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 2020 Benefit Year Enrollee-Level EDGE Limited Data Sets: Methods, Decisions and Notes on Data Use

Date: August 15, 2022

Introduction

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

This memo describes the development and processing of the 2020 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 2020 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 2020 enrollee-level EDGE data files to create the LDS sample, identifies any changes that were made for the LDS sample, and provides suggestions as to how to use certain data elements.

2020 LDS Sample Counts (Unredacted).[5] There are a total of 28,639,876 unique SYSIDs[6] in the 2020 LDS sample enrollment file, 21,352,459 unique SYSIDs with unredacted medical claims, 17,607,048 unique SYSIDs with pharmacy claims, and 1,175,739 unique SYSIDs with unredacted supplemental diagnoses. There are 50,785,962 observations in the enrollment file, 554,424,178 observations in the unredacted medical file, 234,372,979 observations in the pharmacy file, and 7,048,186 observations in the unredacted supplemental file.[7, 8]

---

[1] Consistent with section 1321(c)(1) of the ACA, 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 at 17486 -17490 (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 HIPAA 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 2020 benefit year, HHS operated the risk adjustment program in all states and the District of Columbia. See 45 CFR 153.20 for the definition of "risk adjustment covered plan".
[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.
[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 claimrecord.
[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.

Age and Sex. Consistent with previous LDS samples, we excluded data for enrollees above age 99 as of December 31, 2020, and enrollees with a sex field that indicates "unknown." Additionally, 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 previous LDS samples, the 2020 LDS sample includes the enrollment length – days and enrollment length – months fields. Enrollment dates were excluded in the LDS files due to privacy concerns related to potential identification of 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 Cost Sharing Reduction (CSR) variant identifiers. Consistent with previous LDS samples, 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 2020 LDS sample. Similarly, for all enrollees in American Indian and Alaska Native (AI/AN) CSR 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.

SYSIDs. SYSIDs are system-generated random numbers used to link the unique enrollee records across files. Beginning with the 2017 benefit year enrollee-level EDGE data, SYSIDs were generated such that individuals enrolled with the same issuer across benefit years could be associated in the next year's dataset even if the individuals changed plans. As such, an individual enrollee with data in more than one of the available EDGE 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 Coverage Type. Consistent with previous LDS samples except the 2016 benefit year LDS, the 2020 LDS includes the market coverage type field. We extracted this data field (individual or small group, including for enrollees in merged market states[9]) 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. We did not have enrollees' market (individual, small group) for the 2016 benefit year EDGE claims dataset.

Claim ID. Consistent with previous LDS samples except the 2016 benefit year LDS, the 2020 LDS includes a Claim ID field. The unit of observation in the medical claims data is an individual line item, which describes a service or item being billed to insurance. Claims are composed of one or more line items. This claim ID field allows for reliably combining line items into complete medical 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. Please note for the 2016 EDGE LDS file there was no claim ID field.

[9] For benefit year 2020, Massachusetts and Vermont are treated as having have a merged market for purposes of the HHS-operated risk adjustment program. See https://www.regtap.info/uploads/library/RA_GuidanceMergedMarkets2017_030118_5CR_030118.pdf..

Redaction of Substance Use Disorder Claims for Certain Entities

To comply with the Substance Abuse and Mental Health Services Administration (SAMHSA) 42 CFR Part 2 requirements, Substance Use Disorder (SUD) claims will be redacted for requestors who do not demonstrate that they meet the requirements for receiving such data for scientific research under 42 CFR 2.52.

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. Current Procedural Terminology (CPT®)/Healthcare Common Procedures Coding System (HCPCS) procedure codes are 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.

Similar to previous data years, 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.

2020 LDS Sample Counts (Redacted). After SUD claims redaction in the medical and supplemental files, there were a total of 21,313,986 unique SYSIDs with medical claims and 1,154,176 unique SYSIDs with supplemental diagnoses. There were 543,782,904 observations in the redacted medical file, and 6,799,081 observations in the redacted supplemental file. In the accompanying table, we provide an overview of the impact of SUD redaction on the sample counts.

	2020 LDS Sample		**SUD Redacted 2020 LDS Sample**	
	Observations	**Unique Sysid**	**Observations**	**Unique Sysid**
RARECALE	50,785,962	28,639,876	*No change*	*No change*
RARECALM	554,424,178	21,352,459	543,782,904	21,313,986
RARECALP	234,372,979	17,607,048	*No change*	*No change*
RARECALS	7,048,186	1,175,739	6,799,081	1,154,176

[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.