MAINTAIN AND EXPAND
THE HEALTHCARE COST AND UTILIZATION PROJECT (HCUP)
Contract No. HHSA-290-2006-00009-C

QUARTERLY DATA PROTOCOLS
DEL #1325.04G

THIS IS A REFERENCE DOCUMENT FOR AHRQ HCUP RFP # 12-10001 AND IS AN ADAPTION OF THE ORIGINAL

Do not for cite or distribute this document other than in response to the RFP

April 15, 2011
(revised)

Prepared by:
Dan Whalen

Submitted to:
Linda Simpson, Contracting Specialist
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

Submitted by:
Thomson Reuters
5425 Hollister Ave, Suite 140
Santa Barbara, CA 93111

TABLE OF CONTENTS

INDEX OF FIGURES

INTRODUCTION

In an effort to create more timely information, the Healthcare Cost and Utilization Project (HCUP) is exploring the collection and processing of quarterly data. To achieve this goal, AHRQ plans to leverage its relationships with HCUP Partners to obtain quarterly data that becomes available shortly after a quarter's end. This report describes the procedures and protocols for processing quarterly data into standardized HCUP data files.

HCUP quarterly data files will supplement the HCUP databases, and are not intended to replace annual files. For most Partners, quarterly data is released before completion of quality checks and omit later resubmissions from hospitals and other facilities. Consequently, quarterly data will not be aggregated or otherwise used to construct annual HCUP data. Because quarterly data is available sooner than annual data, it will help create information more quickly than is currently possible. As described in the *Rapid Cycle Projection Plan* (HCUP Deliverable #1325.04A),[1] quarterly data will address four purposes:

1. HCUP Projections
2. Data Mining
3. Ad hoc resource for monitoring outbreak responses and interventions
4. Process improvement.

PROJECTIONS AND DATA MINING

HCUP Projections and Data Mining are intended to provide timely information on trends in hospitalizations. This work is motivated by a desire to rapidly identify utilization changes and create "real time" estimates for researchers and policy makers by filling the year gap from the release of the NIS to the actual year (i.e., the 2008 NIS was released in June 2010). HCUP Projections and Data Mining are intended as supplemental information activities that augment and enhance HCUP data. Estimates from annual Nationwide Inpatient Sample (NIS) files will remain the "gold standard" for hospital-based health care statistics.

AD HOC RESOURCE

Recent HCUP work assessing Rotavirus data have shown the usefulness of recent general purpose data. Much of the Rotavirus data collected were simply quarterly versions of the HCUP Partner's annual data. The analyses drawn from this data were both useful and thought-provoking and pointed towards the need for a general-purpose, ad hoc resource. Quarterly data will fill this need.

PROCESS IMPROVEMENT

Finally, quarterly data can be used for prototyping process improvements. As described in the *Quarterly Data Infrastructure Report* (Deliverable # 825.21B),[2] current annual data processing and quality assessment practices do not scale well for application to more frequent quarterly

[1] http://www.hcup-us.ahrq.gov/team/pd/deliv2/deliverable1325_04A_REVISED.pdf

[2] http://www.hcup-us.ahrq.gov/team/pd/deliv2/deliverable825_21Afinal.pdf

data. Applying the current protocols to quarterly data would require nearly one-half of the HCUP budget. The procedures and protocols described in this report illustrate a process more streamlined and efficient than that used for annual data. Lessons learned from implementing these practices will form the basis for improving the protocols for annual data files.

WORK FLOW FOR QUARTERLY DATA

The process of creating HCUP data files from quarterly data is a series of related and interlocking steps as illustrated in the Figure 1. A specific member of the processing team is expected to direct each step, but other team members may also participate. To ensure both accuracy and thoroughness, a checklist is associated with each step. The steps and checklists are described in the sections that follow.

Formally, this process consists of 12 steps:

1. Data Request
2. Data Receipt
3. Documentation and Specifications Review
4. Source Data Load
5. Initial Process Meeting
6. Release Processing Specs
7. Process State Data
8. Finalize Files
9. Processing Reviews
10. Final Process Meeting
11. Data Delivery
12. Wrap-Up.

STAFFING

For processing quarterly data, we will adapt the team approach used for annual data. Teams will consist of:

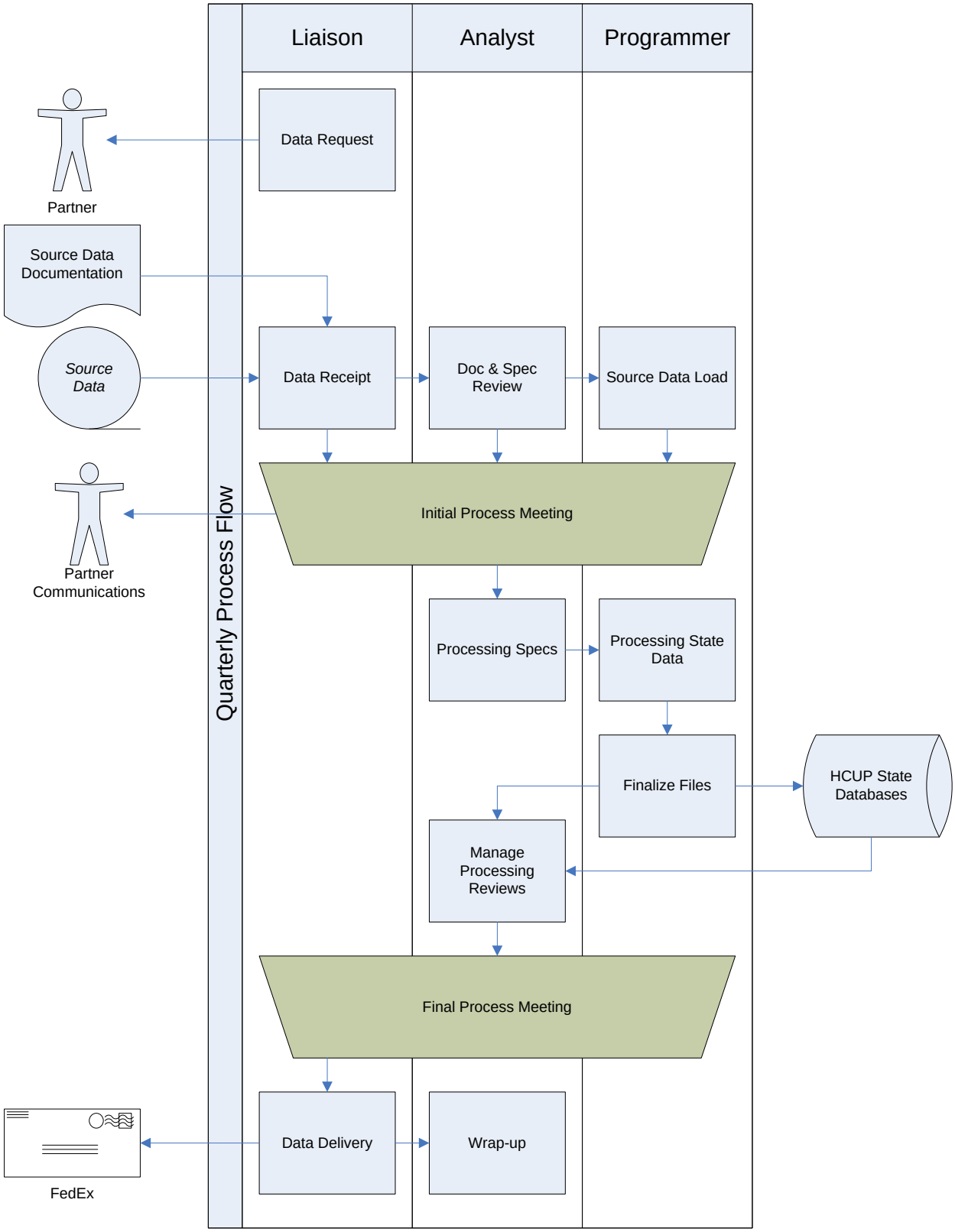
- An Analyst for managing processing, writing processing specifications, and assessing data
- A Programmer for implementing processing specifications and ensuring correct data handling
- A Liaison for coordinating Partner contacts and ensuring compliance with all data agreements and MOAs.

Liaisons from annual processing teams will serve as Liaisons for quarterly processing, ensuring a single point of contact for Partners. We believe this will simplify Partner interactions and facilitate communications. The Analyst and Programmer for quarterly data, however, will differ from the annual processing Analyst and Programmer.

In addition, staffing includes a Task Lead and a Data Acquisition Lead. The Task Lead will to oversee all processing and ensure consistent treatment of data. The Data Acquisition Lead will coordinate Liaisons and all Partner contacts.

With an initial complement of five states, a single team is sufficient. If and when additional states begin supplying quarterly data, additional teams may be added to ensure rapid completion of processing.

Figure 1. Work Flow for Processing Quarterly Data



TASK TRACKING

Progress will be tracked with a collection of linked Microsoft Excel workbooks. Task tracking for quarterly data differs slightly from task tracking for annual data because the work flow is different. Consequently, a different tracking system is needed. Rather than develop a new Web-based system (as is used for annual processing) that we would later modify as we refine our procedures, we plan to use a simpler workbook- based system as a working prototype. We believe his will be more economical for the limited number of states supplying quarterly data. If the number of participating states grows to 12 or more, however, a Web-based database system should be developed based on the workbook prototypes.

There will be one workbook for each state and data year. These state-data year workbooks will include a worksheet for each available data type, plus a summary page to track progress for all data types. A coordinating workbook – extracting information from each state workbook – will be used to manage all quarterly processing. The coordinating workbook will generate weekly and monthly status reports and also accommodate other, as yet undefined, summaries that help monitor progress. An example of the tracking information in the workbooks is found in Appendix A of this report.

TIMING OF WORK

The flow of work is designed to complete data processing within 28 days while allowing sufficient time for careful and thorough reviews of source data, programming logs, and processed data. To insure steady progress toward that completion, procedures set interim target dates for several of the sub-steps. For example, the *Initial Process Meeting* – described below – should occur within seven days (t=7) of data receipt. The Final Process Meeting should occur within 21 days (t=21) of data receipt. These target dates are based on the date on which data are received (t=0) and automatically set in the processing checklist. Checklist items are shown below with the description of each workflow step.

DATA REQUEST

The Liaison directs the data request step. Prior to making the data request, the Liaison should review the MOA along with data acquisition notes. Items to address include:

- Additional data elements to request
- Data elements no longer available
- Formal application (if needed)
- Other Partner- specific requirements.

The request for quarterly data will be interactive. With each data request, we will ask the Partner about known data problems, including the following:

- Are there any known or expected data issues?
- Are there problems with any types of data such as for particular payers?
- Are data for any hospitals omitted or excluded?
- Were there any hospital openings or closings during the quarter?

This information will be useful later for assessing data quality and completeness and also for informing the annual processing teams of important additions, deletions, and changes.

Data Request Checklist

- ☐ Checklist workbook created for the data
- ☐ Obtain and review source documentation
- ☐ MOA reviewed and updated as appropriate
- ☐ Data questions asked of the Partner
- ☐ Application completed (if needed)
- ☐ Check request (if needed)
- ☐ Data request sent
- ☐ Logged in tracking system

DATA RECEIPT

When the data receipt step is complete, the source data will be available for processing, documentation will be available, and all team members will be notified. Moreover, source media will be delivered to the Data Coordinator for secure storage.

For tracking purposes, the data received date is day zero (t=0).

Data Receipt Checklist

- ☐ Data saved to server
- ☐ Source data securely and properly stored
- ☐ Documentation saved to Halfreski
- ☐ MOA updated as appropriate
- ☐ MOA Worksheet reviewed and updated
- ☐ Data receipt notifications sent
- ☐ Logged in tracking system

DOCUMENTATION AND SPECIFICATIONS REVIEW

The Analyst is responsible for documentation and specifications review. During this step, the Analyst will begin making updates to the *Processing Specifications*.

Documentation and Specification Review Checklist

- ☐ Review previous Processing Specifications
- ☐ List blank/missing data elements from previous time period
- ☐ Review source documentation
- ☐ Update data load instructions
- ☐ Logged in tracking system

SOURCE DATA LOAD

The Programmer directs and performs the source data load step. Unless instructed otherwise in the *Processing Specifications*, all data elements should be loaded into SAS data sets with standard HCUP names. Unwanted data elements and variables of low quality will be dropped during later steps. New data elements as well as data elements of unknown type should be named as descriptively as possible, but with an "`x_`" prefix (i.e., `x_ethnicity27` or `x_provider`).

Data quality checks from this step are described below:

1. The Hospital Change Check uncovers
    - missing hospitals
    - new hospitals
    - hospitals with significantly declining discharge counts.
2. The *MOA Check* identifies
    - data elements approved on the MOA but not found of the source data
    - data elements found on the source data but not approved on the MOA.
3. The Variable Change Check indicates
    - data elements not found on the current file but available in the previous file
    - new data elements
        - complete frequencies for variables with fewer than 25 unique values
        - most frequent values for character variables with more than 25+ unique values
        - means and distribution statistics for numeric variables with than 25+ unique values.
4. The Source Data Check lists
    - frequencies for standard categorical variables compared with frequencies from the previous quarter
    - variables with excessive missing values.

Data quality checks are needed for the next step: the *Initial Process Meeting*.

Source Data Load Checklist

- ☐ SAS load data elements/fields
- ☐ Review logs and observation counts
- ☐ Run Hospital Change Check
- ☐ Run MOA Check
- ☐ Run Variable Change Check
- ☐ Run Source Data Check
- ☐ Logged in tracking system

INITIAL PROCESS MEETING

The Analyst directs the *Initial Process Meeting*, although each team member has an important role, and participation by the entire team is important. Each team member summarizes notes and findings from the previous quarter and records potential problems and issues to address

before processing is complete. These issues and items must be resolved before the *Final Process Meeting*. In addition,

- The Liaison will share any data and hospital issues conveyed by the Partner
- The Programmer will share findings from the data quality checks
- The Analyst will coordinate a schedule for processing and reviews.

After the meeting, the Analyst will report the schedule along with significant issues to the Task Lead.

Teams should maintain a goal of holding the Initial Process Meeting within seven days of data receipt (t=7).

Initial Process Meeting Checklist

Liaison

- ☐ Summarize data acquisition notes
- ☐ Review new and discontinued data elements
- ☐ Review opened and closed hospitals
- ☐ Explain other data notes from the Partner
- ☐ Record issues/items for Final Process Meeting

Programmer

- ☐ Summarize processing notes
- ☐ Review new and missing variables (from data)
- ☐ Review new and missing hospitals (from data)
- ☐ Review value distributions for standard variables
- ☐ Record issues/items for Final Process Meeting

Analyst

- ☐ Summarize notes from previous processing
- ☐ Describe final files and deliveries
- ☐ Specify any needed ad hoc statistics
- ☐ Record issues/items for Final Process Meeting
- ☐ Coordinate processing and review schedule
- ☐ Report schedule and significant issues to Task Lead
- ☐ Logged in tracking system

RELEASE PROCESSING SPECIFICATIONS

The Analyst directs and performs the *Processing Specifications* step. Using source documentation, the *Source Data Check* Report, and ad hoc statistics, the Analyst drafts the *Processing Specifications*. The Analyst and Programmer can often expect several iterations of ad hoc analyses to explore and resolve data issues and questions.

Processing Specs Checklist

- ☐ Review source documentation
- ☐ Review Source Data Check Report
- ☐ Review any ad hoc statistics
- ☐ Documentation notes specify data type (IP, AS, ED)
- ☐ Changes indicate additions to, or replacement of current note
- ☐ Update Processing Specifications and save on Halfreski
- ☐ Logged in tracking system

PROCESS STATE DATA

Processing state data will be directed by the Programmer. The core processing steps of standardizing and mapping data elements and adding data tools such as DRGs and CCS variables occur during this step. Data quality tests will be reviewed by the Analyst, who may request additional statistics and further quality tests as needed...

Process State Data Checklist

- ☐ Verify that state-specific processing code is still required
- ☐ Standardize data and add data tools
- ☐ Run data quality tests
- ☐ Analyst review of data quality tests
- ☐ Additional statistics and quality tests as needed
- ☐ Logged in tracking system

FINALIZE FILES

The Programmer directs and performs the file finalization step, dividing the processed data into appropriate final file types and destinations. Output from the *Final MOA Check* will be sent to the Liaison for the Data Review.

Finalize Files Checklist

- ☐ Create final HCUP database files
- ☐ Verify observation/record counts
- ☐ Run Final MOA Check on finalized files
- ☐ Logged in tracking system

PROCESSING REVIEWS

The Analyst manages the processing review step to make certain that each review is completed in a timely manner.

Processing Reviews Checklist

- ☐ Programmer Review
- ☐ Analyst Review
- ☐ Data Review
- ☐ Reviews logged in tracking system

FINAL PROCESS MEETING

The Analyst directs the *Final Process Meeting*, although all team members have significant responsibilities. Each team member summarizes the review for which they were responsible. The Programmer summarizes the peer Programmer Review – and reviews the issues/items identified during the Initial Process Meeting. The Analyst has overall responsibility for ensuring that all issues are resolved, and also to report significant findings to the Task Lead.

Teams should maintain a goal of holding the Final Process Meeting within 21 days of data receipt (t=21).

Final Process Meeting Checklist

Programmer

- ☐ Summarize Programmer Review
- ☐ Verify that all Programmer issues/items were addressed

Liaison

- ☐ Summarize Data Review
- ☐ Verify that all Liaison issues/items were addressed

Analyst

- ☐ Resolve any outstanding Liaison or Programmer issues
- ☐ Summarize Analyst Review
- ☐ Verify that ALL issues/items were adequately addressed
- ☐ Report significant findings to Task Lead
- ☐ Log meeting in tracking system

WRAP-UP

The Analyst directs the wrap-up step. The purpose of this step is to ensure that all HCUP quarterly data files were delivered and documented, and that all work was tracked.

Teams should maintain a goal to complete all processing steps within 28 days of data receipt (t=28).

Wrap-Up Checklist

- ☐ Update HCUPwiki with data acquisition and with processing notes
- ☐ Final Processing Specs saved on Halfreski
- ☐ HCUP-US documentation updated
- ☐ Verify that deliverable copies were sent (as appropriate)
- ☐ All activities logged in tracking system

DATA PROTOCOLS

File and directory names will closely follow standard HCUP naming conventions. Throughout this section, names will use the following shorthand:

- **ST** for the state abbreviation (i.e., CA, MN, or NY)
- **DTYPE** for the type of database
  - QSID for quarterly inpatient data
  - QSASD for quarterly ambulatory surgery (AS) data
  - QSEDD for quarterly emergency department (ED) data
- **YYYY** for the data year (i.e., 2010)
- **Q** for the calendar quarter of the data year (i.e., 1, 2, 3, or 4)
- **FILE** for the type of file (i.e., core, ddev, or cdetail)

SOURCE DATA PROTOCOLS

The steps, protocols, and procedures that follow provide a general description of the data management plan. All data (source data and processed HCUP data) are treated as confidential and managed in accordance with HCUP and Thomson Reuters security policies. Additional details on HCUP security are available in the report *HCUP Security Policies and Procedures*.

Source Data Requests and Data Receipt

Source data is formally requested from participating Partners each calendar year. Liaisons note expected data availability dates for each Partner for follow-up contacts and reminders. For all source data we:

- Record the data receipt in the *HCUP Data Check-in System*
- Assign a Tracking ID Number (TIDN) to the source data media
- Copy data to the source data directory on the quarterly data processing server
- Deposit the source data media in locked, restricted-access storage
- Register the data date of receipt for tracking.

Destruction protocols for source data are described below under the heading “Source Data Destruction.”

Source Data Storage and Backup

Source data is stored separately from processing data and the final analysis files to simplify maintenance and data destruction. On the specified server, source data is saved in the "qsrcdata" directory organized first by year and quarter, then by state, and finally by data type

(i.e., .\qsrcdata**YYYY**q**Q\ST\TYPE**). For example, Minnesota Q1-2010 source data will be saved in the directories:

- .\qsrcdata\2010q1\MN\ip – for inpatient data
- .\qsrcdata\2010q1\MN\as – for ambulatory surgery (AS) data
- .\qsrcdata\2010q1\MN\ed – for emergency department (ED) data

Source data file names will reflect the naming conventions of the Partner; this simplifies Partner communications when resolving data issues.

Incremental backups (changes) for each data year occur Monday through Thursday. Tapes for these backups are reused weekly. In addition, complete backups are performed every Friday.

- Friday backup tapes are generally retained for approximately four weeks
- Weekly backups from the last Friday of each month are retained for up to three years as part of the Monthly backup
- Monthly backups for March, June, September, and December are retained for up to five years as the Quarterly backup.

Destruction protocols for source data backups are described below under the heading “Source Data Destruction.”

Source Data Destruction

The Quarterly Data Processing Group will notify AHRQ when processing of a data year is complete. At AHRQ’s direction,

- All source data for the data year will be deleted from the processing server
- All backup tapes (Monthly and Quarterly) for the data year will be erased and reused
- All source media for the data year will be securely destroyed by a certified vendor.

Quarterly source data will be exiled by data year after all annual data for the data year is received.

PROCESSED DATA PROTOCOLS

The steps, protocols, and procedures that follow provide a general description of the data management plan. All data, both source data and processed HCUP data, are treated as confidential and managed in accordance with HCUP and Thomson Reuters security policies. Additional details on HCUP security are available in the report *HCUP Security Policies and Procedures*.

Creating uniformly- formatted state data files is the key objective for processing quarterly data. Data processing and final files and data elements will be comparable with annual files and data elements.

Quarterly Data Files

The structure of quarterly state data files – QSID, QSASD, and QSEDD – will be simpler than the structure of annual state databases. The quarterly data files will consist of only two or three files:

- Core – the basic intramural data elements

- Data Development – confidential and sensitive data elements such as full dates and (encrypted) patient IDs
- Charges (optional) – detailed charges, sometimes referred to line item details.

Unlike databases for annual data, quarterly databases will not incorporate separate files for diagnoses and procedures, nor will they include data elements for condition severity or a severity file. It should be further noted that quarterly data will not be available from the HCUP Central Distributor, so no quarterly Central Distributor files will be created.

HCUP Data Storage and Backups

HCUP quarterly data is stored on the processing server, Source data is saved in the "qdata" directory organized first by state, then by year and quarter, and finally by data type (i.e., .\qsrcdata**ST\YYYY**q**Q\DTYPE**). For example, Minnesota Q1-2010 data will be saved in the directories:

- .\qdata\MN\2010q1\qsid – for inpatient data
- .\qdata\MN\2010q1\qsasd – for ambulatory surgery (AS) data
- .\qdata\MN\2010q1\qsedd – for emergency department (ED) data

The final files will use the general form **ST_DTYPE_YYYY**q**Q_FILE**. Using Minnesota 2010 Q1 QSID files as an example:

- Core file: `mn_qsid_2010q1_core`
- Data Development file: `mn_qsid_2010q1_ddev`
- Detailed Charges file: `mn_qsid_2010q1_cdetail`.

Incremental backups (changes) for each data year occur Monday through Thursday. Tapes for these backups reused weekly. In addition, complete backups are performed every Friday.

- Friday backup tapes are generally retained for approximately four weeks
- Weekly backups from the last Friday of each month are retained for up to three years as the Monthly backup
- Monthly backups for March, June, September, and December are retained for up to five years as the Quarterly backup.

Destruction protocols for source data backups are described below under the heading “Source Data Destruction.”

Data Elements

Data elements in quarterly data files are largely a subset of data elements found in the annual HCUP databases. The quarterly files omit many derived data elements, such as AHRQ co-morbidity indicators and disease staging variables. In addition to basic discharge data elements, including NUBC- defined information, data elements were selected if they were

- Necessary for PSI calculations
- Needed for comparing quarterly data to annual data
- Useful for common analysis tasks.

To check for omissions, we compared the final list to the data used for the recent HCUP Rotavirus Analysis.

A complete list of the data elements for the quarterly data files is found in Appendix B. The list of data elements in future years may change at AHRQ's discretion.

Delivery of Processed Data

When processing is complete and the final data files pass review, the data is ready for delivery. Data delivery will occur by State, data file, and quarter as the data become available. Deliveries will be included with Friday HCUP deliveries. Files on DVD discs will be encrypted according to data security standards in place at the time of delivery. The delivery will include two DVDs:

- DVD #1: Core File and CDetail file (if available)
- DVD #2: DDev file.

Encryption passwords will be delivered through a separate delivery stream.

In addition, the final quarterly data will be stored on the analysis server. For annual data, this is a task performed separately from processing. The final data will be saved to:

- .\states\qsid\2010 – for all quarters of 2010 inpatient data
- .\states\qsasd\2010 – for all quarters of 2010 AS data
- .\states\qsedd\2010 – for all quarters of 2010 ED data

The final files will use the general form **ST_DTYPE_YYYY**q**Q_FILE**. Continuing our example, the final Minnesota core inpatient data will be named `mn_qsid_2010q1_core`.

QUARTERLY DATA PROCESSING PROGRAMS

The architecture for quarterly data divides the process into efficient and logical steps that encourage thoughtful reviews and assessment of the data. Generally, transformative steps are followed by analysis steps. The programs in this structure are completely new; for purposes of process improvement, the programs and modules from the annual processor are not used. The result is a more efficient organization and arrangement of programs, which will be easier to maintain and update.

DATA FLOW FOR PROCESSING QUARTERLY DATA

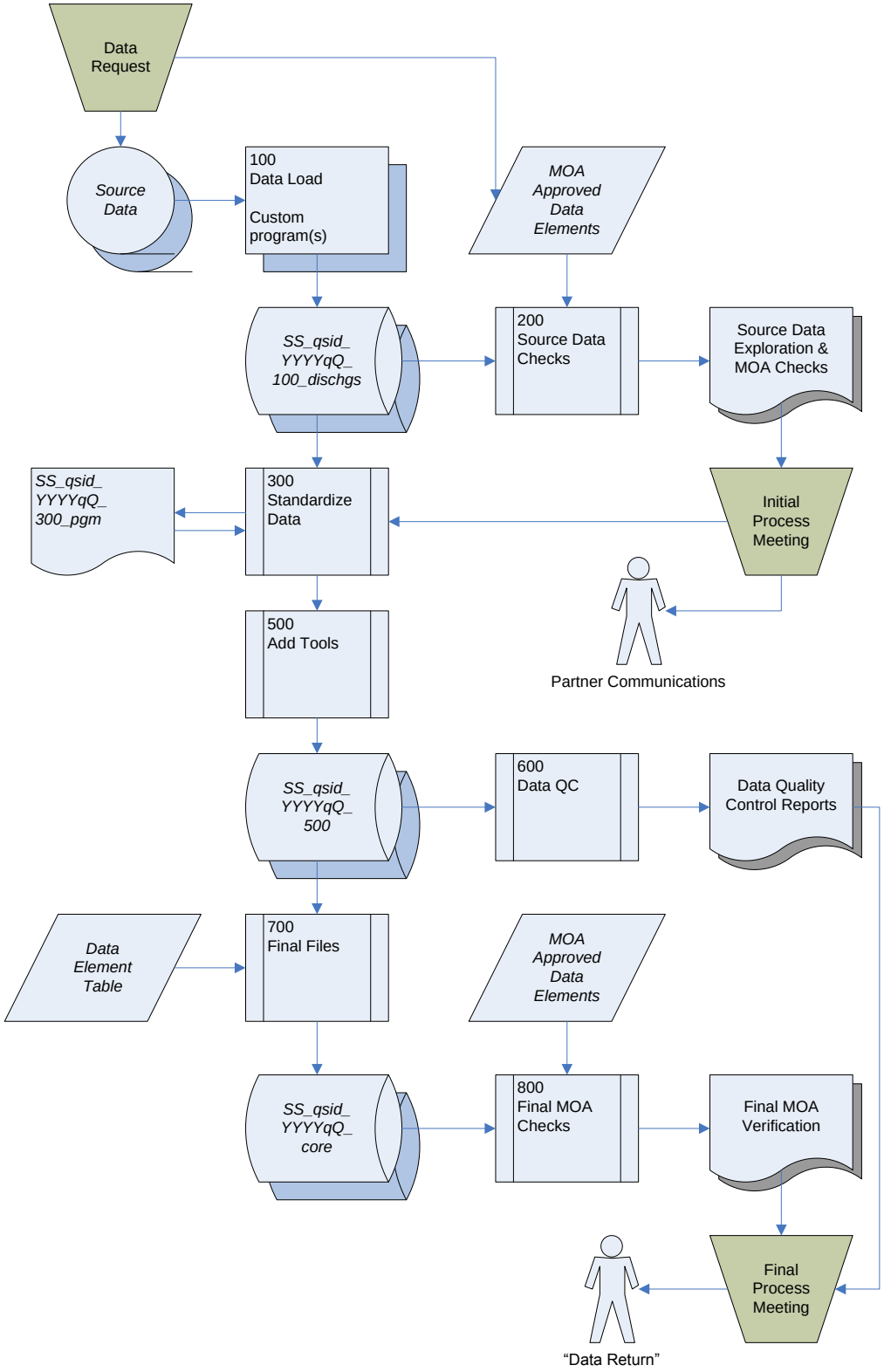
This section illustrates the data flow and describes how the modular programming components fit together to process quarterly data. The overall protocol is divided into seven numbered steps:

- 100 - Data Load
- 200 - Source Data Checks
- 300 - Standardize Data
- 500 - Add Tools
- 600 - Data Quality Control
- 700 - Final Files
- 800 - Final MOA Checks

Steps 100, 300, 500, and 700 involve some form of data processing or transformation. Steps 200, 600, and 800 test or report on data created in a previous step (there is no 400 reporting step because the transformation steps 300 and 500 are set up in a single series of programs). The process is designed in a modular fashion to easily accommodate changes. Most modules will be designed to function independently from other modules. This will facilitate adding and removing modules as the need arises. Figure 2 illustrates the relationship of the modules and the overall flow of data.

Team meetings to examine data quality and discuss data issues are built in toward the beginning and near the end of the process.

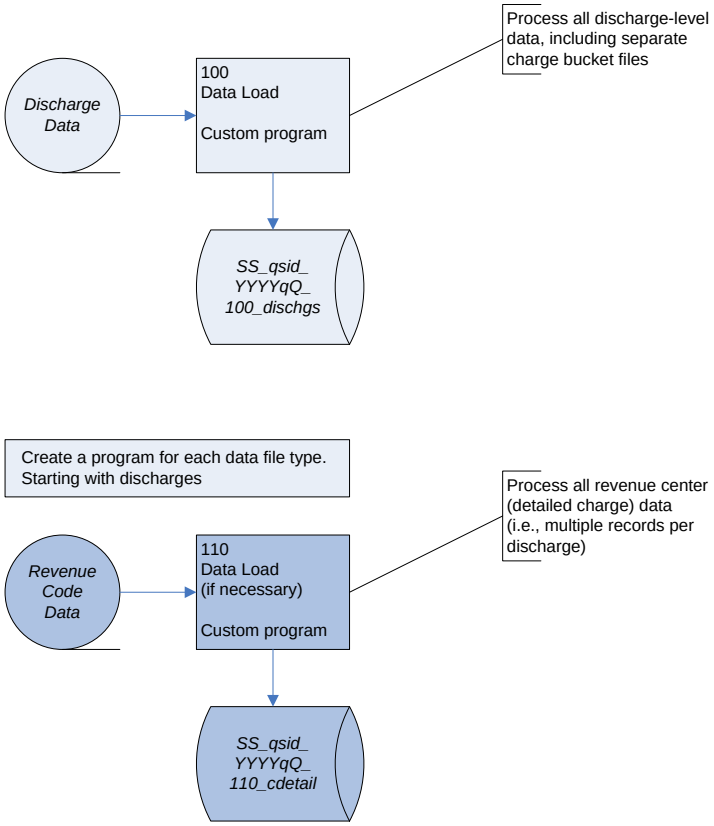
Figure 2. Overall Data Flow for Processing Quarterly Data



Initial Data Load (Step 100)

The first step is loading source data into SAS data sets. This is a manual process. Starting with a minimal program template, the programmer will either: (1) SAS load data elements from raw source data, or (2) rename data elements from source SAS data. Figure 3 illustrates the data flow for this step. The detailed charge data portion of the diagram is a darker blue to indicate that not all states provide detailed charge data with revenue codes.

Figure 3. Data Flow, Database Load



The programmer will account for all data elements provided from the Partner. For that reason, the data load will include all data elements available in the source data. The resulting data set will include two types of data elements:

1. HCUP data elements – dates and other numeric data will be loaded as SAS dates and numbers; we will omit the elaborate data loads used in the annual processor. Those methods provide little value and are holdovers from early HCUP days when data quality was much lower than today.

   All HCUP data elements will be loaded with an "`i`*" prefix, for example*

   - *admission date as* `iadate`
   - *hospital ID as* `idshospid`
   - *total hospital charge as* `itotchg_x`

2. Unwanted, unknown, and uncertain data elements – this will include data elements of unknown quality as well as unexpected and unwanted data elements.

All unwanted and unknown data elements will be loaded with an "x" prefix.

The step 100 program should load all discharge-level data into a single discharge-level data set. If detailed (revenue center) charge data is provided, it should be loaded with a separate program (step 110) into a separate line item-level data set.

Source Data Checks (Step 200)

The first report program takes the input data from step 100 and creates four separate data checks. This step is illustrated in Figure 4 – detailed charges are dark blue to indicate their optional nature.

1. *MOA Check* – compares the data to the MOA and reports data elements that are
    - listed on the MOA but not found in the data
    - found in the data but not listed on the MOA
2. *Data Element Check* – compares the current source data with previous source data and reports
    - new data elements (available in the current data but not found in the previous data)
    - missing data elements (not found in the current data but available in the previous data)
3. *Source Data Check* – compares current data values to values from the previous quarter, for common, essential data elements such as discharge status and payer to identify new and discontinued values
4. *Hospital Check* – compares hospital IDs found in the current and previous time period's data to highlight new and missing hospitals.

The *MOA Check* and the *Data Element Check* examine similar situations from different perspectives. The two checks are designed to function as complements to ensure that the final data include no unapproved data elements. Figure 4 illustrates the data flow for this step.

Each team member holds responsibility for assessing at least one of the data checks. The results of those assessments are shared with all of the processing team at the *Initial Process Meeting*.

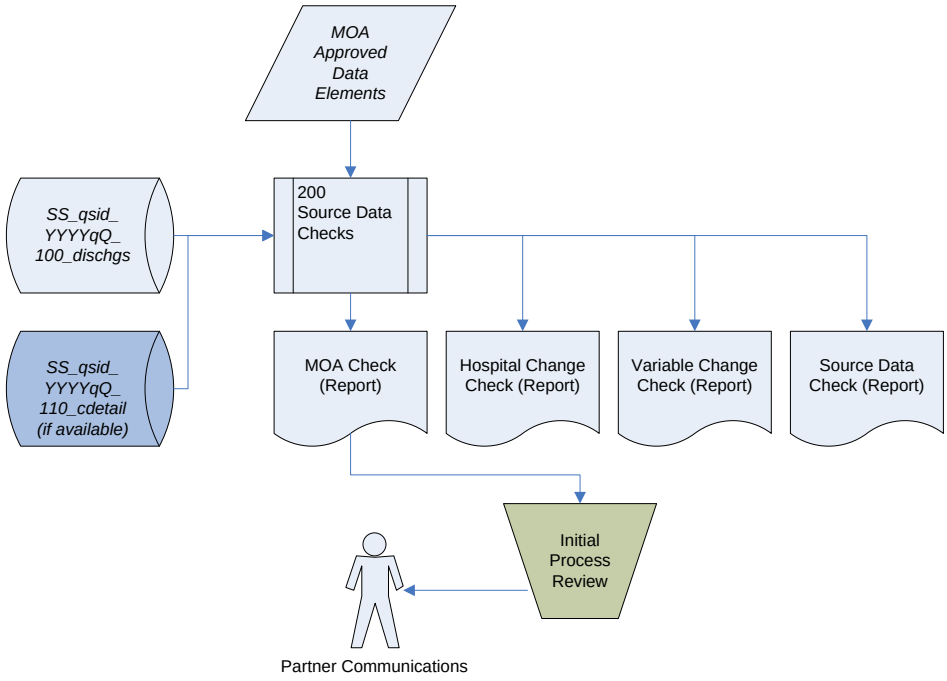
In addition, file construction will be consistent from year to year. Quarterly data processing will ensure that:

- Data elements processed are consistent with the Partner's MOA
- Source data codes are consistent with documented values
- Derived data elements are consistent with HCUP standards.

Initial Process Meeting

The *Initial Process Meeting* occurs after the Source Data Checks are available. This meeting is intended as a communications platform for the team to discuss and review issues that may affect the data. Processing teams will hold the *Initial Process Meeting* within seven days of data receipt. The meeting is described in greater detail above, within the "Work Flow for Quarterly Data" section.

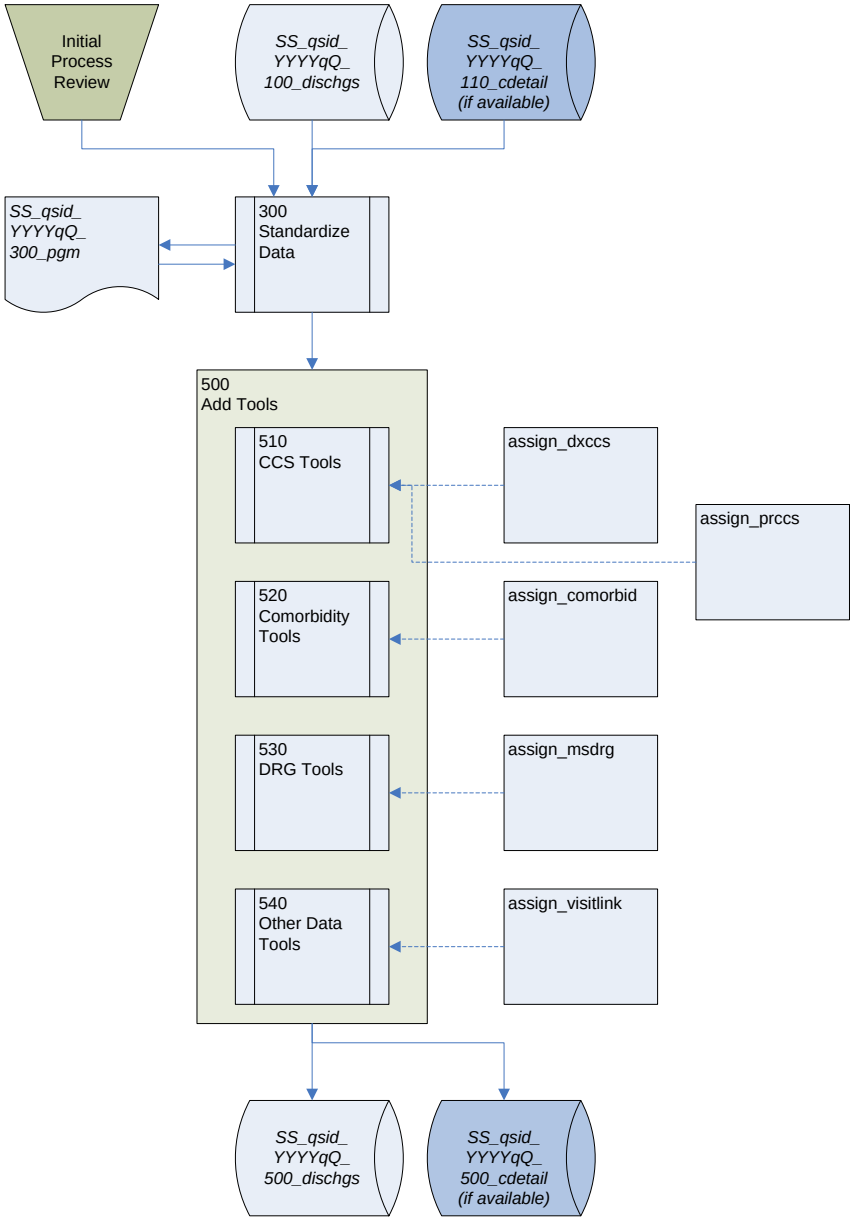
Figure 4. Data Flow, Source Data Checks



Core Data Processing (Steps 300 and 500)

The core processing model for quarterly data consists of a series of loosely coupled, highly cohesive tasks running serially. Programming calls to each task specify a data set on which the task will operate. When complete, the task returns a modified data set that serves an input to the next task. This is computationally less efficient than the model currently used for processing annual data. Operationally, however, this model simplifies the work necessary to add, delete, or modify processing components because the components will be designed with little interdependence. Building separate modules for each operation virtually eliminates the inter-dependence and coupling that binds the annual processor. Metaphorically, this is a relay model passing data from task to task until the "race" is finished. These steps are illustrated in Figure 5 – detailed charges are shown in dark blue to indicate their optional nature.

Figure 5. Data Flow, Core Processing



Core data processing encompasses two separate steps: standardizing basic data elements and adding data elements and tools.

Standardizing Basic Data Elements (Step 300)

Standardizing data elements reflects the mapping (or re-coding) of data elements from state-specific values into HCUP specific values. Processing in this step affects data elements available, in some form, in the data. Some examples are:

- Discharge disposition – `dispub04` from `idisp`
- Payer – `pay1` from `ipay1` and `pay2` from `ipay2`
- ZIP Code – `zip` from `izip`.

The standardization step will clean the data to make it more usable. An example might involve setting invalid values or values that signify “unknown” or “unavailable” to blank or missing. Finally, this step will automatically perform data cleanup to:

- Compress arrays (vectors) of diagnoses and procedures to remove blanks
- Separate ICD-9 procedures from CPT-4 procedures.

All mappings will be specified by the Analyst in the Processing Specs and implemented by the Programmer in a state-specific “PGM” module.

Adding Data Tools (Step 500)

The second facet of core data processing enhances the data by adding derived data elements and data tools. Among the data tools implemented with the first iteration of the quarterly data processor are:

- Single- and multi-level diagnosis CCS (`dxccs1` and `dxmccs1`)
- Single- and multi-level procedure CCS (`prccs1` and `prmccs1`)
- MS-DRG and MDC (`drg` and `mdc`)
- Revisit variables (`visitLink` and `daysToEvent`)
- Transfer indicators (`tran_in` and `tran_out`).

While co-morbidity data elements are not part of the first iterations of quarterly data[3], Figure 5 does present a co-morbidity module. This inclusion is intended to illustrate the flexible nature of the processor: additional modules, such as co-morbidity, are easy to add.

Testing Data Quality (Step 600)

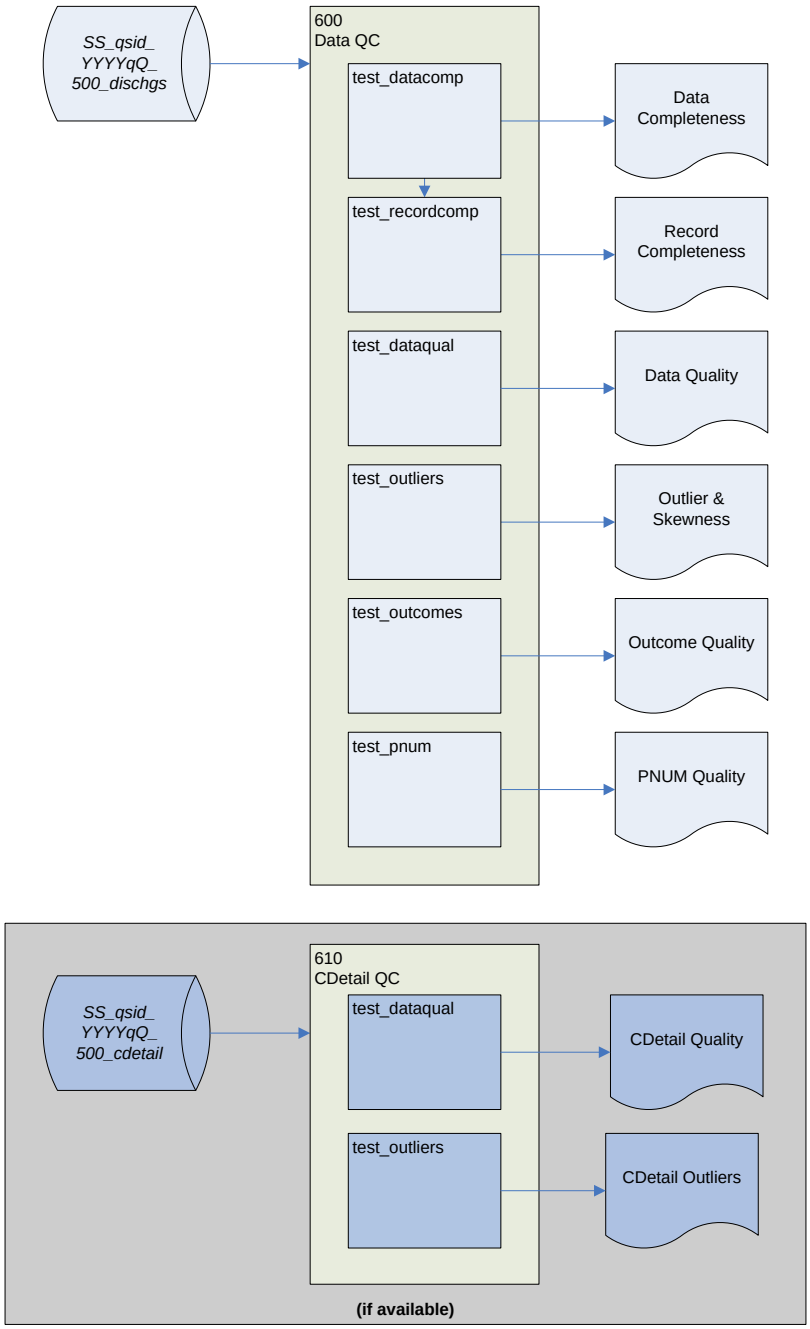
The usefulness of quarterly data depends on its completeness and quality. Complete data will capture nearly all hospital events and quality data will include few, if any, errors. The final report program takes the processed data from steps 300 and 500 to create five or six data quality reports:

1. Data Completeness
2. Record Completeness
3. Data Quality Report
4. Outlier & Skewness Report
5. Outcome Quality Report
6. PNUM Quality Report.

This step is illustrated in Figure 6. Separate data quality reports for detailed charges appear in dark blue to indicate their optional nature.

[3]Refer to Appendix A for a complete list of data elements.

Figure 6. Data Flow, Quality Control



Record Completeness Tests

Record completeness tests focus on discharge (record) counts with the objective of identifying incomplete files. To do this, we compare the number of discharges in the current quarterly data file to those in previously- received data. This comparison data is not an exact or perfect fit because it references record counts from a different time period. However, we expect most quarter-to-quarter differences to be small: growth in the number of NIS discharges in recent

years has been between 0.5 and 1.0 percent. In addition, in recent years, overall U.S. population growth has averaged less than 1 percent annually. Even in quickly- growing states, such as Nevada, the growth rate has averaged less than 4 percent over the past decade.

For assessment against previously- received data, we look for differences between the data submitted by the Partner for the previous quarter and the same quarter from the prior year, usually extracted from the most recently- available SID file. Where data usually exhibit a trend – increasing or decreasing – the previous quarter should be the more relevant comparison, while comparisons to the same quarter of the prior year are more appropriate for data with seasonal variations.

We evaluate record completeness tests across three categories:

1. Record counts by discharge month: same quarter prior year, as expected based on the number of days in the month
2. Record counts by age group: previous quarter, same quarter prior year
3. Record counts by hospital: previous quarter, same quarter prior year.

Data Completeness Tests

Measures of data completeness focus on identifying data elements with unusually large percentages of missing or invalid values. The purpose of these tests is to discover incomplete data by identifying variables with abnormal levels of missing values. Comparisons will be made to inner-decile ranges (between the 10^{th} and 90^{th} percentiles), based on measures from recent years' data. Where norms for missing values are very small, less than 1 percent, the range of acceptable values will be set at 1 percent.

Examined data elements will include:

- Admission date (`adate`)
- Admission type (`atype`)
- Admission point of origin (`pointoforiginub04`)
- Discharge disposition (`dispub04`)
- Patient age (`age`)
- Patient gender (`female`)
- Patient race (`race`)
- Patient ZIP Code (`zip`)
- Principal diagnosis (`dx1`)
- Primary procedure (`pr1`)
- Total charge (`totchg`).

Data Quality Tests

Data quality tests assess average values of common descriptive variables to highlight miscoded data. Comparisons are made to inner-decile ranges, based on measures from recent years and all available states.

Data quality tests will examine the mean values of these data elements:

- Patient gender (`female`)
- Patient age (`age`)
- In-hospital mortality indicator (`died`)

- Emergency department services indicator (hcup_ed)
- Neonatal/maternity indicator (neomat)
- Number of diagnoses (ndx)
- Number of procedures (npr)
- Procedure #1 days from admission (prday1)
- Procedure #2 days from admission (prday2)
- Procedure #3 days from admission (prday3)

For detailed charge data, data quality tests will examine the mean values of:

- Charge detail amount (charge)
- Service day (servday).

Tests of Outliers and Skewness

The outlier tests assess the distribution of select numeric variables to highlight extreme values. For example, the distributions of length of stay (los) and total charge, because these have a lower bound (zero), are usually skewed. However, the amount of skewness that is normal and reasonable has not been previously explored within these data.

The outlier tests are designed to highlight abnormal distributions of data values. Similar to record completeness tests, comparisons will be made to inter-decile ranges from recent years and all available states. Outlier tests will be performed for age and los:

- The difference between the mean and median value
- The percentage of records with values more than two standard deviations above the mean
- Kurtosis, a measure of "peakedness" for the probability distribution, where a high kurtosis value indicates that more of the variance is the result of infrequent extreme deviations.

For detailed charge data, outlier tests will be performed for charge detail amount (charge).

Outcome Quality

Quality measures examine common outcomes, such as average length of stay (ALOS), in-hospital mortality rates, and average total charges. We contrast nine statistics for the ALOS from the current quarterly data file to previously- received data from the same Partner, as well as comparisons to NIS estimates. Together these measures should thoroughly examine the data quality in terms of consistency, accuracy, and reasonableness.

Tests will include:

- Aggregate (overall) statistics
  - Compared to the prior year's average (or rate)
  - Compared to the average (or rate) from same quarter of the previous year
  - Compared to the most recent NIS estimate

  Evaluation: t tests (alpha of 1%).
- Statistics by payer (pay1)
  - Compared to prior year's statistics
  - Compared to the same quarter of the previous year's statistic

- Compared to the most recent NIS (or NEDS) estimates

Evaluation: proportional distribution-based chi square statistics (alpha of 1%).

- Statistics by hospital (`dshospid`)
  - Compared to prior year's statistics
  - Compared to the same quarter of the previous year's statistic

  Evaluation: proportional distribution-based chi square statistics (alpha of 1%).
- Statistics by discharge month (`dmonth`)
  - Compared to prior year's statistics
  - Compared to the same quarter of the previous year's statistic
  - Compared to the most recent NIS (or NEDS) estimates

  Evaluation: proportional distribution-based chi square statistics (alpha of 1%).

PNUM Quality

Patient identifiers (`PNUM`) are not available from all HCUP Partners, so this final report will not be available for all states. When it is available, PNUM tests will evaluate the quality and continuity of PNUM values. Comparisons will be made to inner-decile ranges (between the 10th and 90th percentiles), based on measures from recent years' data. Where norms for missing values are very small, less than 1 percent, the range of acceptable values will be set at 1 percent.

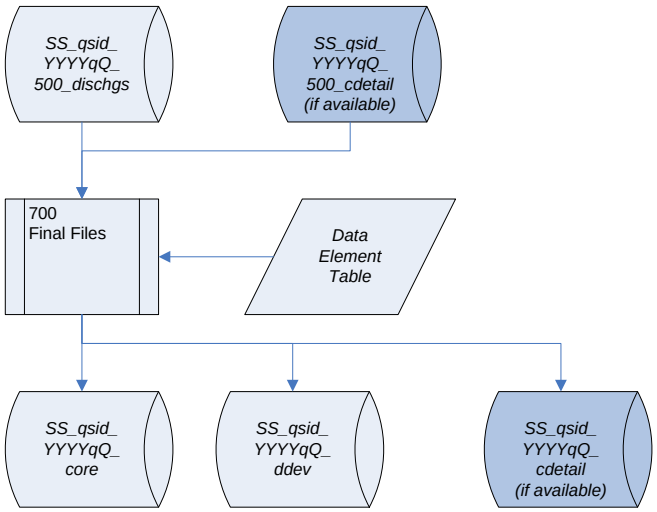
Tests will include:

- Percent with missing values
- Percent with 40+ events
- Percent ZIP3/DOB/Gender with 2+ PNUMs
- Percent overlap with most recent quarter of data
- Percent overlap with most recent quarter of IP data (test for AS and ED data only)
- Percent overlap with most recent quarter of AS data (test for IP and ED data only)
- Percent overlap with most recent quarter of ED data (test for IP and AS data only).

Creating Final Files (Step 700)

The next step in quarterly data processing divides the standardized data into *core*, *ddev* (data development), and optional *cdetail* (detailed charges) deliverable files. This process is controlled by a data element table indicating which data elements are permissible in each file. Figure 7 illustrates this step. As with other diagrams, dark blue is used for detailed charges to indicate their optional nature.

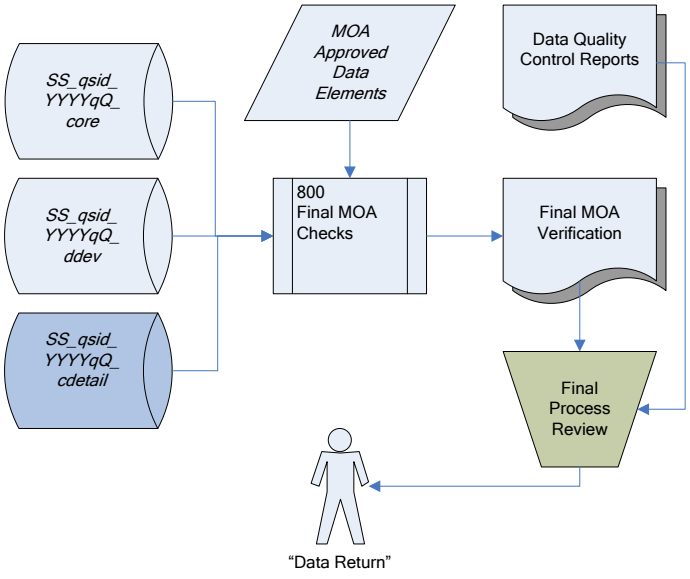
Figure 7. Create Final Deliverable Files



Final MOA Checks (Step 800)

As a closing quality control check, final deliverable files will be compared with the approved data elements from the MOA to ensure that only permitted data elements are included. This step is illustrated in Figure 8.

Figure 8. Final Reviews and Data Delivery



Final Process Meeting

The *Final Process Meeting* occurs after the Data Quality Reports and MOA Verification are available and reviewed. The purpose of this meeting is to verify that processing is complete and correct. Issues identified during the *Initial Process Meeting* must be resolved. Upon completion of the Final Process Meeting, data are ready for delivery to AHRQ and the HCUP Analytic Server.

Processing teams will hold the *Final Process Meeting* within 21 days of data receipt. The meeting is described in greater detail above, within the "Work Flow for Quarterly Data" section.

PROCESSING DATA AND PROGRAMS

Processing data will be organized first by state, then by year and quarter, and finally by data type (i.e., .\qdata**ST\YYYY**q**Q\DTYPE**). Minnesota inpatient data and programs will be saved in the directories:

- .\qdata\MN\2010q1\qsid\pgms – for processing programs

```
mn_qsid_2010q1_100_dataload
mn_qsid_2010q1_200_sourcechks
mn_qsid_2010q1_300_standardize
mn_qsid_2010q1_500_addtools
mn_qsid_2010q1_600_dataqc
mn_qsid_2010q1_700_finalfiles
mn_qsid_2010q1_800_finalchecks
```

- .\qdata\MN\2010q1\qsid\data – for processing data
- .\qdata\MN\2010q1\qsid\data\tests – for data quality test reports
- .\qdata\MN\2010q1\qsid\data\stats – for documentation statistics

Data file names will include an additional qualifier to indicate the specific step that created the data. For example discharge data from the 500 "Add Data Tools" program is named mn_qsid_2010q1_500_discharges.

BACKEND PROCESSING TOOLS

To improve efficiency and reliability, programming code will be divided into cohesive and reasonably compact modules. Design of these modules will minimize "coupling" in that actions performed by each module will modify only the data elements that module was explicitly designed to affect, but it will not modify any other data elements. In computer parlance, modules will be "orthogonal." Modules will be centrally stored on the server in directories of cascading and increasing specificity.

- .\shared\tools – for generally useful HCUP tools
- .\shared\tools\processing – for tools applicable only to data processing
- .\shared\tools\reviews – for tools applicable only for reviews
- .\shared\tools\documentation – for tools applicable to documenting HCUP data files

DOCUMENTATION

Access to quarterly data documentation will be password- protected and accessible only to the HCUP Team. In form, documentation contents will mirror existing HCUP-US documentation of

State Databases. For the quarterly inpatient databases, the HCUP-US page will include a short explanation followed by four sections:

1. Availability and Description of QSID/QSASD/QSEDD Files
2. Descriptions of Data Elements in the QSID/QSASD/QSEDD
3. Load Programs
4. SAS File Information.

AVAILABILITY AND DESCRIPTION OF QSID FILES

Availability of States by Calendar Year and Quarter

- Links to a separate HCUP-US page *qsid_availability*
- A table with columns of states and year-quarter in rows.

File Descriptions

- Links to HCUP-US page *qsid_filedesc.jsp* listing participating states
- Each listed state is a link to the file description page for that state. Because the file composition of quarterly data is very similar to the (annual) intramural files, the description page will focus on data completeness and descriptions of issues important to researchers.

DESCRIPTIONS OF DATA ELEMENTS IN THE QSID

Description for all states across all years and quarters

- Links to the HCUP-US page *qsidvarnote_allstates.jsp*
- List of all quarterly database variables – each listed variable is a link to that variable note

Summary Statistics for all years and quarters with links to the *stats_search.jsp* page – this requires a page update to incorporate quarterly databases.

- A link to the PDF version of the Summary Statistics for a particular calendar year and quarter
- Included under Summary Statistics will be the Data Quality Reports in PDF format

LOAD PROGRAMS

Processing Programs for all states, years, and quarters

- Links to the HCUP-US page *qsidprocprog.jsp*
- A list of all participating states and time periods – each listing links to the Zipped archive of the data processing programs for the specified time period.

SAS FILE INFORMATION

File information for all states across years and quarters

- Links to the HCUP-US page *qfilesearch.jsp* – select by state, year and quarter.

APPENDICES

APPENDIX A – TASK TRACKING

YEAR AND QUARTER TRACKING WORKSHEET

- One workbook per state
- One data type worksheet per state workbook
- Multiple year-quarter sections in each data type worksheet

Guam IP

2011 Q4

	Date	Current Expected Date	Baseline Expected Date	Notes/Reminders for Team
Data Request				
Not Started	#N/A		*(baseline driven from scheduled release date)*	
Sched Data Release		*n/a*	*n/a*	
Data Requested		--	--	
Data Received		--	--	TIDN:
Comments				
Processing				
Not Started	#N/A		*(baseline driven from data received date)*	
Data Load Instructions		--	--	
Data Load Complete		--	--	
Initial Meeting		--	--	
Finalize Specs		--	--	
Process Data		--	--	
Finalize Files		--	--	
Reviews Complete		--	--	
Final Meeting		--	--	
Data Delivery		--	--	
Processing Complete		--	--	
Comments				

top of page

APPENDIX B – QUARTERLY DATA ELEMENTS AND FILES

CORE FILE DATA ELEMENTS

Variable Names	Description	Def	Value	Value label	Comments
qkey	*Record identifier*	*8*	*15(n)*	*Record identifier (coded as SSyyyyQTnnnnnnn, where SS = State FIPs code, yyyy = data year, Q = quarter, T = data type 1 (inpatient), 2 (ambulatory surgery), 3 (emergency department), and nnnnnnn = record counter)*	*Assigned during processing. QKEY does not include a hospital identifier.*
age	Age in years at admission	3	0-124	Age in years	
ageday	Age in days (when < 1 year)	3	0 - 364	Age in days	
amonth	Admission month	3	1-12	Admit month	
atype	Admission type	3	1	NUBC definitions	Code deliveries under Urgent (2)
cptn	CPT-4/HCPCS procedures	$5	5(a)	CPT or HCPCS Procedure Codes	n=15
cptccsn	Clinical Classifications Software (CCS): CPT/HCPCS procedure classification	3	1 - 231	CCS Proc Class	n=15
daystoevent	Randomized sequencing variable for calculating days betwee hospital events associated with a unique patient (visitlink)	8	9(n)	Assigned number	
died	Died during hospitalization	3	n	T/F indicator	Recoded from dispub04
dispub04	Disposition of patient, UB04 standard coding	3	nn	NUBC definitions	
dqtr	Discharge quarter	3	n	Derived from ddate	

Variable Names	Description	Def	Value	Value label	Comments
drg	DRG in use on discharge date	3	3(n)	DRG value	
drgver	DRG grouper version used on discharge date	3	N	DRG grouper version used on discharge date	
dshospid	Data source hospital number	$13	13(a)	Data source hospital number	
dxccsn	Clinical Classifications Software (CCS): diagnosis classification	4	nnnn	CCS Diagnosis Class	n=15
dxmccsn	Multi-Level CCS: principal diagnosis / diagnosis n	$11	nn.nn.nn.nn	Multi-Level CCS value	The DX and E code variables will be 11 characters "nn.nn.nn.nn" with zero padding for single digit numbers - example 09.01.02.02. Unused levels should be blank.
dxpoan	Diagnosis n, present on admission indicator	$1	Y	NUBC definitions	n=15
dxn	Diagnosis	$5	annnn	ICD-9 diagnosis codes	n=15
e_ccsn	Clinical Classifications Software (CCS): E code classification	3	nnnn	CCS E-code Class	n=15
e_mccsn	Multi-Level CCS: E Code n	$11	nn.nn.nn.nn	Multi-Level CCS value	The DX and E code variables will be 11 characters "nn.nn.nn.nn" with zero padding for single digit numbers - example 09.01.02.02. Unused levels should be blank.
ecoden	External cause of injury code	$5	E code	ICD-9 E codes	n=15
e_poan	E Code n, present on admission indicator	$1	Y	NUBC definitions	n=15
female	Indicator of sex	3	n	T/F indicator	
hcup_ed	HCUP emergency department indicator	3	n	HCUP definitions	
hospst	Hospital state	$2	aa	Hospital state postal code	Coded for all observations
los	Length of stay, cleaned	4	0 - 365	Days	Same day stays coded as 0.

Variable Names	Description	Def	Value	Value label	Comments
mdc	MDC in use on discharge date	3	nn	MDC value	Coded for all observations.
medincstq	Median household income state quartile for patient ZIP Code	3	n	State Income Quartile	Quartiles are individually defined for each state. Median Income Quartile for ZIP Codes.
ncpt	Number of CPT/HCPCS procedures for this discharge	3	0 - 15	Number of CPT procedures	
ndx	Number of diagnoses for this discharge	3	0 - 15	Number of diagnoses	
necode	Number of E codes on this record	3	0 - 6	Number of E codes	
neomat	Neonatal/maternal discharge	3	n	HCUP definitions	Based on diagnosis and procedure screens.
npr	Number of procedures for this discharge	3	0 - 15	Number of procedures	
nrevcd	Number of revenue codes for this discharge	3	0 - 25	Number of revenue codes	
pay1	Expected primary payer, uniform	3	n	HCUP definitions	
pay2	Expected secondary payer, uniform	3	n	HCUP definitions	
pl_cbsa	Patient location: Core Based Statistical Area (CBSA)	3	n	Core-Based Statistical Area	
pl_nchs2006	NCHS Urban-Rural Code, 2006	3	n	NCHS Urban-Rural Code	
pnum_r	Person number (re-identified)	5	9(n)	Person number	
pointoforiginub04	Point of origin for admission or visit, UB-04 standard coding	$1	a	NUBC definitions	
prccsn	Clinical Classifications Software (CCS): procedure classification	3	1 - 231	CCS Proc Class	n=15
prdayn	Day of procedure	4	5(n)	Days before/after Admission	

Variable Names	Description	Def	Value	Value label	Comments
prmccsn	Multi-Level CCS: principal procedure / procedure n	$8	nn.nn.nn	Multi-Level CCS value	
prn	Procedure	$4	nnnn	ICD-9 procedure codes	n=15
pstco2	Patient state/county FIPS code	4	nnnnn	State/county FIPS code	From provided PSTCO if available; otherwise derived from ZIP Code
race	Race	3	n	HCUP definitions	Generally, Hispanic is coded as a separate variable so that it can be combined with various races. Perhaps we should follow this convention and replace the “Hispanic” category with a separate Hispanic variable.
revcdn	Line item revenue code	$4	4(a)	NUBC revenue codes	From line item detail.
revchgn	Detailed charges for revenue code n (as received from source)	6	7(n).nn	Charge	From line item detail.
totchg	Total charges, cleaned	6	25 - 1 million	Total Charge (rounded)	
tran_in	Transfer into hospital	3	n	HCUP definitions	
tran_out	Transfer out of hospital	3	n	HCUP definitions	
visitlink	Linkage variable for all events associated with a unique patient	8	9(n)	Assigned number	
year	Discharge year	3	4(n)	Derived from ddate	
zip_s	Patient ZIP Code (synthetic)	$5	5(a)	HCUP definitions	ZIP Codes are primarily used to indicate geographical location so set foreign codes or identifications and homeless to missing.
zipinc_qrtl	Median household income national quartile for patient ZIP Code	3	n	National Income Quartile	

DDEV DATA ELEMENTS

Variable Names	Description	Def	Value	Value label	Comments
qkey	*Record identifier*	*8*	*15(n)*	*Record identifier (coded as SSyyyyQTnnnnnnn, where SS = State FIPs code, yyyy = data year, Q = quarter, T = data type 1 (inpatient), 2 (ambulatory surgery), 3 (emergency department), and nnnnnnn = record counter)*	*Assigned during processing. QKEY does not include a hospital identifier.*
adate	Admission date	6	YYMMDD	Date of admission	Confidential data development file
ddate	Discharge date	6	YYMMDD	Date of discharge	Confidential data development file
medinc	Median household income for patient ZIP Code	6	6(n)	Median Income amount	Confidential data development file
pnum	Person number	5	9(n)	Person number	Confidential data development file
zip	Patient ZIP Code (actual)	$5	5(a)	HCUP definitions	Confidential data development file

CDETAIL DATA ELEMENTS

Variable Names	Description	Def	Value	Value label	Comments
qkey	*Record identifier*	*8*	*15(n)*	*Record identifier (coded as SSyyyyQTnnnnnnn, where SS = State FIPs code, yyyy = data year, Q = quarter, T = data type 1 (inpatient), 2 (ambulatory surgery), 3 (emergency department), and nnnnnnn = record counter)*	*Assigned during HCUP processing. qkey does not include a hospital identifier.*
charge	Detailed charges for revenue code (as received from source)	8	7(n).nn	Charges 1-15	
cpthcpcs	Line item CPT-4/HCPCS procedure code (as received from source)	5	5(a)	CPT or HCPCS Procedure Code	
cptmod1	Line item CPT-4/HCPCS first modifier (as received from source)	2	2(a)	CPT Modifier	
cptmod2	Line item CPT-4/HCPCS second modifier (as received from source)	2	2(a)	CPT Modifier	
rate	Line item rate (as received from source)	8	4(n).nn	Dollars	
revcode	Line item revenue code	$4	4(a)	Revenue Codes 1-15	
servday	Line item days from admission date	8	4(n)	Days	Line item detail. Days from admission date (SERVDAY) is calculated by subtracting the line item service date provided by the data source from the admission date. If SERVDAY is less than -4 or greater than LOS + 3, then set it to invalid (.).