis with real data.
3. Compare the overall survival rate with real data.

1. Show a Close Association Between Synthea's Breast Cancer Staging Incidences with Real Data.

Synthea's breast cancer module assigns patients staging based on the TNM staging system in which stages are split into subcategories. For instance, Stage I has subcategories of IA and IB [44]. However, the data we obtained about the incidence and rates of staging in the real population is based on different categorizations. These categories include macro-stages and location-based metrics [45]. Macro-staging is simply converting substages to its overall stage, such as Stage IA and IB being subsets of Stage I. Location-based staging describes where in the body the breast cancer has metastasized and includes the stages localized, regional, and distant. Because Synthea's breast cancer staging system was not directly compatible with the staging system of the SEERS data, we had to convert Synthea's output to match it. Based on the Susan G. Komen Organization's comparisons between the different staging categorizations, we were able to make the following conversions from Synthea's TNM staging to the staging of the real data in Table 9 [45].

Table 9: Conversion from Synthea Staging to Location and Macro Staging

Synthea's Staging Output	Location Staging	Macro Staging
Stage IA	Localized	Stage I
Stage IB		
Stage IIA		Stage II
Stage IIB	Regional	
Stage IIIA		Stage III
Stage IIIB		
Stage IIIC		
Stage IV	Distant	Stage IV

Based on the data output by Synthea, we were able to convert Synthea's staging to Location-Based and macro staging. In a 1000 population simulation, in which patients received all treatment they needed, we generated and aggregated the following staging data in Table 10.

Table 10: Synthea's Output of Staging Incidence and Survival Rates of Fully Insured Breast Cancer Patients

Synthea Output Data											
Synthea's Staging Output				Conversion to Macro Staging				Conversion to Location Staging			
Stage	# of Cases	# Survivors	# Deaths	Macro Stage	# of Cases	# Survivors	# Deaths	Location Stage	# of Cases	# Survivors	# Deaths
Stage IA	467	463	4	Stage I	479	475	4	Localized	666	645	21
Stage IB	12	12	0								
Stage IIA	187	170	17	Stage II	303	282	21				
Stage IIB	116	112	4					Regional	282	230	52
Stage IIIA	141	97	44	Stage III	166	118	48				
Stage IIIB	0	0	0								
Stage IIIC	25	21	4								
Stage IV	52	16	36	Stage IV	52	16	36	Distant	52	16	36
Overall	1000	891	109		1000	891	109		1000	891	109

Table 10 displays the incidence of each of the stages as defined by Synthea as well as their conversions to Location and Macro staging. It also includes the number of survivors and death of each stage-at-diagnosis. With the Synthea output data in hand, the next step was to compare the number of cases of individual staging and survival rates with real data.

In comparing the percentage of cases per stage, we used the location staging method because of the consistent data verifying what its incidence should be. With the data found, we were able to create the following chart describing the real-world incidences of stages in Table 11.

Table 11: Percentage of Incidences of Real-World Location Stages

	Real Data	
Location Stage	**Percent Per Stage**	**Survival Rate**
Localized (Stage I - IIA)	62.00%	98.80%
Regional (Stage IIB - IIIC)	30.00%	85.50%
Distant (Stage IV)	6.00%	27.40%
Unknown	2.00%	X
Overall	X	89.90%

Based on these real-world percentages, we needed to compare them to Synthea's percentages-by-stage. We calculated how Synthea's output resulted in the incidence of staging throughout the population by dividing the number of patients with each stage-at-diagnosis by the total number of patients (which was 1000). Through this data analysis, we created the following chart which compares Synthea's data percentages with the real data in Table 12.

Table 12: Comparison of Synthea Output to Real Data of Staging Incidences

Synthea to Real Data Comparison				
	Real Data		Synthea Results	
Location Stage	Percent Per Stage	Survival Rate	Percent Per Stage	Survival Rate
Localized (Stage I - IIA)	62.00%	98.80%	66.60%	96.85%
Regional (Stage IIB - IIIC)	30.00%	85.50%	28.20%	81.56%
Distant (Stage IV)	6.00%	27.40%	5.20%	30.77%
Overall	X	89.90%	X	89.10%

In comparing Synthea's data with the real data, we found that there was a consistent alignment between the distribution of stages-at-diagnosis. The greatest difference comes from the Localized stage where the real data shows that 62% of patients should be diagnosed with it while 66.6% of Synthea patients were. This marks the most extreme difference at 4.6%. The percentage of Regional Staging cases was very consistent with only a difference of 1.8%. Distant staging was the most accurate with a percentage of cases difference of only 0.8%. A chart displaying the comparison of Synthea's staging incidences with the real-world incidence is displayed in Figure 7.

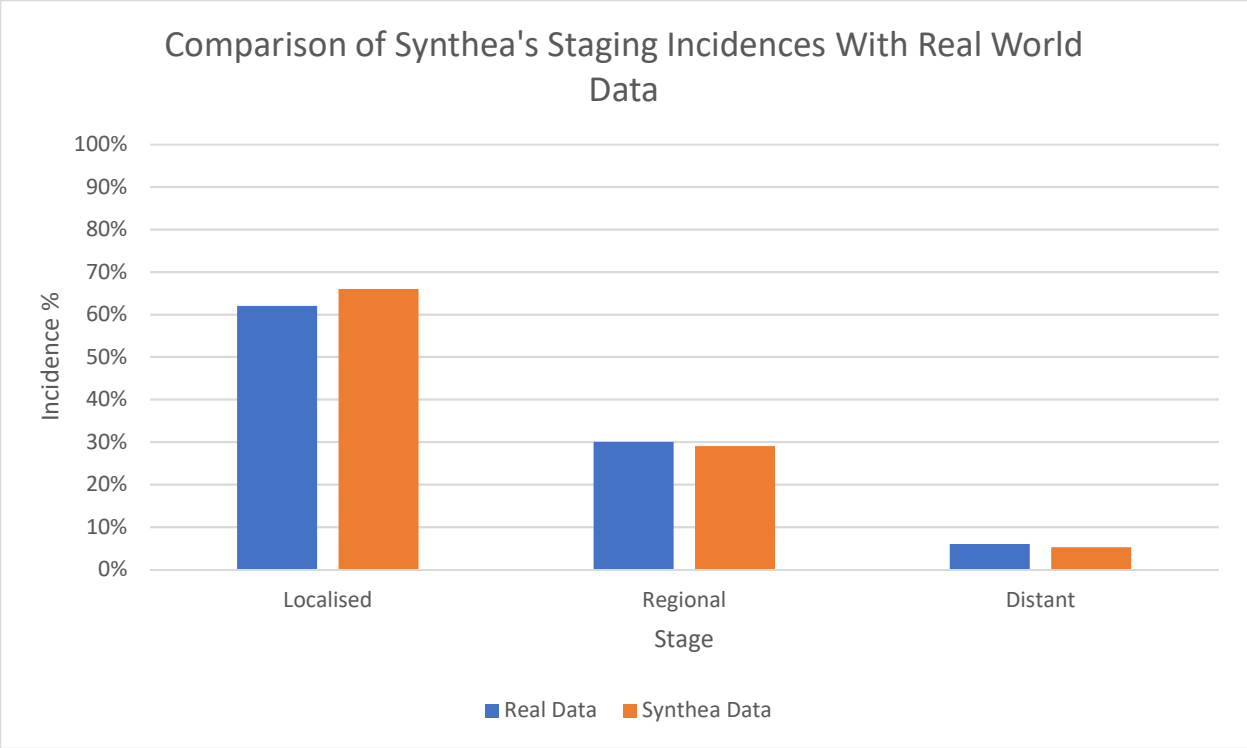

Figure 7: Comparison of Staging Incidences in Synthea and Real Data

2. Comparing Survival Rates By Stage-at-Diagnosis

Using the same Synthea dataset, we were also able to compare Synthea's outputs about survival rates with the information that the SEERS dataset held. Again, by using the data output by Synthea in Table 2, we were able to detail the survival rates of each stage output by Synthea. With the SEERS data we obtained, we were able to compare the percentages of Synthea's survival rates, based on 1000 patients, with the real-world ones. The percentage breakdown data is displayed in Table 13.

Table 13: Comparison of Survival Rates by Stage between Synthea and Real Data

Synthea to Real Data Comparison				
	Real Data		Synthea Results	
	Percent Per Stage	Survival Rate	Percent Per Stage	Survival Rate
Stage I	X	99.00%	47.90%	99.16%
Stage II	X	93.00%	30.30%	93.07%
Stage III	X	72.00%	16.60%	71.08%
Stage IV	X	22.00%	5.20%	30.77%
Localized (Stage I - IIa)	62.00%	98.80%	66.60%	96.85%
Regional (Stage IIb - IIIc)	30.00%	85.50%	28.20%	81.56%
Distant (Stage IV)	6.00%	27.40%	5.20%	30.77%
Overall	X	89.90%	X	89.10%

As you can see, we were able to compare Synthea's data with the real data across two metrics of staging: both Location Staging and Macro-Staging. The survival rate comparisons of Synthea's stages I – III were very consistent the real data, all varying by less than 1%. However, Stage IV's survival rates differed by 8.77% with Synthea outputting a 30.77% survival rate and the real data displaying a 22% survival rate. In comparing the Location staging, Localized was consistent within 1.95% and Regional was consistent within 3.94%. Interestingly, the Distant Stage, which is essentially the same as Stage IV, has a difference of only 3.33% between Synthea and the real data. This exact same metric, defined as Stage IV, has the 8.77% difference mentioned above. It is possible in this case, then, that different sources have different measurements for staging. Most are consistent, however, with the exception of Stage IV and Distant. Overall, Synthea produced very consistent survival rate by stage outputs with the real data.

3. Comparing the Overall Survival Rate

Again using Table 5 and the same 1000 patient simulations, it is clear to see that the overall survival rate output by Synthea is highly consistent with the real data. In the real data, it is shown that the overall survival rate of a population with breast cancer is 89.90%. Synthea's output is just 0.8% different at 89.10%.

We also included a year-by-year breakdown of how Synthea's overall breast cancer survival rates compared with the SEER data, as seen in Figure 8. Synthea's inputs are based on modern data, which is reflected by the fact that Synthea's outputs match up with the SEER data in the most recent time frame. Although there is overlap from 1990-2004, we expect that this is a coincidence because Synthea does not change based on the year and is exclusively based on recent survival rate data. Most likely, the discrepancy is caused by the fact that, at each of the year-ranges, there is a smaller sample size than the 1000 patient overall population, leading to more easily skewed data.

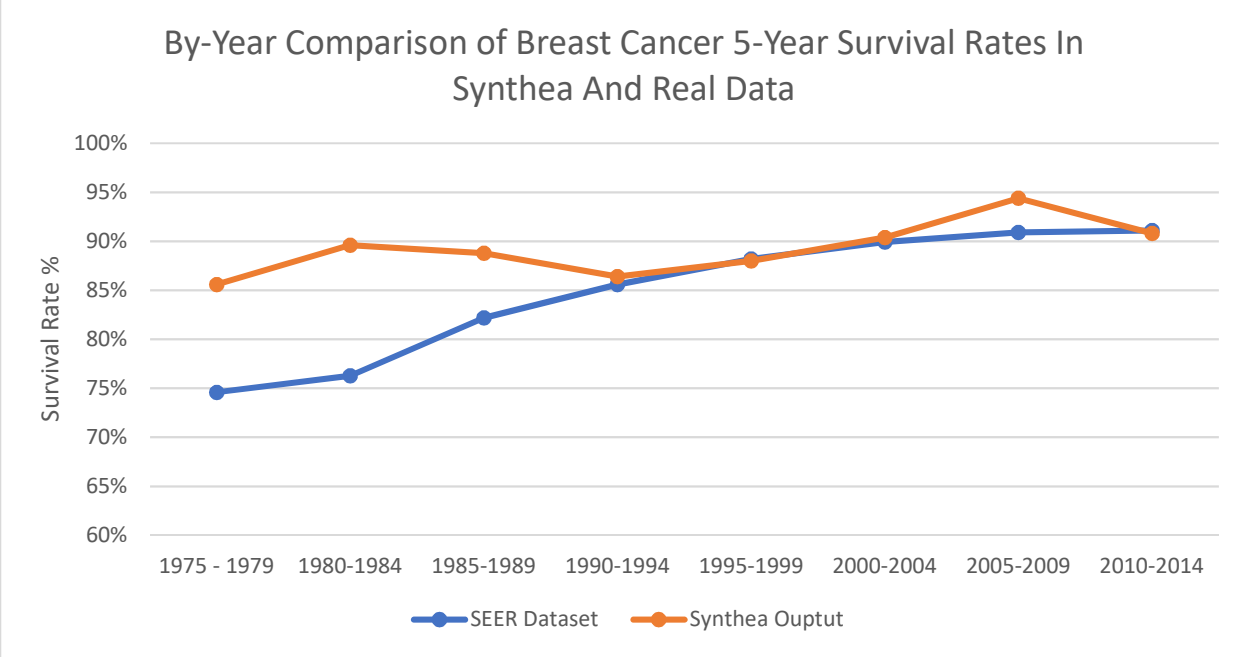

Figure 8: Yearly Survival Rates in Synthea Compared to SEER Data

Compare Synthea's Loss-Of-Care Impacts

To implement the impact that loss-of-care has on a breast cancer patient's prognosis, we first needed to find data showing the survival rate of uninsured breast cancer patients. The study, *Breast Cancer Stage Variation and Survival in association with insurance status and socioeconomic factors in US women aged 18-64 years Old*, provides information on the survival rates of uninsured breast cancer patients, based on an analysis of SEERS data. The study found that uninsured breast cancer patients were 60% more likely to die than insured ones. While the overall survival rate of breast cancer patients is 89.9%, the study found that the survival rate of uninsured patients was 80.4%.

For Synthea to output data consistent with a real-world survival rate for uninsured patients of 80.4%, we needed to implement a feature where loss-of-care increased a person's likelihood of death. As described in the methodology, when a person has untreated treatments, they will have a certain percent chance of death.

With the implementation of this potential death due to loss-of-care feature, we needed to tune the probability of death until it resulted in an overall survival rate of approximately 80.4%. We started by simply giving a patient a 60% chance of death if some treatment went uncovered, however this resulted in far too low survival rates of around 56%. By continually tuning and altering this percent of death number, and testing simulation outputs, we eventually tuned the data to output an overall survival rate of 81% for uninsured patients. Table 14 shows the comparison of the uninsured patients Synthea outputted in a 1000 population simulation with the real data that it should reflect.

Table 14: Comparison of Uninsured Patient Survival Rates Between Synthea and Real Data

Synthea to Real Data Comparison (Uninsured Patients)		
	Real Data Survival Rate	Synthea Data Survival Rate
Stage I	X	93.21%
Stage II	X	85.40%
Stage III	X	62.79%
Stage IV	X	23.44%
Localized (Stage I - IIA)	X	90.94%
Regional (Stage IIB - IIIc)	X	71.96%
Distant (Stage IV)	X	23.44%
Overall	80.40%	81.00%

With a difference of just 0.6% between Synthea's uninsured survival rate and the real data's survival rate, there is clearly a close association between the two.

Compare Synthea's Tuned Health Insurance Levels

The final step in showing that our hypotheses, and the data they are based on, are accurate, was ensuring that the distribution of health insurance reflected real-world levels. To achieve this goal, we needed to find out what pricing for private insurance would result in the correct insurance outputs. We started with a baseline of a $1000 deductible, $55 copay, and $1000 monthly premium. Because people receive private insurance based on their incomes and its affordability, we needed to tune these expenses until Synthea's output reflected the real data.

When tuning the data, we found that we did not have to change anything regarding Medicare's insurance incidence rates. Because Medicare qualification is based on 1.33 * the federal poverty rate, with the poverty rate set at $11000 per year, there was already a set baseline for its cost. In addition, because each person's individual incomes is generated based on real economic probability and data, it was only natural that the Medicare incidence, based on each person's income, would follow reality.

The only data point we had to tune was the cost of private insurance. Through the process, we settled on its deductible costing $1000, its copay set at $55 and a monthly premium of $1400. In evaluating the percentage distributions of how much insurance coverage was available during the simulation, we used the payer_transitions.csv output file which lists every person's insurance company for every year. We simply summed up

the total number of years of coverage for each insurance and used it as a divisor for the number of years that each of the three insurance categories had (Private/Medicaid/No Insurance).

We first started by gathering this data in Table 15, which displays each of these categories' number of years of use.

Table 15: Number of Years Each Insurance Category Was Utilized in Synthea

Synthea Output	
Insurance	**Number of Years of Insurance Use**
Private Insurance	72850
Medicaid	9758
No Insurance	2017
Total Years Lived	84625

The next step was converting this data to percentages and comparing it to the real data in Table 16.

Table 16: Comparison of Percent of Utilization of Insurance Categories Between Synthea and Real Data

	Real Data	Synthea Data
Insurance	**Percentage breakdown of each insurance status**	**Percentage breakdown of each insurance status**
Private Insurance	86.00%	86.09%
Medicaid	11.60%	11.53%
No Insurance	2.40%	2.38%

As you can see, Synthea's outputs are all within 0.1% of the real data for each insurance category.

5.2 Result 2: The Survival Rate of Uninsured Patients is Lower than Insured Patients.

With Synthea's outputs shown to reflect real data, we were able to make conclusions based on Synthea's data with a reasonable degree of certainty of their accuracy. Our hypothesis to test was whether there is a significant decrease in the survival rates of uninsured breast cancer patients verses insured ones. The statistical formatting to analyze this hypothesis is as follows:

- $H_0 : \mu_I <= \mu_U$
- $H_1 : \mu_I > \mu_U$
    - *Where μ_I is the average survival rate of insured breast cancer patients and μ_U is the average survival rate of uninsured breast cancer Patients.*

To test this hypothesis and either reject or fail to reject the null hypothesis, we used a two-sample one-tailed proportion p-test. The two samples were from Synthea datasets that were previously generated: The population where every patient had insurance and the population where no patients had insurance. The overall survival rates output by Synthea for each was as follows:

- Insured breast cancer patient survival rate ($\boldsymbol{\mu_I}$): **89.10%**
- Uninsured breast cancer patient survival rate ($\boldsymbol{\mu_U}$): **81.00%**

Table 17: Comparison of Overall Survival Rates of Uninsured and Insured Breast Cancer Populations in Synthea

Synthea Insured Vs. Uninsured Survival Rates		
	Insured Patients	**Uninsured Patients**
# Survivors	891	810
# Deaths	109	190
Total # Patients	1000	1000
Overall Survival Rate	89.10%	81.00%

Figure 9 displays comparisons of the survival rates by stage of insured and uninsured patients in Synthea. Although each bar comparison may seem relatively similar, there are actually percentage gaps of around 10% for each survival rate comparison.

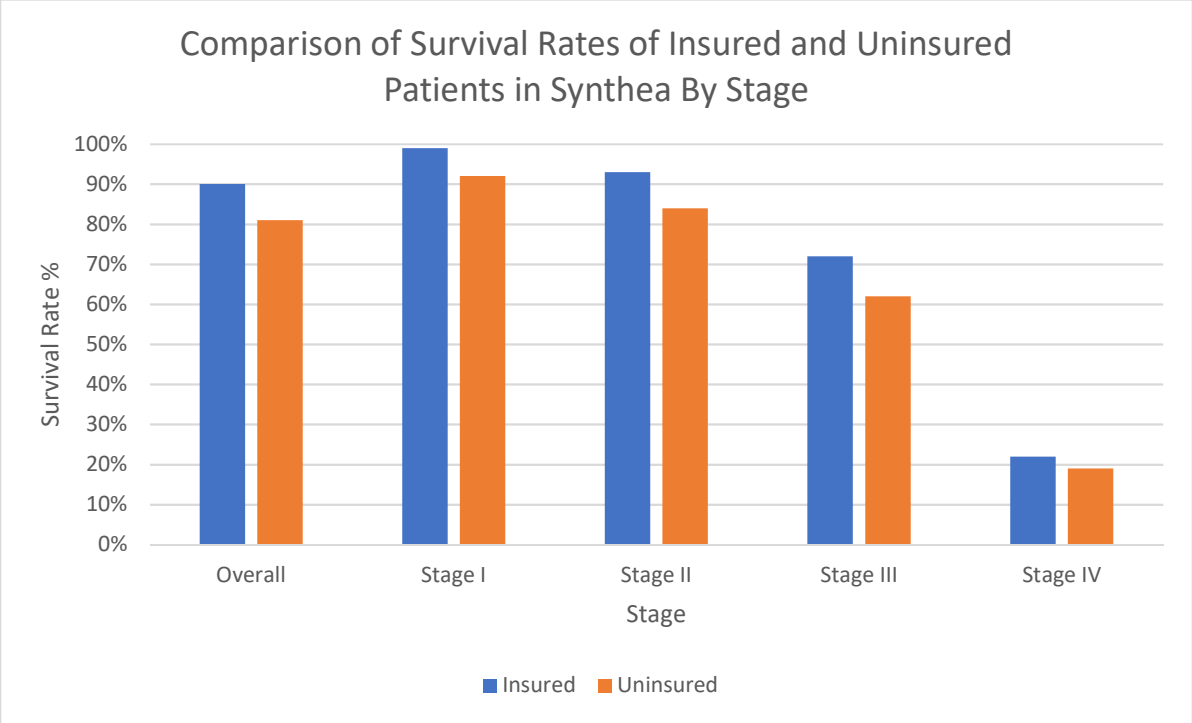

Figure 9: Comparison of Survival Rates By Stage of Insured and Uninsured Patients in Synthea

To complete a statistical analysis with a two-sample one-tailed proportion p-test, we used the following formulas, using a Z-Score as our test statistic. Our statistical test was based on a comparison between the overall survival rate of insured and uninsured breast cancer patients.

- The first step was to calculate the pooled sample proportion (ρ):
  - $\rho = \frac{\mu_I \times n_I + \mu_U \times n_U}{n_I + n_U}$
    - *Where n_I is the sample size of the insured population and n_U is the sample size of the uninsured population.*
  - $\rho = \frac{0.891 \times 1000 + 0.810 \times 1000}{1000 + 1000}$
  - $\rho = 0.8505$
- Next, we needed to calculate the Standard Error (SE):
  - $SE = \sqrt{\rho \times (1 - \rho) \times [1/n_I + 1/n_U]}$
  - $SE = \sqrt{0.8505 \times (0.1495) \times [1/1000 + 1/1000]}$
  - $SE = 0.01595$
- Finally, we calculated the test-statistic Z-Score (z):
  - $z = \frac{\mu_I - \mu_U}{SE}$
  - $z = \frac{0.891 - 0.81}{0.01595}$
  - $z = 5.0783$

With a Z-Score test statistic of 5.0783, we can locate it along the standard normal distribution, providing us with a p-value of less than 0.00001. Because the p-value is less than the significance level of 0.01, we are able to reject the null hypothesis and conclude that **breast cancer patients with health insurance have a higher survival rate than those without insurance.**

6 Conclusion

The goal of this MQP was to implement health insurance and loss-of-care modules in Synthea. Although we chose to limit our initial scope to breast cancer patients, we have laid the groundwork for further research in this area for more diseases, populations, and insurance types. We also limited our research to analyzing survival rates, however future research should expand to analyze the impacts of loss-of-care due to healthcare and insurance expenses to Quality of Life metrics. Quality of Life allows for the analysis of a range of burdens and impacts of conditions and loss-of-care as opposed to simply death. We initially planned for Quality of Life to be the focus of this MQP, however we found that survival rates made more sense for breast cancer patients so we could compare our data with reality. We even implemented much of the necessary Quality of Life code in Synthea so, for a continuation of this research, it could be expanded upon and used.

Another aspect of the MQP included generating populations with levels of insurance that reflect real life. We detailed this in the Methodology and Results sections, but left it as a contribution to Synthea rather than as a part of our hypotheses. Our implementation of insurance levels was shown to be reflective of real life. These insurance levels could be used for simulating populations to test changing insurance variables to analyze what policies and pricing could affect the overall health of the population. It would be fascinating to analyze how factors such as doubling or halving the cost of insurance or qualifications for Medicaid could impact health. Or, what if you made Medicaid free and universal? Or turned private insurance into a utility good?

Synthea outputs the revenues and expenses of every insurance company in the simulation, so you could even analyze the economic impact and cost-effectiveness. Lots of fascinating research opportunity is available here – made possible by Synthea and the ability to simulate and tweak insurance variables to see its impacts.

We have taken the first steps in implementing and analyzing health insurance and its impacts in Synthea. Simulating simply health data, as Synthea did before, is highly useful because of the barriers to access of real data. For more realistic outputs, Synthea needed health insurance because of its integral role in the healthcare system and health policy of the United States. We hope that this MQP, especially its contributions to the capabilities of Synthea, can be expanded on for complete analyses of different insurance costs and policies so that research and analysis could be further conducted on how health insurance impacts the health of a population.

Bibliography

[1] J. H. Randall R. Bovbjerg, "https://www.urban.org/research/publication/why-health-insurance-important," *Urban Institute,* 2007.

[2] "Healthcare.gov," 30 9 2019. [Online]. Available: https://www.healthcare.gov/health-care-law-protections/.

[3] K. Monica, "Top 5 Challenges to Achieving Healthcare Interoperability," EHR Intelligence, [Online]. Available: https://ehrintelligence.com/news/top-5-challenges-to-achieving-healthcare-interoperability.

[4] I. o. Medicine, "Sharing Clinical Research Data: Workshop Summary," Washington, DC, The National Academies Press.

[5] A. Chen, "Why It's Time to Rethink the Laws That Keep Out Health Data Private," The Verge, 29 1 2019. [Online]. Available: https://www.theverge.com/2019/1/29/18197541/health-data-privacy-hipaa-policy-business-science.

[6] S. D. C.-K. V. M. J. R. K. G. J. Thacker SB, "Measuring the public's health," *Public Health Rep.,* 2006.

[7] J. Pitts, "The Purpose of Health Insurance," Health Guidance, 13 8 2019. [Online]. Available: https://www.healthguidance.org/entry/11241/1/the-purpose-of-health-insurance.html.

[8] "Does Health Insurance Affect Health Care Utilization and Health?," The National Bureau of Economic Research, [Online]. Available: https://www.nber.org/aginghealth/spring04/w10365.html.

[9] J. S. Gordon, "A Short History of American Medical Insurance," *Imprimis,* vol. 47, no. 9, September 2018.

[10] A. Pennza, "FAQ - How Much Does Individual Health Insurance Cost?," People Keep, 7 November 2018. [Online]. Available: https://www.peoplekeep.com/blog/bid/97380/faq-how-much-does-individual-health-insurance-cost. [Accessed 3 September 2019].

[11] E. H. a. J. C. B. Edward R. Berchick, "Health Insurance Coverage in the United States: 2017," Census Bureau, Suitland, Maryland, 2017.

[12] I. o. M. (. C. o. A. t. H. o. t. P. i. t. 2. Century, "The Health Care Delivery System," in *The Future of the Public's Health in the 21st Century*, Washington D.C., National Academies Press, 2002, p. Chapter 5.

[13] L. M. R. K. S. P. a. K. A. H. Xiaoling Niu, "Cancer survival disparities by health insurance status," *Cacner Medicine,* vol. 2, no. 3, 8 May 2013.

[14] X. W. M. D. V. H. J. M. C. X. M. M. K. J. J. P. Christine D. Hsu, "Breast cancer stage variation and survival in association with insurance status and sociodemographic factors in US women 18 to 64 years old," *Cancer,* vol. 123, no. 16, pp. 3125-3131, 25 April 2017.

[15] "Key Facts about the Uninsured Population," Kaiser Family Foundation, 7 12 2018. [Online]. Available: https://www.kff.org/uninsured/fact-sheet/key-facts-about-the-uninsured-population/.

[16] Kaiser Family Foundation, "Health Insurance Coverage of the Total Population," Kaiser Family Foundation, 2017. [Online]. Available: https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D.

[17] Centers for Medicare & Medicaid Services, "CMS' Program History," 5 August 2019. [Online]. Available: https://www.cms.gov/About-CMS/Agency-information/History/. [Accessed 7 September 2019].

[18] National Academy of Social Insurance, "What is the History of Medicare?," National Academy of Social Insurance, [Online]. Available: https://www.nasi.org/learn/medicare/history. [Accessed 3 September 2019].

[19] S. Anderson, "A brief history of Medicare in America," Medicare Resources, 1 September 2019. [Online]. Available:

https://www.medicareresources.org/basic-medicare-information/brief-history-of-medicare/. [Accessed 10 September 2019].

[20] M. Koba, "Medicare and Medicaid: CNBC Explains," CNBC, 1 September 2011. [Online]. Available: https://www.cnbc.com/id/43992654. [Accessed 12 September 2019].

[21] J. Skowronski, "A state-by-state guide to Medicaid: Do I qualify?," Policy Genius, 26 January 2018. [Online]. Available: https://www.policygenius.com/blog/a-state-by-state-guide-to-medicaid/. [Accessed 20 September 2019].

[22] P. Berry, "Medicare Resource Center," AARP, 25 September 2019. [Online]. Available: https://www.aarp.org/health/medicare-insurance/info-04-2011/medicare-eligibility.html. [Accessed 28 September 2019].

[23] Wikipedia, "Mortality rate," 20 August 2019. [Online]. Available: https://en.wikipedia.org/wiki/Mortality_rate. [Accessed 10 September 2019].

[24] D. J. Brailer, "Interoperability: The Key To The Future Health Care System," *Health Affairs,* vol. 24, no. 1, 2005.

[25] U.S. Department of Health and Human Services, "Summary of the HIPAA Privacy Rule," 26 July 2013. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html. [Accessed 20 September 2019].

[26] K. O. A. D. Rachel Garfield, "The Uninsured and the ACA: A Primer - Key Facts about Health Insurance and the Uninsured amidst Changes to the Affordable Care Act," Kaiser Family Foundation, 25 January 2019. [Online]. Available: https://www.kff.org/report-section/the-uninsured-and-the-aca-a-primer-key-facts-about-health-insurance-and-the-uninsured-amidst-changes-to-the-affordable-care-act-how-does-lack-of-insurance-affect-access-to-care/. [Accessed 19 September 2019].

[27] BreastCancer.org, "https://www.breastcancer.org/symptoms/understand_bc/risk/understanding, " BreastCancer.org, 20 December 2018. [Online]. Available: https://www.breastcancer.org/symptoms/understand_bc/risk/understanding. [Accessed 15 September 2019].

[28] BreastCancer.org, "Breast Cancer Staging," 23 July 2019. [Online]. Available: https://www.breastcancer.org/symptoms/diagnosis/staging. [Accessed 23 September 2019].

[29] National Cancer Institute, "Overview of the SEER Program," Surveillance, Epidemiology, and End Results (SEER) Program, [Online]. Available: https://seer.cancer.gov/about/overview.html. [Accessed 15 September 2019].

[30] National Program of Cancer Registries, "United States Cancer Statistics (USCS)," Centers for Disease Control and Prevention, 14 May 2019. [Online]. Available:

https://www.cdc.gov/cancer/uscs/public-use/index.htm?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcancer%2Fnpcr%2Fpublic-use%2Findex.htm. [Accessed 15 September 2019].

[31] J. F. M. Christopher JL Murray MD, "Health Metrics and Evaluation: Strengthening the Science," *The Lancet,* vol. 371, no. 9619, 2008.

[32] H. D. Woolhandler S, "The Relationship of Health Insurance and Mortality: Is Lack of Insurance Deadly?," *Annals of Internal Medicine,* vol. 167, no. 6, p. 424–431, September 2017.

[33] R. Nuzum, "Federal and State Health Policy," The Commonwealth Fund, [Online]. Available: https://www.commonwealthfund.org/programs/federal-and-state-health-policy. [Accessed 20 September 2019].

[34] R. M. a. D. G. Gonzalez, "Infant Mortality Rate as a Measure of a Country's Health: A Robust Method to Improve Reliability and Coparability," *Demography,* vol. 54, 2017.

[35] M. Ghoncheh, "Incidence and Mortality and Epidemiology of Breast Cancer in the World," *Asian Pacific Journal of Cancer Prevention,* vol. 17, 2016.

[36] D. LeSueur, "5 Reasons Healthcare Data Is Unique and Difficult to Measure," Health Catalyst, 12 June 2014. [Online]. Available:

https://www.healthcatalyst.com/insights/5-reasons-healthcare-data-is-difficult-to-measure. [Accessed 21 September 2019].

[37] Office for Civil Rights, "Your Rights Under HIPAA," U.S. Department of Health and Human Services, 1 Februrary 2017. [Online]. Available: https://www.hhs.gov/hipaa/for-individuals/guidance-materials-for-consumers/index.html. [Accessed 16 September 2019].

[38] The Healthcare Information and Management Systems Society, "What is Interoperability?," HIMSS, [Online]. Available: https://www.himss.org/library/interoperability-standards/what-is-interoperability. [Accessed 15 September 2019].

[39] R. G. Parrish, "Measuring Population Health Outcomes," *Preventing Chronic Disease,* vol. 7, 2010.

[40] e. a. Christine D. Hsu, "Breast cancer stage variation and survival in association with insurance status and sociodemographic factors in US women 18 to 64 years old," *Cancer,* vol. 123, no. 17, 2017.

[41] BreastCancer.org, "U.S. Breast Cancer Statistics," BreastCancer.org, 13 February 2019. [Online]. Available: https://www.breastcancer.org/symptoms/understand_bc/statistics. [Accessed 12 September 2019].

[42] The Surveillance, Epidemiology, and End Results (SEER) Program, "About the SEER Program," National Cacner Institute, [Online]. Available: https://seer.cancer.gov/about/. [Accessed 12 September 2019].

[43] Centers for Disease Control and Prevention, "Cancer Data and Statistics," Centers for Disease Control and Prevention, 9 July 2019. [Online]. Available: https://www.cdc.gov/cancer/dcpc/data/index.htm. [Accessed 15 September 2019].

[44] Cancer.net, "Breast Cacner: Stages," Conquer Cancer Foundation, July 2019. [Online]. Available: https://www.cancer.net/cancer-types/breast-cancer/stages. [Accessed 16 September 2019].

[45] Susan G. Komen, "Breast Cancer Staging and Stages," Susan G Komen, 8 Septmber 2019. [Online]. Available: https://ww5.komen.org/BreastCancer/breastcancerstagesandstagingbefore2018.html. [Accessed 16 September 2019].

[46] N. Sahni, "Rethinking Healthcare Labor," *The New England Journal of Medicine,* vol. 365, no. 15, 13 October 2011.