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

Date: May 18, 2018

Subject: Enrollee-level EDGE Dataset for Research Requests

Enrollee-level EDGE Dataset for Research Requests

In the 2018 HHS Notice of Benefit and Payment Parameters final rule,[1] CMS finalized collecting masked enrollee-level data from the External Data Gathering Environment or "EDGE" servers[2] (referred to as "enrollee-level EDGE data") beginning with the 2016 benefit year to recalibrate the risk adjustment model, inform development of the Actuarial Value (AV) Calculator and methodology, and calibrate other HHS programs in the individual and small group markets. In this rule, we also discussed making this data publicly available to help governmental entities and independent researchers better understand these markets.

Beginning in the Fall of 2018, we anticipate interested parties could request the 2016 benefit year dataset through the CMS research data request process.[3] Requestors will be required to submit a research purpose statement and sign a data use agreement (DUA) to ensure that the data will be used for the stated research purpose only. We will provide more information on requesting the data in future guidance.

In anticipation of future data requests, this document provides a description of the 2016 enrollee-level EDGE dataset and a list of data elements that CMS will make available for research requests. We welcome stakeholder feedback on the data elements to be released. Comments should be submitted with the subject "Enrollee-level EDGE Dataset for Research Requests" to RARIpaymentoperations@cms.hhs.gov by June 18, 2018.

Data Safeguards

The enrollee-level EDGE data collected by CMS includes an enrollment file, a medical claims file, a pharmacy claims file, and a supplemental diagnosis file[4] for risk adjustment-covered plans in the HHS-operated risk adjustment program.[5] CMS does not collect enrollee-identifiable elements to

---

[1] See HHS Notice of Benefit and Payment Parameters for 2018 Final Rule (December 22, 2016) 81 FR 94058, 94101.

[2] Consistent with 45 C.F.R. § 153.700, in States where HHS is operating the risk adjustment program, issuers must submit enrollment, claims, and encounter data for risk adjustment-covered plans in the individual and small group markets through the EDGE servers. Issuers upload enrollee, pharmaceutical claim, medical claim, and supplemental diagnosis information from their systems to an issuer-owned and controlled EDGE server.

[3] To be made available through the CMS data request center, https://www.resdac.org/cms-data/request/cms-data-request-center.

[4] The supplemental diagnosis file includes supplemental diagnosis codes that issuers submitted to the EDGE servers for the HHS risk adjustment program. Supplemental diagnosis codes are usually found during chart review by the issuer subsequent to medical billing, and therefore may not appear in the medical claims file.

[5] HHS operated the risk adjustment program in all states except Massachusetts for the 2016 benefit year.

safeguard enrollee privacy and issuers' proprietary information. Additionally, we do not collect the masked enrollee identifiers that issuers submit to the EDGE servers, nor do we collect information in the dataset that could be grouped to potentially identify an individual or issuers such as plan and issuer identifiers, premiums, rating areas, or a State. CMS created a de-identified randomly generated enrollee number to link enrollees across the claims files prior to extracting the enrollee-level data from the EDGE servers. The randomly generated enrollee number cannot be decrypted by CMS to identify enrollees or issuers.

Dataset Available for Research Requests

The 2016 benefit year enrollee-level EDGE dataset that CMS will make available for research requests will include an enrollment file, a medical claims file, a pharmacy claims file, and a supplemental diagnoses file. As discussed above, enrollees can be linked across files using a randomly generated number that links files related to a particular enrollee across files (data element labeled SysID). The randomly generated enrollee number cannot be decrypted by CMS to identify the enrollees or issuers. Additionally, the dataset available for research requests will exclude dates and will censor age for enrollees with ages 90 or older to age of 90 to be deemed de-identified in accordance with Health Insurance Portability and Accountability Act of 1996 (HIPAA) requirements under 45 C.F.R. § 164.514(b)(2). The dataset will include categorical variables for metal level and the cost-sharing reduction variant of the enrollee's plan.[6] We anticipate making additional years of the dataset available for public requests in the future as additional years of enrollee-level EDGE data become available.

The 2016 benefit year enrollee-level EDGE dataset available for research requests will include data for approximately 30 million enrollees in the individual and small group market risk adjustment-covered plans in State markets where HHS operated the risk adjustment program.[7] The dataset will exclude seven issuers' data based on identified discrepancies in their 2016 benefit year EDGE data submission; the errors reflected in those discrepancies excluded approximately 496,000 enrollees from the 2016 benefit year enrollee-level EDGE data collected by CMS, and therefore these enrollees' data will also be excluded in the dataset available for research requests. Additionally, any data anomalies will be removed from the dataset available for research requests. Such anomalies include: enrollees with ages greater than 99, enrollees with sex field as "unknown," and enrollees with inconsistent diagnoses given their age and sex (e.g., enrollees with ages less than 2 with a pregnancy diagnosis present). Certain randomly generated enrollee IDs appeared in the claims files but were not included in the enrollment file. CMS will also remove the claims associated with such enrollees from the dataset available for research requests, as these claims will not be useful absent the data elements from the enrollment file. CMS was unable to verify the source of these enrollee IDs because CMS only collects masked data from the issuers' EDGE servers and is unable to decrypt the randomly generated enrollee numbers.

Tables 1, 2, 3, and 4 below provide the data elements that CMS would include in the 2016 benefit year enrollee-level EDGE dataset, which would be made available upon request for research.

[6] We note that CMS did not collect the individual or small group market identifier in the 2016 benefit year enrollee-level EDGE dataset, and therefore will not be available in the 2016 publicly available dataset. We anticipate collecting individual or small group market identifier in the 2017 benefit year enrollee-level EDGE dataset.
[7] *See, supra* note 5.

Description of Data Elements

Table 1: Enrollment Data Elements

Data Element	Description	Data Type	Details
SysID	System generated random number used to link the unique enrollee records across files.	String	
Enrollee Age	Age of the enrollee as of December 31, 2016.	Integer	Censored to 90 for enrollees of age 90 or older.
Age at Start of Enrollment	Age of the enrollee at the start of the enrollment period. Derived from the enrollee date of birth and the enrollment start date in the specified benefit year.	Integer	Censored to 90 for enrollees of age 90 or older.
Enrollee Sex	Sex of enrollee.	String	M = Male F = Female
Enrollment Length - Number of Months	Number of months the enrollment period is active in the specified benefit year. Derived from Enrollment Start and End Dates. ("Enrollment Length – Number of Days" divided by 30 days and rounded to two decimal places)	Decimal	
Enrollment Length - Number of Days	Number of days the enrollment period is active in the specified benefit year. Derived from Enrollment Start and End Dates.	Integer	
Metal Level	The Metal level of the plan in the specified benefit year.	String	P = Platinum G = Gold S = Silver B = Bronze C = Catastrophic
CSR Variant	The cost-sharing reduction variant of the plan in the specified benefit year.	String	00 = Non-Exchange variation 01 = Exchange qualified health plan (QHP) variation (standard plan) 02 = Zero cost-sharing plan variation 03 = Limited cost-sharing plan variation 04 = 73% AV silver plan variation 05 = 87% AV silver plan variation 06 = 94% AV silver plan variation 32 = 100% AV state Medicaid expansion private plan or cost-sharing wrap plan variation 36 = 94% AV or above state Medicaid expansion private plan or cost-sharing wrap plan variation If the 2 digit variant does not exist for the plan, Variant = 'XX' is used.

Table 2: Medical Claims Data Elements

Data Element	Description	Data Type	Details
SysID	System generated random number used to link the unique enrollee records across files.	String	
Form Type Code	Describes claim form type as professional or institutional.	String	'I' = Institutional; 'P' = Professional
Bill Type Code	The code indicating a specific type of bill as reported on institutional claims only.	String	Values should comply with X12 industry standards. If value is not applicable, then the value is empty.
Diagnosis Codes	Code value for the diagnosis code as determined by classification of International Classification of Diseases.	String	Values must comply with X12 industry standards. Does not include a decimal. For medical claims with multiple diagnosis codes, dx codes will be separated with '-'
Discharge Status Code	The facility discharge status of the enrollee.	String	Values must comply with X12 industry standards.
Allowed Total Amount	Total amount allowed for this claim.	Decimal	
Policy Paid Total Amount	Total paid amount for this claim	Decimal	
Derived Service Claim Indicator	Indicator used to distinguish between fee-for-service claims and claims covered under capitation.	String	'Y' = Derived (capitated Service); 'N' = Actual (Fee-For-Service)
Claim Line Sequence Number	Unique number generated to represent service(s) submitted on the claim.	Integer	
Revenue Code	Describes the revenue center in which the service was provided.	String	Values must comply with X12 industry standards. If value is not applicable, then the value should be empty.
Service Code Qualifier	A code that identifies the source of the procedure code (CPT or HCPCS)	String	01 – Dental service codes; 03 - Healthcare Common Procedure Coding System (HCPCS) or Current Procedural Terminology (CPT) If value is not applicable, then the value should be empty.
Service Code	A procedure code that identifies the service rendered: CPT or HCPCS.	String	Values must comply with X12 industry standards. If value is not applicable, then the value should be empty.
Service Code Modifiers	A 2-digit code that may be billed with a CPT/HCPCS service code.	String	Values must comply with X12 industry standards. If value is not applicable, then the value should be empty. For medical claims with multiple service code modifiers, codes will be separated with "-".
Place of Service	A code that identifies where the service was rendered.	String	Values must comply with X12 industry standards. If value is not applicable, then the value should be empty.
Amount Allowed	Total amount allowed by plan.	Decimal	
Amount Paid	Total amount paid, or derived, by plan.	Decimal	
Derived Amount Indicator	Indicator used to distinguish between fee-for-service claims and claims covered under capitation.	String	'Y' = Derived (Capitated Service) 'N' = Actual (Fee-For-Service)

Table 3: Pharmacy Claims Data Elements

Data Element	Description	Data Type	Details
SysID	System generated random number used to link the unique enrollee records across files.	String	
Product/Service ID	Unique ID of the product or service dispensed using the National Drug Code (NDC).	String	
Fill Number	Code identifying whether the prescription is an original (0) or refill (1-999).	Integer	
Dispensing Status	Indicates if the prescription was a partial fill (P) or the completion of a partial fill (C).	String	C = Completion of a partial fill; P = Partial fill A blank implies a complete fill at the time dispensed. If value is not applicable, then the value should be empty.
Total Allowed Cost	Represents the sum of allowed charges for ingredient cost, dispensing fee, and sales tax.	Decimal	
Plan Paid Amount	The total cost of the product/service paid by the plan.	Decimal	
Derived Amount Indicator	Indicator used to distinguish between fee-for-service claims and claims covered under capitation.	String	'Y' = Derived (Capitated Service); 'N' = Actual (Fee-For-Service)

Table 4: Supplemental Claims Data Elements

Data Element	Description	Data Type	Details
SysID	System generated random number used to link the unique enrollee records across files.	String	
Claim Type	Identifies if claim is add or delete claim.	String	'A' = Add; 'D' = Delete Note: Diagnoses should be counted if the count of those diagnosis codes appearing in the medical claims and the supplemental add claims subtracted for the supplemental delete claims is greater than 0.
Supplemental Diagnosis Codes	Code value for the Diagnosis Code as determined by classification of International Classification of Diseases (ICD)	String	Values should comply with X12 industry standards. Explicit decimal is not required. For Supplemental claims with multiple diagnosis codes, dx codes will be separated with "-".