Medicare Part C and Part D Reporting Requirements
Data Validation Procedure Manual

Appendix H: Data Extraction and Sampling Instructions

Prepared by:
Centers for Medicare & Medicaid Services
Center for Medicare
Medicare Drug Benefit and C & D Data Group

Last Updated: December 2022

Table of Contents

List of Exhibits

1 OVERVIEW

The purpose of this document is to provide guidance to reviewers regarding drawing and evaluating census and/or sample files to support validation of Part C and Part D reporting sections.

This document describes guidelines and methodologies for extracting sponsoring organizations' data for data validation review. Two methods of data extraction are available to data validation contractors (reviewers). The first method is referred to as the census. For example, extracting all records used in the calculation of data elements for a specific reporting section would constitute extracting a census of data. When possible, reviewers should attempt to extract the full census. Extracting the census will enable the reviewer to determine with the greatest precision whether reporting sections were submitted accurately. The second method used for data extraction is a random sample. The random sample is a subset of the census data. If extraction of the census proves to be too burdensome due to the size or complexity of the data for a specific reporting section, a sample of records should be extracted instead.

The use of one or both extraction methods described above is key for reviewers as they validate the quality of the data used to calculate Part C and Part D reporting sections. Examples of characteristics evaluated using the census data include appropriate date ranges, appropriate data inclusions and exclusions, the correctness of data values, and the handling of missing values. When extracting a census is not practical, the use of a large enough random sample can accomplish the same goals, although the reviewer will need to rely on statistically valid estimates rather than evaluating the entire population. For both methods, reviewers must examine source data, as a means of verifying that the organization's underlying data are correct: for example, reviewing customer service call logs or member letters to verify that grievances were properly categorized as grievances.

The reviewer will determine whether supervision is required while the sponsoring organization extracts census and/or sample files. It is also left to the reviewer's discretion as to the feasibility of the sponsoring organization extracting census and/or sample files before, during, or after the on-site or virtual visit.

However, it is mandatory that reviewers follow the instructions in this document. If the sponsoring organization's staff is extracting the data, it is highly recommended that the reviewer supervise the data extraction process to ensure these instructions are followed correctly. If the reviewer is unable to supervise the data extraction process, the reviewer should obtain documentation from the sponsoring organization describing how the extraction process was performed. For example, if a random sample is extracted, the reviewer should request and validate the programming code used to extract the sample data. If a full census is extracted, the reviewer should validate that the record counts match between the census extraction and the source and final stage data files.

2 CONCEPTUAL FRAMEWORK FOR DATA EXTRACTION

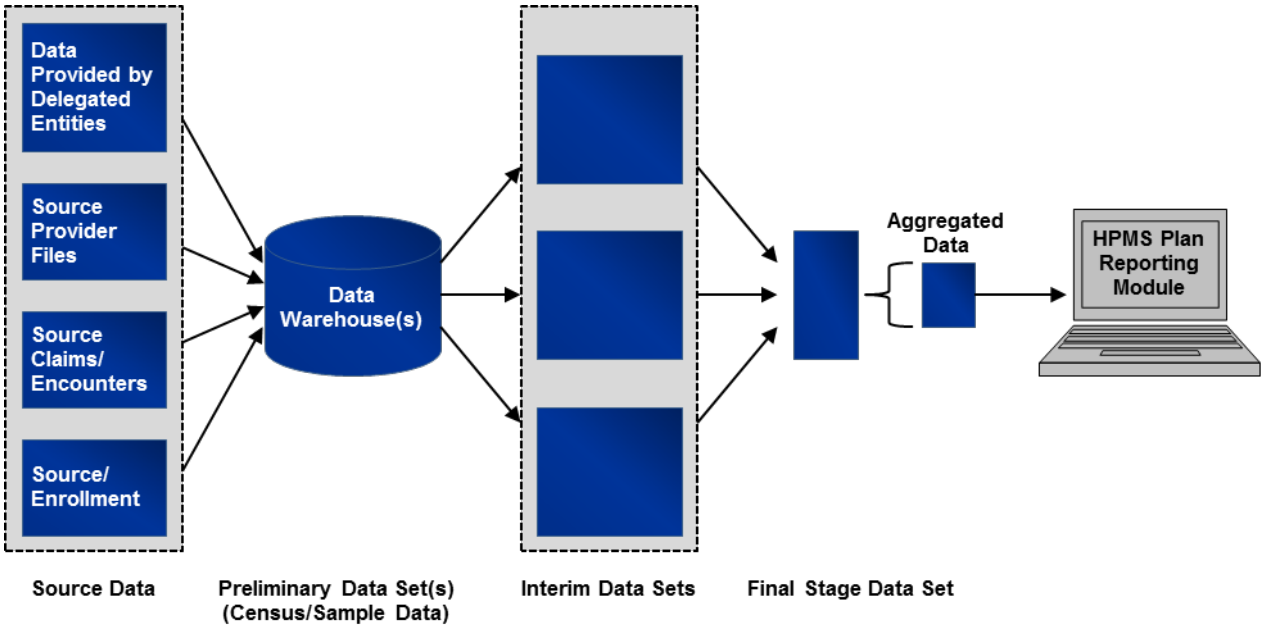
Throughout the document, there are several closely related terms used to describe different data types. Exhibit 1 provides data type definitions. All terms are applicable to both census and random sample methods.

Exhibit 1: Data Type Definitions

Data Type	Definition	Examples
Source Data	The initial source for all data used to create all subsequent databases that will be used to report on each reporting section.	Claims adjudication data, provider files, customer service call logs, enrollment files
Preliminary (Census or Sample) Data	Data stored in a data warehouse that has been cleaned and prepared for analysis.	Databases created from cleaned source data
Interim Data	Data sets stored in the data warehouse and joined to other data sets residing in the data warehouse.	Grievance data set joined with enrollment data for each beneficiary filing a grievance
Final Stage Data	The last detailed data set before calculation of counts or sums for each reporting section.	Data set containing a row for each grievance filed, with fields such as member ID, type of grievance, date filed, date resolved, etc.

Exhibit 2 shows conceptually how sponsoring organizations create aggregated data for submission into the Health Plan Management System (HPMS) and where data extraction is incorporated into the data validation review process.

Exhibit 2: Conceptual Framework for Data Extraction



While actual reporting approaches vary significantly from organization to organization, and even between reporting sections, the general reporting approach can be described as follows:

1. Source data reside on operational systems, such as claims adjudication systems, providerfiles, enrollment files, and data systems maintained by delegated entities.
2. Source data are often uploaded to an analytic data warehouse where data are cleaned and putinto database structures to support analysis.
3. The data in the warehouse are extracted to create an interim data set, which oftencontains manipulated and merged data.
4. Data from interim data sets are combined into a final stage data set.
5. This final stage data set is aggregated to create sums and counts, which are then entered into the HPMS Plan Reporting Module.

The data extraction process produces at least two (and for some reporting sections it could be three) sets of validation data for each reporting section. The first comes from the endpoint of the calculation, and the second is a corresponding set of extracts drawn from a data or analytic warehouse or operational system, which produces underlying data. The third is a sample extract of source data (e.g., customer service call logs) that underlie the data residing in a data warehouse. If interim data sets are produced, the reviewer may consider extracting these data to ensure that data sets have been joined properly. Details on extracting and evaluating the data are outlined in the next section.

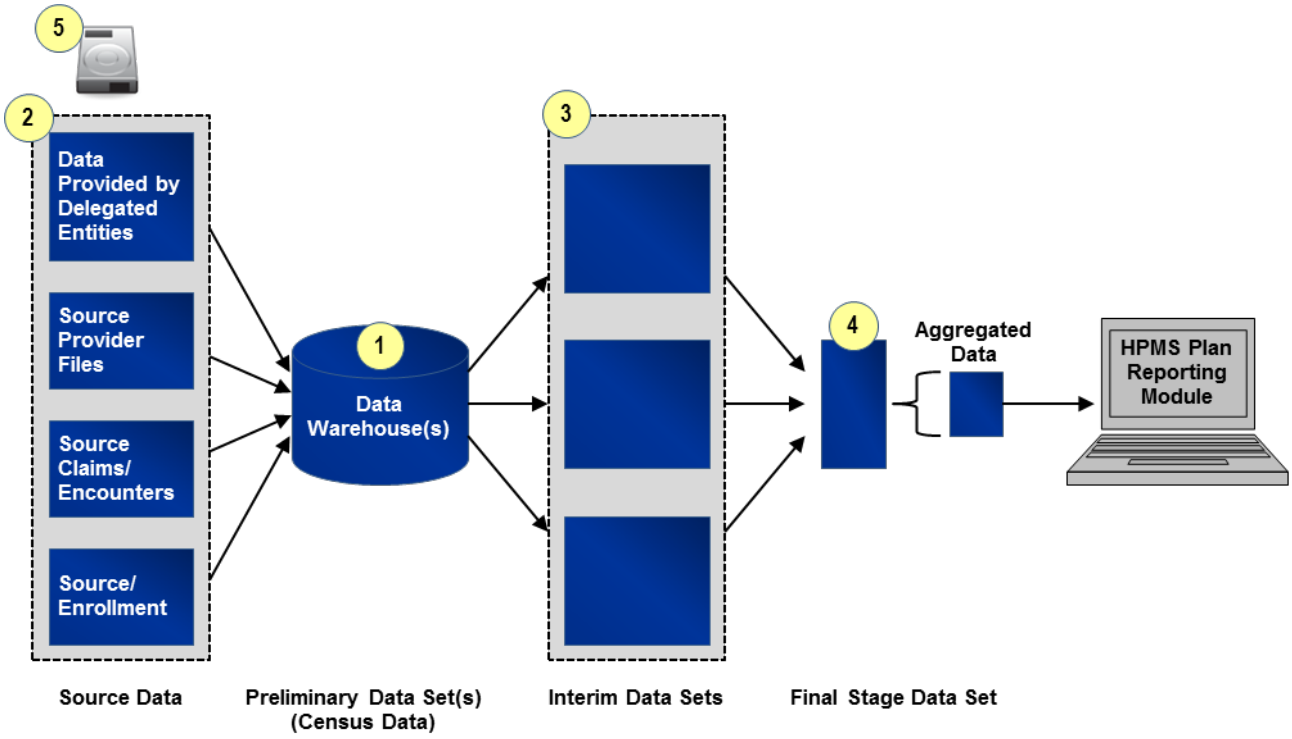
3 DATA EXTRACTION PROCESS DETAIL

3.1 EXTRACTION OF THE CENSUS

Data extraction of the full census will be conducted at the organization's contract level, and in some instances, at the plan level. Extraction of a full census will provide the reviewer with the most precise evaluation of how accurately an organization reports its Part C or Part D data. Extracting the full census is

the most straightforward of the two data extraction methods. The process illustrated in Exhibit 3 applies to all reporting sections where it is deemed practical to extract a census.

Exhibit 3: Application of the Census Process



1. ***Identify and Extract "Preliminary Census Data Set(s)":*** "Preliminary Census Data Set(s)" will include all files containing records extracted from one of the originating data source(s) (e.g., organization's internal data warehouse, enrollment system). The "Preliminary Census Data Set(s)" will include all fields referenced in the programming code used to calculate the reporting section. To identify appropriate source data and fields and date ranges for the "Preliminary Census Data Set(s)," the reviewer will refer to the source/programming code, saved data queries, data dictionaries,analysis plans, etc. provided by the organization.

2. ***Identify and Extract Source Data:*** For some reporting sections within some organizations, the source data may already have been extracted from the source systems (e.g., enrollment system) aspart of step one described above. For other reporting sections such as grievances or coverage determinations and exceptions, the reviewer will need to extract and examine data sources such as customer service call logs or pharmacy claim files. This step is important because it allows the reviewer the opportunity to validate that the underlying data is correct and was accurately uploaded or entered into the data warehouse. See Section 3.3 for instructions on example source data and sample sizes.

3. ***Identify and Extract "Interim Census Data Set(s)" (If applicable):*** Where applicable, the reviewer will identify "Interim Census Data Sets," that is, data sets that have undergone a cleaning process after initial entry into a data warehouse and before being joined to create the "Final Stage Data Set(s)." All "Interim Census Data Sets" should be identified and clearly labeled so that the relationship between data extracts is identified and distinguishable.

4. ***Identify and Extract "Final Stage Data Set(s)":*** The reviewer will identify the last clean and detailed (line-item level) data set used prior to aggregating counts and sums for the data reporting section. This is the cleanest and last line-item level file before data aggregation for entry into HPMS and is referred to as the "Final Stage Data Set." Note that in some cases, multiple "Final Stage Data Sets" will be identified.

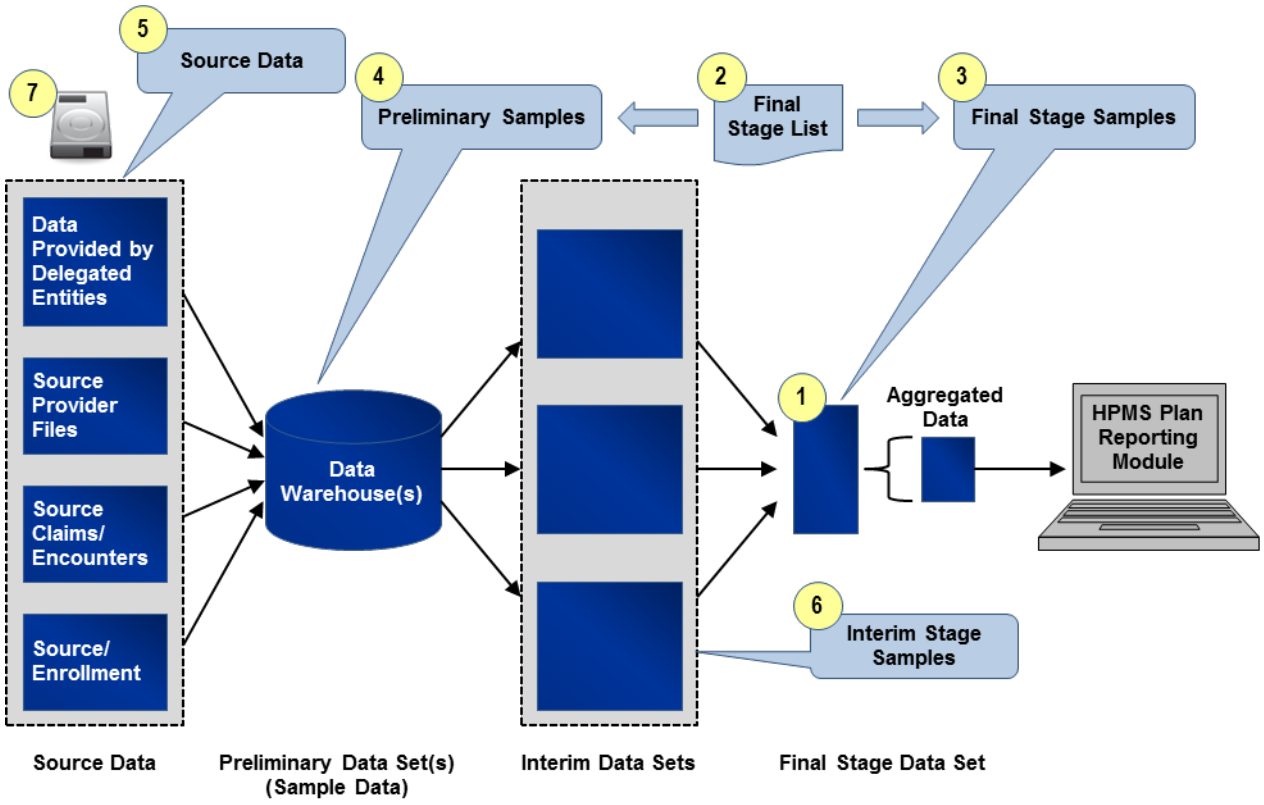
5. ***Write and Encrypt Data to Secure Storage Device:*** The organization will transfer all data files collected to a secure storage device. Organizations undergoing review should coordinate with reviewers to ensure that the organization's security software does not interfere with data transfer. Files requested before or after the on-site or virtual visit can be transferred via a secure web portal or by other methods that comply with regulations governing secure storage and transfer of Protected Health Information (PHI). See Section 4.2 for instructions on the file format.

3.2 EXTRACTION OF THE SAMPLE DATA

In general, sampling will be conducted at the organization's contract level, but for some reporting sections sampling will take place at the plan benefit package (PBP) level. In cases where organizations have multiple contracts that use the same data sources and processes for each contract, only one sample is required. This one sample must be randomly drawn from pooled data from all contracts so that it is representative of the systems and processes across the contracts. For organizations with multiple contracts, where data sources and processes differ among contracts, separate samples are required for each unique contract. Based on information obtained during the review, the reviewer will determine whether the data sources are the same and processes are standardized across an organization's multiple contracts; this will aid in determining whether one or more samples need to be drawn. It is the responsibility of the reviewer to determine the appropriate sample size for each reporting section which is why including a statistician on the reviewer team is required. For guidance on the minimum sample size, see Section 4.1. The above description for extracting a sample from a contract or several contracts applies in exactly the same way for reporting sections that are reported at the PBP level, where sampling is required for a PBP or several PBPs.

Drawing the sample data follows the same six-step process for each reporting section. Details on each step of the process are outlined and illustrated at a high level in Exhibit 4.

Exhibit 4:Application of Sampling Process



1. ***Identify "Final Stage Data Set(s)":*** The reviewer will identify the last clean and detailed (line-item level) data set used prior to aggregating counts and sums for the reporting section. This is the cleanest and last line-item level file before data aggregation for entry into HPMS and is referred to as the "Final Stage Data Set." As with the process of extracting the census, in some cases, multiple "Final Stage Data Sets" will be identified.

2. ***Draw Random Sample to Create "Final Stage List":*** The reviewer will work with a knowledgeable organization resource to draw a random list of distinct sampling units (e.g., member IDs, Provider IDs) from the appropriate "Final Stage Data Set(s)." This list is called the "Final Stage List" and is required for extracting the source and final stage sample data. Reviewers should use standard statistical practices when determining sample sizes. Sampling units and sample size for the "Final Stage List" will vary by reporting section. In cases where there are multiple "Final Stage Data Sets," the reviewer will ensure that the "Final Stage List" is representative of all the "Final Stage Data Sets."

   Generally, the selection of the "Final Stage List" should be pulled using simple random sampling. For guidance on these methods, see Section 4.1. The reviewer may apply more complex approaches if needed (stratified samples, for example). Determination of the appropriate size and type of random sample must follow sound statistical principles and be well documented.

3. ***Create "Final Stage Sample(s)":*** Using the "Final Stage List," the organization will provide the reviewer a "Final Stage Sample." The "Final Stage Sample" will be extracted from the "Final Stage Data Set" and will include all records associated with the identified sampling units in the "Final Stage List." The "Final Stage Sample" will contain all fields from the "Final Stage Data Set." In cases where there are multiple "Final Stage Data Sets," there will be multiple "Final Stage Samples."

As an example, the Grievances "Final Stage Sample" will include all records and fields in the "Final Stage Data Set" associated with the distinct Case IDs identified in the Grievances "Final Stage List."

4. ***Create "Preliminary Sample(s)":*** Using the "Final Stage List," the organization will provide for the reviewer one or more "Preliminary Samples." Each "Preliminary Sample" will be a file containing records extracted from one of the originating data source(s) (e.g., organization's internal data warehouse, enrollment system), and it will include all records within the reporting period(s) associated with the identified sampled units in the "Final Stage List." The "Preliminary Sample(s)" will include all fields referenced in the programming code used to calculate the reporting section. To identify appropriate originating data sources and fields for the "Preliminary Sample(s)," the reviewer will refer to the source/programming code, saved data queries, data dictionaries, analysis plans, or other documentation provided by the organization.

   As an example, the Special Needs Plans (SNPs) Care Management reporting section may have at least two "Preliminary Samples." One will consist of all claims from the reporting period associated with the distinct Member IDs identified in the SNPs "Final Stage List." The second will consist of all enrollment records in the reporting period associated with these Member IDs.

   Note: *The actual number of records in the "Final Stage Sample(s)" and "Preliminary Sample(s)" will vary, and in many cases, it will be substantially larger than the "Final Stage List" sample size. For example, the "Final Stage List" of Member IDs for the SNPs reporting section will likely result in "Preliminary Samples" of more than the total number of Member IDs because of multiple claims and enrollment records for each member.*

   Note: *If the data are the same as the "Final Stage Data Set," the "Final Stage Sample" will be sufficient.*

5. ***Extract Source Data:*** Similar to the census process, reporting sections and within some organizations, the source data may have already been extracted as part of step four described above. If not, sample source data will need to be extracted in order to validate that the underlying data are correct and was accurately uploaded or entered into the data warehouse. See Section 3.3 for instructions on example source data and sample sizes.

6. ***Create "Interim Stage Sample(s)":*** Where applicable, the reviewer will identify "Interim Stage Data Sets", that is, data sets that have undergone a cleaning process after initial entry into a source system and before being joined to create the "Final Stage Data Set(s)." The reviewer will apply the same methodology for extraction of the "Preliminary Sample" as described in Step 4. All "Interim Stage Samples" should be identified and clearly labeled so that the relationship between data extracts are identified and distinguishable.

7. ***Write and Encrypt Data to Secure Storage Device:*** The organization will transfer all data files to a secure storage device. Organizations undergoing review should coordinate with reviewers to ensure that the organization's security software does not interfere with data transfer. Files requested before or after the on-site or virtual visit can be transferred via a secure web portal or by other methods that comply with regulations governing the secure storage and transfer of Protected Health Information (PHI). See Section 4.2 for instructions on the file format.

3.3 REQUIREMENT FOR EXTRACTION AND REVIEW OF SOURCE DATA

Review and validation of source data ensures that organizations are accurate in the numbers being reported by ensuring that CMS regulations and guidance are being followed and that the underlying data is correct and has been uploaded correctly. Reviewers must review this data and sponsoring organizations and delegated entities must make available necessary source data files and documents. Exhibit 5 provides guidance on examples of source data needed for each reporting section. A sample size of 30 is required unless that sample size is not available. There may be other source data as this list is not all inclusive. It is expected that the reviewer will pull the sample from across the different data sources versus pulling all source data from one location. The source data should represent a random sub-sample of the data underlying the census/sample records pulled from the data warehouse. While the sample size may not allow for the greatest precision in detecting an error, it will serve as an additional verification step. For some of these reporting sections, the source data are the same data that are included in the data warehouse and therefore this new requirement should not affect those reporting sections and the reviewer should continue using the census or sample sizes previously recommended.

Exhibit 5: Requirement for Extraction and Review of Source Data

Reporting Section	Example Source Data
Part C Reporting Sections	
Grievances	Customer Service Call Logs Member Letters Case Notes
Organization Determinations / Reconsiderations	Adjudicated Claims Customer Service Call Logs Member/Provider/Authorized Representative Requests
Special Needs Plans (SNPs) Care Management	Enrollment Files A and B Electronic or paper copies of the completed health risk assessment tool, evidence of communication (facsimile, e-mail, letter, etc.) with providers for verification of care (reports from specialists, copies of medical records, copies of medical histories, etc.), the OASIS assessment tool for beneficiaries receiving home care, or the MDS assessment tool for beneficiaries in long-term care facilities (Data Elements C and D
Part D Reporting Sections	
Medication Therapy Management (MTM) Programs	Claims Files (to confirm changes to drug therapy) Evidence of communication (e.g., prescriber letters to confirm medication reviews, prescriber interventions, and changes to drug therapy) Member Files (to confirm LTC residency) MTM Program files (if separate from Member Files) (to confirm targeting criteria, enrollment, opt out dates, reasons)
Grievances	Customer Service Call Logs Member Letters Case Notes
Coverage Determinations/ Redeterminations	Adjudicated Claims Customer Service Call Logs Member/Provider/Authorized Representative Requests Case Notes
Improving Drug Utilization Review Controls	Rejected Claims Adjudicated Claims Case Notes/Member files Customer Service Call Logs Member/Provider/Authorized Representative Requests

3.4 EVALUATING THE DATA

The reviewer will use each reporting section's full census or samples from source, interim, and final stage data sets to validate against the applicable Part C and/or Part D reporting requirements. Specific validation checks requiring census or sample data are included in Validation Standard 2 in the "Data Validation Standards" and the "Findings Data Collection Form (FDCF)." Validation Standard 2 is reproduced below in Exhibit 6. The validation of all criteria except for meeting deadlines will be conducted using the extracted data.

Exhibit 6: Validation Standards Applicable to Extracted Data

VALIDATION STANDARDS
2. A review of source documents (e.g., programming code, spreadsheet formulas, analysis plans, saved data queries, file layouts, process flows) and census or sample data, whichever is applicable, indicates that data elements for each reporting section are accurately identified, processed, and calculated. Criteria for Validating Reporting Section Criteria (Refer to reporting section criteria section below): • The appropriate date range(s) for the reporting period(s) is captured. • Data are assigned at the applicable level (e.g., plan benefit package or contract level). • Appropriate deadlines are met for reporting data. • Terms used are properly defined per CMS regulations, guidance, Reporting Requirements and Technical Specifications. • The number of expected counts (e.g., number of members, claims, grievances, procedures) are verified; ranges of data fields are verified; all calculations (e.g., derived data fields) are verified; missing data has been properly addressed; reporting output matches corresponding source documents (e.g., programming code, saved queries, analysis plans); version control of reported data elements is appropriately applied; QA checks/thresholds are applied to detect outlier or erroneous data prior to data submission.

As specified in the "Data Validation Standards" and "FDCF", reviewers should evaluate the data in conjunction with the programming code, spreadsheet formulas, analysis plans, saved data queries, file layouts, and process flows provided by the organization. The reviewer should evaluate the data submissions for overall data accuracy for missing information, invalid fields, implausible fields (range checks), demographic errors, or other errors causing linkage or data aggregation failures. All results of the data validation findings should be recorded in the HPMS Plan Reporting Data Validation Module (PRDVM) (and "FDCF", if used), including the number and percentage of errors or variance from HPMS-filed data found when examining the source data. For purposes of recording results in the PRDVM and FDCF (if used), an error is any discrepancy that either impacted the number of events reported or has the potential to impact the number of events reported in future reporting periods. These errors must be reported in the "Review Results" area of the PRDVM and FDCF and include the sample size selected for the source data.

Important Note: Sample data can be used to validate individual records (e.g., to validate the values of specific data elements), however, the total counts or sums for data elements cannot be determined without using the census data. The following are examples of how sample data can be used to validate specific data elements:

- To verify that calculations are performed correctly
- To ensure date ranges are correct
- To evaluate whether specific records have been filtered or categorized properly
- To verify that any manual manipulation of the source and final stage data is accurate
- To evaluate missing data and the impact on the calculation of derived data fields

- To verify that the organization is properly defining terms per CMS regulations, guidance, Reporting Requirements and Technical Specifications

4 ADDITIONAL GUIDANCE

4.1 SAMPLING GUIDANCE WHEN SELECTING A RANDOM SAMPLE

The calculation of each data element requires the organization to pull data from key data sources. The validation samples will reflect the same process but will be limited to relatively small samples of data.

Conceptually, selecting a simple random sample follows this process:

1. Use a pseudo-random number generator (e.g., SAS ranuni function or MS Excel's Random Number Generator in the Data Analysis dialog box) to assign a uniform random number to each record in the key data source.[2]

2. Sort the records by the new random number, from lowest value to highest value.

3. After identifying sample size (n), write the key fields from the first n records of the sorted key data source to a new file.

Alternate Approach: Organizations using SAS for standard calculation may opt to use Proc SURVEYSELECT.

In cases where reviewers need to extract a random sample, Exhibit 7 provides guidance on the proper sample units and the minimum sample sizes for each reporting section. As mentioned above, reviewers should use sound statistical principles when determining the appropriate sample size.

Exhibit 7: Sampling Units and Minimum Sample Size for "Final Stage List"

Reporting Section	Sampling Unit	Sample Size1F[3]
Part C		
Grievances	Case ID	150
Organization Determinations/Reconsiderations	Case ID	150
Special Needs Plans (SNPs) Care Management	Member ID	205
Part D		
Grievances	Case ID	150
Coverage Determinations/Redeterminations	Case ID	150
Improving Drug Utilization Review Controls	Member ID	205

Note: The Medication Therapy Management Program reporting section is removed from this table because there should be no requirement to sample the data in lieu of a census file since the beneficiary upload file can serve as the census file.

[2] *Note: Random number generators require seed numbers as input, but often have options to use the system clock as a seed. It is recommended that the organization key in a literal number as a seed, to assure the sample can be replicated if necessary; these seeds could change from year to year but should be documented.*

[3] *Depending on the size of the organization, some reporting sections will have populations that are smaller than the recommended sample size. In these cases, the entire population will be used for selecting the "Final Stage List."*

4.2 FILE REQUIREMENTS FOR DATA TRANSFER TO REVIEWER

The organization must write all data files to tab-delimited or comma-delimited text files with variable names in the first row, and transfer these files to the reviewer's secure storage device. The organization must also provide the reviewer a file layout or data dictionary for the data files in either Word documents or Excel spreadsheets on the same secure storage device. Naming conventions should be consistent between files and their corresponding layout (e.g., if a sample for Part C Grievances is extracted and labeled "PartCGrievanceSample.txt", the corresponding layout should be named PartCGrievanceLayout.doc). An example file layout is illustrated in Exhibit 8.

Exhibit 8: Example File Layout

Name	Description	Data Type/Length	Data Values	Calculation
M_ID	Member ID	Character (16)		Unique counts
DOR	Grievance Date of Receipt	Date MMDD YYYY		Date
M Status	Member Status	Numeric (2)	1=Enrolled; 2=Disenrolled	

4.3 DATA SECURITY

The organization is responsible for ensuring that it has established mutually agreeable methods for sharing proprietary and/or secure (PHI/PII) information with the reviewer and that the reviewer complies with all HIPAA privacy and security requirements.