Synthetic Healthcare Database for Research – SyH-DR

A Synthetic Nationally Representative All-Payer Claims Database

SAMPLING, WEIGHTING, AND SYNTHETIZATION METHODOLOGIES

AHRQ Publication No. 22-0039-2-EF
May 2022

AHRQ
Agency for Healthcare Research and Quality

TABLE OF CONTENTS

SAMPLING DESIGN

Purpose and Overview of Sampling Design

The purpose of sampling for this database was to create a relatively compact, nationally representative dataset of healthcare enrollees while providing sufficient analytic capacity to meet the analytic needs of researchers at granular levels such as race/ethnicity, sex, age group, and insurance source within each state. There were two primary goals of sampling. First, since the SyH-DR source data for each payer covered different proportions of their respective populations, the source data was sampled so that the SyH-DR has a roughly equal proportion of persons covered by each payer, with respect to the populations covered by the given payer. Second, certain subgroups were oversampled so that typically rare subgroups would have sufficient sample size for analysis, providing maximum analytic utility for researchers

Sampling was performed independently for each payer.

Sampling Frame

The sampling frames consisted of the following files:

- **2016 Medicare Sample Enrollment File**. The Centers for Medicare & Medicaid Services (CMS) provided a Medicare sample file that was based on a simple random sample with a single sampling rate of 20%.[1]

- **2016 Medicaid Enrollment File**. CMS provided a file that included all Medicaid enrollees.

- **2016 Commercial Sample Enrollment File**. Commercial insurance plans provided a 20% simple random sample of commercial data covering about 30% of the commercially insured population, so the file contained about 6% of the population.

One goal of the sampling was to create a database where the data for each payer represents a roughly equal proportion of persons covered by that payer in the population. The overall sampling rate used was 6% because the commercial data source had data for the lowest percentage of the population, 6%.

Table 1 provides, for each payer, the estimated number of persons in the entire U.S. population, the number of persons in the source files, and the final sample size in the SyH-DR. The numbers in column (2) represent the number of persons in the source files. Column (1) is the

[1] Specifically, the 5% random sample is created using the following method: “The 5% random sample consists of people who had a Medicare HIC number equal to the Claim Account Number (CAN) plus Beneficiary Identity Code (BIC) (HIC=CAN+BIC) where the last two digits of the CAN are in the set {05, 20, 45, 70, 95}.” The 20% random sample is created using the same method but uses the values of the last two digits of the CAN so that a 20% random sample is chosen.

Source: Centers for Medicare & Medicaid Services, Chronic Conditions Data Warehouse. (2019). *CCW Medicare administrative data user guide* (Version 3.6). https://www.ccwdata.org/documents/10280/19002246/ccw-medicare-data-user-guide.pdf

estimated population with insurance coverage for that payer. Column (3) represents the final number of persons in the SyH-DR for each payer.

Table 1: Population, Sample Frame, and Sample Sizes by Payer

Source	(Estimated) Population Size (1)	Source File Size (2)	Sample Size (3)
Medicare	60,785,720	12,157,144	3,570,105
Medicaid	97,782,330	97,782,330	5,771,393
Commercial	179,952,463	9,573,472	9,494,289
Total	338,520,513	119,512,946	18,835,787

Implications of Data Quality in the Medicaid File

Due to data quality issues in the Transformed Medicaid Statistical Information System (T-MSIS) Medicaid data, several criteria were applied to the sampling frame:

- Excluded persons with missing or invalid sex or age values, where invalid age was negative or non-numeric and invalid sex was any value other than “M” or “F.”

- Excluded any persons not residing in the 50 states or the District of Columbia.

Precision Goals for Sampling

The general strategy for sampling is to produce a nationally representative dataset that enables sufficient precision estimates for the domains of interest. Domains of interest are based on state-level, age, race and ethnicity, and coverage qualification reason categories. For each payer (Medicaid, Medicare, and commercial), domains were established by crossing the submitting state with three characteristics classifications: age category (0–18, 19–64, 65+), race/ethnicity category (“non-Hispanic Black,” “Hispanic,” “Other”), and reason for coverage category (specified in more detail below). A target sample size of 1,000 persons was chosen for domain cells because this is a reasonable and standard minimum sample size that researchers use for analysis.

The precision goals for sampling were to produce a 3.1% margin of error of the 95% confidence interval for a proportion of 50% for each domain. An example of a proportion of 50% for a domain might be “50% of Hispanics aged 18–64 in New Mexico have used prescriptions.” If the sample size for Hispanics aged 18–64 in New Mexico is larger than 1,000, the 95% confidence interval for this proportion would be narrower than 46.9% to 53.1%. If the sample size is met at

a proportion of 50% where confidence intervals are the widest, then it will also be met at any other proportion where confidence intervals are tighter.

With the overall 6% sampling rate, this means that as long as the domain-level population size is greater than 16,700, a stratified sample design with an equal selection probability will yield a sample that would meet the targeted precision goals. There are, however, domains with population sizes less than 16,700. For these domains, oversampling was performed to meet the minimum sample size threshold of 1,000, while ensuring that the sampling rate was not too high. The highest sampling rate for the sample design was set as 20%. With this plan, we had at least 1,000 people sampled from each key domain where population sizes are equal to or higher than 5,000. For the remaining small domains, although we selected a sample with a sampling rate of 20%, we do not expect that the sample would support domain-level analysis for those domains with population size less than 5,000.

Sampling Procedures

For the Medicare and Medicaid databases, the sampling frame was stratified by state and sorted by ZIP code, race/ethnicity, age group, reason for coverage, and sex.

An overall 6% sampling rate was applied to all three payers' populations. Because the source Medicare file represented 20% of the Medicare population, the 30% subsample rate was applied to the source Medicare file (6% divided by 20%, which equals 30%). For Medicaid, the source data represented the entire population of beneficiaries, so the sample rate of 6% was applied to the source Medicaid file. The source files from the commercial insurance plans already represented about 6% of the population, so it was not sampled further.

Sample Selection

The process for determining the sampling rate by domain was as follows: First, an equal systematic sample of the source files was extracted to produce subsamples with 6% of the population. As stated before, this translates to a 30% sampling rate for Medicare and a 6% sampling rate for Medicaid. Systematic samples were drawn from the Medicare and Medicaid extract at these rates. Strata consist of all records that shared the same submitting state. Sort order is determined by age category, race/ethnicity, eligibility group, and sex; and the sequencing is serpentine. Serpentine sorting reverses the sort order as each boundary is crossed for higher-level sorting data elements, thus helping ensure that adjacent records are similar with respect to as many sorting data elements as possible. For example, if three data elements each with three categories (Low, Medium, High) are used for sorting, then the resulting order would be that shown in **Table 2**. With this sorting procedure, a resultant sample would maintain the properties of a stratified sampling, homogeneous within each stratum, with sorting data elements as implicit stratification data elements.

Table 2: Example of Serpentine Sorting

Data element 1	Data element 2	Data element 3
Low	Low	Low
		Medium
		High
	Medium	High
		Medium
		Low
	High	Low
		Medium
		High
Medium	High	High
		Medium
		Low
	Medium	Low
		Medium
		High
	Low	High
		Medium

		Low
High	Low	Low
		Medium
		High
	Medium	High
		Medium
		Low
	High	Low
		Medium
		High

From this initial sample, tabulations of state crossed individually with three domain classifiers—age group, race/ethnicity, and coverage qualification reason—were generated. For Medicare and Medicaid, three age groups were used: 0–18, 19–64, 65 and older. Also, for Medicare and Medicaid, race/ethnicity was organized as “Hispanic,” “Black (non-Hispanic),” and “Other,” as derived from race codes in the source files. For Medicare, the coverage qualification reasons were “End-stage renal disease (ESRD),” “Disability without ESRD,” and “Old age.” For Medicaid, the coverage qualification reasons were “Child,” “Adult,” “Disabled,” “Aged,” “Expansion,” or “Other,” and these were determined based on a mapping of the reason for enrollment code data element. Next, the population size was computed for each of these domains (for Medicare, this was an estimate from the extract, i.e., 5 × 20% source file sample size) and the initial sample size. For each crossing of state and domain category, we computed the measure of size (MOS) as:

(Medicaid) If $n_h < 1000$, $MOS_t = \min \{(1000 / N_h) / .06, .20 / .06\}$

(Medicare) If $n_h < 1000$, $MOS_t = \min \{(1000 / N_h) / .3, .20 / .3\}$

Otherwise (i.e., $n_h \geq 1000$), $MOS_t = 1$,

MOS_t – Measure of size for domain type t:

- t: 1 = Age, 2 = Race/ethnicity, or 3 = Coverage qualification reason

n_h – Initial sample size for domain

N_h – Population size for domain.

Thus, every person record in the extract had three applicable measures of size (MOS_1, MOS_2, MOS_3)—one for each of the domain types. Each person was then assigned an overall MOS, which was equal to the maximum of these three measures. The final sample was drawn using systematic sampling, where the overall MOS was the probability that the person was sampled. That is, the probability of sampling was proportional to size so that persons with a higher final MOS had a higher probability of being sampled.

Sample Summary

Medicaid Sampling

- Almost all targeted domains (50 states by three age groups, by three race/ethnicity groups, by six coverage qualification categories) have sample sizes higher than the minimum threshold, 1000.

The Medicaid sampling counts and rates for each domain are presented in **Table 3** and **Table 4** below.

Table 3: Frame, Sample Counts, and Sample Rate by Domain Categories for Medicaid Sampling

Domain	Domain Category	Population Size	Frame Count	Sample Count	Frame Sample Rate	Population Sample Rate
Race	*Black*	16,891,513	16,891,513	1,016,435	6.02%	6.02%
	Hispanic	19,482,059	19,482,059	1,169,532	6.00%	6.00%
	Other	59,816,308	59,816,308	3,585,426	5.99%	5.99%
Age	*0–18*	42,329,690	42,329,690	2,536,969	5.99%	5.99%
	19–64	45,921,768	45,921,768	2,753,407	6.00%	6.00%
Coverage Qualification Reason	*Children*	39,490,311	39,490,311	2,366,324	5.99%	5.99%
	Adult	14,610,382	14,610,382	875,619	5.99%	5.99%

Domain	Domain Category	Population Size	Frame Count	Sample Count	Frame Sample Rate	Population Sample Rate
	Disabled	10,331,353	10,331,353	619,932	6.00%	6.00%
	Aged	7,120,005	7,120,005	428,962	6.02%	6.02%
	Expansion	18,585,087	18,585,087	1,114,194	6.00%	6.00%
	Other	6,052,742	6,052,742	366,362	6.05%	6.05%
Total		96,189,880	96,189,880	5,771,393	6.00%	6.00%

Table 4: Realized Sample Counts by Domain for Medicaid Sampling

Domain Type	Frame Size Count	# Cells	Sample Count	
			Mean	Min
Race	*5000+*	131	44,036	996
	1000–4999	4	655	203
	0–1000	3	6	0
Age	*5000+*	150	38,475	1,011
	1000–4999			
	0–1000			
Reason	*5000+*	256	22,520	996
	1000–4999	9	580	249
	0–1000	25	45	0

Medicare Sampling

- A majority of targeted domains (50 states by three age groups, by three race/ethnicity groups, by three coverage qualification categories) have sample sizes higher than the minimum threshold, 1000.

The Medicare sampling counts and rates for each domain are presented in **Table 5** and **Table 6** below:

Table 5: Frame, Sample Counts, and Sample Rate by Domain Categories for Medicare Sampling

Domain	Domain Category	Population Size	Frame Count	Sample Count	Frame Sample Rate	Population Sample Rate
Race	*Black*	6,377,505	1,275,501	389,327	30.52%	6.10%
	Hispanic	1,612,815	322,563	109,386	33.91%	6.78%
	Other	51,977,160	10,395,432	3,099,336	29.81%	5.96%
Age	*0–18*	2,290	458	456	99.56%	19.91%
	19–64	9,154,835	1,830,967	564,627	30.84%	6.17%
	65+	50,810,355	10,162,071	3,032,966	29.85%	5.97%
Coverage Qualification Reason	*Aged*	50,817,810	10,163,562	3,032,883	29.84%	5.97%
	Disability	8,994,225	1,798,845	541,484	30.10%	6.02%
	ESRD	155,445	31,089	23,662	76.11%	15.22%
Total		59,967,480	11,993,496	3,598,029	30.00%	6.00%

Table 6: Realized Sample Counts by Domain for Medicare Sampling

Domain Type	Frame Size Count	# Cells	Sample Count	
			Mean	Min
Race	*5000+*	118	30,132	992
	1000–4999	20	522	215
	0–1000	15	115	34
Age	*5000+*	102	34,973	1,103
	1000–4999	–	–	–
	0–1000	48	9.5	1
Reason	*5000+*	110	32,293	992
	1000–4999	27	515	245
	0–1000	16	96	29

WEIGHTING METHODOLOGY

Purpose and Overview of Weighting

The purpose of weighting the persons in this database is to create a *nationally representative* healthcare database for research.

The U.S. population can be partitioned into eight (possibly overlapping) subpopulations:

1. Medicare population, covering those who were enrolled in Medicare.

2. Medicaid population, covering those who were enrolled in Medicaid.

3. Medicare/Medicaid dual-enrolled population, covering those who were dual-eligible and enrolled in both Medicare and Medicaid.

4. CHIP-eligible population, covering children who were enrolled in CHIP.

5. TRICARE-eligible population, covering those who were enrolled in TRICARE.

6. VA healthcare-eligible population, covering those who were enrolled in VA health care.

7. Commercially insured population, covering those who had commercial health insurance such as ACA market exchange, employer-based, direct-purchase, and federal employee coverage.

8. Uninsured population, covering those who did not have health insurance.

Target Population Represented in SyH-DR

The target population covered by the SyH-DR includes those who were insured either by a government program (Medicare, Medicaid, or CHIP) or commercial health insurance at any point during 2016, thus covering subpopulations (1), (2), (3), (4), and (7), as defined above. The SyH-DR includes records of all persons who were selected in the sampling process.

The SyH-DR includes a representative sample of all persons in the enrollment files to provide researchers with the ability to draw useful population-based estimates.

Those who were insured solely by TRICARE (5) or VA (6) are not included in the SyH-DR. Therefore, we did not attempt to represent them in the weighting process, based on the understanding that their health conditions, diagnosis patterns (distribution and comorbidity status), and treatments (as shown by procedure codes) might be substantially different from those covered by commercial or public health plans (Medicaid and Medicare). Moreover, TRICARE and VA healthcare beneficiaries have access to hospital facilities that are available only to them, and the treatments provided by these hospitals may be inconsistent with those provided to enrollees with other types of coverage.

Weighting Analysis

The purpose of weighting is to account for the selection probabilities of sample units in the dataset so that each unit can properly represent units in the sampling frame. A subsequent weighting adjustment is designed to address coverage gaps in the sampling frame against the target population. A calibration adjustment was used to match the weighted totals of units in the source files to benchmark values obtained from the American Community Survey (ACS) for counts of persons, and from the Healthcare Cost and Utilization Project (HCUP) data for claims counts.

For the SyH-DR to be a representative sample of the target population, the analysis weights were constructed in two steps:

1. Developed a base sample weight that was the inverse of the probability of selection into the sample.

2. Conducted a calibration adjustment of the base weight to match the weighted totals with the population control totals from ACS and HCUP.

For Step 1, the probabilities of selection reflect the sampling process.

Prior to calibration adjustments, the weighted totals (i.e., just using sample weights) were substantially different from the population control totals from ACS by age and sex (see **Table 7**). Weighted encounter counts by diagnosis were also compared with the control totals from HCUP (see Appendix A, Exhibits A.1 through Exhibit A.3). Calibration adjustments using a raking algorithm were used to align these values.

Table 7: Comparison of Source File Estimates to Controls from ACS

Source File	Data Element	Category	Control Estimate (from ACS)	Weighted Totals (with Base Sample Weight)	Difference from Control
Medicare	*Age*	*0–64*	7,809,356	8,700,679	11.40%
		65+	43,146,979	47,970,478	11.20%
	Sex	*Female*	28,104,291	31,062,770	10.50%
		Male	22,852,044	25,608,387	12.10%
Medicaid	*Age*	*0–17*	28,273,909	35,506,482	25.58%
		18-64	25,352,260	39,456,782	55.63%
		65+	6,247,157	7,054,098	12.92%
	Sex	*Female*	32,575,000	45,274,661	38.99%
		Male	27,298,326	36,742,702	34.60%
Commercial	*Age*	*0–17*	42,808,485	32,049,397	-25.10%
		18–64	139,503,428	110,129,504	-21.10%
		65+	5,340,313	2,662,400	-50.15%
	Sex	*Female*	94,473,874	71,744,792	-24.06%

		Male	93,178,352	73,096,509	-21.55%

Note: Estimates are pro-rated by the proportion of months enrolled in 2016. For example, if someone is enrolled for only 10 months in the year, they would be counted towards totals for {Weight} x 10/12.

For calibration adjustments, we used control totals from two sources:

- 2016 ACS five-year population estimates, on a person level
- 2016 HCUP (NIS and NEDS files), on an encounter level

The ACS controls were computed on a person-level basis from ZIP code-level summary files produced by the Census Bureau. The Census Bureau estimates used for these controls were created by using weighted estimates (via ACS weights) that were based on ACS respondents' responses about the type of coverage they or other household members were enrolled in *at the time they were interviewed*. Respondents were specifically asked to confirm whether one or more coverages applied to them, including Medicare and Medicaid. For commercial coverage, respondents were asked separately whether they had employee-sponsored insurance (ESI), or direct purchase insurance. Each respondent was allowed to identify multiple coverages they and other household members had.

Control totals were computed for commercial coverage for enrollees under age 65 using the sums of estimates for ESI and direct purchase insurance (potentially, the same person could have both, which resulted in an overcount, but this was expected to occur rarely). For enrollees 65 and over, some ACS-reported commercial coverage is supplemental to Medicare coverage as retiree or Medigap insurance. Because enrollees with supplemental coverage are not included in the commercial data and are already represented in Medicare data (i.e., it would be duplicative to represent them on commercial coverage data), ACS control totals were produced in a way that attempts to exclude enrollees with commercial coverage that is supplemental to Medicare. In particular, the commercial data represents anyone who receives insurance through an employer, either from their own employer or as a spouse or dependent. Control totals were produced to align as closely as possible to this population. To achieve this, commercial control totals were counted only for respondents that are reported on ACS with:

(ESI or Direct Purchase) and not Medicare

-OR-

ESI and Worked 20+ hours/week

-OR-

Had spouse with ESI coverage[2]

To avoid duplicative counting of respondents reporting both direct purchase and employer-based coverage, ACS commercial controls for the age 65 and over population were computed from public use micro-records rather than census tabulations used for the under 65 population. However, as this microdata did not include ZIP codes, geographic summarization was only made to state level rather than the ZIP-3 level used for under 65 enrollees.

Because some ZIP codes shown in the ACS files were not present in the source file samples, or represented subpopulations too small for adjustment, five-digit ZIP code areas were summarized to a three-digit ZIP code level. So, the crossing of three-digit ZIP code (where available and to state level otherwise), age group, and sex formed the basis for the person-level raking domains.

Additionally, because ACS estimates are made on a point-in-time basis (i.e., respondents were asked about current coverage), to compare these estimates to source file person records, the records must be prorated by the proportion of the year in which the represented persons were covered. For example, for someone who was enrolled in Medicaid for only one month in a year, the probability that that person would have indicated this enrollment when surveyed by ACS (which conducts interviews throughout the year) would be 1 in 12.

HCUP controls were developed based on microdata compiled in the 2016 National Inpatient Sample (NIS) and the National Emergency Department Sample (NEDS) files. Both were organized on an encounter-level basis with no person-level identifier. In turn, the controls generated were specific to encounter-level estimates.

Among the encounter-level data available on the HCUP files were the diagnoses assigned to the person for the encounter. Each encounter was categorized by the primary diagnosis code. The category assigned for each HCUP-reported encounter was based on the first three characters of the ICD-10 code, except in cases where that category (based on the three-digit ICD-10 code) represents less than 0.25% of all encounters, in which case the category was reassigned to be the two-digit ICD-10 code. Weighted (by HCUP weight) tabulations were then made (separately for NIS and NEDS) to produce HCUP control totals for each diagnosis category, age, and primary payer (Medicare, Medicaid, or commercial), by demographic cell consisting of age category (0–17, 18–26, 27–44, 45–64, 65+) and sex crosstabs.

To apply the HCUP controls to the sampled files, inpatient (IP) and emergency department (ED) claims were separately assigned to diagnosis groups using the collapsing rules developed in the HCUP summarization process. For each person, the number of IP and ED claims falling under each of the diagnosis groups was tabulated. Important considerations are that a person record will have non-zero values for at most just a few diagnosis categories and that IP and ED claims were tabulated distinctly.

[2] Spouse with ESI coverage was determined if ACS respondent was married and another person in the same household was married, age 55 and over, worked 20 hours a week or more, and had ESI coverage.

The process of adjusting weights so that the person- and encounter-level tabulations conformed to the control totals is called raking. A standard raking procedure called iterative proportional fitting was used. The raking controls from ACS were generally organized by demographic cells that consisted of the crossing of three-digit ZIP code, age group (0–17, 18–64, and 65+), and sex. The raking controls from HCUP were organized by demographic cells that consisted of crosstabs of age group (0–17, 18–26, 27–44, 45–64, 65+) and sex. However, there were multiple sets of these controls, each specific to a combination of type of claim (IP or ED) and diagnosis group. Generally, there were more than 500 different sets of HCUP controls, with each set consisting of control totals for 10 demographic groups (a combination of five age groups and two sex groups) for each diagnosis. Raking was conducted in passes, separately for each of the three source files. In each pass, we ran through a set of control totals:

- One for ACS—with person-record enrollment prorated by proportion of year covered.
- 500+ for diagnosis groups in HCUP NIS and NEDS combined.

After each pass, the weighted sums of source file sample records in each demographic cell were computed (for ACS, a combination of three-digit ZIP code, three age groups, and sex; and for HCUP, a combination of five age groups and sex for each diagnosis). These weight totals were compared to the controls, and for each cell the weight adjustment was computed as:

$$Adjustment_{R,D,G} = \frac{Control_{D,V,R}}{\sum_{i \in D} Weight_{R-1,i} \cdot E_G}$$

where i represents the persons in the demographic ***D***, R is the rake number, D is the demographic cell for person i, E_G is the encounters for diagnosis group G under analysis (the fraction of months in the year that the person was enrolled in health insurance coverage as reported in the ACS, or the number of encounters the person had with a primary diagnosis in diagnosis G for HCUP).

These computed adjustments were then applied back to the weight prior to this pass:

$$Weight_{R,i} = Adjustment_{R,D,G} \cdot Weight_{R-1,i}$$

Thus, for each raking pass, the weight assigned to a person was adjusted once to the applicable ACS control total and possibly several more times for each IP and ED encounter in the corresponding claims data that the person was shown to have had. Raking continued for multiple cycles until convergence was reached.

Modifications to the above process were made for each payer, depending on the specific characteristics of the availability and distributions of demographic information:

- **For Medicare**, the bulk of enrollees were in the 65+ age group; the other age groups, particularly the youngest, ages 0–17, were insufficiently populated to allow for effective raking. For this reason, all under-65 age groups were collapsed into a single age group, 0–64. Also, three-digit ZIP areas that had a total estimated Medicare enrollment of fewer than 1,000 persons were assigned to a nationwide catch-all category.

- **For Medicaid**, the reporting of ZIP code information varied considerably by submitting state, and for some states there were no populated ZIP code values. To address this issue, all submitting states that had ZIP code non-missing rates of 98% or greater were identified and a raking was done at three-digit ZIP code level. For the remaining submitting states, we combined all records for the state into a single geographic area (representing the entire state); thus, the state was used for raking instead of the three-digit ZIP code. Additionally, the Medicaid data we received did not include records submitted by Arkansas. Since the SyH-DR was designed to provide national representation over state representation, all person records in the West South Central geographic area (which comprises Arkansas, Louisiana, Oklahoma, and Texas) were collapsed into a single replacement area. Moreover, the 65+ age group was not sufficiently populated to allow for effective raking. To avoid unduly high weight variations, which were likely to occur in small cells, for ACS raking the three-digit ZIP code levels were collapsed at the state level for persons 65 and over.

- **For commercial**, some three-digit ZIP codes were shown in ACS to have commercially insured persons for which no records were found in the commercial sample. Thus, we were not able to align the overall ACS totals with the commercial data. These three-digit ZIP codes without records only occurred in 14 states. For these states, state was used for raking instead of the three-digit ZIP code. State instead of three-digit ZIP code was also used for the 65+ population, as noted previously.

Because the commercial source files did not include state codes, a three-digit ZIP code-to-state mapping was used. This table was generated using SAS's embedded ZIP code-to-state equivalence table. Generally, for each distinct three-digit ZIP code, all included five-digit ZIP codes map to the same state. There were, however, a few cases where the five-digit ZIP codes that shared the same first three digits mapped to different states. In these situations, the three-digit ZIP code was mapped to the state.

Initial attempts at raking also showed that having five age groups for HCUP raking led to extreme weight assignments (because of random fluctuations within small cells). Therefore, the age groups 18–26, 27–44, and 45–64 were collapsed into one age group: 18–64.

Weights in the SyH-DR

The weighting process resulted in a single final weight that was created for each enrollment record. The resulting weights by age group, sex, eligibility source, and race[3] are presented, grouped by payer, in Appendix A, Exhibits A.4 through Exhibit A.6. The final weight in the SyH-DR allows for weighted estimates of person-level characteristics and hospital service utilization to track closely to national estimates and key domains defined by key variables listed above. This is true for the data elements directly used for benchmark values. In addition, if key data elements of interest for analyses are closely related to data elements used for benchmarking, such estimates may also be approximately unbiased.

[3] The commercial data does not include eligibility source or race.

SYNTHETIZATION METHODOLOGY

Purpose and Overview of Synthetization

The SyH-DR is a partially synthesized database in two regards. First, all person-level data elements were not synthesized, although they may differ from the source data due to de-identification steps. Second, in claims-level data, some data elements were partially synthesized—in the sense that at some level of aggregation, synthesized values were identical to the original values (e.g. synthetic diagnoses were identical to original diagnoses at the level of the diagnosis category), some data elements were fully synthesized, and some data elements were not synthesized. We use the term "retained data elements" to refer to unsynthesized data elements and the components of partially synthesized data elements that were identical to their original values. In all cases, synthetic data elements were created to resemble the marginal distribution of the original data elements.

Synthetic data elements were generated from imputation models that used retained data elements as covariates to predict the values of these data elements. Imputation allows relationships between synthetic data elements and retained data elements to be preserved, to the extent that imputation models are able to capture such relationships.

The choice of retained and synthetic data elements was made taking into account the disclosure risk of retaining a data element, the analytic importance of the data element for health research, and data use considerations stipulated by data providers.

Synthetization was performed at the claim level. Therefore, the claims in the SyH-DR are the original claims with some data elements partially or fully replaced by synthetic values. The partially synthesized data elements retain the following original portion of the data element:

- ICD-10-CM diagnosis codes: First three characters
- ICD-10-PCS procedure codes: Clinical Classifications Software (CCS) for ICD-10-PCS
- CPT and HCPCS procedure codes: Clinical Classifications Software for Services and Procedures (CCS-Services and Procedures)
- Generic Drug Names: Therapeutic class from the Cerner Multum drug, herbal, and nutraceutical database

Table 8 lists whether each data element in the inpatient, outpatient, and pharmacy files is retained (subject to de-identification procedures), masked, partially synthesized, or fully synthesized. *Note that the SyH-DR contains only generic drug names, so Multum therapeutic classes are not included.*

Table 8: List of Partially Synthesized, Fully Synthesized, Masked, and Retained Data Elements

	Inpatient and Outpatient Files	Pharmacy Files
Retained	Person weight	Person weight
	Claim type code	Claim line number
	Service begin date	Fill date
	Service end date	
	Length of stay	
	Type of bill code	
Masked	Person ID	Person ID
	Facility ID	Pharmacy claim number
	Claim control number	Claim control number
Partially Synthesized	Primary diagnosis code	Generic drug name
	Diagnosis codes 1–25	
	ICD procedure codes 1–25	
	CPT procedure codes 1–35	
Fully Synthesized	Attending physician specialty	Plan paid amount
	Admission type	Total charge amount
	Discharge status	
	Plan paid amount	

	Total charge amount	

The following sections describe the methodology for the synthetization of each synthesized data element.

Inpatient and Outpatient Files

Diagnoses

The Inpatient and Outpatient files contain diagnosis codes, almost all of which are ICD-10-CM codes in the source files. The SyH-DR reports up to 25 diagnosis codes for each claim, although the number of diagnosis codes varies by payer and file type. **Table 9** lists the maximum number of diagnosis codes reported in each file.

Table 9: Maximum Number of Diagnosis Codes by Payer and Setting

Payer	Inpatient File	Outpatient File
Medicaid	12	2
Medicare	25	25
Commercial	11	11

ICD-10-CM codes range between three and seven characters. Each code begins with a letter, followed typically by two numbers (although the second number may, in rare cases, be a letter). There are more than 50,900 unique ICD-10-CM diagnoses recorded across all claims in the source files.

Diagnosis codes were partially synthesized. Diagnosis codes in each claim from the source files were replaced with synthetic diagnosis codes, where the synthetic codes belonged to the same diagnosis category as the original diagnosis code, and where a diagnosis category is the first three characters of the ICD-10-CM code. In other words, diagnosis codes in the SyH-DR preserve the first three characters of the ICD codes observed in a claim in the source files.

In the ICD-10-CM coding system, diagnosis categories describe the general type of disease or injury. For example, diagnosis category A01 describes “typhoid and paratyphoid fevers.” Granular diagnoses in this category include typhoid fever, unspecified (A01.00), typhoid meningitis (A01.01), typhoid fever with heart involvement (A01.02), and so on. Diagnosis categories are mutually exclusive; that is, each diagnosis belongs to exactly one diagnosis category. ICD-10-CM diagnoses observed in the source files were grouped into 1,925 diagnosis categories.

Pre-processing

To synthesize diagnoses, services files were first prepared by removing diagnosis codes that do not begin with a letter. Diagnosis codes that do not begin with a letter are sometimes typographical errors but are most often ICD-9 codes. These codes make up about 1.2% and

1.8% of all diagnosis codes in the Medicaid Outpatient and Inpatient files, respectively, and are a trivial proportion for the commercial and Medicare files. Then, diagnosis codes were deduplicated by person. Multiple claims might be observed for each person, and diagnosis codes might recur across claims—for example, if a person had repeated hospital visits to treat a persistent condition. The purpose of deduplicating diagnosis codes by person was to ensure that each unique diagnosis code for a given person was replaced by exactly one synthetic diagnosis code. Hence, if a diagnosis code was observed in multiple claims for a given person, the same synthetic diagnosis code would be generated across all these claims. In other words, patterns of recurring diagnosis codes across claims were preserved in the Beta version, even if the codes themselves were different.

Modeling

Synthetic diagnosis codes were generated by selecting a code from the set of diagnosis codes belonging to the same diagnosis category as the original diagnosis code. The probability of selecting each code was given by a model that used as predictors the age and sex of the person, as well as the claim type (inpatient, outpatient, or ED) and all diagnosis categories that were observed in that claim. Specifically, a binary classification model was estimated for each diagnosis, using all claims that contained a diagnosis from that diagnosis category. For example, a model for diagnosis A01.00 would be trained using all claims with a diagnosis in category A01, with the goal of predicting whether the diagnosis in a claim was A01.00 or not (i.e., some other diagnosis from category A01). Gradient boosting models were used for the classification task.

Once a model was trained, predicted probabilities of a person having that diagnosis on a claim were generated. The predicted probabilities were then calibrated such that the mean predicted probability across all claims was equal to the actual observed prevalence of that diagnosis in the Alpha data files. This process was repeated for all diagnoses in that category. Finally, a synthetic diagnosis was drawn from the set of diagnoses in that category with probabilities proportionate to the calibrated probabilities. Note that because the selection of synthetic diagnoses is probabilistic, each run of the Beta files produces a different set of synthetic diagnoses.

Post-processing

After the synthetic diagnoses were drawn, they were then post-processed for inclusion in the SyH-DR. As part of this post-processing, imputation was performed for claims with missing primary diagnosis codes (PRMRY_DX_CD). We impute claims with missing PRMRY_DX_CD using the value of the synthesized ICD_DX_CD_1. Missing PRMRY_DX_CD was observed in about 0.12% of claims in the source commercial data files (inpatient and outpatient), about 0.47% of the source Medicaid Inpatient file, and about 3.57% of the source Medicaid Outpatient file. No missingness for PRMRY_DX_CD was observed in the Medicare data files. For Medicaid claims with observed PRMRY_DX_CD values, PRMRY_DX_CD is always identical to ICD_DX_CD_1. For the commercial files, PRMRY_DX_CD is different from ICD_DX_CD_1; users should therefore note that imputed PRMRY_DX_CD may not be an accurate representation of the missing primary diagnosis code. Users who do not wish to use imputed PRMRY_DX_CD may use the PRMRY_DX_IMPUTED flag to identify such claims.

Finally, the index numbers of the diagnosis codes (e.g., ICD_DX_CD_1, ICD_DX_CD_2, … ICD_DX_CD_25) do not have any clinical significance. As such, no effort was made to preserve the original ordering or index numbers of the diagnoses, and users should not assign any analytic meaning to the order of the diagnoses. To enhance user accessibility of the diagnosis codes, they were moved up to lower number spots if diagnosis code data elements initially had no value or were removed because they did not begin with a letter (see pre-processing step, above). For example, consider a claim with a primary diagnosis code and diagnosis codes 1, 2, and 4, but a diagnosis code 3 with a value that was removed. In this case, diagnosis codes 1 through 4 would be shifted to lower number spots so that diagnosis code data elements 1 through 3 would be populated and diagnosis code 4 would be unpopulated.

ICD Procedures

The source files report up to 25 ICD procedure codes for each claim, although the number of procedure codes varies by payer and file type. **Table 10** lists the maximum number of procedure codes reported in each file:

Table 11: Maximum Number of ICD Procedure Codes by Payer and Setting

Payer	Inpatient File	Outpatient File
Medicaid	6	-
Medicare	25	25
Commercial	6	6

ICD-10-PCS codes consist of seven alphanumeric characters, with the first character describing a section (e.g., medical and surgical, obstetrics, etc.). There are more than 24,800 unique ICD-10-PCS codes recorded across all claims in the source files.

ICD-10 procedure codes were partially synthesized. As with diagnoses, procedure codes in each claim from the source files were replaced with synthetic procedure codes in the SyH-DR, where the synthetic codes belonged to the same procedure category as the original diagnosis code. Procedure categorization was performed using the Clinical Classifications Software (CCS) for ICD-10-PCS. The CCS for ICD-10-PCS categorizes ICD-10 procedure codes into 224 mutually exclusive, clinically meaningful categories. All 224 categories were observed in the source files.

As an example, suppose ICD-10 procedure code 00964JZ, “Removal of Synthetic Substitute from Cerebral Ventricle, Percutaneous Endoscopic Approach,” was observed in the original data. The procedure is categorized in CCS procedure category 2, “Insertion; replacement; or removal of extracranial ventricular shunt.” This procedure would be replaced by another

procedure in the same category in the Beta file, for example 0W110JG, “Bypass Cranial Cavity to Peritoneal Cavity with Synthetic Substitute, Open Approach.”

As with diagnosis synthetization, the probability of selecting each procedure code was given by a model. The model for procedures used as predictors the age and sex of the person, as well as the claim type (inpatient, outpatient, or ED), all diagnosis categories, and all ICD-10 procedure categories that were observed in that claim. In other words, the model predicts the likelihood of observing a procedure on the claim by using information about person demographics and other clinical information from the claim, such as broad diagnosis and procedure categories. Generated probabilities were then calibrated to match the prevalence of the procedure from the source files, and a synthetic procedure was drawn from the set of procedures in that CCS category with probabilities proportionate to the calibrated probabilities.

Notes

About 3.28 million procedure codes were observed across all claims, of which about 1.5% could not be mapped to a CCS procedure category. These codes could not be mapped because they were not valid ICD-10 procedure codes (e.g., did not have seven characters). Almost all of these codes were in the Medicaid files. Codes that could not be mapped were omitted from the SyH-DR.

Multiple instances of a given procedure were sometimes observed in a single claim. A given procedure might also be observed across multiple claims for a given person. We understood these patterns to mean that the person underwent a procedure multiple times during a hospital stay or returned to the hospital multiple times for the same procedure. In the SyH-DR, these patterns were preserved by replacing each unique procedure for a given person in the source files with exactly one synthetic procedure. Finally, as with diagnosis codes, the index numbers of the procedure codes (e.g., ICD_PRCDR_CD_1, ICD_PRCDR_CD_2, … ICD_PRCDR_CD_25) do not have any clinical significance. As such, no effort was made to preserve the original ordering or index numbers of the procedures.

CPT Procedures

CPT procedure codes are five-digit codes that describe tests, surgeries, evaluations, and other medical procedures. The Alpha data files report up to 35 CPT procedure codes for each claim, with the exception of the Medicaid Inpatient file, which does not contain any CPT procedure codes. The synthetization methodology for CPT procedures was similar to that of ICD procedures, except that procedure categorization was performed using CCS for Services and Procedures. About 11,000 unique CPT codes were observed in the source files, grouped into 242 mutually exclusive CCS categories. The model for CPT procedures used as predictors the age and sex of the person, as well as the claim type (inpatient, outpatient, or ED), all diagnosis categories, and all CPT procedure categories that were observed in that claim.

Notes

About 103.9 million CPT procedure codes were observed across all claims, of which about 1.9% could not be mapped to a CCS procedure category. These codes could not be mapped because

they were not valid CPT procedure codes (e.g., they did not have five digits, or they were HCPCS Level II codes). About 51% of these codes were in the Medicaid Outpatient file, 46% were in the Medicare files, and 3% were in the commercial files. Codes that could not be mapped were omitted from the SyH-DR. Like diagnosis and ICD procedure codes, the order of CPT procedure codes was not preserved, and synthetic CPT procedure codes were placed in order, starting at CPT procedure code 1, with no gaps.

Attending Physician Specialty

The attending physician specialty describes the CMS specialty code corresponding to the attending physician. Attending physician specialty is a fully synthesized data element. Synthetization of this data element was treated as a multiclass classification problem; that is, one specialty from the set of 108 possible specialties was selected.

The probability of selecting an attending physician specialty for a claim was given by a multiclass classification model that used as predictors the age and sex of the person, as well as the diagnosis category of the primary diagnosis for that claim. A separate model was estimated for each payer and claim type (inpatient, outpatient, ED), for a total of nine models. Gradient boosting models were used for classification. Provisional synthetic attending physician specialties were generated by randomly drawing an attending physician specialty from the set of specialties, based on modeled probabilities.

Notes

A significant proportion of claims from the Medicaid and Medicare files had missing attending physician specialty in the source files. About 95% of claims from the Medicare Inpatient file did not have an attending physician specialty, as did 25% of claims from the Outpatient file. About 55% of claims from Medicaid Inpatient and Outpatient files did not have an attending physician specialty. If attending physician specialty was missing in a claim from the source files, specialty was also set to be missing in the SyH-DR.

Admission Type

The admission type code indicates the type and priority of an inpatient admission associated with the service. Admission type is a fully synthesized data element. Synthetization of this data element was treated as a multiclass classification problem; that is, one admission type code from the set of six possible codes (emergency, urgent, elective, newborn, trauma center, or unknown) was selected.

The probability of selecting an admission type for a claim was given by a multiclass classification model that used as predictors the age and sex of the person, as well as the diagnosis category of the primary diagnosis for that claim. Only inpatient files contained admission type codes. A single model was trained for all claims, regardless of payer.

Notes

Only a small proportion (4.6%) of admission type codes were missing (none in the Medicare file, 10% in the commercial file, and 7.3% in the Medicaid file). Because missingness was relatively limited for this data element, synthetic admission types were imputed for inpatient claims where admission type codes were missing in the source files.

Discharge Status

The discharge status describes the status of the person as of the service end date for a claim. Discharge status is a fully synthesized data element. Synthetization of this data element was treated as a multiclass classification problem; that is, one discharge status from the set of possible discharge statuses was selected. The sizes of the sets (i.e., the number of unique discharge statuses) range from 36 to 44, depending on the payer.

The probability of selecting a discharge status for a claim was given by a multiclass classification model that used as predictors the age and sex of the person, as well as the claim type (inpatient, outpatient, or ED), the subsequent claim type (i.e., the claim type for the next claim observed for that person, or "LAST" if no further claims were observed), a flag for whether the claim overlapped with another claim, and all diagnosis categories that were observed in that claim. A separate model was estimated for each payer. Gradient boosting models were used for classification. In addition, for Medicaid Inpatient and Medicare claims, a binary classification model was estimated to predict whether the person expired (having a discharge status of either 20, 40, 41, or 42) or not for the final claim observed for that person.

Provisional synthetic discharge statuses were generated by randomly drawing a discharge status from the set of statuses, based on modeled probabilities. Then, two edits were made after provisional synthetic discharge statuses were generated. Because it would be implausible for the person to have expired and then have subsequent claims, provisional synthetic values were edited to ensure that such scenarios did not occur (that is, ensuring that a status of "person expired" could only happen in the final claim). A second set of edits was made for Medicaid Inpatient and Medicare claims to align provisional synthetic discharge statuses for final claims with model predictions of whether the person expired or not. For example, if the person was predicted to have expired based on the binary classification model, but the multiclass classification model assigned a status code other than codes 20, 40, 41, or 42, then the synthetic status code was edited to be 20 (person expired) to be consistent with the prediction that the person expired.

Notes

The Medicaid Outpatient source file did not have values for discharge status, so discharge status was not synthesized for this file. About 3.4% of claims in the Medicaid Inpatient source file were missing discharge status, so discharge statuses were imputed for these claims in the SyH-DR. Finally, the commercial data files did not include any discharge statuses indicating the person expired. Rather, the data provider recoded all such statuses to "missing." Thus, when a missing discharge status was observed, it was not possible to know whether the status was truly missing, or it was recoded from "person expired." All missing discharge statuses in the commercial data files were first recoded to "00" (unknown value), prior to training the classification model.

Plan Paid Amount and Total Charge Amount

Plan paid amount is available in the source files for all three payers, whereas total charge amount is only available for Medicare and Medicaid. These two variables are fully synthesized variables. Because these two variables are jointly synthesized, we describe the methodology for both variables in this section.

The plan paid amount for each claim was modeled using the person's age and sex, as well as the length of stay for the claim and the counts of each procedure observed on the claim. Conceptually, the model expresses the idea that there is a base cost for each hospital visit that depends on the person's age and sex; the cost then increases with each day of stay as well as each additional procedure performed during the stay, with each procedure having a procedure-specific cost (i.e., each procedure has its own coefficient in the model).

To manage computational cost, the procedures included in the regression model were limited to the 4,000 most frequent procedures for inpatient claims, 1,200 most frequent for outpatient claims, and 300 most frequent for ED claims. These frequency cutoffs were chosen such that at least 95% of total procedures observed across all claims for each claim type were included. Models were estimated for each payer crossed with claim type (i.e., a separate model for commercial inpatient, commercial outpatient, commercial ED, Medicare inpatient, and so on). Note that the models were estimated using the original source files, but predicted values were based on synthetic procedures. That is, the coefficients for each procedure, or the cost per procedure, were based on source files' cost and procedure data; these coefficients were then applied to procedures observed on synthetic claims to obtain synthetic plan paid amounts.

The models yielded a predicted plan paid amount for each claim, conditional on the predictors. In other words, two claims with identical values for each of the predictors would have identical predicted plan paid amounts. To create synthetic values, we used the predictive mean matching method. Each claim (we call this the recipient claim) was matched to 50 claims from the same primary diagnosis category that have predicted plan paid amounts that are closest to—i.e., having the smallest absolute difference from—the predicted plan paid amount for that claim. A claim was then randomly selected from these 50 claims; we call this randomly selected claim the donor claim. The actual plan paid amount and total charged amount (if available) for the donor claim were used as the synthetic plan paid amount and total charged amount for the recipient claim. Finally, synthetic values were edited such that if a claim was originally missing a plan paid amount or a total charge amount, the respective value would be omitted. That is, if the actual plan paid amount or total charged amount for a donor claim was missing, then their synthetic counterparts would also be missing.

Pharmacy Claims Files

Generic Drug Name

The source pharmacy claims files contain National Drug Codes (NDCs) for each claim. An NDC describes the drug filled in each claim, and each NDC has 11 characters. NDCs were converted to generic drugs using a crosswalk drawn from the Cerner Multum drug, herbal, and nutraceutical database. For example, NDC 50580-600-02 (tablet, film coated Tylenol Regular

Strength) is mapped to the generic drug acetaminophen. Each generic drug is identified by a Multum drug ID. Because the crosswalk can contain multiple drug ID mappings for an NDC in different years, we first removed older mappings of the same NDCs, which resulted in 187,513 unique NDC code–drug ID mappings. This file was merged with the source files to get the set of all drug IDs. We observed around 3,200 unique drug IDs being mapped to the NDCs in the source files.

Drug IDs were partially synthesized. Drug IDs in each claim from the source file were replaced with synthetic drug IDs in the SyH-DR, where the synthetic IDs belonged to the same therapeutic class as the original drug ID. In other words, the drug IDs in the SyH-DR preserve the therapeutic classes observed in a claim in the source files.

In the Multum therapeutic class coding system, therapeutic classes describe the general type of drug. For example, therapeutic class 40 describes “cardiovascular agents.” Granular sub-classes in this category include anti-arrhythmic agents, angiotensin-converting enzyme (ACE) inhibitors, beta-adrenergic blocking agents, vasodilators, diuretics, and so on. Therapeutic classes have a hierarchical structure, with some classes having sub-classes or sub-sub-classes. Each drug ID was mapped to the most granular class level available for that drug. That is, if the Multum database reported a sub-sub-class for a drug, it was mapped to the sub-sub-class; if the Multum database only reported a class for a drug, it was mapped to the class. This implies that it is possible for claims to have an unknown therapeutic class if they mapped to a valid sub-class or sub-sub-class. For expositional purposes, we call all classes, including sub-classes and sub-sub-classes, “therapeutic classes.”

Drugs may be associated with more than one therapeutic class. For the purpose of synthetization, if a drug was associated with multiple therapeutic classes, the combination of therapeutic classes was considered its own therapeutic class. Therapeutic classes recoded in this manner were therefore mutually exclusive; that is, each drug ID belonged to exactly one therapeutic class. Drug IDs observed in the source files were grouped into 513 therapeutic classes.

Users should note that since the synthetization was done combining all three payer types, there may be some drug IDs that appear in the SyH-DR for a payer type (e.g., commercial) that are not present in the corresponding source file.

Pre-processing

To synthesize drug IDs, NDCs that were missing or contained any non-numeric characters were first removed. Such NDCs constituted around 0.00097% for Medicare and 2% for Medicaid. We then combined the pharmacy claim files from all three payers into a single file for synthetization. Next, we merged the files with the drug ID–NDCs crosswalk. Note that not all NDCs were present in the crosswalk. Across all three payers, about 2.4% of NDCs could not be mapped to a drug ID. (These proportions vary by payer: about 1% for Medicare, 1.8% for commercial, and 5.8% for Medicaid.)

Next, the data file was merged with the drug ID–therapeutic class crosswalk. Then, the file was deduplicated by drug ID for each person. Multiple claims might be observed for each person

corresponding to the same drug ID. For example, if a person has an initial prescription and subsequent refills for ACE inhibitors for a cardiovascular condition, there would be multiple rows with the same drug ID for the same person. The purpose of deduplicating drug IDs by person was to ensure that each unique drug ID for a given person was replaced by exactly one drug ID. Hence, if a drug ID was observed in multiple claims for a given person, the same synthetic drug ID would be generated across all these claims. This preserves the patterns of recurring drug IDs across claims in the SyH-DR.

Modeling

Synthetic drug IDs were generated by selecting a value from the set of drug IDs belonging to the same therapeutic class as the original drug ID. The probability of selecting each drug was given by a model that used as predictors the age and sex of the person, as well as all therapeutic classes that were observed for that person. Specifically, a binary classification model was estimated for each drug, using all claims that contained a drug from that therapeutic class. For example, a model for drug ID D00001 (i.e., Acyclovir) would be trained using all claims with a drug in therapeutic class 229 (purine nucleosides), with the goal of predicting whether the drug ID for a claim was D00001 or not (i.e., some other drug from category 229). Gradient boosting models were used for the classification task.

Once a model was trained, predicted probabilities of a person getting that drug on a pharmacy claim were generated. These probabilities were then calibrated such that the mean predicted probability was equal to the actual observed frequency of that drug in the source files. This process was repeated for all drugs in that category. Finally, a synthetic drug ID was drawn from the set of drugs in that therapeutic class with probabilities proportional to the calibrated probabilities. Note that because the selection of synthetic drugs is probabilistic, each run of the SyH-DR produces a different set of synthetic drugs.

Post-processing

After the synthetic drugs were drawn, they were then post-processed for inclusion in the SyH-DR. Synthetic drug IDs were replaced by the names of the generic drugs. NDCs that were not in the NDC–drug ID crosswalk were assigned “Unknown Generic Drug” as the drug name.

Total Paid Amount and Total Charge Amount

The source pharmacy claim files contain two cost data elements: total charge amount and plan paid amount. Along with the drug IDs, we also synthesized these two cost data elements for each claim. The costs in each claim from the source files were replaced with corresponding synthetic costs in the SyH-DR. If costs were missing for a given claim in the source files, no synthetic costs were produced for those claims in the SyH-DR. The data was grouped by payer, age, and synthetic drug ID to produce the synthesized drug costs.

Pre-processing

To synthesize the drug costs, we started with the final data file in the synthetic drug ID process. A similar data cleaning procedure was followed to get a de-duplicated file, by the cost data

element (plan paid amount or total charge amount) and synthetic drug ID combination for each person. Multiple claims might be observed for each person corresponding to the same drug ID with the same cost. The purpose of deduplicating costs by drug ID–person combination was to ensure that each unique drug cost for a given person was replaced by exactly one drug cost. In other words, if a person were to refill a drug multiple times over the course of the year, it is expected that the cost of the drug would be consistent each time, instead of varying for each refill. In some cases, drug costs varied slightly (by a few cents) across claims for a given person. The drug costs were averaged for the same drug for a person to get one cost observation per drug ID for each person.

Drug cost imputation

Synthetic drug cost data elements were generated from the empirical distribution of the claim-level average of the cost data elements for a given drug. The distribution is derived separately for each payer type, drug ID, and age group combination. That is, to generate the synthetic cost for a drug for a given person, drug costs were subset for that drug to those for persons in the same age group and for the same payer. From that subset of drug costs, one value was drawn, with the probability of a particular cost value proportional to the frequency of that cost in that payer type–age group subset. Therefore, each synthetic drug cost in the SyH-DR is an actual drug cost observed for that drug in the source files, for some person in the same age group.

Users may notice that cost synthetization was performed differently for inpatient and outpatient claims and pharmacy claims. Differences in the cost synthetization methodology were motivated by differences in the data structure of the inpatient and outpatient and pharmacy claims files. In the inpatient and outpatient files, each claims record includes a bundle of services and procedures, which was assigned a single plan paid amount and total charge amount. Because costs were not itemized, more extensive modeling had to be performed for the services files to synthesize costs. In contrast, the pharmacy files included a cost for each individual drug (except for drugs linked to a service claim in the Medicare file; the pharmacy files did not include cost data for these drugs). Since a distribution of costs was available for each individual drug, cost synthetization for drugs was more straightforward.

Masking Identifiers Methodology

The following identifiers on the person, services (inpatient and outpatient), and pharmacy files were masked: person ID, Medicaid beneficiary ID, claim control number, facility ID, and pharmacy claim number.

A unique nine-digit value was randomly assigned to each value of an identifier. A number drawn from a uniform distribution between 0 and 1 was first assigned to each identifier. Within each payer and identifier type (e.g., Medicare person ID), identifiers were ordered by their assigned numbers, and then assigned sequential values (1, 2, 3, …) according to their order. Finally, numerical prefixes were appended to each assigned value to distinguish different types of identifiers.

The type of identifier can be ascertained using the leading digits of the nine-digit value, as shown in **Table 11**.

Table 11: Masked Identifier Starting Characters by Data Element and Payer

Identifier	Payer		
	Commercial	**Medicare**	**Medicaid**
Person ID	Starting with 10	Starting with 30	Starting with 50
Medicaid Beneficiary ID			Starting with 51
Facility ID	Starting with 13	Starting with 33	Starting with 53
Claim Control Number	Starting with 15 or 16	Starting with 35, 36, or 37	Starting with 55 or 56
Pharmacy Control Number	Starting with 2	Starting with 4	Starting with 6

APPENDIX A: ADDITIONAL TABLES

Table A.1: Pre-raking Weight Comparison to HCUP Control Totals: Commercial Data

Source	DX_Category	Group	Level	Control File Estimate	Pre-raking Weighted Totals	Difference from Control
Commercial	*ED_J06*	*Age*	*0–17*	317,067	186,385	-41.20%
	ED_J06		*18–64*	284,867	229,274	-19.50%
	ED_J06		*65+*	9,945	3,724	-62.60%
	ED_J06	*Sex*	*Female*	337,141	229,953	-31.80%
	ED_J06		*Male*	274,738	189,430	-31.10%
	ED_M54	*Age*	*0–17*	47,580	28,379	-40.40%
	ED_M54		*18–64*	1,050,839	637,144	-39.40%
	ED_M54		*65+*	59,706	14,304	-76.00%

Source	DX_Category	Group	Level	Control File Estimate	Pre-raking Weighted Totals	Difference from Control
	ED_M54	*Sex*	*Female*	656,478	371,841	-43.40%
	ED_M54		*Male*	501,647	307,986	-38.60%
	ED_N39	*Age*	*0–17*	79,286	56,059	-29.30%
	ED_N39		*18–64*	450,869	322,340	-28.50%
	ED_N39		*65+*	50,458	9,966	-80.20%
	ED_N39	*Sex*	*Female*	502,668	340,676	-32.20%
	ED_N39		*Male*	77,945	47,690	-38.80%
	ED_R07	*Age*	*0–17*	81,799	57,171	-30.10%
	ED_R07		*18-64*	1,747,231	1,203,162	-31.10%
	ED_R07		*65+*	114,123	28,180	-75.30%
	ED_R07	*Sex*	*Female*	1,078,650	695,782	-35.50%
	ED_R07		*Male*	864,504	592,731	-31.40%
	ED_R10	*Age*	*0–17*	312,663	220,451	-29.50%
	ED_R10		*18–64*	1,765,817	1,170,427	-33.70%
	ED_R10		*65+*	57,867	16,093	-72.20%
	ED_R10	*Sex*	*Female*	1,431,943	924,855	-35.40%
	ED_R10		*Male*	704,403	482,115	-31.60%

Table A.2: Pre-raking Weight Comparison to HCUP Control Totals: Medicaid Data

Source	DX_Category	Group	Level	Control File Estimate	Pre-raking Weighted Totals	Difference from Control
Medicaid	*ED_J06*	*Age*	*0–17*	1,321,628	898,099	-32.05%
	ED_J06		*18–64*	398,142	288,182	-27.62%
	ED_J06		*65+*	2,781	1,774	-36.21%
	ED_J06	*Sex*	*Female*	924,775	649,673	-29.75%
	ED_J06		*Male*	797,777	538,383	-32.51%
	ED_M54	*Age*	*0–17*	76,193	58,575	-23.12%
	ED_M54		*18–64*	1,009,817	597,517	-40.83%
	ED_M54		*65+*	9,228	5,734	-37.86%
	ED_M54	*Sex*	*Female*	677,413	425,996	-37.11%
	ED_M54		*Male*	417,825	235,829	-43.56%
	ED_N39	*Age*	*0–17*	211,944	146,199	-31.02%
	ED_N39		*18–64*	546,444	355,174	-35.00%
	ED_N39		*65+*	10,824	7,614	-29.66%
	ED_N39	*Sex*	*Female*	699,294	467,460	-33.15%
	ED_N39		*Male*	69,918	41,528	-40.60%
	ED_R07	*Age*	*0–17*	148,086	100,755	-31.96%
	ED_R07		*18–64*	1,148,728	594,940	-48.21%

Source	DX_Category	Group	Level	Control File Estimate	Pre-raking Weighted Totals	Difference from Control
	ED_R07		*65+*	19,966	10,633	-46.74%
	ED_R07	*Sex*	*Female*	772,101	428,182	-44.54%
	ED_R07		*Male*	544,680	278,146	-48.93%
	ED_R10	*Age*	*0–17*	495,887	327,823	-33.89%
	ED_R10		*18-64*	1,614,109	863,802	-46.48%
	ED_R10		*65+*	13,499	7,604	-43.67%
	ED_R10	*Sex*	*Female*	1,505,242	845,844	-43.81%
	ED_R10		*Male*	618,252	353,384	-42.84%
	ED_J06	*Age*	*0–17*	1,321,628	684,160	-48.20%
	ED_J06		*18–64*	398,142	258,518	-35.10%
	ED_J06		*65+*	2,781	758	-72.70%
	ED_J06	*Sex*	*Female*	924,775	527,656	-42.90%
	ED_J06		*Male*	797,777	415,779	-47.90%
	ED_M54	*Age*	*0–17*	76,193	50,597	-33.60%
	ED_M54		*18–64*	1,009,817	517,756	-48.70%
	ED_M54		*65+*	9,228	3,442	-62.70%
	ED_M54	*Sex*	*Female*	677,413	378,651	-44.10%
	ED_M54		*Male*	417,825	193,144	-53.80%

Source	DX_Category	Group	Level	Control File Estimate	Pre-raking Weighted Totals	Difference from Control
	ED_N39	*Age*	*0–17*	211,944	116,566	-45.00%
	ED_N39		*18–64*	546,444	323,402	-40.80%
	ED_N39		*65+*	10,824	3,899	-64.00%
	ED_N39	*Sex*	*Female*	699,294	409,147	-41.50%
	ED_N39		*Male*	69,918	34,720	-50.30%
	ED_R07	*Age*	*0–17*	148,086	93,703	-36.70%
	ED_R07		*18–64*	1,148,728	495,084	-56.90%
	ED_R07		*65+*	19,966	6,175	-69.10%
	ED_R07	*Sex*	*Female*	772,101	370,809	-52.00%
	ED_R07		*Male*	544,680	224,154	-58.80%
	ED_R10	*Age*	*0–17*	495,887	278,834	-43.80%
	ED_R10		*18–64*	1,614,109	751,214	-53.50%
	ED_R10		*65+*	13,499	4,637	-65.60%
	ED_R10	*Sex*	*Female*	1,505,242	749,373	-50.20%
	ED_R10		*Male*	618,252	285,313	-53.90%

Table A.3: Pre-raking Weight Comparison to HCUP Control Totals: Medicare Data

Source	DX_Category	Group	Level	Control File Estimate	Pre-Raking Weighted Totals	Difference from Control
Medicare	*ED_J06*	*Age*	*0–64*	86,323	70,928	-17.80%
	ED_J06		*65+*	96,052	95,675	-0.40%
	ED_J06	*Sex*	*Female*	112,803	105,044	-6.90%
	ED_J06		*Male*	69,572	61,559	-11.50%
	ED_M54	*Age*	*0–64*	343,537	259,818	-24.40%
	ED_M54		*65+*	492,290	453,113	-8.00%
	ED_M54	*Sex*	*Female*	505,678	436,525	-13.70%
	ED_M54		*Male*	330,148	276,406	-16.30%
	ED_N39	*Age*	*0–64*	160,068	121,198	-24.30%
	ED_N39		*65+*	709,047	587,645	-17.10%
	ED_N39	*Sex*	*Female*	640,574	521,974	-18.50%
	ED_N39		*Male*	228,541	186,870	-18.20%
	ED_R07	*Age*	*0–64*	510,333	300,982	-41.00%
	ED_R07		*65+*	1,010,085	731,432	-27.60%
	ED_R07	*Sex*	*Female*	880,977	609,182	-30.90%
	ED_R07		*Male*	639,442	423,231	-33.80%
	ED_R10	*Age*	*0–64*	453,335	282,600	-37.70%

Source	DX_Category	Group	Level	Control File Estimate	Pre-Raking Weighted Totals	Difference from Control
	ED_R10		*65+*	538,822	454,039	-15.70%
	ED_R10	*Sex*	*Female*	630,925	465,575	-26.20%
	ED_R10		*Male*	361,232	271,064	-25.00%

Table A.4: Post-raking Weight Summary: Commercial Data

Source	Group	Level	Estimate	Weight Mean	Weight Minimum	Weight Medium	Weight Maximum	Weight Coefficient of Variation
Commercial	*Age*	*0–17*	51,247,859	24.65	1.01	23.31	950.2	35.15%
		18–64	168,324,728	23.27	1	22.59	999.8	28.14%
		65+	6,712,891	36.75	1.01	34.24	953.4	50.09%
	Sex	*Female*	113,717,238	24.22	1	22.79	999.8	33.25%
		Male	112,568,239	23.45	1	22.18	843.4	31.16%

Table A.5: Post-raking Weight Summary: Medicaid Data

Source	Group	Level	Estimate	Weight Mean	Weight Minimum	Weight Medium	Weight Maximum	Weight Coefficient of Variation
Medicaid	*Age*	*0–17*	32,156,670	13.28	1	12.32	643.9	29.79%
		18–64	30,631,279	10.68	1	9.39	730.9	50.99%
		65+	7,076,976	14.71	1	14.34	171	25.28%
	Sex	*Female*	37,986,428	11.92	1	11.61	730.9	40.49%

		Male	31,878,498	12.33	1	11.83	693.6	41.44%
	Eligibility	*A: CHILDREN*	31,154,622	13.17	1	12.32	643.9	30.57%
		B: ADULT	9,188,845	10.49	1	9.2	693.6	48.04%
		C: DISABLED	7,662,270	12.36	1	11.18	574.6	56.69%
		D: AGED	6,377,250	14.87	1	14.37	116.8	23.08%
		E: EXPANSION	11,018,253	9.89	1	9.06	730.9	43.82%
		F: OTHER	4,463,686	12.18	1	11.45	729.8	45.52%
	Race	*Unknown*	13,612,139	12.39	1	11.77	730.9	40.43%
		Black	12,802,812	12.6	1	11.93	626.4	40.75%
		Other	239,661	12.69	1	11.96	205.9	40.90%
		Asian	3,191,714	10.67	1	10.26	463.6	32.82%
		Hispanic	13,582,754	11.61	1	12.24	498.7	39.37%
		American Indian	877,903	12.32	1	11.46	226.6	44.61%

Table A.6: Post-raking Weight Summary: Medicare Data

Source	Group	Level	Estimate	Weight Mean	Weight Minimum	Weight Medium	Weight Maximum	Weight Coefficient of Variation
Medicare	*Age*	*0–17*	64,095	165.2	1.81	4.6	1000	196.20%
		18–64	8,140,491	14.44	1	14.36	196.6	30.56%
		65+	45,275,326	15.06	1.5	14.72	211.4	13.05%

Source	Group	Level	Estimate	Weight Mean	Weight Minimum	Weight Medium	Weight Maximum	Weight Coefficient of Variation
	Sex	*Female*	29,434,069	15.09	1	14.73	1000	28.22%
		Male	24,045,843	14.84	1	14.67	1000	32.08%
	Eligibility	*Aged*	45,281,572	15.06	1.5	14.72	211.4	13.01%
		Aged & Dual Eligibility	7,990,486	14.78	1	14.44	196.6	27.57%
		ESRD	166,679	9.98	1	4.73	1000	551.20%
		ESRD & Dual Eligibility	41,174	5.92	1	4.77	65.9	61.57%
	Race	*Unknown*	877,457	14.89	1.01	14.6	1000	58.38%
		White	42,830,700	15.05	1	14.71	1000	28.66%
		Black	5,790,160	14.91	1	14.8	1000	31.32%
		Other	1,048,330	14.91	1.06	14.58	1000	34.37%
		Asian	1,222,971	14.87	1.65	14.54	303.1	16.00%
		Hispanic	1,456,650	13.57	1.01	14.61	1000	42.51%
		American Indian	253,646	14.94	1.59	14.84	79.2	24.04%