Medicare Part C and Part D Measure
Data Extraction and Sampling Instructions for Data Validation Contractors

August 26, 2010

DRAFT

Prepared by:
Centers for Medicare & Medicaid Services
Center for Drug and Health Plan Choice

TABLE OF CONTENTS

1.0 OVERVIEW 1

2.0 CONCEPTUAL FRAMEWORK FOR DATA EXTRACTION 2

3.0 DATA EXTRACTION PROCESS DETAIL 3

3.1 Extraction of the Census 3
3.2 Extraction of the Sample Data 4
3.3 Evaluating the Data 6

4.0 APPENDIX 7

4.1 Sampling Guidance 7
4.2 File Requirements for Data Transfer to Reviewer 8
4.3 Data Security 8

1.0 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 measures.

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 measure 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 reported measures 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 measure, a sample of records should be extracted instead.

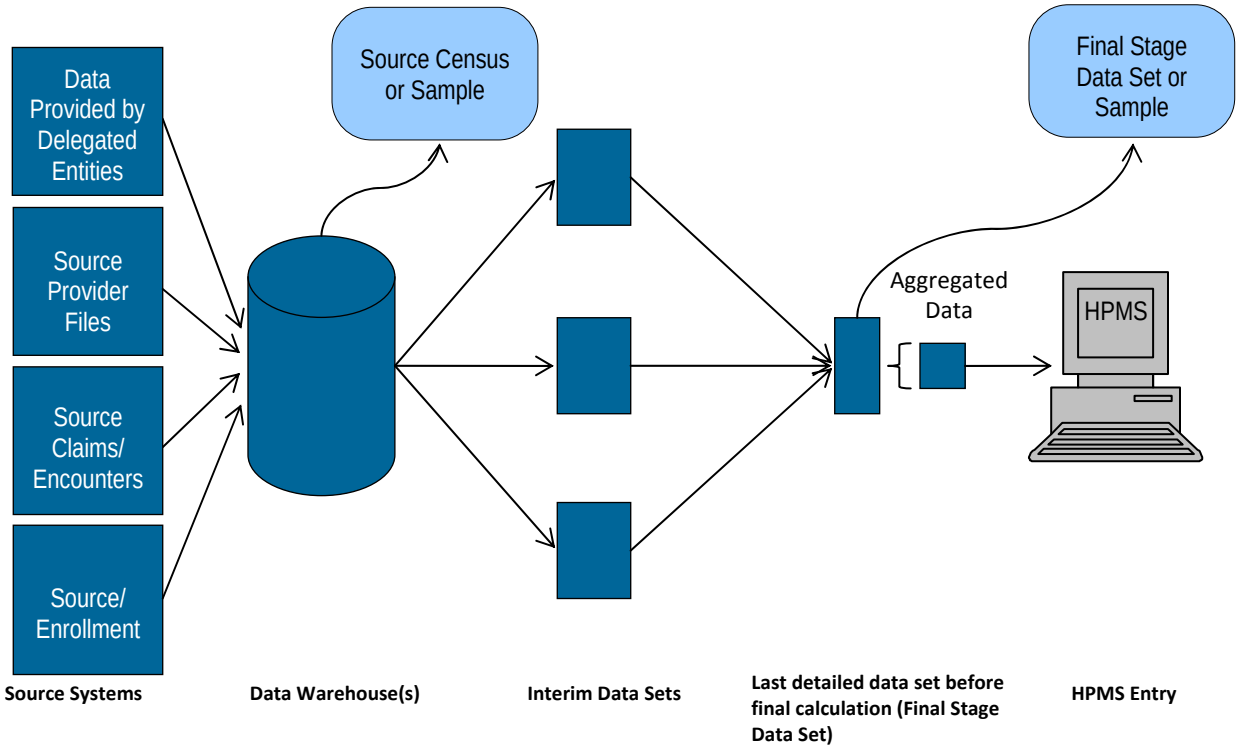
The use of one or both of the extraction methods described above are key for reviewers as they validate the quality of the data used to calculate Part C and Part D measures. Examples of characteristics evaluated using the census data include appropriate date ranges, appropriate data inclusions and exclusions, correctness of data values, and 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.

The reviewer will determine whether or not 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 visit.

2.0 CONCEPTUAL FRAMEWORK FOR DATA EXTRACTION

Figure 1 below shows conceptually how sponsoring organizations create aggregated data for submission into HPMS and where data extraction is incorporated into the data validation review process.

Figure 1: Conceptual Framework for Data Extraction



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

1. Original data resides on operational systems, such as claims adjudication systems, provider files, enrollment files, and data systems maintained by delegated entities.
2. Many organizations have analytic warehouses where data is cleansed and put into database structures to support analysis.
3. Measure calculation begins with a series of extracts, which are manipulated and merged, creating interim data sets.
4. Data from interim steps are combined into a detailed data set.
5. This detailed data set is aggregated to create sums and counts, which are then entered into HPMS.

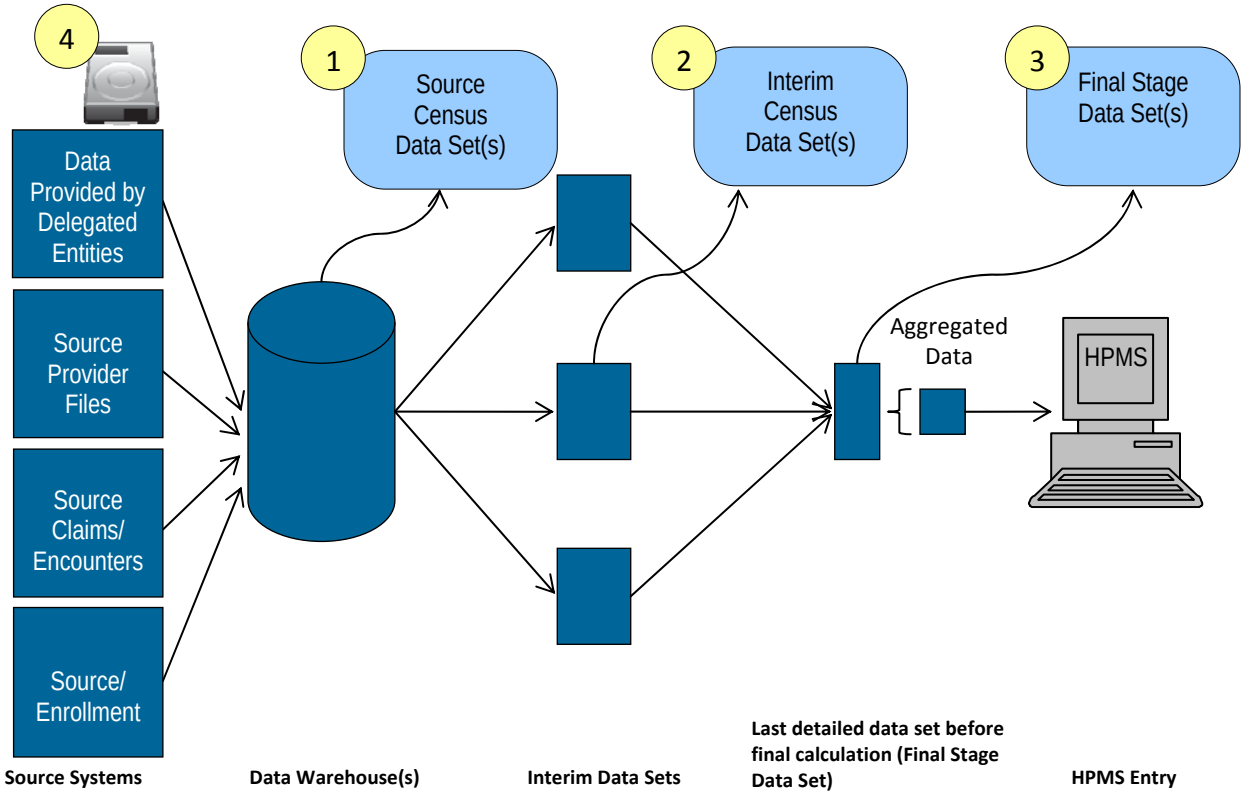
The data extraction process produces at least two sets of validation data for each measure. The first comes from the endpoint of the calculation, and the second is a corresponding set of extracts drawn from source data (e.g. data warehouse or operational systems which produce underlying data). If necessary, the reviewer may request to review source documents that have been used to enter data into a data warehouse or operational system (e.g. a call log for grievances). 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.

3.0 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. Extraction of a full census will provide the reviewer with the most precise evaluation of how accurately an organization reports their Part C or Part D data. Extracting the full census is the most straightforward of the two data extraction methods. The process illustrated in Figure 2 applies to all measures where it is deemed practical to extract a census.

Figure 2: Application of Sampling Process



1. ***Identify and Extract "Source Census Data Set(s)":*** "Source 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 "Source Census Data Set(s)" will include all fields referenced in the programming code used to calculate the measure. To identify appropriate originating data sources and fields and date ranges for the "Source 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 "Interim Census Data Set(s)" (Optional):***
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 source system 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.

3. ***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 measure.

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.

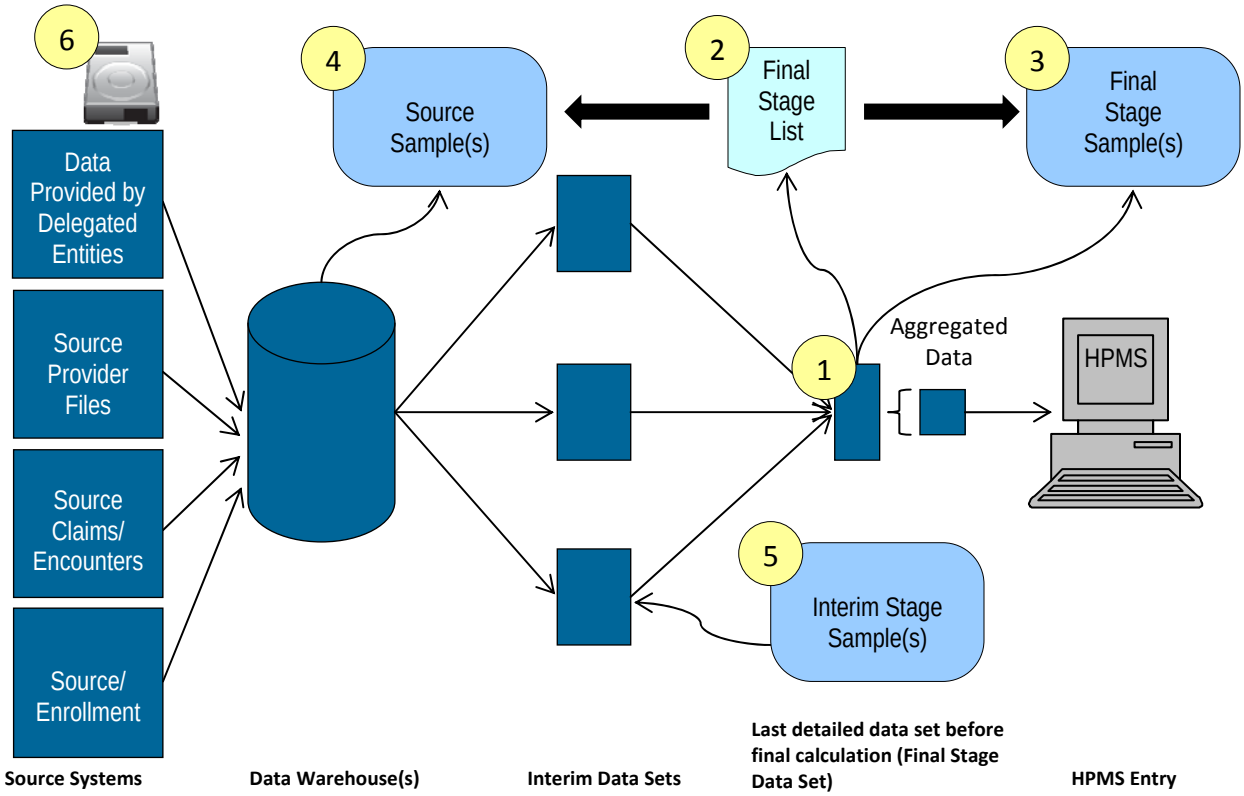
4. ***Write and Encrypt Data to Secure Storage Device:*** The organization will transfer all data files collected to a secure storage device. Organizations undergoing review should coordinate with reviewers to ensure that the organization's security software does not interfere with data transfer. Files requested before or after the on-site visit can be transferred via a secure web portal or by other methods that comply with regulations governing secure storage and transfer of Personal Health Information (PHI). See Section 4.2 in the Appendix 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. 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 measure. For guidance on sample size, see Section 4.1 in the Appendix.

Drawing the sample data follows the same six-step process for each measure. Details on each step of the process are outlined and illustrated at a high-level in Figure 3.

Figure 3: 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 data measure. 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 measure. In cases where there are multiple "Final Stage Data Sets," the reviewer will assure 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 in the Appendix. 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 Benefit Utilization "Final Stage Sample" will include all records and fields in the "Final Stage Data Set" associated with the distinct Member IDs identified in the Benefit Utilization "Final Stage List."

4. ***Create "Source Sample(s)":*** Using the "Final Stage List," the organization will provide for the reviewer one or more "Source Samples." Each "Source 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 "Source Sample(s)" will include all fields referenced in the programming code used to calculate the measure. To identify appropriate originating data sources and fields for the "Source 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 Procedure Frequency measure may have at least two "Source Samples." One will consist of all claims from the reporting period associated with the distinct Member IDs identified in the Procedure Frequency "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 "Source 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 Procedure Frequency measure will likely result in "Source Samples" of more than the total number of Member IDs because of multiple claims and enrollment records for each member.*

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

5. ***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 "Source Sample" as described in Step 4. All "Interim Stage Samples" should be identified and clearly labeled so that the relationship between data extracts is identified and distinguishable.

6. ***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 visit can be transferred via a secure web portal or by other methods that comply with regulations governing secure storage and transfer of Personal Health Information (PHI). See Section 4.2 in the Appendix for instructions on the file format.

3.3 Evaluating the Data

The reviewer will use each measure'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. Validation Standard 2 is reproduced below. The validation of all criteria except for meeting deadlines will be conducted using the extracted data.

Figure 4: 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, if applicable, indicates that data elements for each measure are accurately identified, processed, and calculated. Criteria for Validating Measure-Specific Criteria (Refer to measure-specific 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 (e.g., quarterly). • Terms used are properly defined per CMS regulations, guidance and Reporting Requirements 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 Findings Data Collection Form, 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 Findings Data Collection Form.

4.0 APPENDIX

4.1 Sampling Guidance

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.[1]

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, Table 1 provides guidance on the proper sample units and the minimum sample sizes for each measure. As mentioned above, reviewers should use sound statistical principles when determining the appropriate sample size. Refer to item 2 in the Supporting Statement's statistical section for further guidance on determining the appropriate sample size.

Table 1: Sampling Units and Minimum Sample Size for "Final Stage List"

Measure	Sampling Unit	Sample Size[2]
Part C		
Benefit Utilization	Member ID	205
Procedure Frequency	Member ID	205
Serious Reportable Adverse Events (SRAEs)	Member ID	205
Provider Network Adequacy	Provider ID	150
Grievances	Case ID	150
Organization Determinations/Reconsiderations	Case ID	150
Employer Group Plan Sponsors	N/A	N/A
Plan Oversight of Agents	Agent ID	150
Special Needs Plans (SNPs) Care Management	Member ID	205
Part D		
Retail, Home Infusion, and LTC Pharmacy Access	N/A	N/A
Medication Therapy Management Programs (MTMP)	Member ID	205
Grievances	Case ID	150
Coverage Determinations/Exceptions	Case ID	150
Appeals	Case ID	150

[1] 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.

[2] Depending on the size of the organization, some measures 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."

Measure	Sampling Unit	Sample Size[2]
Long-Term Care (LTC) Utilization	Claim ID	150
Employer/Union-Sponsored Group Health Plan Sponsors	N/A	N/A
Plan Oversight of Agents	Agent ID	150

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 Table 2.

Table 2: Example File Layout

Name	Description	Data Type/Length	Data Values	Calculation
M_ID	Member ID	Character (16)		Unique counts
DOR	Grievance Date of Receipt	DateMMDDYYYY		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.