DEPARTMENT OF HEALTH & HUMAN SERVICES
Centers for Medicare & Medicaid Services
Center for Consumer Information & Insurance Oversight
200 Independence Avenue SW
Washington, DC 20201

Subject: Creation of the 2017 and 2018 Benefit Year Enrollee-Level EDGE Limited Data Sets: Methods, Decisions and Notes on Data Use

Date: December 4, 2020

Introduction

This memo documents the methods, decisions, and key steps taken in creating the 2017 and 2018 benefit year enrollee-level External Data Gathering Environment (EDGE) limited data sets (LDS). Because the data used to create the LDS was collected for recalibration of the models used in the HHS-operated risk adjustment program established under section 1343 of the Patient Protection and Affordable Care Act (PPACA),[1] the Centers for Medicare & Medicaid Services (CMS) needed to establish the elements of data handling and analytic file construction. Specifically, CMS needed to explain the elements and file constructs that required attention or departed from the procedures already in place with the other data used for the HHS-operated risk adjustment program (MarketScan® data). In the 2020 Payment Notice, CMS finalized a policy to create and make available, on an annual basis, enrollee-level EDGE data as a LDS file for qualified requestors who seek these data for research purposes.[2] CMS has made this LDS available beginning with the 2016 benefit year.[3]

This memo describes the development and processing of the 2017 and 2018 benefit year enrollee-level EDGE data extract to create the enrollee-level EDGE LDS files ("LDS sample"). The enrollee-level EDGE extract data files were produced by creating an extract from the issuers' EDGE servers reflecting 2017 and 2018 benefit year data for risk adjustment covered plans in the individual and small group (or merged) markets, where HHS operated the risk adjustment program.[4] Files were extracted consisting of four components: an enrollment file (RARECALE), medical claims (RARECALM), pharmaceutical claims (RARECALP), and supplemental claims (RARECALS). The rest of this document outlines how CMS used the 2017 and 2018 enrollee-level EDGE data files to create the LDS sample, identify any changes that were made for the LDS sample, and provide suggestions as to how to use certain data elements.

2017 LDS Sample Counts (Unredacted).[5] There are a total of 30,940,022 unique SYSIDs[6] in the 2017 unredacted LDS sample enrollment file, 22,169,499 unique SYSIDs with medical claims, 18,730,269 unique SYSIDs with pharmacy claims, and 904,955 unique SYSIDs with supplemental diagnoses. There are 48,265,770 observations in the unredacted enrollment file, 583,678,933 observations in the unredacted

[1] Consistent with section 1321(c)(1) of the PPACA, HHS is responsible for operating the risk adjustment program on behalf of any state that elects not to do so.
[2] Patient Protection and Affordable Care Act; HHS Notice of Benefit and Payment Parameters for 2020, Final Rule (2020 Payment Notice), 84 FR 17454 (April 25, 2019). Available at: https://www.federalregister.gov/documents/2019/04/25/2019-08017/patient-protection-and-affordable-care-act-hhs-notice-of-benefit-and-payment-parameters-for-2020. For the definition of "limited data set," see 45 CFR 164.514(e).
[3] For more details, see here: https://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/LimitedDataSets/index.htm.
[4] For the 2017 and 2018 benefit years, HHS operated the risk adjustment program in all states and the District of Columbia.
[5] See the below discussion about the *Redaction of Substance Use Disorder Claims for Certain Entities* regarding the LDS sample counts in the redacted LDS files.
[6] SYSIDs are system-generated random numbers used to link the unique enrollee records across files.

medical file, 256,156,159 observations in the unredacted pharmacy file, and 7,131,971 observations in the unredacted supplemental file.[7]

2018 LDS Sample Counts (Unredacted).[8] There are a total of 29,609,700 unique SYSIDs in the 2018 unredacted LDS sample enrollment file, 21,442,989 unique SYSIDs with medical claims, 18,461,557 unique SYSIDs with pharmacy claims, and 1,084,873 unique SYSIDs with supplemental diagnoses. There are 42,532,405 observations in the unredacted enrollment file, 570,866,925 observations in the unredacted medical file, 245,558,612 observations in the unredacted pharmacy file, and 8,054,966 observations in the unredacted supplemental file.

Age and Sex. Consistent with the 2016 LDS sample, we have excluded data for enrollees above age 99 and enrollees with a sex field that indicates "unknown." Additionally, for the 2017 and 2018 LDS sample files, we censored the age data field to 89 for enrollees with ages greater than 89. That is, the age for enrollees age 89 and above is listed as 89.

Months of enrollment. Consistent with the 2016 LDS sample, the 2017 and 2018 LDS samples include the enrollment length – days and enrollment length – months fields. Enrollment dates were excluded in the LDS files due to privacy concerns related to identifying enrollees' dates of birth in combination with the age variable. The enrollment length – months field was calculated as the EDGE data element "Enrollment Length – Days" divided by 30 days and rounded to two decimal places.

Metal and CSR variant identifiers. Consistent with the 2016 LDS sample, to prevent the identification of enrollees in plans that had small sample sizes in certain combinations of metal and CSR levels, we excluded data for enrollees older than 30 years of age in catastrophic plans in the LDS sample. Similarly, for all enrollees in American Indian and Alaska Native (AI/AN) cost-sharing reduction plan variants (limited cost-sharing or zero cost-sharing plans), Medicaid expansion private plans, or cost-sharing wrap plan variations, we replaced the CSR data field with an "11" and a missing value for the metal level data field.

New in 2017 and 2018 LDS Samples

SYSIDs. SYSIDs are system-generated random numbers used to link the unique enrollee records across files. Beginning with the 2017 and 2018 benefit years, SYSIDs were generated such that individuals enrolled with the same issuer both years could be associated in the next year's dataset even if they changed plans. As such, an individual enrollee with data in both the 2017 and 2018 LDS samples who was enrolled with the same issuer will have the same SYSID in each sample. That individual's SYSID will not be the same in the 2016 LDS sample, if they also have data in that benefit year.

Market Indicator. We did not have enrollees' market (individual, small group) for the 2016 benefit year EDGE claims dataset. We extracted this data field (individual or small group, including for enrollees in merged market states) beginning with the 2017 benefit year enrollee-level EDGE extract. Issuers associated enrollees in merged market states to either the individual or small group market based on the type of coverage sold.

Claims. Unlike the 2016 EDGE extract, the 2017 and 2018 enrollee-level EDGE extracts include a claim identifier field. The unit of observation in the medical claims data is an individual line item, which

[7] An enrollee may have more than one observation in the enrollment file for separate enrollment records submitted to the EDGE servers, and multiple observations in the claims file for each separate service and claim record.
[8] See the below discussion about the *Redaction of Substance Use Disorder Claims for Certain Entities* regarding the LDS sample counts in the redacted LDS files.

describes a service or item being billed to insurance. Claims are composed of one or more line items. This claim identifier field allows for reliably combining line items into complete claims. Pharmacy claims do not have a claim identifier field and are not composed of more than one line item. The unit of observation in the pharmacy claims is a drug fill.

Duplicate 2018 Records. The 2018 enrollee-level EDGE extract includes duplicate records for some enrollees. These duplicates occur in all file types: enrollment, medical claims, pharmacy claims, and supplemental diagnoses. For recalibration, particularly for the medical and pharmacy claims to avoid double-counting of allowed charges, we excluded records where all fields identically matched another record. For the 2018 LDS sample, we left the duplicates in to provide researchers with the original, unedited dataset to reflect how the data was received, but researchers may want to exclude identical, duplicate records in all 2018 files. In addition to the records that are exact duplicates, we also note that a very small number of records in the 2018 medical claims file are near identical duplicates, with the order of the modifiers listed in the service code modifier field being the only difference between two records. While there may still be some duplicates, the 2017 LDS sample does not have the same duplicate issues as the 2018 LDS sample.

Redaction of Substance Use Disorder Claims for Certain Entities

In order to comply with Substance Abuse and Mental Health Services Administration (SAMHSA) 42 CFR Part 2 requirements, Substance Use Disorder (SUD) claims will be redacted for requestors who are not covered entities or business associates as defined by HIPAA.[9]

For the redacted LDS sample medical file, a claim with any of the relevant SUD ICD-10 diagnosis codes[10] was identified and excluded. Given that each header/line for a claim has the same ICD-10 diagnosis codes, this rule effectively excludes the entire claim. CPT/HCPCS procedure codes are also at the claim line level in the medical file. If any claim lines include one of the SUD CPT/HCPCS procedure codes,[11] the entire claim was excluded for the SUD redacted sample. All diagnosis codes in the supplemental diagnoses file were excluded for redacted claims. Additionally, we redacted SUD ICD-10 diagnosis codes in the supplemental diagnoses file, and also redacted the associated claims in the medical file if the supplemental diagnosis was to be added, rather than deleted, from the claim. Finally, we note that issuers do not submit DRGs or ICD-10 procedure codes to their EDGE servers, and therefore, SUD redaction based on such codes was not necessary.[12]

Similar to 2016, we did not redact SUD drugs from the prescription drug data in the LDS sample, as the use of a SUD drug does not imply the patient is being treated for SUD due to the prevalence of off-label prescribing.

2017 LDS Sample Counts (Redacted). After SUD claims redaction in the medical and supplemental files, there were a total of 22,125,709 unique SYSIDs with medical claims and 890,192 unique SYSIDs with supplemental diagnoses. There were 572,254,233 observations in the redacted medical file, and

[9] See 45 CFR 2.52; see also 45 CFR 160.103 for HIPAA definitions of "covered entity" and "business associate."
[10] The list of ICD-10 diagnosis and CPT/HCPCS codes excluded are available here: https://www.resdac.org/articles/redaction-substance-abuse-claims.
[11] Ibid.
[12] Beginning with the 2017 data, the methodology for SUD redaction from EDGE LDS samples incorporates the new claim identifier field. CMS did not extract the claim identifier as part of the 2016 enrollee-level EDGE data extract.

6,929,016 observations in the redacted supplemental file. In the accompanying table, we provide an overview of the impact of SUD redaction on the claims and total allowed amounts.

	2017 LDS Sample		**SUD Redacted 2017 LDS Sample**	
	Observations	**Unique Sysid**	**Observations**	**Unique Sysid**
RARECALE	48,265,770	30,940,022	*No change*	*No change*
RARECALM	583,678,933	22,169,499	572,254,233	22,125,709
RARECALP	256,156,159	18,730,269	*No change*	*No change*
RARECALS	7,131,971	904,955	6,929,016	890,192

2018 LDS Sample Counts (Redacted). After SUD claims redaction in the medical and supplemental files, there were a total of 21,400,116 unique SYSIDs with medical claims and 1,068,808 unique SYSIDs with supplemental diagnoses. There were 560,476,701 observations in the redacted medical file, and 7,828,882 observations in the redacted supplemental file. In the accompanying table, we provide an overview of the impact of SUD redaction on the claims and total allowed amounts.

	2018 LDS Sample		**SUD Redacted 2018 LDS Sample**	
	Observations	**Unique Sysid**	**Observations**	**Unique Sysid**
RARECALE	42,532,405	29,609,700	*No change*	*No change*
RARECALM	570,866,925	21,442,989	560,476,701	21,400,116
RARECALP	245,558,612	18,461,557	*No change*	*No change*
RARECALS	8,054,966	1,084,873	7,828,882	1,068,808