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

CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES
CENTER FOR CONSUMER INFORMATION
& INSURANCE OVERSIGHT

Subject: Creation of the 2016 Benefit Year Enrollee-Level EDGE Limited Data Set: Methods, Decisions and Notes on Use

Introduction

This memo documents the methods, decisions, and key steps taken in creating the 2016 benefit year enrollee-level EDGE limited data set (LDS). Because the dataset 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 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 limited data set file for qualified requestors who seek these data for research purposes.[2]

This memo describes the development and processing of the 2016 benefit year enrollee-level EDGE data extract for the HHS-operated risk adjustment program recalibration ("recalibration sample") and the creation of the enrollee-level EDGE LDS files ("LDS sample"). Since the HHS-operated risk adjustment program has used MarketScan® data for its model recalibration since the 2014 benefit year, we structured the enrollee-level EDGE extract data and recalibration sample to align with the MarketScan® data structure where possible to allow for ease in combining the data.

LDS Sample Selection

The enrollee-level EDGE extract data files were produced by an creating an extract from the issuers' EDGE servers reflecting 2016 benefit year data for risk adjustment covered plans in the individual and small group markets, in states where HHS operated the risk adjustment program.[3] Files were extracted in July 2017, 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 2016 enrollee-level EDGE data files and addressed differences between the enrollee-level EDGE data and MarketScan® data for recalibration of the risk adjustment models used in the HHS-operated program. We also identify any changes that were made for the LDS sample and provide suggestions as to how to use certain data elements.

[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 the 2016 benefit year, HHS operated the risk adjustment program in all states except for Massachusetts. As such, this 2016 benefit year dataset does not reflect any data from issuers of risk adjustment covered plans in Massachusetts. See 45 CFR 153.20 for a definition of "risk adjustment covered plan."

LDS Sample Counts. There are a total of 30,634,593 unique SYSIDs[4] in the 2016 LDS sample enrollment file, 21,487,252 unique SYSIDs with medical claims, 17,751,316 unique SYSIDs with pharmacy claims, and 827,879 unique SYSIDs with supplemental diagnoses. There are 44,461,673 observations in the enrollment file, 542,498,817 observations in the medical file, 230,572,809 observations in the pharmacy file, and 8,165,088 observations in the supplemental file.[5]

Orphan enrollment and claims. The data extract specified that "orphan claims" – those that cannot be linked to a specific person – should not be included in the file. Despite this, a large number of SYIDs appeared in the medical claims and pharmaceutical claims files but were absent from the enrollment file. Orphan claims, without corresponding enrollment records, were excluded in the recalibration and LDS samples. Our EDGE contractor later found that certain issuers did not clear temporary staging files, which led to SYSIDs being mismatched between enrollment and claims files. While we excluded orphaned claims records, we had not at first excluded the orphaned enrollment records (i.e., enrollment records that cannot be linked to any claims even though claims may have been billed), as these were identified by our EDGE contractor later on. We had also not at first excluded enrollment records that can be linked to medical claims but cannot be linked to any pharmacy claims even though pharmacy claims may have been billed. Our contractor was able to identify the list of orphaned enrollment records, and associated claims records, which we excluded from the recalibration sample and the LDS sample enrollment files.

Duplicate records. The enrollee-level EDGE extract files contained some records that were exact duplicates of each other, across all fields in the enrollment, medical claims, and pharmacy claims files. For recalibration, we did not use the duplicate enrollment records. Each person was identified by SYSID, so that two records with the same SYSID and the same enrollment (timing and plan type) were treated as a single individual. For the LDS sample, we left the duplicates in. Researchers can choose whether to treat them as multiple enrollments or duplicate records of a single enrollment. In addition to the records that are exact duplicates, there are other cases of overlapping enrollment periods. Because we did not have all the fields necessary to identify truly duplicative claims records, we retained all suspected duplicate records in the claims files in the LDS sample.

Months of enrollment. The LDS sample includes the enrollment length – days and enrollment length – months fields. Enrollment dates were excluded in the LDS files due to privacy concerns of identifying enrollees' date 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.

Capitation. Since the risk adjustment models used in the HHS-operated program are used to evaluate enrollees' expenditures, the risk adjustment model recalibration sample requires meaningful and comparable cost (allowed charges) data. In MarketScan® recalibration data, individuals were excluded if they were enrolled in a capitated plan. Enrollees' plan type information was not available in the enrollee-level EDGE extract, but the presence of capitated claims is a reasonable proxy for it. The "derived claim indicator" field in the medical claims and pharmaceutical claims distinguishes between fee-for-service (FFS) claims and claims covered under capitation. For the recalibration sample, a sample exclusion was applied to exclude enrollees with any claims paid on a capitated basis. For analytical purposes, data users might want to include enrollees with derived claims, for example, to assess the prevalence of certain diagnoses. Therefore, enrollees with derived claims are included in the LDS sample.

[4] SYSIDs are system-generated random numbers used to link the unique enrollee records across files.

[5] 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.

Plan type. With MarketScan® data, we were able to restrict the sample to beneficiaries enrolled in preferred provider organization (PPO) and other FFS plan types. However, we did not have the required information in the 2016 benefit year EDGE data extract to accomplish this for the recalibration sample.[6] Therefore, for the LDS sample file, plan type is not restricted to PPO and other FFS plan types.

Mental health and prescription drug coverage. The MarketScan® analytic file excluded any individuals enrolled in plans that lacked coverage for mental health or prescription drugs. This exclusion was not necessary for the recalibration sample because all risk adjustment covered plans are required to provide both types of coverage under the essential health benefits (EHB) regulations.[7]

Age and sex. Consistent with the MarketScan® population, we restricted the analytic sample for the 2019 and 2020 benefit years' risk adjustment model recalibrations to include only enrollees younger than 65.[8] However, in the raw enrollment file, 790,581 records (2.5%) indicated an enrollee age of 65 or greater; most of these (528,307) were exactly age 65. It is likely that many of these enrollees were transitioning to Medicare or remained in small group plans through their employers. Because the population in the EDGE data extract does include some enrollees age 65 and up, we have included data for enrollees age 65-99 in the LDS sample. In both the recalibration and LDS samples, we have excluded data for enrollees above age 99, as this seemed potentially indicative of a data error; there were 175 total enrollees above age 99 in the raw 2016 enrollment file, 63 of whom were age 116.

Additionally, for the LDS sample files, we censored the age data field to 89 for enrollees ages greater than 89. That is, the age for enrollees age 89 and above is listed as 89.

The EDGE data extract includes a few individuals with a sex field that indicates "unknown." These were primarily infants, and these records were excluded from the recalibration and LDS samples.

Metal and CSR variant identifiers. The enrollee-level EDGE data extract included a data field identifying the enrollee's plan metal level and CSR variant for the 2016 benefit year. To prevent the identification of enrollees in plans that had small sample sizes in certain combinations of metal and CSR levels, we made certain changes in the LDS sample. First, we excluded data for enrollees older than 30 years of age in catastrophic plans in the LDS sample. Second, 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 did not identify the plan type to avoid identifying the small sample size of enrollees and plans, but instead indicated an 11 (limited cost-sharing, zero cost-sharing, Medicaid expansion private or cost-sharing wrap plans) in the CSR data field and provided a missing value for the metal level data field.

Additional exclusions carried over from MarketScan®. In addition to the dataset exclusions described above, we applied several other exclusions for the recalibration and LDS samples identical to those used for the MarketScan® data: (1) we excluded enrollees if age was greater than 1 and any newborn birth diagnosis was present; (2) we excluded enrollees if sex was male and any pregnancy diagnosis was

[6] We did not have enrollees' market (individual, small group) for the 2016 benefit year EDGE claims dataset either. We will have this data field (individual or small group, including for enrollees in merged market states) beginning for the 2017 benefit year EDGE claims recalibration dataset.

[7] See 45 CFR 156.110.

[8] In the EDGE data, age is defined as of the end of the year (December 31, 2016), whereas in MarketScan® age was defined as of the final month of each person's enrollment. In both cases, age is rounded down to the nearest integer; for example, someone who is 59 years and 363 days old would be counted as 59.

present; (3) we excluded enrollees if age was less than 2 and any pregnancy diagnosis was present.[9] All of these exclusions are based on filtered claims, to ensure that the diagnoses that we use to restrict the sample are all valid. These exclusions were also applied to the LDS sample.

Claims

The enrollee-level EDGE extract for the 2016 benefit year did not include the data fields necessary to uniquely identify claims. As a result, we were not able to identify duplicates in the claims files. This mattered for two reasons. First, we assumed all records in the claims files were accurate and processed them for the purposes of model recalibration. Second, we had to impute claim identifiers in order to aggregate the allowed amounts for each SYSID.[10] This imputed claim identifier is not included in the LDS sample. The enrollee-level EDGE extract did include a variable that numbers line items within claims (claim_seq), which is included in the LDS sample.

Each record in the medical claims file represents a line item on a claim, and in order to process this data at the claims level, we needed to develop a method to "aggregate up" the line items that constitute each claim. We determined that line items that have identical values in all of the following fields must be from the same claim: 1) SYSID (randomly generated person/plan identifier), 2) form type code (an institutional/professional claim identifier), 3) service start date, 4) service end date, 5) claim paid date, 6) total allowed amount (from the claim header), and 7) total paid amount (claim header). It is extremely unlikely that two distinct claims would have identical values in all seven of these fields.

Expenditures. There are two possible ways to determine the allowed amount and paid amount for each claim. Each record includes the allowed amount and paid amount for the claim line only (referred to as "allowed amount by plan" and "amount paid by plan"), as well as the header allowed amount and paid amount for the entire claim. So, for example, a claim with four lines might look like this:

Claim line	Header allowed amount	Header paid amount	Claim line allowed amount	Claim line paid amount
1	500	400	100	100
2	500	400	200	150
3	500	400	160	120
4	500	400	40	30

We used the header allowed amount to construct total expenditures for each individual in the recalibration sample, consolidating all lines for a claim into one record in order to avoid double-counting.

We found that the sum of the claim line allowed amounts did not always equal the header amount. This may be because of our inability to definitively identify claims. However, we determined that using the header amounts would be more reliable than summing the claim lines.

Claims with a $1.00 header allowed amount. In the medical claims data, there were a large number of claims with a header allowed amount of exactly $1.00. When we examined these $1.00 EDGE claims, we

[9] For the 2020 benefit year recalibration, we excluded enrollees if age was less than 8 and pregnancy diagnosis was present.

[10] For the 2017 benefit year, we have included a unique claim identifier field, a hashed claim identifier, in the data extract. The claim identifier is a random hashed number assigned for each set of service line items associated with each claim, and cannot be used to identify the enrollee, plan, or medical record. Including this claim identifier will allow data users to associate all service line items under the same claim, and also permit more rigorous checks of data quality.

determined that they were for legitimate services – office visits, tests, labs, etc. – and most likely reflected bundling of services with $1.00 as a placeholder header allowed amount. The actual cost of the service was most likely captured in another one of the patient's claims. For these reasons, we retained these claim amounts and their associated diagnoses for the purposes of the recalibration sample. By contrast, all claims flagged as capitated in the raw EDGE data extract were excluded from the recalibration sample. As previously mentioned, we have included the derived claims as part of the LDS sample files.

Supplemental claims. The supplemental file lacks the information necessary to filter claims, such as bill type codes. However, all supplemental claims should generally be treated as valid since the supplemental claims file is already filtered to be associated with a valid medical claim. We were unable to determine why some supplemental claims did not match any SYSIDs in the data; therefore, these unmatched claims were excluded from the LDS sample. For other supplemental claims, when we tried to match the supplemental claim to its original claim based on the SYSID, start and end dates, the most common problem was that the start and end dates did not match any claims for the matching SYSID. We disregarded mismatches where the supplemental claims matched a SYSID in the data but did not match any claim start and end dates. However, we also included such supplemental claims without matched start and end dates in the medical claims file in the LDS sample.

To determine whether any given diagnosis is present for an individual enrollee, we used a diagnosis counting method that works as follows: we created a flag outside of the sample dataset, which is triggered (set equal to 1) if the count of diagnoses from the medical claims file plus the count of adds from the supplemental file minus the number of deletes from the supplemental file is greater than zero. If this quantity is less than or equal to zero, then the flag is not triggered.

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.[11] For the redacted sample medical file, the claim header or line with header having any of the relevant SUD ICD-10 diagnosis codes[12] 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 at the claim line level in the medical file. A claim line with any of the relevant SUD CPT/HCPCS procedure codes[13] was identified and excluded for the SUD redacted sample. As noted above, there is no variable in the 2016 enrollee-level EDGE extract to link a group of header or service lines into a single claim. Therefore, we used the imputed claim identifier (see discussion above in the claims section), and excluded header and line items with the same imputed claim identifier if line items within a claim included SUD CPT/HCPCS codes for the redacted LDS sample. We redacted SUD ICD-10 diagnosis codes from the supplemental diagnoses file, but did not subsequently adjust the redaction of SUD claims in the medical file due to the inability to link the supplemental diagnoses to medical claims. 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.

After SUD claims redaction in the medical and supplemental files, there were a total of 21,445,954 unique SYSIDs with medical claims and 810,773 unique SYSIDs with supplemental diagnoses. There were 530,398,775 observations in the medical file, and 7,629,232 observations in the supplemental file. In

[11] See 45 CFR 2.52; see also 45 CFR 160.103 for HIPAA definitions of "covered entity" and "business associate."
[12] The list of ICD-10 diagnosis and CPT/HCPCS codes excluded are available here: https://www.resdac.org/articles/redaction-substance-abuse-claims.
[13] Ibid.

the accompanying table, we provide an overview of the impact of SUD redaction on the claims and total allowed amounts.

	2016 LDS Sample		**SUD Redacted 2016 LDS Sample**	
	Observations	**Unique Sysid**	**Observations**	**Unique Sysid**
RARECALE	44,461,673	30,634,593	*No change*	*No change*
RARECALM	542,498,817	21,487,252	530,398,775	21,445,954
RARECALP	230,572,809	17,751,316	*No change*	*No change*
RARECALS	8,165,088	827,879	7,629,232	810,773