CENTERS FOR MEDICARE & MEDICAID SERVICES

Synthetic Medicare Enrollment, Fee-for-Service Claims, and Prescription Drug Event Data Public Use File (Synthetic Medicare Claims PUFs)

USER GUIDE

May 2023

User Guide

Centers for Medicare and Medicaid Services (CMS)
Synthetic Medicare Claims PUFs

List of Tables

List of Figures

1 Introduction

1.1 Scope and Purpose

The Centers for Medicare and Medicaid Services (CMS) Synthetic Medicare Enrollment, Fee-for-Service Claims, and Prescription Drug Event Data Public Use File (Synthetic Medicare Claims PUFs) is a synthetic dataset representing enrollment information and healthcare claims for 8,671 Medicare beneficiaries between the ages of 0 and 110. Access to real enrollment and claims data is restricted to protect the privacy of beneficiaries. However, since synthetic data are realistic-but-not-real data, the typical privacy and security restrictions do not apply, and the data can be released publicly without restrictions.

CMS created this synthetic dataset to allow interested parties to gain familiarity with using Medicare claims data while protecting beneficiary privacy. The synthetic data are available in CMS' Research Identifiable File (RIF) format, meaning that even though they are not tied to any real patient data, they mimic the real claims data that CMS makes available to researchers. While these files can increase user's knowledge of claims data and skill analyzing such data, they have very limited inferential research value and should not be used draw conclusions about Medicare beneficiaries due to the synthetic processes used to create the files.

The synthetic dataset is publicly accessible and can be downloaded directly from https://data.cms.gov/collection/synthetic-medicare-enrollment-fee-for-service-claims-and-prescription-drug-event. The data are packaged in a ZIP file and can be extracted using any common utility that supports that file type. Each data file is plain text and can be viewed in any text editor.

The purpose of this document is to serve as a technical reference for users of CMS' Synthetic Medicare Claims PUFs. The document is organized as follows:

- Section 2 provides background information on synthetic data generation and describes the Synthea modeling architecture, modules, and data resources
- Section 3 summarizes the attributes of this public release version of the dataset
- Section 4 presents a comparison of the attributes of the synthetic data versus real data
- Section 5 is a review of use cases that leverage the CMS synthetic Medicare claims dataset
- Section 6 enumerates the fixed and random values of the data fields in the RIF files
- Section 7 covers the acronyms and glossary

1.2 Disclaimer

The U.S. government does not assume any legal liability or responsibility for the accuracy, completeness, or usefulness of the dataset and shall not be liable for any consequential, incidental, or indirect damages claimed to be suffered as a result of use of the data. Please note also that the data in this public release is subject to change without prior notice.

2 Synthetic Data Primer

2.1 Understanding Synthetic Data

Synthetic data are realistic-but-not-real data that can be used or shared without the privacy and security risks associated with real health data. It is generated either from models based on aggregated statistics (e.g., modeling and simulation without direct access to any individual data points) or models abstracted from sensitive data (e.g., machine learning models that were trained from, but do not preserve, individual data records).

Synthetic data are not deidentified data. Deidentified data are often modified from real data points using methodologies such as masking or deleting fields and introducing noise. However, deidentification does not guarantee privacy or eliminate risk.[1]

Synthetic data has been widely used as a safe alternative to deidentified data with the advantage that there are no individual sensitive records underneath any synthetic records that can ever be reidentified.[2] However, because the models used to simulate synthetic claims are flawed, as all models are, synthetic data will not standup to a rigorous comparison with a real data set.

Nevertheless, synthetic clinical data sets are generally useful in a variety of use cases including software testing and validation (e.g., developing databases or health apps, including privacy and security testing), education (especially in Health IT), academic research preparation, feasibility assessments, and algorithm validation. This synthetic data should not be used for clinical discovery and scientific inference.

2.2 Synthea™

Synthea is an open-source synthetic patient generator that models the medical history of synthetic patients. Its mission is to create high-quality synthetic, realistic-but-not-real, patient data and associated health records covering every aspect of healthcare. Patients are generated independently via Monte Carlo processes[3] over probabilistic disease models represented with modules.

Synthea was used to generate this claims dataset because it: (a) provides fully synthetic output based only on publicly available data; (b) facilitates transparency and continuous improvement of clinical workflow and disease progression models; (c) covers beneficiaries entire lifetime of health problems and diseases; and (d) enables scalable collaborative development among experts in a broad range of clinical and technical backgrounds. For further details please refer to the

[1] Gallagher, Thomas, Kudakwashe Dube, and Scott McLachlan. "Ethical issues in secondary use of personal health information." *IEEE Future Directions: Technology Policy & Ethics.* (May 2018) http://sites.ieee.org/futuredirections/tech-policy-ethics/may2018/ethical-issues-in-secondary-use-of-personal-health-information/

[2] Walonoski, Jason, et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record." *Journal of the American Medical Informatics Association*, 25(3) (Jul. 2018): 230-238. https://doi.org/10.1093/jamia/ocx079

[3] Probabilistic modeling using Monte Carlo method: https://en.wikipedia.org/wiki/Monte_Carlo_method

paper by Walonoski et al. (2018).[4] Please refer also to the Synthea project Wiki[5] for more information.

To help users gain insight into the strengths and weaknesses of this synthetic data set and whether it can be used for their particular use cases, a comparison of real data with this synthetic dataset is presented in Section 4, and sample use-cases are examined in Section 5.

2.2.1 Architecture and Data Sources

Synthea generates synthetic patient records using an agent-based modeling approach[6] that simulates the patient's exposure to probabilistic disease modules that are created using clinical care maps and publicly available disease incidence and prevalence data.

Each synthetic patient is generated independently, as they progress from birth to death, through modular representations of various diseases and conditions.

Each patient goes through every module in the system. When a patient dies or the simulation reaches the specified end date, that patient record can be exported in several different formats.

Figure 2-1 illustrates the Synthea concept of operations. The clinical disease models are constructed using clinical care maps and publicly available data on disease incidence and prevalence. The clinical care maps are based on clinical practice guidelines gathered from peer-reviewed journals and published reports by medical specialty societies.

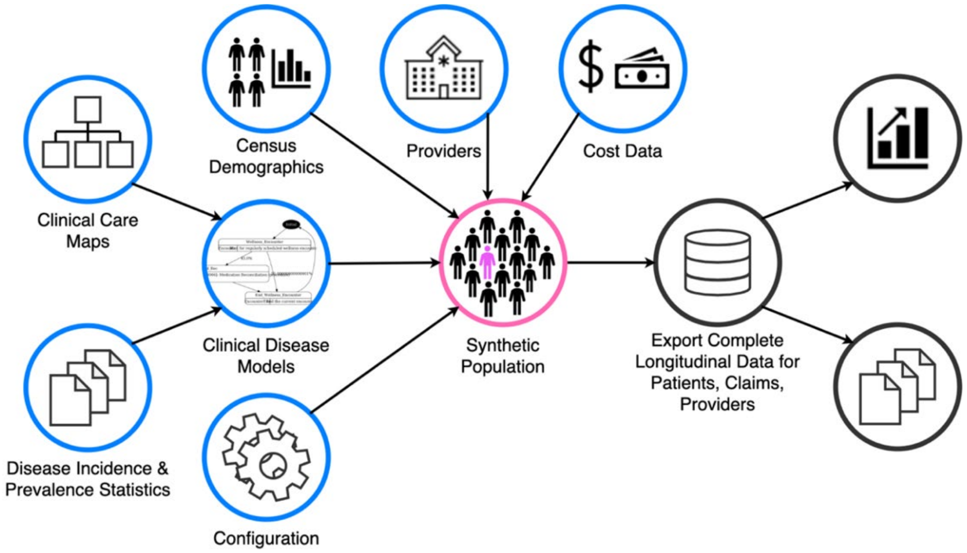


Figure 2-1. Synthea Concept of Operations

[4] Walonoski, Jason, et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record." *Journal of the American Medical Informatics Association*, 25(3) (Jul. 2018): 230-238. https://doi.org/10.1093/jamia/ocx079

[5] Synthea Project Wiki. https://github.com/synthetichealth/synthea/wiki

[6] Definition of "agent-based modeling approach": https://en.wikipedia.org/wiki/Agent-based_model

Disease incidence and prevalence data are derived from publicly available statistics from the Center for Disease Control and Prevention (CDC), National Institutes of Health (NIH), and peer-reviewed literature. These statistics are coupled with census demographics data in the developing Synthea modules to drive the disease progression and treatment models.

Synthea uses demographic data from the 2010 US Census[7] and can generate representative populations for any town or city in the United States. These include county and subcounty demographic distributions of gender, race, and age groups. Synthea also uses education level attainment and income level distributions from the 2010-2014 American Community Survey 5-Year Estimates. Finally, Synthea uses Social Determinants of Health county-level data derived from the 2018 Social Vulnerability Index and the 2021 Community Health Rankings. For further information, please refer to the Demographic Data documentation in the Synthea project Wiki.[8]

Providers include labs, hospitals, clinics, treatment centers and other facilities that offer healthcare services and receive reimbursements from government and private sector payers. Information on facility name, type of Medicare services, accreditation, ownership, and addresses are obtained from CMS Provider of Services files.[9] ZIP code data are derived from publicly available files.

The cost of healthcare services is modeled in Synthea based on data from public sources such as the Healthcare Cost and Utilization Project[10] and National Library of Medicine (NLM).[11] Please note that healthcare costs vary widely.[12] Synthea models cost data as a set of comma-separated value (CSV) files,[13] where each row of each file specifies a cost value for a good or service.

Payer data are derived from publicly available files. Financial characteristics of payers such as premiums, copays, and coinsurance are not accurate and only a single plan is modeled for each insurer.

[7] United States Census Bureau. https://www.census.gov

[8] Synthea Project Wiki. https://github.com/synthetichealth/synthea/wiki/Default-Demographic-Data

[9] CMS Provider of Services Files. https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Provider-of-Services

[10] Healthcare Cost and Utilization Project. https://hcupnet.ahrq.gov/

[11] National Library of Medicine. https://www.ncbi.nlm.nih.gov

[12] For example, Hop, M. Jenda, et al. (2014)[12] indicated that total healthcare cost per burn patient in high-income countries ranged from $704 to $717,306 with a median of $44,024. Interpretation of the modeled healthcare costs should recognize this high variability.

[13] Synthea cost data in CSV files. https://github.com/synthetichealth/synthea/tree/master/src/main/resources/costs

3 CMS Synthetic Claims Data

3.1 Scope

This public release of the CMS synthetic claims data consists of 8,671 synthetic beneficiaries. Table 3-1 shows a profile of the data by claim type and time period.

Table 3-1. Profile of Synthetic Claims Data

Claim Type	Count	%	Distinct Procedures	Period Covered
Nursing Facility	12,526	.66	27	January 8, 2015 – March 1, 2023
Outpatient	574,861	30	107	March 6, 2015 – March 2, 2023
Carrier	1,120,655	59	38	March 6, 2015 – March 2, 2023
Hospice	12,088	.64	13	November 18, 2014 – February 27, 2023
Durable Medical Equipment	103,798	5.4	40	January 8, 2015 – March 2, 2023
Inpatient	58,030	3.1	106	February 25, 2015 – March 2, 2023
Home Health Agency	6,215	.32	22	March 9, 2015 – March 1, 2023
Total	1,888,173			
			Distinct Medications	
Part D	515,520		14,627	March 8, 2015 – March 3, 2023

3.2 Limitations

Not all events in Synthea are exported in the RIF format. This is primarily because Synthea is modeling clinical activities, and only those activities which have corresponding billable codes are included.

The RIF data format requires ICD-10-CM, HCPCS, NDC, and other code systems. Synthea natively uses SNOMED-CT, RxNorm, LOINC, and CVX code systems. Synthea maps its code systems to RIF's code systems using a set of mapping files. Not all codes used by Synthea have defined mappings to the corresponding RIF code system. We use the term "mappable" for a code that can be mapped from Synthea to the corresponding RIF code system and "non-mappable" for codes that cannot. Codes that are "non-mappable" are assumed to be unbillable, so are not included in the synthetic claims.

In most cases, the Synthea RIF exporter filters data to ensure that only events with mappable code are exported. Table 3-2 describes the filters for non-exportable events.

Table 3-2. Filter for Non-Exportable Events

Claim Type	Events not exported under any of the following conditions
Outpatient or Inpatient	Missing any of the following mappable items: • Reason • Procedure that occurred during the encounter • Diagnosis that was made during or prior to the encounter Procedures with non-mappable codes
Carrier, Home Health, or Hospice	Missing mappable diagnosis made during or prior to the encounter
Durable Medical Equipment	Missing both of the following mappable items: • Diagnosis made during or prior to the encounter • Device or supply line item
Hospice	Missing any of the following items: • Procedure that occurred during the encounter • Diagnosis that was made during or prior to the encounter
Prescription	Medication with missing mappable code

Synthea does not model values for all RIF file fields; Section 6 lists all the fields that are populated with fixed or randomly selected values.

3.3 Eligibility

While Synthea simulates health events from birth to death, not all those events will be eligible for CMS coverage. Only patients who are aged 65 or above, have end-stage renal disease (ESRD), or are disabled are eligible for Medicare benefits.

To ensure that the patient population includes eligible individuals, Synthea was configured to only generate patients that are alive and met eligibility criteria at the end of the simulation. In addition, Synthea was configured to only export up to 10 years of patient history (it could be less if only a portion of a beneficiary's 10-year history was eligible for Medicare).

3.3.1 Plan Details

Synthea was configured to model fee-for-service Medicare claims and Part D prescription claims.

3.3.2 Part D Enrollment

The RIF exporter adds simulation of Part D plan enrollment. For this data set, Synthea was configured with 10 Part D plans, each with five identical plan benefit packages.

Approximately 70%[14] of patients will be enrolled in a Part D plan for any given year. Twenty percent (20%) of patients will change plans during open enrollment and 1% will change plans at other times of the year.

Medication claims are only exported for the periods when a patient is enrolled in a Part D plan.

[14] Medicare Part D Coverage and Cost in 2019. https://www.kff.org/medicare/issue-brief/10-things-to-know-about-medicare-part-d-coverage-and-costs-in-2019/

3.3.3 Part C Enrollment

The RIF exporter also adds simulation of Part C plan enrollment, but this enrollment does not affect exported claims, all of which are fee-for-service.

For this data set, Synthea was configured with 10 Part C plans, each with five identical plan benefit packages.

Approximately 58%[15] of patients will be enrolled in a Part C plan for any given year. Twenty percent (20%) of patients will change plans during open enrollment and 1% will change plans at other times of the year. Part C and D enrollment and changes in enrollment are independent (not correlated) for any given patient.

3.4 Data Format

The data included in this release conforms to the data format specification as described in the following documents:

- Medicare Beneficiary Summary File (MBSF) Base with Medicare Part A, B, C, and D[16] defines each of the data fields and the included code values used in the beneficiary file.
- Medicare Fee-for-Service (FFS) Claims[17] defines each of the data fields and the included code values used in the claim files.
- Medicare Part D Event Drug Characteristics[18] defines each of the data fields and the included code values used in the Part D claim file.
- Fields defined in the MBSF, FFS, and PDE codebooks[17,18,19] have both a "long name" and a "short name." The synthetic data uses the "long name" a majority of the time but also uses the "short name."

The files are row based, with each line in the file representing one row of data. The first row of each file contains column headers, subsequent rows contain data. Within each row, columns (or fields) are separated by the pipe "|" character. The data consists of 19 separate files described in Table 3-3.

[15] Medicare Enrollment Update and Key Trends. https://www.kff.org/medicare/issue-brief/medicare-advantage-in-2021-enrollment-update-and-key-trends/

[16] CODEBOOK: Medicare Beneficiary Summary File (MBSF) Base with Medicare Part A, B, C, and D FEBRUARY 2021 │ VERSION 1.4. https://www2.ccwdata.org/documents/10280/19022436/codebook-mbsf-abcd.pdf

[17] CODEBOOK: Medicare Fee-For-Service (FFS) Claims (for Version L) APRIL 2022 │ VERSION 1.8. https://www2.ccwdata.org/documents/10280/19022436/codebook-ffs-claims.pdf

[18] CODEBOOK: Medicare Part D Event (PDE)/Drug Characteristics APRIL 2021 │ VERSION 1.3. https://www2.ccwdata.org/documents/10280/19022436/codebook-pde.pdf

Table 3-3. Synthetic Medicare Claims Public Use File (PUFs)

#	Filename	Description
1-11	beneficiary_2015.csv to beneficiary_2023.csv	each contains one row for each beneficiary that captures the state of the beneficiary at the end of the corresponding year
12	carrier.csv	contains one or more rows for each carrier claim, one claim may include multiple line items, and each is a separate row in the file
13	dme.csv	contains one or more rows for each durable medical equipment claim, one claim may include multiple line items, and each is a separate row in the file
14	hha.csv	contains one or more rows for each home health claim, one claim may include multiple line items, and each is a separate row in the file
15	hospice.csv	contains one or more rows for each hospice claim, one claim may include multiple line items, and each is a separate row in the file
16	inpatient.csv	contains one or more rows for each inpatient claim, one claim may include multiple line items, and each is a separate row in the file
17	outpatient.csv	contains one or more rows for each outpatient claim, one claim may include multiple line items, and each is a separate row in the file
18	pde.csv	contains one row for each Part D claim
19	snf.csv	contains one or more rows for each skilled nursing facility claim, one claim may include multiple line items, and each is a separate row in the file

For linking data across files, each file contains a `BENE_ID` column whose value is a unique identifier for a particular synthetic beneficiary. A complete history for any given beneficiary can be assembled by extracting all rows with the same value of the `BENE_ID` column from each of the files.

3.5 Accessing the Data

This section describes the process for obtaining the data, viewing it, and working with it.

3.5.1 Downloading

The data can be downloaded directly from https://data.cms.gov/collection/synthetic-medicare-enrollment-fee-for-service-claims-and-prescription-drug-event.

3.5.2 Extracting

The data are packaged in a ZIP file and can be extracted using any common utility that supports that file type.

3.5.3 Viewing

Each data file is plain text and can be viewed with text editors that can open large files. Spreadsheet applications can import the data as "CSV" with the field separator changed from a comma or tab to the pipe character, '|'. Spreadsheet applications may automatically transform some data, e.g., by removing leading zeroes from ZIP codes, that may negatively impact use of the data.

3.5.4 Transformation and Analysis

Examples of how to read and analyze the data with Jupyter Notebooks are provided for the data comparisons in Section 4 and use-case analyses in Section 5. The Jupyter Notebooks can be found at https://github.com/synthetichealth/rif-analysis.

4 Comparison of Synthetic and Real Claims Data

MITRE compared the synthetic data against Medicare enrollment, FFS claims, and PDE data dating back to 2015. Most of the specific analyses that follow limit the comparison of each data set to calendar year 2022. This calendar limitation was applied to make the results easier to visualize, limit the effects of year-to-year variation in the results, and improve query and analysis performance.

For this synthetic data release, real data was **not** used to train or construct the synthetic data; it was only used for comparison and validation purposes.

The synthetic data was generated using standard Synthea methodology with the RIF exporter described in Section 2.2 and Section 3.3. The version of Synthea used was the “master” branch, commit “40cb89b” within version 3.1, and the data was generated on 2 March 2023.

As discussed in Section 2.1, synthetic data has limitations as the modeling approach, input data, and assumptions are inherently subject to errors and will not withstand rigorous comparison with a real data set. However, the purpose of this comparison is to provide users with insights into the strengths and weaknesses of this synthetic data set and whether it can be used for their particular use cases.

The results of this comparison and the use case examples presented in Section 5.2, show how the synthetic data performs at replicating certain features of the real data, and each user should evaluate whether the synthetic data are suitable for their purposes.

4.1 Demographics

Synthetic data was generated for beneficiaries in all 50 states plus the District of Columbia. At “birth”, each synthetic beneficiary was assigned to a county within a state and county-level demographic data on age, race, gender, and education level distributions was used to weight randomly assigned values of those attributes prior to starting the simulation.

Both the real and the synthetic data were analyzed to look at beneficiary distributions by state, age, gender, and race.

Figure 4-1 shows that the distribution of beneficiaries by state is similar for the synthetic and real data. The real data are based on actual Medicare enrollment in each state, and the synthetic data are based upon estimates of eligible beneficiaries from Census Bureau data.

Real distributions are a slightly distorted reflection of the general Census Bureau data, because the real data includes international beneficiaries or beneficiaries in U.S. territories.

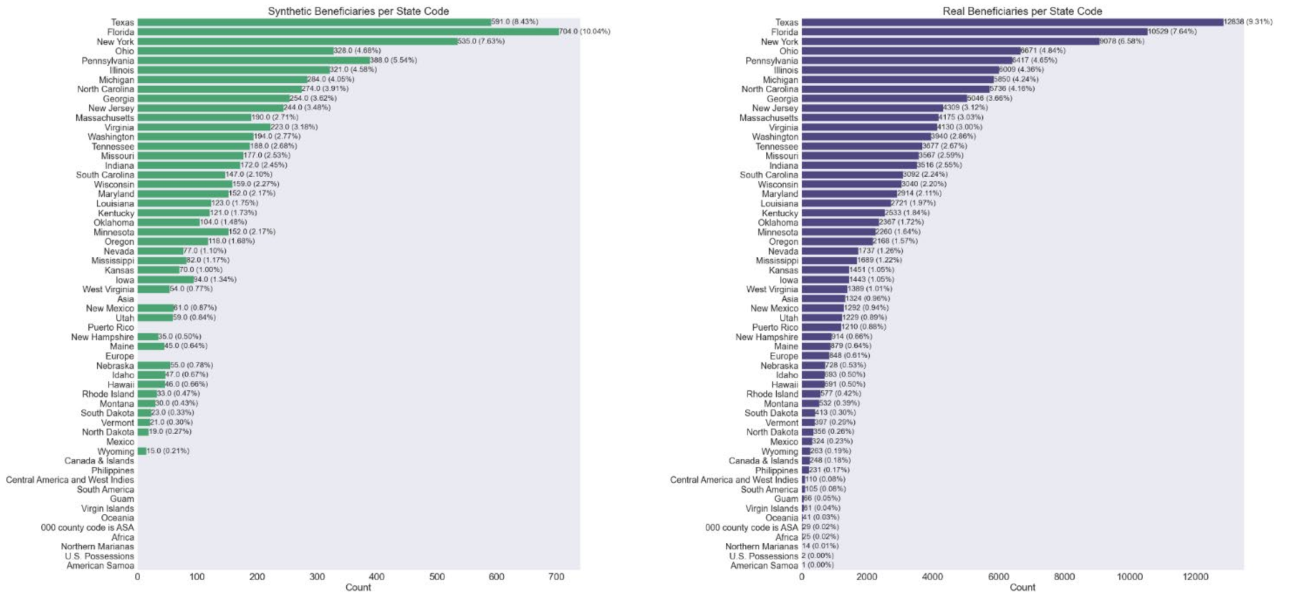


Figure 4-1. Beneficiaries Per State Code

Figure 4-2 shows beneficiary age distribution of both synthetic and real data.

The age distribution from real data reflects the diversity of the real beneficiary population. There are some beneficiaries younger than 65 due to qualifying eligibility criteria, but there is a large spike of beneficiaries beginning at age 65 and trailing off on the upper end of the spectrum.

The synthetic data has a steeper drop-off of beneficiaries around the 80-year mark, but also slightly over-represents the disabled population under the age of 20. The latter change is intentional, in order that the small sample of 8,671 synthetic people includes statistically rare qualified disabled beneficiaries present in the over 60 million individuals enrolled in Medicare.

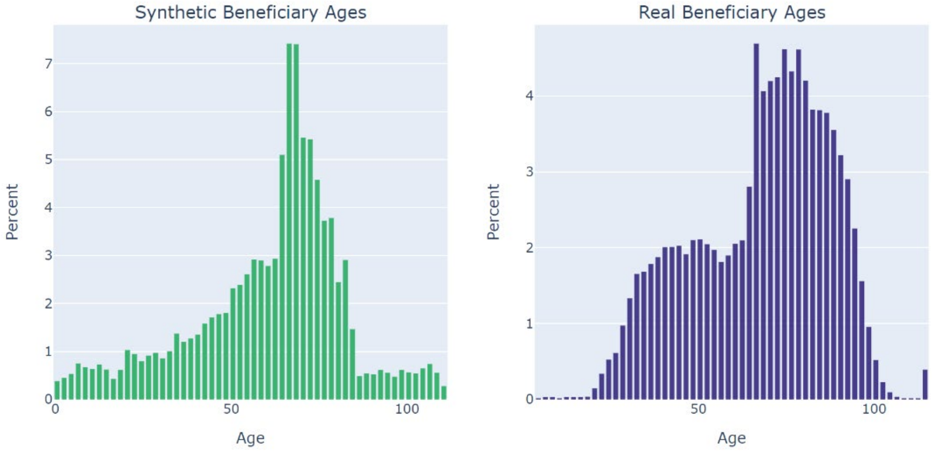


Figure 4-2. Beneficiary Age Distributions

Figure 4-3 illustrates the total number of synthetic beneficiaries by year. Each year additional synthetic people eligible for Medicare enroll as beneficiaries as they qualify either by age, end-stage renal disease, or disability criteria.

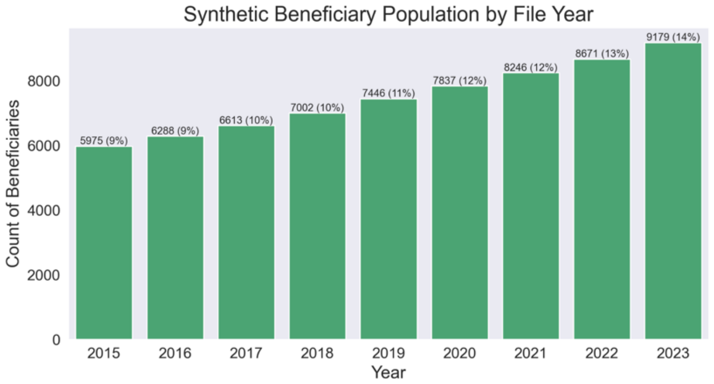


Figure 4-3. Synthetic Beneficiary Population by File Year

Figure 4-4 shows the range and variance of beneficiaries' race by age. The real data illustrates the diversity of the beneficiary population. The synthetic beneficiary population show slightly tighter distribution around age 65, but also has more outliers in the under 20 population reflective of an emphasis on beneficiaries eligible through disability.

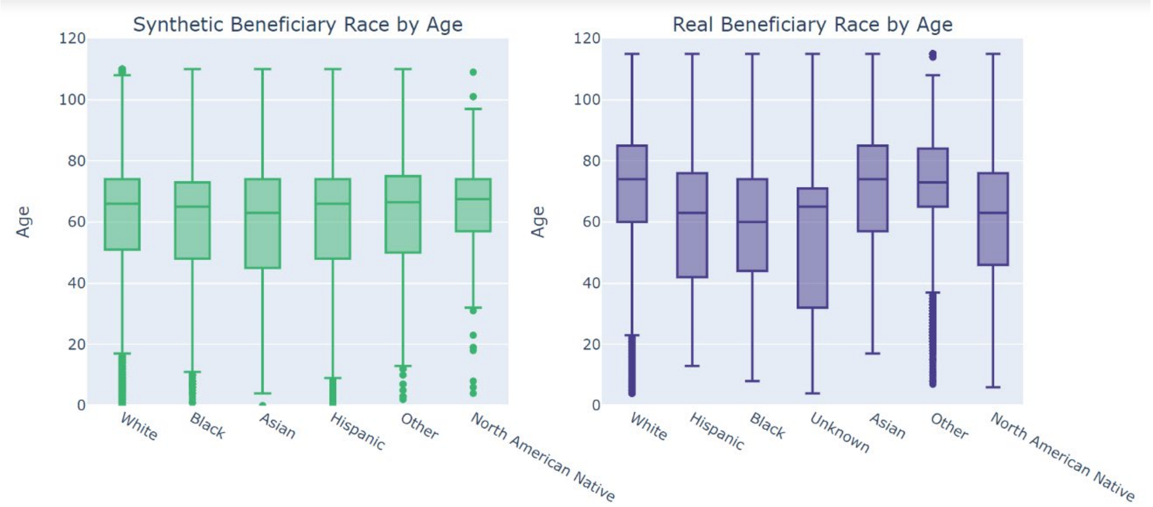


Figure 4-3. Beneficiary Race by Age

Figure 4-5 shows that the beneficiary gender distribution of both synthetic beneficiaries and real beneficiaries both have slightly more male beneficiaries than female.

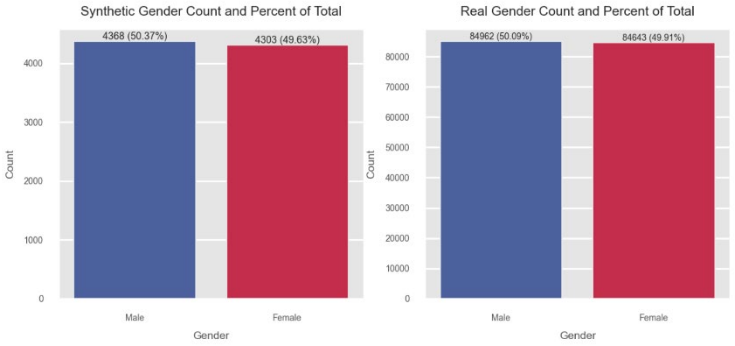


Figure 4-4. Beneficiary Gender

Figure 4-6 shows that distributions of beneficiary gender by race are similar. The synthetic data reflects the estimates from the Census Bureau. As one might anticipate, the results are similar, but not the same. In particular, the synthetic data has a higher representation of minorities than the real data sample.

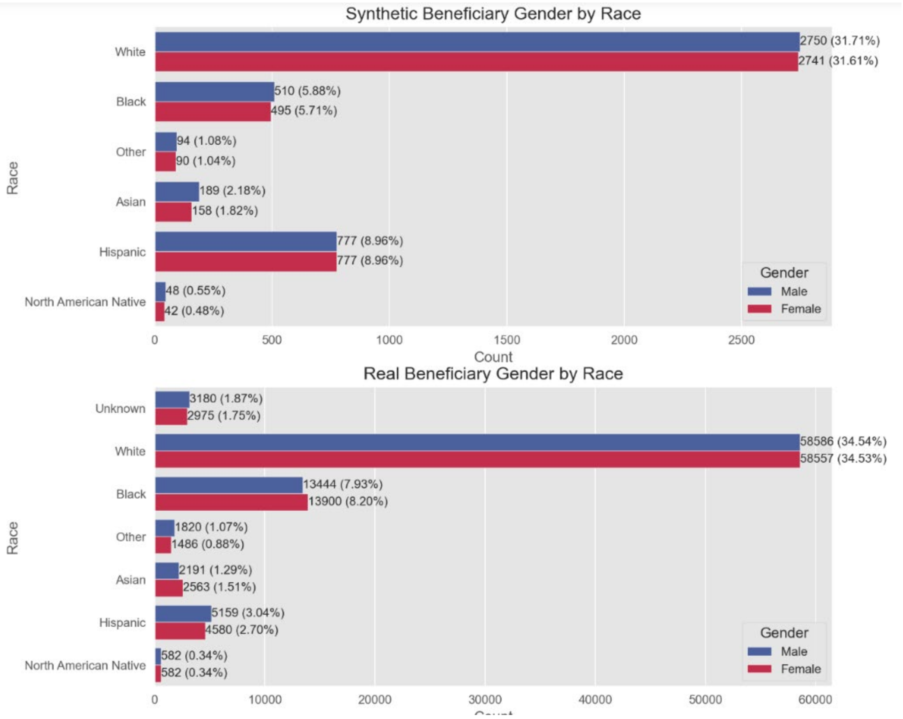


Figure 4-5. Beneficiary Gender by Race

4.2 Number of Claims per Beneficiary by Service Type

Figure 4-7 shows the distribution of number of claims per beneficiary by service type, namely: carrier, inpatient, outpatient, durable medical equipment (DME), skilled nursing facility (SNF), hospice, home health agency (HHA), and prescription drug events (PDE). This analysis was limited to claims within calendar year 2022. The synthetic claims data pertain to 8,671 synthetic beneficiaries. The real data includes claims from 13,290 living beneficiaries in 2022[19].

Both datasets show that a majority of claim types have zero claims per beneficiary. This is because while beneficiaries are enrolled in Part C, some do not have any fee-for-service claims. In addition, each beneficiary enrolled in fee-for-service does not utilize every service type each year. For example, only a small percentage of beneficiaries receive a Hospice or HHA service each year.

Notable differences between the synthetic data and real data include the length of the tails on the X-axis (total claims), as well as SNF, hospice, and HHA claims, where the synthetic data shows an overwhelming majority of synthetic beneficiaries have zero claims, and only a tiny fraction have a single annual claim.

[19] For comparison purposes, the real data was queried to only contain living beneficiaries with enrollment data having a reference year of 2022 and at least one claim. This sample was selected using the "tablesample" clause and "system" sampling method in PostgreSQL. Queries using "tablesample system" randomly selects rows from a database table, based on the percentage of rows specified (inpatient = 1%, outpatient = 1%, SNF = 1%, HHA = 1%, hospice = 1%, DME = 1%, carrier = 1%). Sample size was restricted due to database performance.

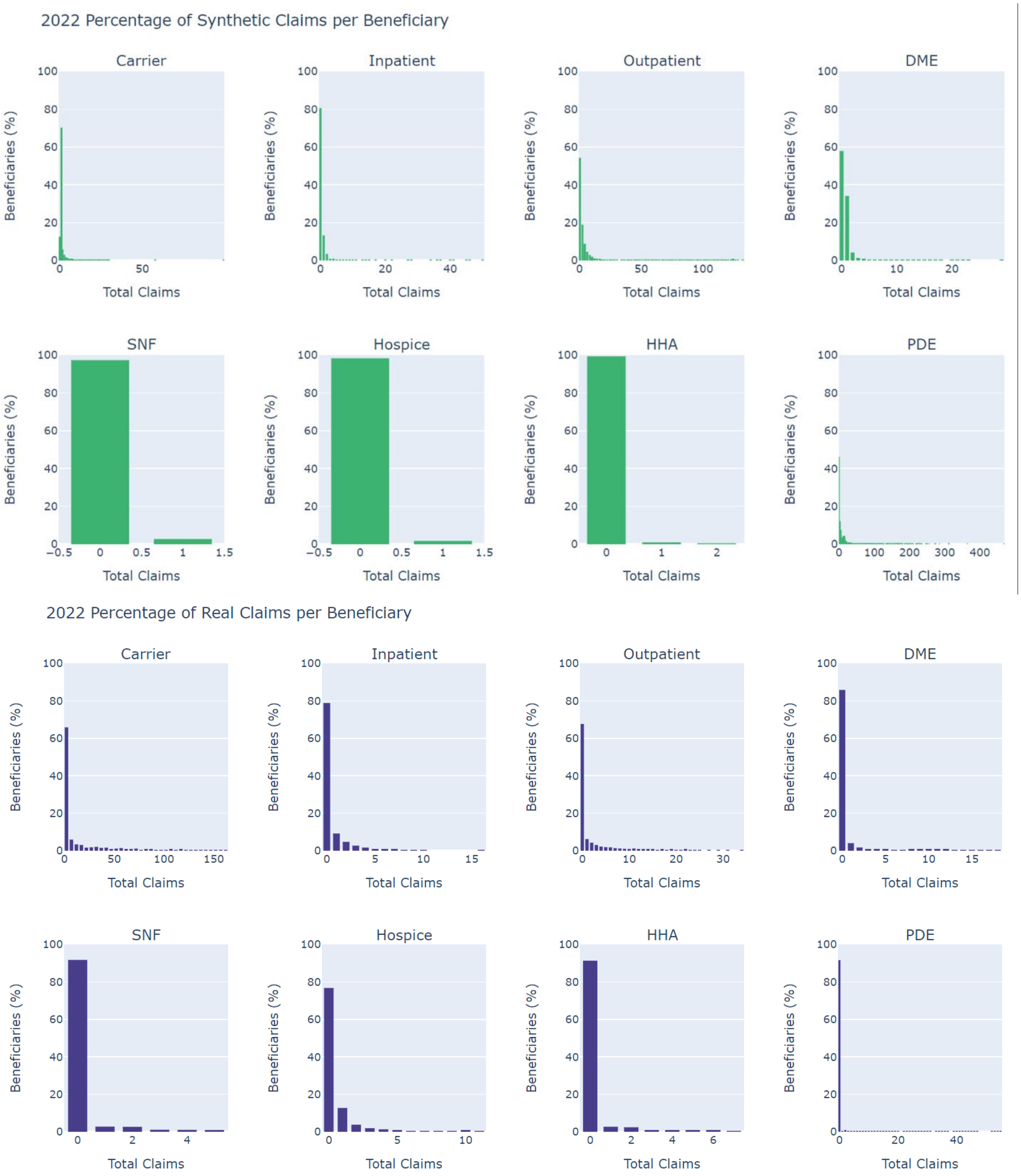


Figure 4-6. Percentage of Synthetic and Real Claims per Beneficiary by Service Type

4.3 Claimants by Service Type

Table 4-1 shows the number of synthetic and real claimants per service type. Hospice and home health agency claimants are significantly under-represented in the synthetic claims compared to real claims data.

Table 4-1. 2022 Claimants by Service Type

Service Type	Synthetic (N=8,671)	Real (N=169,353)
Skilled Nursing Facility	235 (1.2%)	14,808 (5.3%)
Outpatient	6,053 (31%)	55,572 (20%)
Carrier	7,581 (39%)	85,106 (30%)
Hospice	150 (.77%)	41,842 (15%)
Durable Medical Equipment	3,653 (18%)	26,827 (9.6%)
Inpatient	1,684 (8.6%)	35,955 (12%)
Home Health Agency	61 (.31%)	16,703 (6.0%)
Total	19,417 (100%)	276,813 (100%)

4.4 Payments

Figure 4-8 shows the distribution of claims payments by service type, namely: carrier, inpatient, outpatient, durable medical equipment (DME), skilled nursing facility (SNF), hospice, home health agency (HHA), and prescription drug events (PDE). The synthetic claims data pertain to 8,671 synthetic beneficiaries. The real data includes claims from 2,400,475 beneficiaries in 2022[20].

There are notable differences between the synthetic and real data. Inpatient claims do not match at first glance, but the bulk of the claims are situated approximately between $1K and $20K, and that portion of the distributions are similar. Other claim payment distributions share the same shape, such as outpatient or HHA claims, but the scale on the payment amount axis does not match. Some of the claim types in the synthetic data have extremely long tails with very high payment amounts (e.g., inpatient), while others such as DME, have compact distributions with very low payment amounts (i.e., double digit reimbursement).

The real data also included claims with negative payment amounts (not shown), a phenomenon not modeled in the synthetic data. Overall, the payment amounts in the synthetic data do not closely model reality. This is due to the approximations used as cost inputs (described in Section 2), and variation in line items per claim.

[20] This sample was selected using the “tablesample” clause and “system” sampling method in PostgreSQL. Queries using “tablesample system” randomly selects rows from a database table, based on the percentage of rows specified (inpatient = 1%, outpatient = 1%, SNF = 1%, HHA = 1%, hospice = 1%, DME = 1%, carrier = 0.1%). Sample size was restricted due to database performance.

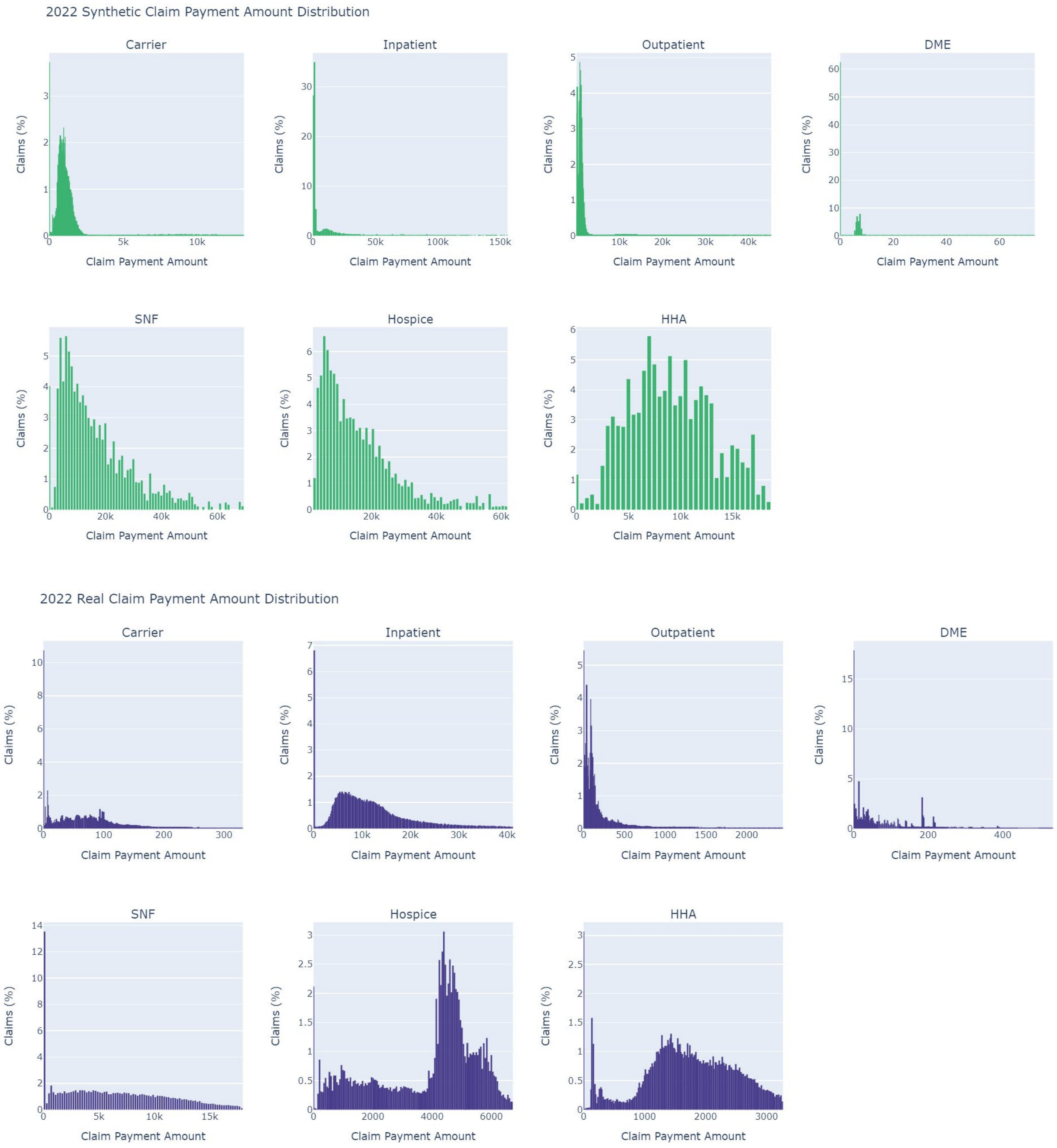
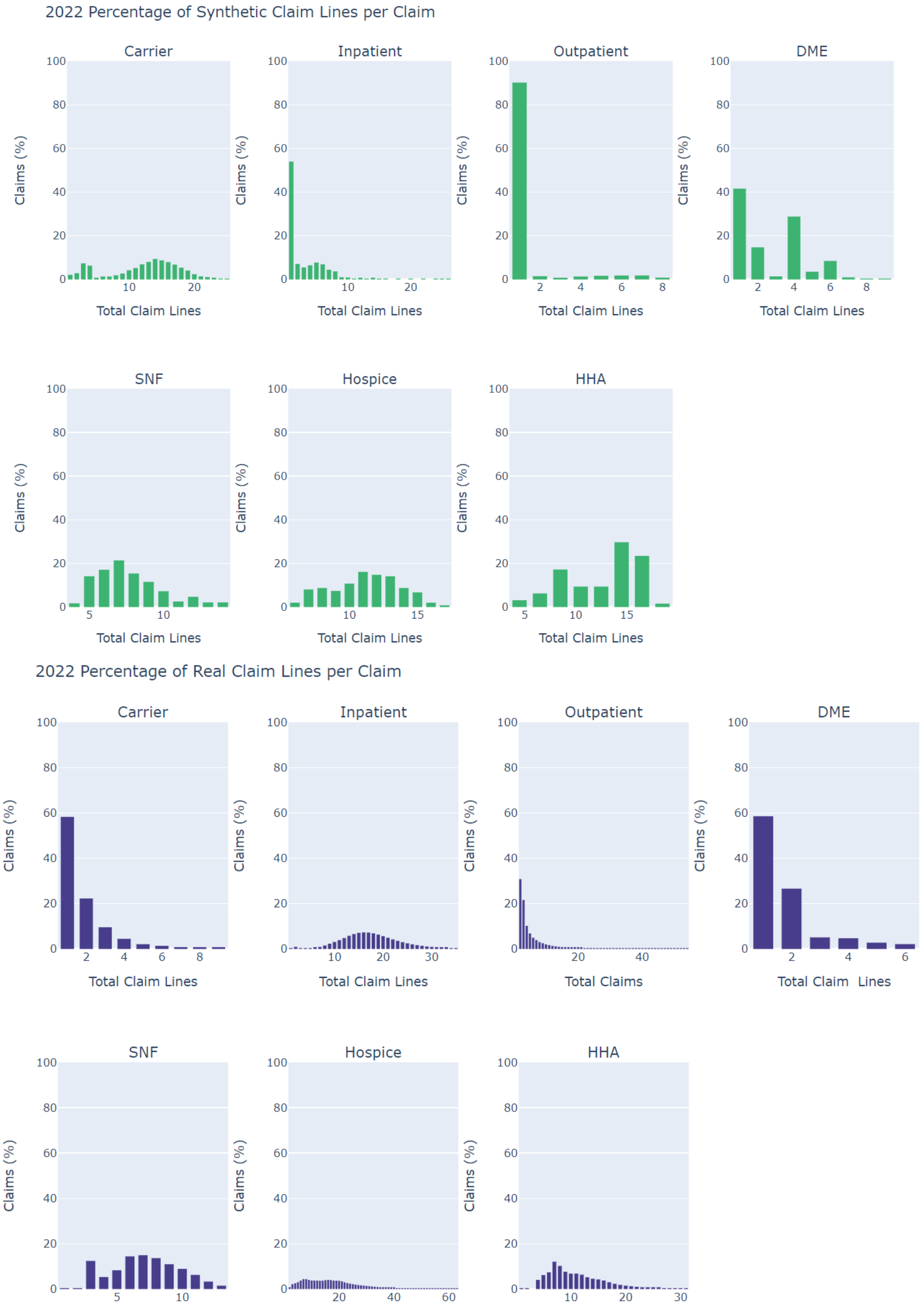


Figure 4-7. **Percentage of Synthetic and Real Claim Payment Amounts**

4.5 Line Items Count Distributions by Service Type

Figure 4-9 shows a comparison of the number of line items per claim for a given service type. The analysis was restricted to the year 2022. This comparison shows the most discrepancy between the real data and the synthetic data.

The number of line items per claim in each data set have very different distributions. Some of the synthetic distributions are skewed towards one line item per claim (e.g., inpatient and outpatient), but with a long tail of results, while the real distributions can sometimes center on different points (e.g., inpatient). DME, SNF, Hospice, and HHA are vaguely approximate, with the real data having much longer tails in most cases.

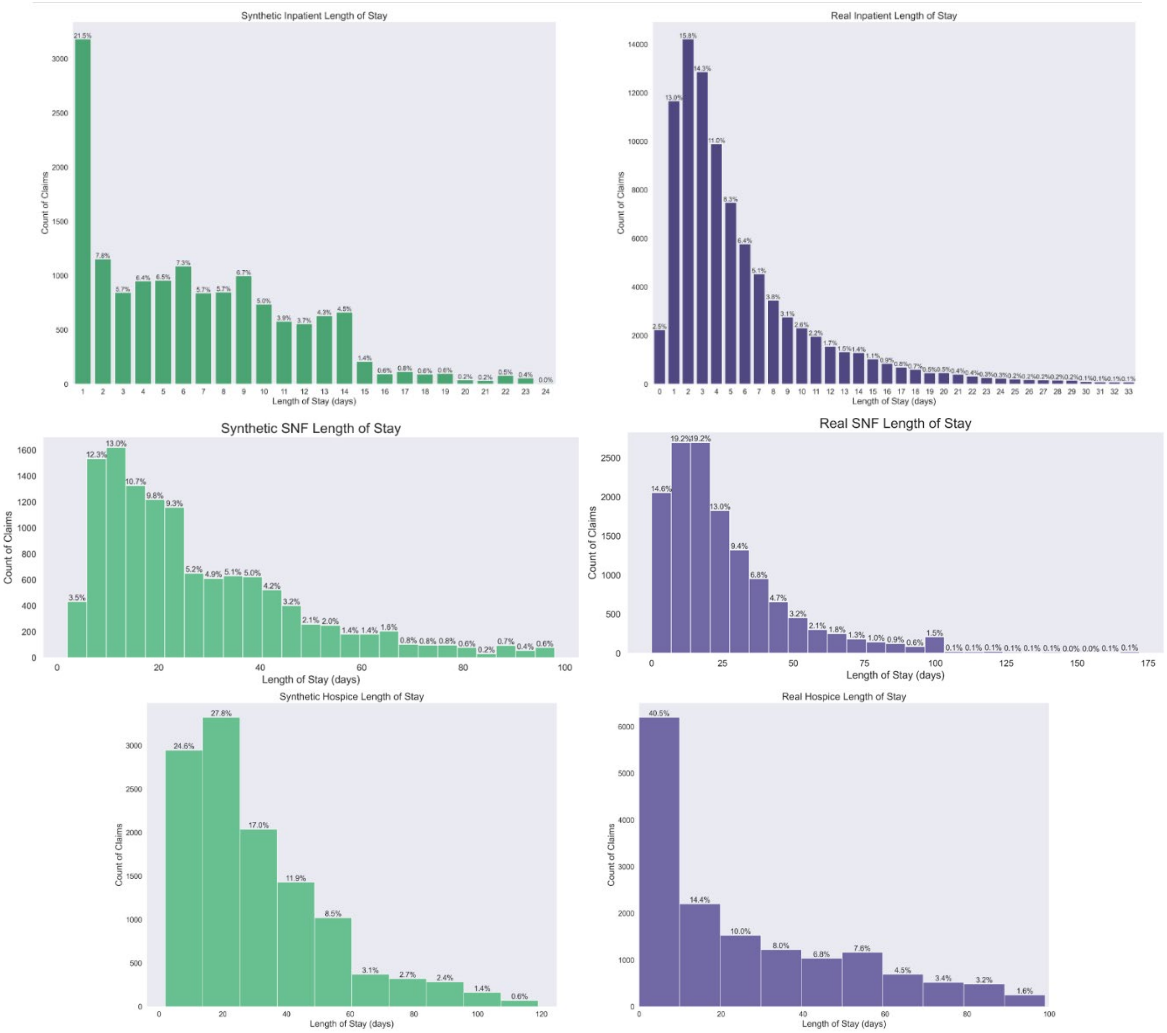

2022 Percentage of Synthetic Claim Lines per Claim

2022 Percentage of Real Claim Lines per Claim

Figure 4-8. 2021 Percentage of Real and Synthetic Claim Lines per Claim

4.6 Inpatient, SNF and Hospice Length of Stay Distributions

Figure 4-10 illustrates the Length of Stay (LOS) distributions for inpatient, SNF, and hospice claims. Again, this analysis was restricted to 2022 data. These synthetic LOS do not match the real data. The real inpatient LOS also shows most claims to be less than a week in length, with what appears to be an exponential distribution out to 30 days. On the other hand, the synthetic inpatient LOS has a longer more gradual curve, with an increased number of single-day inpatient visits. With SNF claims, both the synthetic and real LOS appear to be exponential distributions with a majority of claims being 30 days or less, and a tail out to 100 days or more. The real SNF LOS also has rare outliers out to 300 days. For Hospice, both the synthetic and real LOS appear to be exponential distributions with most claims being 40 days or less, with tails out to 100 days or more.



Synthetic Inpatient Length of Stay

Real Inpatient Length of Stay

Synthetic SNF Length of Stay

Real SNF Length of Stay

Synthetic Hospice Length of Stay

Real Hospice Length of Stay

Figure 4-9. Inpatient, SNF and Hospice Length of Stay Use Cases

5 Use Case Examples

Based upon a literature review of frequent claims data research topics, comorbidity analysis and hospital readmissions were selected for additional analysis. In practice both use cases are often paired with a specific hypothesis, disease, or other study objective — but here we strip these use cases down to their most basic form to illustrate how the synthetic data might be used and to provide some insight into how synthetic claims compare with real claims.

5.1 Comorbidity Analysis

In this analysis we performed Association Rule Mining[21] to detect similar comorbidity relationships between synthetic data and a random sample of real data. To do this we looked at ICD-10 diagnosis codes for each beneficiary across all claim service types (e.g., inpatient, carrier, hospice, etc.). When we compared the resulting rules from the real and synthetic data there were few rules in common. One explanation for this is that the synthetic data will typically only use one specific code for a given diagnosis, while the real data often uses a multitude of very specific and granular ICD-10 diagnosis codes representing variation in a disease. For instance, the synthetic data will use one code for diabetes while the real data might have dozens of different diabetes codes indicating various stages of disease.

Therefore, in the next phase of our analysis we grouped ICD-10 diagnosis codes together according to CCW chronic condition definitions[22]. Once we grouped codes into chronic conditions, Association Rule Mining was able to detect more rules in common. What this analysis shows is that the synthetic data does capture some clinical comorbidity relationships present in the real data, but diagnosis codes may not match. The differences in diagnosis codes can be overcome using a method of grouping codes together such as the CCW Chronic Condition definitions.

The Top 10 results shown in Table 5-1 indicate that the synthetic data demonstrates realistic comorbidity relationships. For instance, the rule with the most confidence and support that was learned in both the real and synthetic data was row 1 which roughly translates into “patients with chronic kidney disease, ischemic heart disease, diabetes, hypertension are also likely to have anemia.”

[21] Association Mining Rule: https://en.wikipedia.org/wiki/Association_rule_learning#Confidence

[22] 30 CCW Chronic Conditions Algorithms. https://www2.ccwdata.org/documents/10280/19139421/chr-chronic-condition-algorithms.pdf

Table 5-1. Results of Comorbidity Analysis using Association Rule Mining

#	Antecedents	Consequents	Real Support	Real Confidence	Synthetic Support	Synthetic Confidence
1	(Chronic Kidney Disease, Ischemic Heart Disease, Diabetes, Hypertension)	(Anemia)	0.03	0.04	1.00	0.84
2	(Ischemic Heart Disease, Diabetes, Anemia, Hyperlipidemia)	(Chronic Kidney Disease)	0.02	0.04	1.00	0.44
3	(Chronic Kidney Disease, Ischemic Heart Disease, Anemia, Hyperlipidemia)	(Diabetes)	0.02	0.04	1.00	0.57
4	(Rheumatoid Arthritis/Osteoarthritis, Hyperlipidemia, Anemia)	(Hypertension)	0.02	0.04	1.00	0.74
5	(Chronic Kidney Disease, Ischemic Heart Disease, Hypertension, Hyperlipidemia)	(Anemia)	0.02	0.06	1.00	0.72
6	(Chronic Kidney Disease, Ischemic Heart Disease, Hypertension, Hyperlipidemia)	(Diabetes)	0.02	0.04	1.00	0.47
7	(Ischemic Heart Disease, Diabetes, Hypertension, Hyperlipidemia, Chronic Kidney Disease)	(Anemia)	0.02	0.04	1.00	0.85
8	(Ischemic Heart Disease, Diabetes, Hypertension, Hyperlipidemia, Anemia)	(Chronic Kidney Disease)	0.02	0.04	1.00	0.51
9	(Ischemic Heart Disease, Hypertension, Hyperlipidemia, Chronic Kidney Disease, Anemia)	(Diabetes)	0.02	0.04	1.00	0.56
10	(Chronic Kidney Disease, Ischemic Heart Disease, Hypertension, Hyperlipidemia)	(Diabetes, Anemia)	0.02	0.04	1.00	0.40

5.2 Hospital Readmissions Analysis

We conducted the preliminary steps of a 30-day hospital readmissions analysis using a simple feature selection algorithm. The objective of this analysis was not to predict readmissions directly, that could be done with further algorithm development, instead the goal was compare the relative importance of features across the real and synthetic data sets.

In this analysis we conducted feature selection to predict 30-day hospital readmissions based on Social Determinants of Health (SDOH), age, gender, and chronic conditions. The SDOH data were approximated in both data sets based on the beneficiary's county of residence using county-level data from the publicly available 2018 Social Vulnerability Index[23] and the 2021 County Health Rankings[24].

We used the Chi-Square test of independence, a measure of how related each feature (input variable) is to the value we want to predict (output variable), in this case whether the beneficiary had a 30-day hospital readmission. Typically, a feature selection process attempts to eliminate features that are identical to the prediction variable or are identical to another feature – but these steps were omitted to show the raw results. The results are listed in Table 5-2, where the features are sorted based on the Chi-Square score, the synthetic features are in the left-columns, and the real features are in the right-columns.

In Table 5-2, shows both the features that are related according to Chi-Square and those that are not. Features with no predictive power are marked with an "X" in the Result columns. Among both the synthetic and real data sets Chronic Kidney Disease, Age, Ischemic Heart Disease, and Diabetes were among the top-10 features. However, the Chi-Square scores for the synthetic data are much higher, suggesting that the same features in the synthetic data have a much stronger relation than they do in the reality.

However, there are noticeable weaknesses in the synthetic data as well. The highest-rated feature in the real data was Breast Cancer, but was insignificant and rated nearly last in the synthetic data. Conversely, the synthetic data highly scored features that were insignificant in the real data, such as Non-Alzheimer's Disease, Lung Cancer, and Prostate Cancer.

In both cases, SDOH data did not rank in the top-10 features, but Vehicle Access and Housing Cost burden both had low scores above a non-significant threshold. The low-rating is not a valid interpretation of how SDOH might affect real readmissions and is likely the result of how we randomly assigned the SDOH categories based on county-level. In which case, one interpretation of the results could be that any feature that falls below the SDOH data has little to no predictive power on its own.

Finally, keep in mind that this simple analysis only considered features independently and did not look at the predictive power of combinations of features. The goal was to show a comparison of possible features across the two data sets, and inform users building predictive readmission models which features they may want to select.

[23] Social Vulnerability Index. https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

[24] 2021 County Health Rankings. https://www.countyhealthrankings.org/sites/default/files/media/document/analytic_data2021.csv

Table 5-2. Results of Feature Analysis using Chi-Square

Synthetic Features	Score	P-Value	Result	Real Data Features	Score	P-Value	Result
Chronic Kidney Disease	1782.87	~= 0.0		Breast Cancer	721.51	< 0.0001	
Age	1713.58	~= 0.0		Chronic Kidney Disease	454.78	< 0.0001	
Ischemic Heart Disease	1037.37	< 0.0001		Anemia	377.20	< 0.0001	
Diabetes	1011.42	< 0.0001		Heart Failure and Non-Ischemic Heart Disease	356.69	< 0.0001	
Hyperlipidemia	759.03	< 0.0001		Age	334.91	< 0.0001	
Prostate Cancer	711.78	< 0.0001		Diabetes	212.61	< 0.0001	
Non-Alzheimer's Dementia	674.56	< 0.0001		Atrial Fibrillation and Flutter	178.96	< 0.0001	
Alzheimer's Disease	612.62	< 0.0001		Chronic Obstructive Pulmonary Disease	133.87	< 0.0001	
Atrial Fibrillation and Flutter	541.39	< 0.0001		Ischemic Heart Disease	119.09	< 0.0001	
Lung Cancer	404.25	< 0.0001		Depression, Bipolar, or Other Depressive Mood Disorders	79.12	< 0.0001	
Asthma	359.54	< 0.0001		Hyperlipidemia	52.41	< 0.0001	
Anemia	309.52	< 0.0001		Food Insecurity	42.57	< 0.0001	
Hypertension	262.58	< 0.0001		Acute Myocardial Infarction	33.15	< 0.0001	
Acute Myocardial Infarction	172.85	< 0.0001		Housing Cost Burden	26.53	< 0.0001	
Gender	170.48	< 0.0001		Vehicle Access	25.42	< 0.0001	
Chronic Obstructive Pulmonary Disease	159.39	< 0.0001		Pneumonia, All-cause	22.71	< 0.0001	
Osteoporosis	158.62	< 0.0001		Hypertension	19.71	< 0.0001	
Colorectal Cancer	64.07	< 0.0001		Gender	18.56	< 0.0001	
Vehicle Access	33.63	< 0.0001		Alzheimer's Disease	18.44	< 0.0001	
Hypothyroidism	21.62	< 0.0001		Hypothyroidism	17.63	< 0.0001	
Heart Failure and Non-Ischemic Heart Disease	20.86	< 0.0001		Non-Alzheimer's Dementia	9.60	0.002	X
Depression, Bipolar, or Other Depressive Mood Disorders	18.53	< 0.0001		Rheumatoid Arthritis and Osteoarthritis	5.18	0.023	X
Housing Cost Burden	15.25	< 0.0001		Lung Cancer	2.51	0.112	X
Rheumatoid Arthritis and Osteoarthritis	3.32	< 0.1	X	Colorectal Cancer	1.90	0.167	X
Food Insecurity	1.83	0.176	X	Prostate Cancer	0.62	0.430	X
Breast Cancer	0.54	0.459	X	Asthma	0.62	0.431	X
Pneumonia, All-cause	0.34	0.558	X	Osteoporosis	0.38	0.535	X

Features marked with a Result of “X” indicates no predictive power.

6 Fixed and Random Value Fields

Synthea does not model values for all the RIF file fields. In these cases, each field is assigned a fixed value, or a value randomly taken from a set of allowed values. Additional information can be found in the Random and Fixed Values section of the Synthea RIF Exporter Wiki page.[25] The following subsections show which fields in each file are handled this way. Where a value can be one from a set of allowed values, this is shown as a comma-separated list. Where the field is always empty, this is shown as [Blank].

6.1 Beneficiary

Table 6-1. Beneficiary

Name	Description	Value(s)
1. DUAL_ELGBL_MONS	Months of Dual Eligibility	0
2. ENHANCED_FIVE_PERCENT_FLAG	Enhanced Medicare 5% Sample Indicator	[Blank]
3. ENRL_SRC	Source of Enrollment Data	CME
4. HMO_IND_01 - HMO_IND_12	HMO indicator (1 – January … 12 – December)	[Blank]
5. BENE_HMO_CVRAGE_TOT_MONS	HMO Coverage Count	0
6. PTC_PLAN_TYPE_CD_01 - PTC_PLAN_TYPE_CD_12	Part C Plan Type Code – January to Part C Plan Type Code – December	[Blank]
7. SAMPLE_GROUP	Medicare 1, 5, or 20% strict sample group indicator	[Blank]
8. VALID_DEATH_DT_SW	Valid Date of Death Switch	[Blank]

[25] https://github.com/synthetichealth/synthea/wiki/CMS-BFD-RIF-Export#random-and-fixed-values

6.2 Inpatient

Table 6-2. Inpatient

Name	Description	Value(s)
1. AT_PHYSN_UPIN	Claim Attending Physician UPIN Number	[Blank]
2. BENE_LRD_USED_CNT	Beneficiary LRD Used Count	0
3. CLAIM_QUERY_CODE	Claim Query Code	3
4. CLM_E_POA_IND_SW2 – CLM_E_POA_IND_SW1	Claim Diagnosis E Code II Diagnosis Present on Admission Indicator Code to Claim Diagnosis E Code I Diagnosis Present on Admission Indicator Code	[Blank]
5. CLM_FAC_TYPE_CD	Claim Facility Type Code	1
6. CLM_FREQ_CD	Claim Frequency Code	1
7. CLM_MCO_PD_SW	Claim MCO Paid Switch	0
8. CLM_MDCR_NON_PMT_RSN_CD	Claim Medicare Non Payment Reason Code	[Blank]
9. CLM_NON_UTLZTN_DAYS_CNT	Claim Non Utilization Days Count	0
10. CLM_PASS_THRU_PER_DIEM_AMT	Claim Pass Thru Per Diem Amt	10
11. CLM_PPS_CPTL_DRG_WT_NUM	Claim PPS Capital DRG Weight Number	0
12. CLM_PPS_CPTL_DSPRPRTNT_SHR_AMT	Claim PPS Capital Disproportionate Share Amt	0
13. CLM_PPS_CPTL_EXCPTN_AMT	Claim PPS Capital Exception Amt	0
14. CLM_PPS_CPTL_FSP_AMT	Claim PPS Capital FSP Amt	0
15. CLM_PPS_CPTL_IME_AMT	Claim PPS Capital IME Amt	0
16. CLM_PPS_CPTL_OUTLIER_AMT	Claim PPS Capital Outlier Amt	0
17. CLM_PPS_IND_CD	Claim PPS Indicator Code	[Blank]
18. CLM_PPS_OLD_CPTL_HLD_HRMLS_AMT	Claim PPS Old Capital Hold Harmless Amt	0
19. CLM_SRC_IP_ADMSN_CD	Claim Source Inpatient Admission Code	1, 2, 4, 5
20. CLM_SRVC_CLSFCTN_TYPE_CD	Claim Service Classification Type Code	1
21. CLM_TOT_PPS_CPTL_AMT	Claim Total PPS Capital Amt	0
22. DSH_OP_CLM_VAL_AMT	Operating Disproportionate Share Amount	0
23. FI_CLM_ACTN_CD	FI Claim Action Code	[Blank]
24. FI_CLM_PROC_DT	FI Claim Process Date	[Blank]
25. FI_NUM	FI Number	[Blank]
26. ICD_DGNS_E_CD2 – ICD_DGNS_E_CD12	Claim Diagnosis E Code II	[Blank]
27. IME_OP_CLM_VAL_AMT	Operating Indirect Medical Education (IME) Amount	0
28. NCH_ACTV_OR_CVRD_LVL_CARE_THRU	NCH Active or Covered Level Care Thru Date	[Blank]
29. NCH_BENE_BLOOD_DDCTBL_LBLTY_AM	NCH Beneficiary Blood Deductible Liability Amt	0
30. NCH_BENE_MDCR_BNFTS_EXHTD_DT_I	NCH Beneficiary Medicare Benefits Exhausted Date	[Blank]
31. NCH_BLOOD_PNTS_FRNSHD_QTY	NCH Blood Pints Furnished Quantity	0
32. NCH_CLM_TYPE_CD	NCH Claim Type Code	60

Name	Description	Value(s)
33. NCH_DRG_OUTLIER_APRVD_PMT_AMT	NCH DRG Outlier Approved Payment Amt	0
34. NCH_NEAR_LINE_REC_IDENT_CD	NCH Near Line Record Identification Code	V
35. NCH_PRMRY_PYR_CD	NCH Primary Payer Code	[Blank]
36. NCH_PROFNL_CMPNT_CHRG_AMT	NCH Professional Component Charge	4
37. NCH_VRFD_NCVRD_STAY_FROM_DT	NCH Verified Noncovered Stay from Date	[Blank]
38. NCH_VRFD_NCVRD_STAY_THRU_DT	NCH Verified Noncovered Stay Through Date	[Blank]
39. OP_PHYSN_UPIN	Claim Operating Physician UPIN Number	[Blank]
40. OT_PHYSN_NPI	Claim Other Physician NPI Number	[Blank]
41. OT_PHYSN_UPIN	Claim Other Physician UPIN Number	[Blank]
42. RNDRNG_PHYSN_UPIN	Revenue Center Rendering Physician UPIN	[Blank]

6.3 Outpatient

Table 6-3. Outpatient

Name	Description	Value(s)
1. AT_PHYSN_UPIN	Claim Attending Physician UPIN Number	[Blank]
2. CLAIM_QUERY_CODE	Claim Query Code	3
3. CLM_FAC_TYPE_CD	Claim Facility Type Code	1
4. CLM_FREQ_CD	Claim Frequency Code	1
5. CLM_MCO_PD_SW	Claim MCO Paid Switch	0
6. **CLM_MDCR_NON_PMT_RSN_CD**	Claim Medicare Non Payment Reason Code	[Blank]
7. CLM_OP_BENE_PMT_AMT	Claim Outpatient Beneficiary Payment Amount	0
8. CLM_SRVC_CLSFCTN_TYPE_CD	Claim Service Classification Type Code	3
9. FI_CLM_PROC_DT	FI Claim Process Date	[Blank]
10. FI_NUM	FI Number	[Blank]
11. HCPCS_1ST_MDFR_CD - HCPCS_2ND_MDFR_CD	Revenue Center HCPCS Initial Modifier Code to Revenue Center HCPCS Second Modifier Code	[Blank]
12. ICD_DGNS_E_CD2 - ICD_DGNS_E_CD12	Claim Diagnosis E Code II to Claim Diagnosis E Code XII	[Blank]
13. NCH_BENE_BLOOD_DDCTBL_LBLTY_AM	NCH Beneficiary Blood Deductible Liability Amount	0
14. NCH_BENE_PTB_COINSRNC_AMT	NCH Beneficiary Part B Coinsurance Amount	0, 10, 20
15. NCH_CLM_TYPE_CD	NCH Claim Type Code	40
16. NCH_NEAR_LINE_REC_IDENT_CD	NCH Near Line Record Identification Code	W
17. NCH_PRMRY_PYR_CD	NCH Primary Payer Code	[Blank]
18. NCH_PROFNL_CMPNT_CHRG_AMT	NCH Professional Component Charge	4
19. OP_PHYSN_UPIN	Claim Operating Physician UPIN Number	[Blank]
20. OT_PHYSN_NPI	Claim Other Physician NPI Number	[Blank]
21. OT_PHYSN_UPIN	Claim Other Physician UPIN Number	[Blank]
22. REV_CNTR_1ST_ANSI_CD - REV_CNTR_4TH_ANSI_CD	Revenue Center 1st ANSI Code to Revenue Center 4th ANSI Code	[Blank]
23. REV_CNTR_1ST_MSP_PD_AMT - REV_CNTR_2ND_MSP_PD_AMT	Revenue Center 1st Medicare Secondary Payer Paid Amount to Revenue Center 2nd Medicare Secondary Payer Paid Amount	0
24. REV_CNTR_APC_HIPPS_CD	Revenue Center APC/HIPPS	[Blank]
25. **REV_CNTR_BENE_PMT_AMT**	Revenue Center Beneficiary Payment Amount	0
26. REV_CNTR_BLOOD_DDCTBL_AMT	Revenue Center Blood Deductible Amount	0
27. **REV_CNTR_DSCNT_IND_CD**	Revenue Center Discount Indicator Code	[Blank]

Name	Description	Value(s)
28. REV_CNTR_OTAF_PMT_CD	Revenue Center Obligation to Accept as Full (OTAF) Payment Code	[Blank]
29. **REV_CNTR_PACKG_IND_CD**	Revenue Center Packaging Indicator Code	[Blank]
30. **REV_CNTR_PMT_MTHD_IND_CD**	Revenue Center Payment Method Indicator Code	4
31. **REV_CNTR_STUS_IND_CD**	Revenue Center Status Indicator Code	4
32. REV_CNTR_UNIT_CNT	Revenue Center Unit Count	1
33. RNDRNG_PHYSN_UPIN	Revenue Center Rendering Physician UPIN	[Blank]
34. RSN_VISIT_CD1 - RSN_VISIT_CD3	Reason for Visit Diagnosis Code I to Reason for Visit Diagnosis Code III	[Blank]

6.4 Carrier

Table 6-4. Carrier

Name	Description	Value(s)
1. CARR_CLM_ENTRY_CD	Carrier Claim Entry Code	1
2. CARR_CLM_HCPCS_YR_CD	Carrier Claim HCPCS Year Code	1
3. CARR_CLM_PMT_DNL_CD	Carrier Claim Payment Denial Code	1
4. CARR_CLM_PRVDR_ASGNMT_IND_SW	Carrier Claim Provider Assignment Indicator Switch	A
5. CARR_LINE_ANSTHSA_UNIT_CNT	Carrier Line Anesthesia Unit Count	0, 1
6. CARR_LINE_MTUS_CD	Carrier Line Miles/Time/Units/Services Indicator Code	[Blank]
7. CARR_LINE_PRVDR_TYPE_CD	Carrier Line Provider Type Code	0
8. CARR_LINE_RDCD_PMT_PHYS_ASTN_C	Carrier Line Reduced Payment Physician Assistant Code	0
9. CARR_LINE_RX_NUM	Carrier Line RX Number	[Blank]
10. CLM_CLNCL_TRIL_NUM	Clinical Trial Number	[Blank]
11. CLM_DISP_CD	Claim Disposition Code	1
12. HCPCS_1ST_MDFR_CD - HCPCS_2ND_MDFR_CD	Line HCPCS Initial Modifier Code	[Blank]
13. HPSA_SCRCTY_IND_CD	Carrier Line HPSA/Scarcity Indicator Code	[Blank]
14. LINE_BENE_PMT_AMT	Line Beneficiary Payment Amount	0
15. LINE_BENE_PRMRY_PYR_CD	Line Beneficiary Primary Payer Code	[Blank]
16. LINE_BENE_PRMRY_PYR_PD_AMT	Line Beneficiary Primary Payer Paid Amount	0
17. LINE_CMS_TYPE_SRVC_CD	Line HCFA Type Service Code	1
18. LINE_HCT_HGB_TYPE_CD	Hematocrit/Hemoglobin Test Type Code	R1
19. LINE_ICD_DGNS_VRSN_CD	Line Diagnosis Code Diagnosis Version Code (ICD-9 or ICD-10)	0
20. LINE_PMT_80_100_CD	Line Payment 80%/100% Code	[Blank]
21. LINE_PRCSG_IND_CD	Line Processing Indicator Code	A
22. LINE_SERVICE_DEDUCTIBLE	Line Service Deductible Indicator Switch	[Blank]
23. NCH_CLM_BENE_PMT_AMT	NCH Claim Beneficiary Payment Amount	0
24. NCH_CLM_TYPE_CD	NCH Claim Type Code	71
25. NCH_NEAR_LINE_REC_IDENT_CD	NCH Near Line Record Identification Code	O
26. LINE_REC_IDENT_CD	Line Provider Payment Amount	O
27. PRF_PHYSN_UPIN	Carrier Line Performing UPIN Number	[Blank]
28. PRNCPAL_DGNS_VRSN_CD: 0	Primary Claim Diagnosis Code Diagnosis Version Code (ICD-9 or ICD-10)	0
29. PRTCPTNG_IND_CD	Line Provider Participating Indicator Code	1, 2, 3, 4, 5, 6, 7

6.5 Part D

Table 6-5. Part D

	Description	Value(s)
1. ADJSTMT_DLTN_CD	Adjustment Deletion Code	[Blank]
2. BRND_GNRC_CD	The Brand-Generic Code Reported by the Submitting Plan	B, G
3. CMPND_CD	Compound Code	0
4. CTSTRPHC_CVRG_CD	Catastrophic Coverage Code	[Blank]
5. DRUG_CVRG_STUS_CD	Drug Coverage Status Code	C
6. DSPNSNG_STUS_CD	Dispensing Status Code	[Blank]
7. LICS_AMT	Low Income Cost Sharing Subsidy Amount (LICS)	0
8. NSTD_FRMT_CD	Non-Standard Format Code	[Blank]
9. OTHR_TROOP_AMT	Other Troop Amount	0
10. PLAN_PBP_REC_NUM	Plan PBP Record Number	999
11. PLRO_AMT	Patient Liability Reduction Due to Other Payer Amount (PLRO)	0
12. PRCNG_EXCPTN_CD	Pricing Exception Code	[Blank]
13. PRSCRBR_ID_QLFYR_CD	Prescriber ID Qualifier Code	01
14. RPTD_GAP_DSCNT_NUM	Gap Discount Amount Reported by the Submitting Plan	0
15. RX_ORGN_CD	Prescription Origin Code	0, 3, 4
16. SUBMSN_CLR_CD	Submission Clarification Code	[Blank]

6.6 Durable Medical Equipment

Table 6-6. Durable Medical Equipment

Name	Description	Value(s)
1. CARR_CLM_ENTRY_CD	Carrier Claim Entry Code	1
2. CARR_CLM_HCPCS_YR_CD	Carrier Claim HCPCS Year Code	1
3. CARR_CLM_PMT_DNL_CD	Carrier Claim Payment Denial Code	1
4. CARR_CLM_PRVDR_ASGNMT_IND_SW	Claim Provider Assignment Indicator Switch	A
5. CLM_CLNCL_TRIL_NUM	Clinical Trial Number	[Blank]
6. CLM_DISP_CD	Claim Disposition Code	1
7. DMERC_LINE_MTUS_CD	DMERC Line Miles/Time/ Units/Services Indicator Code	0
8. DMERC_LINE_SCRN_SVGS_AMT	DMERC Line Screen Savings Amount	0
9. DMERC_LINE_SUPPLR_TYPE_CD	DMERC Line Supplier Type Code	0,1,2,3,4,5,6,7,8
10. HCPCS_1ST_MDFR_CD - HCPCS_4TH_MDFR_CD	Line HCPCS Initial Modifier Code to DMERC Line HCPCS Fourth Modifier Code	[Blank]
11. LINE_BENE_PRMRY_PYR_CD	Line Beneficiary Primary Payer Code	[Blank]
12. LINE_BENE_PRMRY_PYR_PD_AMT	Line Beneficiary Primary Payer Paid Amount	1
13. LINE_DME_PRCHS_PRICE_AMT	Line DME Purchase Price Amount	[Blank]
14. LINE_HCT_HGB_TYPE_CD	Hematocrit/Hemoglobin Test Type code	R1
15. LINE_ICD_DGNS_VRSN_CD	Line Diagnosis Code Diagnosis Version Code (ICD-9 or ICD-10)	0
16. LINE_NDC_CD	Line National Drug Code	[Blank]
17. LINE_PMT_80_100_CD	Line Payment 80%/100% Code	[Blank]
18. LINE_PRCSG_IND_CD	Line Processing Indicator Code	A
19. LINE_SERVICE_DEDUCTIBLE	Line Service Deductible Indicator Switch	[Blank]
20. NCH_CLM_BENE_PMT_AMT	NCH Claim Beneficiary Payment Amount	0
21. NCH_CLM_TYPE_CD	NCH Claim Type Code	82
22. NCH_NEAR_LINE_REC_IDENT_CD	NCH Near Line Record Identification Code	M
23. PRNCPAL_DGNS_VRSN_CD	Primary Claim Diagnosis Code Diagnosis Version Code (ICD-9 or ICD-10)	0
24. PRTCPTNG_IND_CD	Line Provider Participating Indicator Code	1,2,3,4,5,6,7

6.7 Home Health Agency

Table 6-7. Home Health Agency

Name	Description	Value(s)
1. AT_PHYSN_UPIN	Claim Attending Physician UPIN Number	[Blank]
2. CLM_FAC_TYPE_CD	Claim Facility Type Code	3
3. CLM_FREQ_CD	Claim Frequency Code	1, 9
4. CLM_HHA_LUPA_IND_CD	Claim HHA Low Utilization Payment Adjustment (LUPA) Indicator Code	[Blank]
5. CLM_HHA_RFRL_CD	Claim HHA Referral Code	[Blank]
6. CLM_MDCR_NON_PMT_RSN_CD	Claim Medicare Non-Payment Reason Code	[Blank]
7. CLM_PPS_IND_CD	Claim PPS Indicator Code	[Blank]
8. CLM_SRVC_CLSFCTN_TYPE_CD	Claim Service Classification Type Code	3
9. FI_CLM_PROC_DT	FI Claim Process Date	[Blank]
10. FI_NUM	FI Number	[Blank]
11. HCPCS_1ST_MDFR_CD – HCPCS_2ND_MDFR_CD	Revenue Center HCPCS Initial Modifier Code to Revenue Center HCPCS Second Modifier Code	[Blank]
12. ICD_DGNS_E_CD2 – ICD_DGNS_E_CD12	Claim Diagnosis E Code II to Claim Diagnosis E Code XII	[Blank]
13. NCH_CLM_TYPE_CD	NCH Claim Type Code	10
14. NCH_NEAR_LINE_REC_IDENT_CD	NCH Near Line Record Identification Code	V, W, U
15. NCH_PRMRY_PYR_CD	NCH Primary Payer Code	[Blank]
16. REV_CNTR_1ST_ANSI_CD	Revenue Center 1st ANSI Code	[Blank]
17. REV_CNTR_APC_HIPPS_CD	Revenue Center APC/HIPPS	[Blank]
18. REV_CNTR_PMT_MTHD_IND_CD	Revenue Center Payment Method Indicator Code	4
19. REV_CNTR_STUS_IND_CD	Revenue Center Status Indicator Code	4
20. RNDRNG_PHYSN_UPIN	Revenue Center Rendering Physician UPIN	[Blank]

6.8 Hospice

Table 6-8. Hospice

Name	Description	Value(s)
1. AT_PHYSN_UPIN	Claim Attending Physician UPIN Number	[Blank]
2. BENE_HOSPC_PRD_CNT	Beneficiary's Hospice Period Count	[Blank]
3. CLM_FAC_TYPE_CD	Claim Facility Type Code	8
4. CLM_FREQ_CD	Claim Frequency Code	1, 9
5. CLM_MDCR_NON_PMT_RSN_CD	Claim Medicare Non-Payment Reason Code	[Blank]
6. CLM_SRVC_CLSFCTN_TYPE_CD	Claim Service Classification Type Code	1
7. FI_CLM_PROC_DT	FI Claim Process Date	[Blank]
8. FI_NUM	FI Number	[Blank]
9. HCPCS_1ST_MDFR_CD - HCPCS_2ND_MDFR_CD	Revenue Center HCPCS Initial Modifier Code to Revenue Center HCPCS Second Modifier Code	[Blank]
10. ICD_DGNS_E_CD2 - ICD_DGNS_E_CD12	Claim Diagnosis E Code II to Claim Diagnosis E Code XII	[Blank]
11. NCH_CLM_TYPE_CD	NCH Claim Type Code	50
12. NCH_NEAR_LINE_REC_IDENT_CD	NCH Near Line Record Identification Code	V
13. NCH_PRMRY_PYR_CD	NCH Primary Payer Code	[Blank]
14. REV_CNTR_BENE_PMT_AMT	Revenue Center Beneficiary Payment Amount	0
15. RNDRNG_PHYSN_UPIN	Revenue Center Rendering Physician UPIN	[Blank]

6.9 Skilled Nursing Facility

Table 6-9. Skilled Nursing Facility

Name	Description	Value(s)
1. AT_PHYSN_UPIN	Claim Attending Physician UPIN Number	[Blank]
2. CLAIM_QUERY_CODE	Claim Query Code	3
3. CLM_FAC_TYPE_CD	Claim Facility Type Code	2
4. CLM_FREQ_CD	Claim Frequency Code	1,9
5. CLM_MCO_PD_SW	Claim MCO Paid Switch	0
6. CLM_MDCR_NON_PMT_RSN_CD	Claim Medicare Non Payment Reason Code	[Blank]
7. CLM_NON_UTLZTN_DAYS_CNT	Claim Non Utilization Days Count	0
8. CLM_PPS_CPTL_DSPRPRTNT_SHR_AMT	Claim PPS Capital Disproportionate Share Amount	0
9. CLM_PPS_CPTL_EXCPTN_AMT	Claim PPS Capital Exception Amount	0
10. CLM_PPS_CPTL_FSP_AMT	Claim PPS Capital FSP Amount	0
11. CLM_PPS_CPTL_IME_AMT	Claim PPS Capital IME Amount	0
12. CLM_PPS_CPTL_OUTLIER_AMT	Claim PPS Capital Outlier Amount	0
13. CLM_PPS_IND_CD	Claim PPS Indicator Code	[Blank]
14. CLM_PPS_OLD_CPTL_HLD_HRMLS_AMT	Claim PPS Old Capital Hold Harmless Amount	0
15. CLM_SRC_IP_ADMSN_CD	Claim Source Inpatient Admission Code	1,2,4
16. CLM_SRVC_CLSFCTN_TYPE_CD	Claim Service classification Type Code	1
17. FI_CLM_ACTN_CD	FI Claim Action Code	[Blank]
18. FI_CLM_PROC_DT	FI Claim Process Date	[Blank]
19. FI_NUM	FI Number	[Blank]
20. ICD_DGNS_E_CD2 - ICD_DGNS_E_CD12	Claim Diagnosis E Code II to Claim Diagnosis E Code XII	[Blank]
21. NCH_ACTV_OR_CVRD_LVL_CARE_THRU	NCH Active or Covered Level Care Thru Date	[Blank]
22. NCH_BENE_BLOOD_DDCTBL_LBLTY_AM	NCH Beneficiary Blood Deductible Liability Amount	0
23. NCH_BENE_MDCR_BNFTS_EXHTD_DT_I	NCH Beneficiary Medicare Benefits Exhausted Date	[Blank]
24. NCH_BLOOD_PNTS_FRNSHD_QTY	NCH Blood Pints Furnished Quantity	0
25. NCH_CLM_TYPE_CD	NCH Claim Type Code	20
26. NCH_NEAR_LINE_REC_IDENT_CD	NCH Near Line Record Identification Code	V
27. NCH_PRMRY_PYR_CD	NCH Primary Payer Code	[Blank]
28. NCH_QLFYD_STAY_FROM_DT	NCH Qualified Stay from Date	[Blank]
29. NCH_QLFYD_STAY_THRU_DT	NCH Qualify Stay Through Date	[Blank]
30. NCH_VRFD_NCVRD_STAY_FROM_DT	NCH Verified Noncovered Stay from Date	[Blank]
31. NCH_VRFD_NCVRD_STAY_THRU_DT	NCH Verified Noncovered Stay Through Date	Blank]
32. OP_PHYSN_UPIN	Claim Operating Physician UPIN Number	[Blank]
33. OT_PHYSN_NPI	Claim Other Physician NPI Number	[Blank]
34. OT_PHYSN_UPIN	Claim Other Physician UPIN Number	[Blank]
35. RNDRNG_PHYSN_UPIN	Revenue Center Rendering Physician UPIN	[Blank]

7 List of Acronyms

Table 7-1. List of Acronyms

ANSI	American National Standards Institute
BFD	Beneficiary FHIR Data
CCW	Chronic Conditions Data Warehouse
CDM	Common Data Model
CMS	Centers for Medicare & Medicaid Services
CSV	Comma-Separated Values
DME	Durable Medical Equipment
ESRD	End Stage Renal Disease
FFS	Fee-for-Service
FHIR	Fast Healthcare Interoperability Resources
FI	Fiscal Intermediaries
FSP	Federal Specific Portion
HAPI	Http Application Programming Interface
HCFA	Health Care Financing Administration
HCPCS	Healthcare Common Procedure Coding System
HHA	Home Health Agency
HL7	Health Level 7
HMO	Health Maintenance Organization
ICD-10	International Classification of Diseases, Tenth Revision
ICD-10-CM	International Classification of Diseases, Tenth Revision, Clinical Modification
ICD-9	International Classification of Diseases, Ninth Revision
IME	Indirect Medical Education
JSON	JavaScript Object Notation
LICS	Low-Income Cost-Sharing Subsidy
LOS	Length of Stay
LRO	Lead Regional Office
LUPA	Low Utilization Payment Adjustment
MBSF	Medicare Beneficiary Summary File
MCO	Managed Care Organization
NCH	National Claims History
NLM	National Library of Medicine
OMOP	Observational Medical Outcomes Partnership
OTAF	Obligated to Accept Field
PDE	Prescription Drug Event
PPS	Prospective Payment System
PUF	Public Use File
RIF	Research Identifiable File
UPIN	Unique Physician Identification Number
SDOH	Social Determinants of Health
SNF	Skilled Nursing Facility
SNOMED-CT	Systematized Nomenclature of Medicine-Clinical Terms

8 Appendix

8.1 Clinical Disease Modules

The Synthea Generic Module Framework[26] allows for the creation of simulation models representing the progression and standards of care for common diseases from a set of predefined states, transition probabilities, and conditional logic. The models are based on publicly available health data, including disease incidence and prevalence statistics sourced from CDC, NIH, and peer-reviewed literature, and clinical practice guidelines sourced from clinical specialty societies or peer-reviewed literature.

Table 8-1 shows the list of current disease modules with hyperlinks to their static diagrams in the Synthea Project Wiki – Module Gallery.[27]

Table 8-1. Subset of Current Disease Modules

Allergic Rhinitis	COPD	Injuries	Rheumatoid-Arthritis
Allergies	Dementia	Lung Cancer	Self-Harm
Appendicitis	Dermatitis	Lupus	Sexual Activity
Asthma	Ear Infections	Med Rec	Sinusitis
Atopy	Epilepsy	Metabolic Syndrome Care	Sore Throat
Attention Deficit Disorder	Female Reproduction	Metabolic Syndrome Disease	Total Joint Replacement
Bronchitis	Fibromyalgia	Opioid Addiction	Urinary Tract Infection
Colorectal Cancer	Food Allergies	Osteoarthritis	Wellness Encounters
Contraceptives	Gout	Osteoporosis	
Contraceptive Maintenance	Homelessness	Pregnancy	

Disease modules either have (a) Companion Guides[28] that provide additional information on the scope, intent, and state transition logic of the module, or (b) include citations and references used in building the module. Some modules have Jupyter Notebooks that present basic analysis of the resulting data in the module validation repository.[29]

Typically, each module is developed in a cycle, starting with clinical research and design of the model, construction of the module, execution of the module to produce a dataset, and analysis of the resulting data, and if necessary, fine-tuning the module to replicate important statistics or features according to the research. The calibration and validation processes are iterative and undergo clinical review but are often limited by inaccessibility to real-world patient-level data. However, population-level aggregate statistical data can be utilized to perform validation of the

[26] Synthea Generic Module Framework. https://github.com/synthetichealth/synthea/wiki/Generic-Module-Framework

[27] Synthea Project Wiki – Module Gallery. https://github.com/synthetichealth/synthea/wiki/Module-Gallery

[28] Synthea Project Wiki – Module Companion Guide. https://github.com/synthetichealth/synthea/wiki/Module-Companion-Guides

[29] Synthea Project Wiki – Module Validation. https://github.com/synthetichealth/module-validation

modeling results that yield realistic synthetic datasets which are iteratively tuned to improve degrees of accuracy.[30]

8.2 Exporting Synthetic Electronic Health Records

Once a simulated patient dies or the simulation reaches the specified end date, that synthetic patient record can be exported to a variety of standard and ad-hoc data formats. Each supported data format is supported by a dedicated code within the Synthea code structure that maps from the internal Synthea data model into whatever is required for a given format including:

- Health Level 7 (HL7®) Fast Healthcare Interoperability Resources (FHIR®[31]
- Comma-separated values (CSV)
- JavaScript Object Notation (JSON)
- Beneficiary FHIR Data (BFD) Research Identifiable Files (RIF)
- Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)

Please refer to the provided hyperlinks for details on each of the foregoing formats.

Synthea generates HL7 FHIR records using the HAPI FHIR[32] library to generate a FHIR bundle for each patient. For example, the FHIR R4 exporter maps a Synthea health record procedure[33] object into a FHIR procedure[34] resources.

Mappings from the internal Synthea data model can be simple or complex depending on the target data format. A variety of transformations may be required such as mapping from one code system to another, creating derived data by performing calculations on Synthea data elements, and applying filters to Synthea data that ensure exported data are appropriate for the target data format.

[30] Walonoski, Jason, et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record." *Journal of the American Medical Informatics Association*, 25(3) (Jul 2018): 230-238. https://doi.org/10.1093/jamia/ocx079

[31] FHIR® is the registered trademark of Health Level Seven International (HL7).

[32] HAPI FHIR Library. https://hapifhir.io

[33] Synthea Health Record Procedure. https://github.com/synthetichealth/synthea/blob/master/src/main/java/org/mitre/synthea/world/concepts/HealthRecord.java

[34] FHIR R4 Resource Procedure. https://www.hl7.org/fhir/procedure.html