Synthetic Healthcare Database for Research (SyH-DR)

A Synthetic Nationally Representative All-Payer Claims Database

INTRODUCTION TO

SYNTHETIC HEALTHCARE DATABASE FOR RESEARCH

AHRQ Publication No. 24-0019-1-EF

December 2023

Table of Contents

INTRODUCTION TO SYNTHETIC HEALTHCARE DATABASE FOR RESEARCH 1
Overview of SyH-DR 1
Roadmap of Document 1
Source Data 1
File Structure of SyH-DR 2
Description of SyH-DR Files 2
Harmonization To Create Uniform All-Payer Database and SyH-DR Data Elements 3
Sample Design for SyH-DR 4
Nationally Representative Person-Level Weights in SyH-DR 4
Synthetization Methodology 4
Identifier Masking Methodology 5
De-Identification Methodology 5
Implications of Synthetization, Identifier Masking, and De-Identification for Research 5
Appendix A. List of SyH-DR data elements by masked, synthesized, and suppressed status 6

INTRODUCTION TO SYNTHETIC HEALTHCARE DATABASE FOR RESEARCH

Overview of SyH-DR

The Synthetic Healthcare Database for Research (SyH-DR) is an all-payer, nationally representative claims database. The database consists of a sample of inpatient, outpatient, and prescription drug claims, including utilization, payment, and enrollment data, for people insured by Medicare, Medicaid, or commercial health insurance in 2016. AHRQ created SyH-DR, in part, as a resource to facilitate improvements to price and quality transparency in healthcare.

SyH-DR is a synthetic database that replicates the structure and statistical properties of the original claims data while protecting privacy and confidentiality of people and institutions. Synthetic data are created by statistically modeling or changing original data so that new values or data elements are generated while maintaining the original data's statistical properties. Additional steps, such as masking, are taken to reduce the risk of identifying people and institutions so that the data may be made publicly available to a broad community of researchers.

Roadmap of Document

This document provides (1) a brief description of the source data and SyH-DR data files available to public users and (2) information on data processing, including harmonization, sampling/weights, synthetization, identifier masking, and de-identification of the source data to produce SyH-DR data. The sections describing each data processing topic are presented in the order they are completed to produce SyH-DR.

Source Data

Table 1 provides the dataset name, year of data, payer, and setting for each of the input data files. Medicare and Medicaid files come from the Centers for Medicare & Medicaid Services (CMS) and commercial data comes from commercial insurance plans.

Table 1: Input Data Files

Input Files	Unit	Year	Payer	Setting
Transformed Medicaid Statistical Information System (T-MSIS) Analytic File (TAF) demographic and eligibility	Person level	2016	Medicaid	NA
TAF claims—inpatient (IP)	Claim level	2016	Medicaid	Inpatient
TAF claims—other files (outpatient/OP*)	Claim level	2016	Medicaid	Outpatient
TAF claims—RX (prescription drug)	Claim level	2016	Medicaid	Pharmacy
Master Beneficiary Summary File (MBSF)—Base A/B/C/D	Person level	2016	Medicare	NA
Medicare fee-for-service claims (IP and OP files)	Claim level	2016	Medicare	Inpatient & outpatient
Medicare encounter records (IP and OP files)	Claim level	2016	Medicare	Inpatient & outpatient

Input Files	Unit	Year	Payer	Setting
Medicare Part D event data	Claim level	2016	Medicare	Pharmacy
Commercial—member	Person level	2016	Commercial	NA
Commercial—inpatient	Claim level	2016	Commercial	Inpatient
Commercial—outpatient	Claim level	2016	Commercial	Outpatient
Commercial—pharmacy	Claim level	2016	Commercial	Pharmacy
CMS Provider of Services (POS) Current file	Provider level	2016	Medicare & Medicaid	NA

*Outpatient includes emergency department (ED) visits.

File Structure of SyH-DR

SyH-DR includes:

- Data in CSV, SAS, or Stata format.
- More than 18 million people with over 30 million hospital service claims (i.e., inpatient and outpatient) and 180 million pharmacy claims.
- Person-level weights to produce nationally representative estimates.
- SyH-DR documentation and tools—including codebook and programs for loading the CSV data into SAS and Stata.

Description of SyH-DR Files

SyH-DR has the following 14 data files available for download. In addition to the data files, Stata and SAS read-in programs are available to load each of the CSV files into the respective software.

Claims Files

- Commercial Inpatient File
- Commercial Outpatient File
- Commercial Person-Level File
- Commercial Pharmacy File
- Medicaid Inpatient File
- Medicaid Outpatient File
- Medicaid Person-Level File
- Medicaid Pharmacy File
- Medicare Inpatient File
- Medicare Outpatient File
- Medicare Person-Level File
- Medicare Pharmacy File

Provider Files

- Medicaid Provider File
- Medicare Provider File

Below, we describe the type of information contained in each of the SyH-DR data files. SyH-DR includes person files, inpatient and outpatient claims files, pharmacy claims files, and provider files (Medicare and Medicaid only).

- **Person files:** These files contain demographic and eligibility information for samples of people (i.e., beneficiaries/members) enrolled by Medicare, Medicaid, and/or commercial payers in the 2016 calendar year. It is important to note that even though a person was enrolled, they may not have used any inpatient, outpatient, or pharmacy services in the calendar year. Thus, the samples may include people who may not have had any claims in the inpatient, outpatient, or pharmacy files. Person demographics include data elements such as age, race, sex, and ZIP Code of residence. Payer eligibility information includes data elements such as program eligibility, Medicare and Medicaid dual eligibility status, and eligibility for pharmacy benefits.
- **Inpatient and outpatient files:** These files contain institutional claims (i.e., inpatient hospitalizations and outpatient visits, including emergency department visits), for sampled people appearing in the person files. The unit of observation in the inpatient files is an inpatient stay record and in the outpatient files, outpatient visit records (claim).
- **Pharmacy files:** The pharmacy files contain prescription claims for sampled people appearing in the person files. The pharmacy files have two types of prescription-related claims, cases where a person received medication in an inpatient hospital setting (i.e., inpatient stays with prescribed medicine) and cases where a person received a prescription drug from a pharmacy. The unit of observation is either an inpatient stay claim (where a person received medicine) or a pharmacy claim (i.e., filled by a pharmacy).
- **Provider files:** These files contain a limited set of hospital characteristics derived from the CMS Provider of Services (POS) Current Files. The records in the provider files can be linked to the claims files using the facility ID. Further, records in these files have been aggregated to reduce re-identification risk of facilities. Note, these files are only available for the Medicare and Medicaid payers and will only include facilities currently certified with CMS.

Harmonization To Create Uniform All-Payer Database and SyH-DR Data Elements

The goal of harmonization is to increase the interoperability of disparate data sources. Standardization techniques such as renaming variables, relabeling characteristics (i.e., formats and typing), and reassigning variable value definitions are applied to each of the variables across the source data files. The final product of harmonization is a set of data files with the same variable names and characteristics and shared coding schema. For example, sex code (present in all source data) was renamed to Sex_Ident_Cd. Further, the values for sex code in the Medicare file were recoded from a numeric system (i.e., 0, 1, 2) to the common coding scheme (i.e., U, M, and F) that are shared by the non-Medicare payer data files (i.e., Medicaid and commercial). The SyH-DR Codebook report contains the final code value definitions for each of the variables in the data files.

In addition to variable standardization, a series of inclusion/exclusion criteria were applied to each data file to harmonize the types of claims included in each source file. For example, only the approved final action claims were retained in the source files. The following criteria were applied to the source data used to develop SyH-DR. The data must be:

- Paid/Approved Final Action Claims,

- From inpatient hospitals (including hospital inpatient and religious non-medical hospital inpatient) or outpatient facilities (including outpatient hospitals, ambulatory surgical centers, and critical access hospitals), and
- Fee-for-service or encounter records.
- In rare cases, claims data can sometimes contain hundreds of revenue or Current Procedure Terminology (CPT) codes on a single claim. To reduce the number of unique codes captured in a claim and, by extension, processing time and resources, the first 35 line items (per unique claim ID) are kept from each of the IP, OP, and RX line/revenue files.

Sample Design for SyH-DR

Source data were sampled with two goals in mind. The first goal was to create a nationally representative all-payer database that roughly included an equal proportion of people covered by each payer (Medicare, Medicaid, Commercial), with respect to the populations covered by the given payer. The second goal was to oversample certain subgroups. This approach ensures the typically rare subgroups would have sufficient sample sizes for analysis and provide maximum analytic utility. A more detailed description of the sampling methods is available in the **SyH-DR Sampling, Weighting, and Synthetization Methodologies** report.

Nationally Representative Person-Level Weights in SyH-DR

SyH-DR includes person-level weights that were designed to produce nationally representative estimates of person-level characteristics and hospital service utilization. Raking methods were used to produce weights that conform demographics and hospital service utilization data of the samples to the populations covered by each payer, using American Community Survey and Healthcare Cost and Utilization Project data, respectively. A more detailed description of the methods used to create the weights is available in the **SyH-DR Sampling, Weighting, and Synthetization Methodologies** report.

Using the Weights To Obtain Nationally Representative Estimates

When accounting for sampling and weighting adjustments, users should use the person weight (PERSON_WGHT) for their analysis to project the demographics and hospital service utilization of the SyH-DR sample to the entire U.S. population with health insurance in 2016. However, when creating estimates representing certain populations, users should check estimates against other available (public or private) data sources.

Synthetization Methodology

SyH-DR is a partially synthesized database, meaning that some data elements were fully synthesized, some were partially synthesized, and some were not synthesized. Synthetization was performed at the claim level. Therefore, the claims in SyH-DR are the original claims with some data elements partially or fully replaced by synthetic values.

All person-level data elements were not synthesized, although they may differ from the source data due to de-identification steps. 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., the first 3-digits [diagnosis category] of the synthetic diagnoses were identical to original diagnoses). In all cases, synthetic data elements were created to resemble the marginal distributions of the original data elements. The complete synthetization methodology is available on the SyH-DR **Sampling, Weighting, and Synthetization Methodologies** report.

Appendix A lists each of the data elements in SyH-DR and whether the variable has been synthesized (partial or full).

Identifier Masking Methodology

Variables that contain source identifier (ID) numbers (e.g., Medicare Beneficiary ID) have each been recoded into a new numbering system. Assigning a new number to each of these variables means users will no longer be able to link the person, facility, or claim record back to the source data. It is important to note that the masked number and source number maintain a 1-to-1 relationship. This ensures that each person, facility, or claim can be uniquely identified (i.e., no two people, facilities, or claims share the same masked ID) and that SyH-DR files are still linkable within the payer type using the masked person ID and facility ID. Appendix A lists each data element in SyH-DR and whether the variable has been masked.

De-Identification Methodology

Using the appropriate risk assessment methods, we identified several variables as high risk (e.g., rare ICD-10 diagnosisDX codes or enrollment indicators that reveal partial date of birth or death) and made their values suppressed in SyH-DR. Several suppression methods were applied to the "high-risk" variables in SyH-DR. Such methods included value aggregation, changing of values to missing, random noise, values reassignment, rounding, record deletion, and top/bottom coding. Appendix A lists each data element in SyH-DR and whether the variable has been suppressed.

Implications of Synthetization, Identifier Masking, and De-Identification for Research

SyH-DR is constructed in a way that balances analytical utility with disclosure protection. Since SyH-DR is partially synthesized, masked, and de-identified, it has some limitations for research compared with identifiable claims data. Users should note the following elements of the data:

- ID variables (i.e., member, provider, and claim IDs) have been masked. The original ID value has been changed to a completely new unique identifier that cannot be linked back to the source data files.
- SyH-DR supports person-level longitudinal analyses, by payer. Within each payer, the person, inpatient, outpatient, and pharmacy files can be linked using the person ID to track a person's hospital visits and pharmacy claims throughout 2016. Further, this linkage allows for the association of medical or pharmacy services with their demographics. It is important to note, because person IDs are specific to a payer, people cannot be linked across payers. It is possible that people may be insured by multiple

payers and therefore present in the data for multiple payers, but this is not known, and those people would have different identifiers for each payer.

- Synthetic values were generated from models trained on the original data. Although the models attempt to capture statistical dependencies between key variables, the synthetic data may not capture statistical relationships among any given set of variables. Users with research questions that span multiple variables, one or more of which are synthesized, are advised to validate results from SyH-DR against other data sources.
- Synthetic diagnoses, procedures, and drugs were generated so that univariate distributions of these variables mimicked distributions in the source data. Due to a stochastic (i.e., random) element in the synthetic data generation process, the frequencies of synthetic diagnoses, procedures, and drugs will differ from those in the source data. While these differences are usually small, they may be pronounced for rare diseases, procedures, or drugs. As such, researchers are advised to exercise caution when using SyH-DR to study rare diseases beyond the diagnosis category level (i.e., with a more granular ICD-10 diagnosis code than the first three characters), rare procedures beyond the Clinical Classifications Software category level, and rare drugs beyond the therapeutic class level. By the same principle, researchers should exercise caution when interpreting results from small domains (e.g., estimating the prevalence of a disease among a certain age group in a ZIP Code).
- SyH-DR cannot be used to study newborns since all claims that directly identified a newborn were removed.
- Age is provided only in bins, so researchers cannot use SyH-DR to analyze events that happen at an exact year of age or for people with an exact birth date. However, the age bins were designed to be consistent with age bins used by the Census Bureau and other data sources so that SyH-DR can be merged with other data sources by age bin.
- ZIP Codes with a very small population were aggregated to a lower number of digits, so not all five-digit ZIP Codes are available.
- A small amount of random noise was added to length of stay, so use of SyH-DR for analyses that rely on an exact length of stay should be done with caution. In addition to the added noise, length of stay has been capped at 60 days.
- SyH-DR has no mortality data for people insured by commercial plans.
- Mortality data via discharge status are available for Medicare and Medicaid, but they are synthesized, and the ZIP Codes were randomly switched to relatively near ZIP Codes for individuals coded as expired.
- Mortality information is not available via enrollment flags.
- The Children's Health Insurance Program (CHIP) indicator applies only to people less than 18 years of age in SyH-DR, so pregnant women with CHIP are not indicated.

Appendix A. List of SyH-DR data elements by masked, synthesized, and suppressed status

Variable Description	Variable Name	Location in SyH-DR (Claims, Person, RX, Provider)	Masked	Synthesized (Full, Partial, No)	Suppressed
Person ID	Person_ID	Claims, Person, RX	Yes	No	No
Person Weight	Person_Wght	Claims, Person, RX	No	No	Yes, rounded to nearest hundredth (i.e., two digits after decimal)
Facility ID	Facility_ID	Claims, Provider	Yes	No	No
Claim ID	Clm_Cntl_Num	Claims	Yes	No	No
Pharmacy Claim ID	Phmcy_Clm_Num	RX	Yes	No	No
Beneficiary ID	Bene_ID		Yes	No	No
Medicaid Submitting State Code	MCaid_Submttg_St_Cd	Claims, Person, RX	No	No	No
Age (Low/High)	Age_Low/Age_High	Person	No	No	Yes, age values top coded to 85
Sex/Gender Code	Sex_Ident_Cd	Person	No	No	No
Race Code	Race_Cd	Person	No	No	Yes, several values aggregated into 'other'
ZIP Code	ZIP_Cd	Person	No	No	Yes, certain 5-digit values have been changed to 2/3 digits. For those who expired in the facility, a new ZIP Code was assigned (Medicare & Medicaid).
County Code	County_FIPS_Cd	Person	No	No	No
State Code	State_Cd	Person	No	No	No
Commercial Insurance Eligibility	Cmrcl_Insrc_XX (XX between 1-12)	Person	No	No	No
CMS Medicare & Medicaid Dual Eligible Status Codes	Dual_Elgbl_XX (XX between 1-12)	Person	No	No	Yes, certain values changed to 1/0 to avoid disclosure of DOB or DOD.

Medicaid CHIP Enrollment Code	Mdcd_Chip_Enrlmt	Person	No	No	Yes, monthly values aggregated into annual indicator. Ages 0-18 only.
Medicaid Enrollment Indicators	Mdcd_Enrolmt_XX (XX between 1-12)	Person	No	No	No
Medicaid HMO Enrollment Indicators	Mdcd_MCO_enrolmt_XX (XX between 1-12)	Person	No	No	No
Medicare Entitlement Indicators	Mdcr_Entlmt_Ind_XX (XX between 1-12)	Person	No	No	Yes, certain values changed to 1/0 to avoid disclosure of DOB or DOD.
Medicare HMO Coverage Indicators	Mdcr_HMO_Cvrg_XX (XX between 1-12)	Person	No	No	Yes, certain values changed to 1/0 to avoid disclosure of DOB or DOD.
Pharmacy Coverage Indicators (Medicare and Commercial Only)	Phrmcy_Cvrg_XX (XX between 1-12)	Person	No	No	(Medicare Only) Yes, certain values changed to 1/0 to avoid disclosure of DOB or DOD.
Reason for Enrollment Code (Medicare and Medicaid Only)	Rsn_Enrlmt_Cd	Person	No	No	Yes, values aggregated into broad categories
Medicaid Restrictive Benefits Indicator	Rstrctd_Bnfts_Cd	Person	No	No	Yes, monthly values aggregated into annual indicator
Admission Type	Admsn_Type	Claims	No	Yes	Yes, newborn admissions removed from SyH-DR.
Attending Specialty	AT_Spclty	Claims	No	Full	No
Claim Type Code	Clm_Type_Cd	Claims	No	No	No
CPT Procedure Codes	CPT_Prcdr_Cd_X (Where X between 1-35)	Claims	No	Partial, assigned code value in same CCS as original claim.	No
Discharge Status	Dschrg_Stus	Claims	No	Full	Yes, certain codes aggregated into ‘other’
ICD Procedure Code	ICD_Prcdr_Cd_X (Where X between 1-25)	Claims	No	Partial, assigned code value in same CCS as source claim.	No

Length of Stay (LOS)	LOS	Claims	No	No	Yes, random noise added post-synthetization. Top coded to 60 days.
Plan Payment Amount	Plan_Pmt_Amt	Claims, RX	No	Full	No
Primary/Secondary ICD-10 Diagnosis (DX) Codes	Prmry_DX_Cd/ ICD_DX_Cd_X (Where X between 1-25)	Claims	No	Partial, first three digits same as source DX code.	Yes, several diagnosis values changed to missing values
Service Begin/End Dates	Srvc_Beg_Date/ Srvc_End_Date	Claims	No	No	No
Type of Bill Code	TOB_Cd	Claims	No	No	No
Total Charged Amount	Tot_Chrg_Amt	Claims, RX	No	Full	No
Fill Date	Fill_DT	RX	No	No	No
Generic Drug Name	Generic_Drug_Name	RX	No	Partial, assigned name in the same Multum Therapeutic Class as source claim.	No
Line Number	Line_Nbr	RX	No	No	No
Provider Category Code	Prvdr_Ctgry_CD	Facility	No	No	Yes, values aggregated into broader categories
Provider Ownership Code	Prvdr_Ownrshp_CD	Facility	No	No	Yes, values aggregated into broader categories
Provider Participation Code	Prvdr_Prtcptn_Cd	Facility	No	No	No

AHRQ Publication No. 24-0019-1-EF
December 2023
www.ahrq.gov