Frequently Asked Questions (FAQs)

January 15, 2013

The Centers for Medicare and Medicaid Services (CMS) *Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use Files (DE-SynPUF)*

1. What is the CMS Linkable 2008–2010 Medicare data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)?

The *CMS Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)* is a set of free downloadable files containing a subset of the data fields contained in inpatient, outpatient, carrier, and prescription drug event (PDE) claims and the beneficiary summary files. The files represent a synthetic 5% sample of 2008 Medicare beneficiaries and their claims from 2008 to 2010. The *CMS Linkable 2008–2010 Medicare DE-SynPUF* was designed to provide data that would be useful for data entrepreneurs for software and application development and research training purposes. The decisions regarding which variables to include in the *DE-SynPUF* were based on suggestions from the CMS claims data experts and data entrepreneurs. Each file contains the same set of variables in each year. The variable names in the *DE-SynPUF* were kept the same as those in the actual Medicare data unless they were significantly coarsened to decrease re-identification risk. In those cases, "SP_" was added to the original variable name for distinguishing.

2. Why was the CMS Linkable 2008–2010 Medicare DE-SynPUF created?

CMS is committed to increasing access to its Medicare claims data. As a first step to meeting the demand for linkable, multiple-year Medicare data, CMS provides the *Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use Files (DE-SynPUF)*.

3. How many years of data does the CMS Linkable 2008–2010 Medicare DE-SynPUF contain?

The *DE-SynPUF* represents a 5% sample of synthetic 2008 Medicare beneficiaries and their synthetic claims from 2008 to 2010.

4. What types of files does the CMS Linkable 2008–2010 Medicare DE-SynPUF contain?

The *DE-SynPUF* contains five types of files: the *CMS Beneficiary Summary DE-SynPUF*, the *CMS Inpatient Claims DE-SynPUF*, the *CMS Outpatient Claims DE-SynPUF*, the *CMS Carrier Claims DE-SynPUF* (also known as the Physician/Supplier Part B Claims file), and the *CMS Prescription Drug Events (PDE) DE-SynPUF*.

5. What is the CMS Beneficiary Summary DE-SynPUF?

The *CMS Beneficiary Summary DE-SynPUF* contains information on demographics, health plan enrollment, chronic conditions, and reimbursement based on a sample of synthetic Medicare beneficiaries. This is a beneficiary-level file (i.e., each record represents a synthetic Medicare beneficiary).

6. What is the CMS Inpatient Claims DE-SynPUF?

The *CMS Inpatient Claims DE-SynPUF* contains synthetic institutional claims for hospital inpatient services provided to the synthetic Medicare beneficiaries in the *CMS Beneficiary Summary DE-SynPUF* in 2008, 2009, and 2010. Each record in an inpatient file pertains to a synthetic inpatient claim.

7. What is the CMS Outpatient Claims DE-SynPUF?

The *CMS Outpatient Claims DE-SynPUF* contains synthetic institutional claims for outpatient services provided to the synthetic Medicare beneficiaries in the *CMS Beneficiary Summary DE-SynPUF* in 2008, 2009, and 2010. Each record in an outpatient file pertains to a synthetic outpatient claim. A synthetic Medicare beneficiary in the *CMS Beneficiary Summary DE-SynPUF* could have any number of outpatient claims, including no claims in a given year.

8. What is the CMS Carrier Claims DE-SynPUF?

The *CMS Carrier Claims DE-SynPUF* contains synthetic noninstitutional claims for physician/supplier services provided to those synthetic Medicare beneficiaries in the *CMS Beneficiary Summary DE-SynPUF* in 2008, 2009, and 2010. Each record in a carrier file pertains to a synthetic physician/supplier claim. A synthetic Medicare beneficiary in the *CMS Beneficiary Summary DE-SynPUF* could have any number of physician/supplier claims, including no claims in a given year.

9. What is the CMS Prescription Drug Events (PDE) DE-SynPUF?

The *CMS Prescription Drug Events (PDE) DE-SynPUF* contains synthetic PDEs provided for those synthetic Medicare beneficiaries in the *CMS Beneficiary Summary DE-SynPUF* in 2008, 2009, and 2010. Each record in this PDE file pertains to a synthetic drug event, but may be considered a claim. A synthetic Medicare beneficiary in the *CMS Beneficiary Summary DE-SynPUF* could have any number of drug events, including no drug events in a given year.

10. Can I know which claims/events belong to the same Medicare beneficiary?

Yes. A unique cryptographic identifier, *DESYNPUF_ID*, identifying the synthetic beneficiaries was provided in each *CMS Linkable 2008–2010 Medicare DE-SynPUF*. Users can link two or more *DE-SynPUF* files using the *DESYNPUF_ID* as the linking key. Note that the *DESYNPUF_ID* was specifically created for the *DE-SynPUF* to identify synthetic beneficiaries. This identifier carries no information about the patient or any patient records, and is provided solely for reference and data processing purposes.

11. Why are some variables in actual Medicare claims not available in the CMS Linkable 2008–2010 Medicare DE-SynPUF?

To protect the privacy of Medicare beneficiaries and provide manageable files, not all variables in the actual data sets were included in the *DE-SynPUF*. However, the *DE-SynPUF* was designed to contain the primary variables of value to data entrepreneurs and researchers.

12. How is the CMS Linkable 2008–2010 Medicare DE-SynPUF different from the 5% CMS standard research sample?

There is no overlap in terms of beneficiaries between the 5% CMS standard research sample and the *CMS Linkable 2008–2010 Medicare DE-SynPUF*. These two 5% samples are disjoint.

Furthermore, the *DE-SynPUF* does not contain the actual data for its source 5% sample but rather synthetic data for the beneficiaries and their claims. Therefore, the data in the *DE-SynPUF* should not be used for inference to the actual Medicare population.

13. What are the limitations of the CMS Linkable 2008–2010 Medicare DE-SynPUF?

Because the *DE-SynPUF* has the structure of Medicare data, these synthetic files suffer the same data limitations, as do Medicare data. The Research Data Assistance Center (ResDAC) offers *ResConnect* to provide researchers with an in-depth explanation of common CMS procedures, files, variables, and utilities.[1] The Chronic Condition Data Warehouse Web site also offers analytic guidance.[2] Users of the file who are not familiar with Medicare data are strongly advised to visit ResDAC or the Chronic Condition Data Warehouse to learn more about Medicare data whenever they have questions about using the *DE-SynPUF*.

Moreover, all variables in the *DE-SynPUF* are imputed, suppressed, and coarsened as part of disclosure treatment. As a result, the *DE-SynPUF* has very limited inferential research utility because of the synthetic process used to generate the data. That is, analyses using the *DE-SynPUF* to draw inferences about Medicare beneficiaries, providers, or the Medicare program will be misleading and often incorrect.

14. Can I use the CMS Linkable 2008–2010 Medicare DE-SynPUF to conduct empirical research and infer the results to the actual Medicare population?

NO. The *DE-SynPUF* preserves the detailed data structure of key variables at both the beneficiary and claim levels. However, the data are fully "synthetic" for disclosure safety. As a result, much of the interdependence and co-variation among variables have been altered. The efforts to reduce the risk of re-identification diminish the analytic utility of the files in the sense that univariate and multivariate statistics, such as conditional correlations and regression coefficients, would not be reliable.

15. Then what can I use the CMS Linkable 2008–2010 Medicare DE-SynPUF for?

The primary purpose of the *DE-SynPUF* is to aid data entrepreneurs and researchers in understanding the complexity and wealth of information in the data, and to develop software applications or research protocols that use the data. The DE-SynPUF itself is not designed to allow for inferences to be made about the actual Medicare beneficiary population.

The analytic utility of the data file differs based on the type and level of analysis being conducted:

- **Demographic**: The *DE-SynPUF* estimates of demographic characteristics (date of birth, date of death, sex, race, state, county) of the beneficiary population match the univariate frequency of the full population of beneficiaries enrolled in Medicare at any time during the 2008 year.
- **Clinical**: The *DE-SynPUF* estimates for clinical variables such as chronic conditions can provide researchers with bounds on how many cases with a specific condition are likely to be in the Medicare claims, which could be used to generate power calculations for a grant application.
- **Economic/financial**: The *DE-SynPUF* estimates for the economic and financial variables provide a *lower bound* for the true estimate of cost for the full population of beneficiaries enrolled in Medicare at any time during the 2008 year and costs for 2009 and 2010 for this 2008 beneficiary example.
- **Multivariate modeling**: The dynamic relationships between variables (demographic, health plan enrollment, clinical, economic/financial, and provider information) were altered, to limit re-identification risk. Therefore, analyses from multivariate modeling should be interpreted with caution. However, the programs and procedures employed in the multivariate modeling will function on the CMS Limited Data Sets or Identifiable Data prior to 2011.

[1] Research Data Assistance Center. ResConnect. http://www.resdac.org/resconnect
[2] Chronic Condition Data Warehouse. "Analytic Guidance." http://www.ccwdata.org/analytic-guidance/index.htm

16. How was the CMS Linkable 2008–2010 Medicare DE-SynPUF created?

The *CMS Linkable 2008–2010 Medicare DE-SynPUF* originated from a disjoint (mutually exclusive from **existing** samples) 5% random sample of beneficiaries from the 100% Beneficiary Summary File for 2008. To exclude any overlap with the beneficiaries in the existing 5% CMS research sample,[3] the beneficiaries in that other sample were excluded, and a 5-in-95 random draw was made with the remaining 95% of beneficiaries. A variety of statistical disclosure limitation techniques were used to protect the confidentiality of Medicare beneficiaries in the *CMS Linkable 2008–2010 Medicare DE-SynPUF*. The *DE-SynPUF* was created by starting with an actual beneficiary as a "seed" for a synthetic beneficiary. Synthetic beneficiaries and their claims are based on actual seed beneficiaries. Disclosure is reduced through multiple deterministically or stochastically applied treatment mechanisms. First, hot decking based procedures are used to find donors for beneficiary-level variables and individual claims. Second, other synthetic processes are used to protect other elements of the data. Disclosure limitation methods used in the process include variable reduction, suppression, substitution, synthesis, date perturbation, and coarsening. Please refer to the *CMS Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) User Manual* for details regarding how *DE-SynPUF* was created.

17. What has been done to protect the privacy of Medicare beneficiaries?

Of paramount importance in the release of the *DE-SynPUF* is the protection of beneficiary confidentiality. To that end, all directly identifiable information has been altered in accordance with HIPAA privacy rules.

The following disclosure limitation methods were also used:

- **Variable Reduction**: Limit the number of variables in each table to a set that is useful and appropriate for development users.
- **Suppression**: Remove records, either beneficiaries or claims, which are rare in the data and have disclosure risk even in a synthetic file by applying appropriate k-anonymity rules based on either population- or sample-specific counts.
- **Substitution**: Alter values of variables by replacing with values from a similar donor record, based on key variables (i.e., conditional on matching certain variables).
- **Imputation**: Alter values of single variables by drawing values from empirical distributions conditioned on key variables. The empirical distributions are first coarsened and truncated, removing potentially identifying values.
- **Date Perturbation**: Alter timelines by changing dates and intervals between events.
- **Coarsening**: Numeric variables (e.g., year of birth or expenditures) coarsened enough to limit disclosure but remain realistically useful.

Please refer to *the CMS Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) User Manual* for more information.

[3] "Differences in How the Medicare 5% Files Are Generated." Technical Brief, ResDAC Publication Number TN-011, March 2007. Research Data Assistance Center, University of Minnesota, Minneapolis, MN.

18. How was provider confidentiality protected?

To improve provider confidentiality, some of the disclosure limitation methods described in FAQ #17 were implemented on provider variables. All codes identifying provider institutions or physicians were altered to reduce the likelihood of their identification. Synthetic providers and physicians are associated with synthetic beneficiaries based solely on the geography of the synthetic beneficiaries they serve. Because these fields are random numbers and characters, with no association to any known ID number, analytic inferences to Medicare providers should not be made when using these variables.

19. What data cleaning steps were performed to obtain the initial 5% Medicare sample?

To mimic the actual Medicare data, no data cleaning steps were performed. As a result, data anomalies (e.g., zero drug costs or zero days supply in *DE-SynPUF* PDE, negative reimbursement payment in beneficiary files) should be expected.

20. How can I learn more about how to use the CMS Linkable 2008–2010 Medicare DE-SynPUF?

Users of the file who are not familiar with Medicare data are strongly advised to visit Research Data Assistance Center (ResDAC)[4] and the Chronic Condition Data Warehouse[5] to learn about Medicare data whenever they have questions about using *DE-SynPUF*. ResDAC provides education and training opportunities to researchers. Please refer to the ResDAC training website for more information.[6] Moreover, ResDAC offers ResConnect to provide researchers with an in-depth explanation of common CMS procedures, files, variables, and utilities.[7] The Chronic Condition Data Warehouse Web site also offers analytic guidance.[8]

21. What is the plan for future data releases?

At this time, it is unknown whether there will be future data releases of the *DE-SynPUF*. Future releases of the *DE-SynPUF* are dependent on user response and feedback to the files. Please send feedback and comments to Research Data Assistance Center (http://www.resdac.org/) via resdac@umn.edu or 1-888-9RESDAC.

22. How may I provide feedback on the CMS Linkable 2008–2010 Medicare DE-SynPUF?

Questions and comments can be submitted to the Research Data Assistance Center (http://www.resdac.org/) via resdac@umn.edu or 1-888-9RESDAC.

[4] http://www.resdac.org/
[5] http://www.ccwdata.org/
[6] Research Data Assistance Center. Training. http://www.resdac.org/training
[7] Research Data Assistance Center. ResConnect. http://www.resdac.org/resconnect
[8] Chronic Condition Data Warehouse. Analytic Guidance. http://www.ccwdata.org/analytic-guidance/index.htm