



# REPORT ON THE FINAL DATA BASE

*Contract #270-96-0007*

## **Volume I: Technical Report**

November 2, 1999

*Submitted to:*

Project Officer: Joan Dilonardo  
SAMHSA Services Administration  
Center for Substance Abuse Treatment (CSAT)  
Rockwall II Building, Suite 740  
5600 Fishers Lane  
Rockville, MD 20852

*Submitted by:*

The MEDSTAT Group, Inc.  
5425 Hollister Avenue  
Suite 140  
Santa Barbara, CA 93111

Unless disclosure is required by the Freedom of Information act 5 U.S.C. 552, as amended, this proposal or quotation includes data that shall not be disclosed outside the Government and shall not be duplicated, used or disclosed - in whole or in part -- for any purpose other than to evaluate this proposal or quotation. If, however, a contract is awarded to this offeror or quoter as a result of -- or in connection with -- the submission of this data, the Government shall have the right to duplicate, use, or disclose the data to the extent provided in the resulting contract. This restriction does not limit the Government's right to use information contained in this data if it is obtained from another source without restriction. The data subject to this restriction are contained in sheets all.

**TABLE OF CONTENTS**

Report on the Final Data Base ..... i

Chapter One: Overview ..... 1

    Importance of Linking Data Bases in State Systems ..... 1

    General Content and Structure of the IDB ..... 2

    Discussion of the Data Base and Data Quality ..... 4

Chapter Two: Data Base Design and Layout..... 7

    Data Sources ..... 7

    Population..... 7

    Overview of Processing Steps ..... 8

    Standardizing Variables and Common Crosswalks..... 9

    Linking ..... 10

Chapter Three: Delaware Processing Documentation ..... 15

    Delaware Medicaid..... 15

    Delaware Mental Health/AOD Data..... 26

Chapter Four: Oklahoma Processing Documentation..... 37

    Oklahoma Medicaid ..... 37

    Oklahoma Mental Health..... 46

Chapter Five: Washington Processing Documentation..... 51

    Overview ..... 51

    Washington Medicaid Data ..... 51

    Mental Health Data..... 56

    Department of Alcohol and Substance Abuse (DASA)..... 65

Chapter Six: Person Summary File..... 73

    Development..... 73

Chapter Seven: Potential Limitations and Tentative Plans for Improvement..... 76

    Overview ..... 76

    Specific Limitations with Respect to Data and Data Base Design ..... 76

    Data Base Design Recommendations and Enhancements..... 79

Chapter Eight: Discussion of Resource Requirements and Utilization..... 81

    Staff Resources ..... 81

    Software..... 82

    Hardware ..... 82

    Opportunities for Efficiencies ..... 83

Chapter Nine: Description of Documentation ..... 85

    Overview ..... 85

    Structure ..... 86

## CHAPTER ONE: OVERVIEW

This document provides a detailed description of the various activities in the development of a data base that integrates mental health services, drug and alcohol services and Medicaid data for three States (Delaware, Oklahoma and Washington). These activities were carried out as part of SAMHSA Contract #270-96-0007. The report is divided into three major sections:

This volume (Volume 1) describes the general concept of the data base and the detailed design including extensive file definitions and information, flow charts and crosswalks. Separate sections detail the State-specific preprocessing activities, and other aspects of the data base. These are intended to help the researcher understand the approach we took in processing data for each State and data source. Since source data received for this project varied in terms of layout and content, each was processed separately according to the contents of the file. This volume also contains appendixes with detailed information on the selection of the population and contents of the final files.

Volume 2 contains all source code used in processing the data, including programs used to load State data into SAS and routines for linking person-level data. These programs can be viewed in any text editor (such as Notepad, BBEdit, or Emacs). This code is primarily of interest to programmers and other parties interested in how files were processed and constructed. Also included in Volume 2 is the users guide, which provides users of the data base with detailed technical information regarding the structure and composition of the data base. Electronic versions of the data dictionary with State-specific documentation are also included as part of this volume.

Volume 3 includes the actual data bases (on CD-ROM), and a brief description of the contents of each file. The person summary files are included on separate CD-ROMs in Volume 3.

### IMPORTANCE OF LINKING DATA BASES IN STATE SYSTEMS

In many States, mental health services (MH), alcohol and other drug abuse services (AOD), and Medicaid services are all operated by different State agencies. Where these services are offered by the same agency, they are often supported with separate information systems. Traditionally, managers and planners in any one State agency do not know much about the services that their clients are receiving from the other two agencies. The development of this Integrated Data Base (IDB) allows for the first time the identification of how clients use services in each of the three systems of care. By linking physical health and MH/AOD services provided by Medicaid with the specialized MH/AOD services provided by MH and by AOD agencies, a more complete picture of the need for services for persons with MH and/or AOD problems can be developed and studied.

The data base that is now in place provides a solid foundation for comprehensive data integration and systematic updating over time. This resource can be used to address questions about the use, cost, and outcomes of publicly funded mental health, alcohol, and other drug treatment, and Medicaid related services at the State level. Data integration across years will enable researchers to track trends and anticipate the need for system changes and changes in State legislation. Already with the 1996 integrated MH/AOD data base it is possible to do comparative cost analyses of populations with and without overlapping MH/AOD problems. Analyses such as these can help provide a truer estimation of client costs for comprehensive services.

*A linked data base increases credibility of findings.* Unlike traditional outcome studies that rely on self-report, with an IDB policy makers can rely on the information directly derived from administrative data bases that provide actual counts and costs in regard to specific types of services and dollars spent. With an IDB that spans multiple years, it is easy to document that clients who used mental health or AOD services

cost fewer Medicaid dollars in healthcare after receiving MH/AOD treatment than they did before treatment.

*The IDB facilitates identification of underserved and high-cost clients.* An IDB makes it possible to identify mutual clients of the MH, AOD, and Medicaid systems and, more specifically, to identify high-cost clients. This kind of information is valuable in making policy decisions about which populations to target for service and how to deal with these clients. It also provides the necessary information for changing State legislation.

*The IDB reduces the cost of MH/AOD treatment systems research.* Studies can be performed at lower cost because researchers can analyze existing information rather than going through the costly process of following a client over time.

*The IDB overcomes some of the methodological barriers to outcome studies.* The IDB provides access to pre-treatment and post-treatment data on the same individuals, allowing measurements of actual differences in costs and utilization prior to treatment and after treatment. It also minimizes legal and ethical problems that are inherent in the use of control groups in outcome studies.

*The IDB allows for efficient and cost-effective cross-system studies.* Without an IDB, cross-system studies are prohibitively expensive in the costs of data retrieval from multiple systems and in the complexities of client follow-up.

What can be done with one year of data is limited compared to what can be done with multi-year data, and the intent of this project is to continue collecting and assembling data. Development of the multi-year prototype will allow for the tracking of trends within and across the mental health, AOD, Medicaid, and eventually criminal justice, child welfare, employment, and other systems.

## **GENERAL CONTENT AND STRUCTURE OF THE IDB**

The IDB has been designed to provide researchers with an opportunity to access a comprehensive source of mental health and substance abuse utilization and expenditures. In constructing the data base, the mental health and alcohol and other drugs (MH/AOD) population was defined as broadly as possible, based on diagnosis, procedure codes, service types and revenue codes. In addition, the data base provides an opportunity to examine all medical services provided by Medicaid to the selected population. This capability enables researchers to better estimate the true cost of serving the MH/AOD population.

The current data base reflects the third iteration of data base design and development for this project, the first iteration being the technical architecture and the second being the prototype data base. Each of these prior generations are described in previous contract deliverables (Task 24 and Task 25, respectively). Feedback on each generation informed refinement and development of the subsequent iteration, culminating in the final data base described here.

This version of the data base includes data for calendar year 1996 and consists of eleven files. Separate versions of the data bases are being produced for each of the three participating States, and the structure of the data bases is identical across all States. Three of the files contain general client and demographic information. The remaining eight files consist of Medicaid or State agency service and treatment information. These files are:

- **Client Core:** includes one record per unique individual user of any Mental Health Service, Alcohol and Drug Service and Medicaid Service. The file contains limited demographic information, indicators for mental health, AOD diagnoses, and service file information.

- Client Detail: contains detailed client information from MH/AOD agency records describing client attributes and conditions at various points during treatment.
- Medicaid Eligibility: includes multiple records per person based on periods of enrollment in the Medicaid program.
- Service Core: includes multiple service records per user. There is one record for each service, based on the date the service was provided. The 'Client Core' and 'Service Core' Files are designed to contain normalized demographic and service information common to Medicaid data, Mental Health Agency data and Alcohol/Drug Abuse Agency data.
- Medicaid Inpatient Services Detail: includes multiple records per user of acute care inpatient hospital stays based on each admission and discharge. This file contains both utilization and expenditure information for Medicaid services.
- Medicaid Long Term Care: includes multiple records per user of long term care services based on admission, discharge and interim billing dates. This file contains inpatient psychiatric care, skilled nursing, and intermediate care facilities.
- Medicaid Pharmacy: includes multiple utilization and expenditure records per person based on prescribed drugs, one record per prescription filled.
- Medicaid Outpatient/Other Service: includes multiple records for Medicaid services not specifically mentioned above. These services include home health, transportation, durable medical equipment, emergency room, clinics, and physician (including outpatient) services to mention a few. This file includes both utilization and expenditure information and is organized by date of service.
- Community Mental Health File: includes multiple records per user based on services provided by community mental health treatment centers.
- Institutional Mental Health File: includes multiple records per user based on admission into a Mental Health Institution.
- Alcohol and Drug Service File: includes multiple records per user based on utilization of these services.

Each of these files contains an encrypted unique client identification number and a unique service identification number. These identifiers enable researchers and other users of the data base to link across files to perform a variety of data analyses. Specific contents for each file are included in the data dictionary found in Appendix A.

## **Function**

Our goal is to provide researchers with a data base that allows for analysis and reporting at the individual file level, across files, or at an aggregate level with the use of the person summary file.

We see these files as an important step forward in characterizing the co-occurring utilization of mental health services, alcohol and substance abuse services and Medicaid services and the costs associated with the delivery of these services.

Another feature of the data base relates to the manner in which person-level information was matched across data sources and unduplicated where necessary. This allows for more reliable estimates of the number of persons served and the cost per person.

## **DISCUSSION OF THE DATA BASE AND DATA QUALITY**

This section describes the general characteristics of the source data used to create the data base, and elaborates the role of the data base.

Prior to construction of the eleven-file data base, the structure of individual State files was standardized to a uniform format. To keep information on each individual State file consistent, key data elements were identified and State data element values were mapped into uniform values. Every attempt was made to develop crosswalks for the State values in a consistent manner, although this was complicated by the nature of some State data elements and values. Crosswalks were reviewed by State participants to ensure that gross errors were not made. All key element crosswalks are included in Appendix B.

Across the board, we found the data from each State to be of relatively high quality both in terms of completeness and validity. As expected, Medicaid data (both eligibility and claims/encounters) were relatively similar for all three States, whereas MH/AOD data varied from data source to data source. This variation is due chiefly to the common functions and reporting requirements for Medicaid systems, whereas State MH/AOD programs and systems are less constrained. As a result, each State's Medicaid system and resulting data is somewhat standardized, whereas the structure of mental health programs and data varies significantly from State to State. We were aware of this variability during the design phase, and made every effort to include data elements that were common to all States and agencies.

While there were some missing data elements in some files, and missing or out-of-range values for records within a file, there was no significant pattern to the missing data. Although it is difficult to quantify the amount of missing data, our preliminary review of the data indicates that the majority of key data elements are present across all States. We successfully developed and tested crosswalks for key data elements (e.g., category of service, provider type) and these data were assigned valid values for the majority of the observations in the data base. Virtually all of the data quality issues identified during construction of the prototype were resolved through discussion with the States and implemented here. Detailed descriptions of the data quality issues can be found in each State summary.

It is important to note that in the course of developing the IDB, we processed more than 40 files accounting for more than fifty-six million observations. Although it is useful to strive for 100% completeness and validity, any data base of this magnitude and complexity derived from such diverse sources will contain at least some missing and incomplete data.

### **Design Considerations**

The project team patterned the content of the data base from three exemplary sources:

The Treatment Episode Data Set (TEDS) was used as a model for Alcohol/Other Drug data elements. Elements from the minimum data set are included in the IDB, primarily in the Program File and Client Core. Although the TEDS file does not contain true service-level detail to use as a model, we have been successful in capturing service detail where available with the service categories (SERVCAT) variable on the Service Core file.

The Mental Health Statistical Improvement Program (MHSIP) Guidelines for Mental Health were used as the model for MH minimum data bases. No corresponding data set for Mental Health, such as the TEDS data set for AOD, exists, however we have included much of the information recommended by the guidelines. In many instances, the information from MHSIP is not well defined, but we have used it as our basis to determine the definitions on the mental health portion of the data set.

The Statistical Medicaid Research Files (SMRF) were created as part of the HCFA MSIS system. Although we were unable to use actual SMRF data, we were able to pattern the Medicaid portion of the data base after the SMRF file layouts for each file type. Consistent with SMRF files, we have created Pharmacy, Inpatient, Other and Eligibility files for each State. In addition, we also created SMRF uniform values for Category of Service, Provider Type and Place of Service.

## **Content**

As mentioned previously, the major function of this data base is to provide an analytic resource for information about services utilization and cost. To this end, the data base serves as an operational data store to assist in conducting basic and applied research. Data have been preserved at the most detailed level possible, while being transformed to standardized values. Where appropriate, “flags” have been added to indicate the presence or absence of certain logical conditions (e.g., under age 18, diagnoses indicate schizophrenia). A full description of the data base can be found in Chapter 2, and State-specific processing details are elaborated in Chapters 3, 4, and 5. Selected components of the data base are described below.

### ***Person Summary File***

During the design process for the prototype data base, we decided to preserve source data (e.g., claims and eligibility information) at the most detailed level possible. In doing so, our goal was to allow for a wide range of potential views and uses of the data base. A potential drawback of this approach is that some of the more common types of analyses may require manipulation of the data structures prior to any programming required for analytical processing.

As a generic example, information regarding outpatient claims is stored in one file whereas inpatient claims and demographic data are stored in two other files. In order to obtain information about persons who have both inpatient and outpatient encounters, data from each file must be merged prior to analysis. Although this may seem inefficient, it is actually more efficient than creating either a single (very large) file or a series of files that include all possible combinations of possible merges (e.g., one summary for inpatient and outpatient, another for inpatient, outpatient and pharmacy).

However, we also realize that some researchers may prefer to work with the data at an aggregate level as a means of expediting analysis. In order to demonstrate such analyses, we developed a person summary file. This file contains a sample of data elements comprising a collection of commonly accessed data elements, aggregated and summarized by person.

After processing of the detailed State files was completed, a summary person-level file was created for each State. This file summarizes service utilization and expenditures for each client and presents the most commonly referenced demographic data including age, gender, and various other client attributes. This file is discussed in more detail in Chapter 6 of this report.

### ***Assignment of Costs***

As a step toward determining the true costs of MH and AOD services, an attempt was made to assign costs to the services provided by both the Medicaid and non-Medicaid State agencies. To capture this information, the first step was to request information from the agencies on reimbursement rates with the somewhat simplistic idea that the given rates could be directly applied according to procedure code or type of treatment provided. We also requested their assistance developing a reasonable method to assign or impute costs for the services. We defined ‘cost’ as the amount burdened by the agency (or other entity) that actually paid for the service. If both charges and payments were available, we requested the payment amount as the appropriate measure of cost.

Although each agency complied with our request, the issue of applying the rates turned out to be far more complicated than first envisioned. For each State, it was necessary to develop individual algorithms for each specific treatment within each general type of service. For instance, within a hospital stay, it was sometimes necessary to impute costs by determining the number of days of care and then applying a per diem rate. For other hospital stays, we may have been given a total amount at the client level. With some types of stays it was necessary to apply weighted fiscal year costs per bed. In some instances the provider's tax identification number determined the application of costs.

In other instances, the amount of missing information was determined to be such that a cost estimate either could not be developed, or any estimate computed would be unusable for research or other purposes. Rates were not applied in these situations, and cost estimates were set to missing.

Although every attempt was made to apply the rates in a reasonable and accurate manner, researchers must be aware of the complexities of applying costs to services. It must be noted that the costs applied to MH/AOD State agency services are not true costs and cannot be treated as such. These costs should be treated as estimates or relative costs, and used only as an approximate measure of what the true costs are.

It should also be noted that we have made no attempt to unduplicate services or payments for services provided by Medicaid and Non-Medicaid State agencies. Therefore MH/AOD State Agency payments cannot be added with Medicaid payments for a total cost of service. However, sufficient information is present to allow researchers so inclined to perform this unduplication if their research goals require it.

Detailed cost information and application of rates to MH/AOD services can be found within each State specific chapter in this report. We encourage the user of the data to read the material in the sections thoroughly to have a better understanding of how the rates were applied.



## CHAPTER TWO: DATA BASE DESIGN AND LAYOUT

### DATA SOURCES

The data bases link Medicaid claims data with MH and AOD data for each of the three States in the IDB: Delaware, Oklahoma, and Washington. Medicaid data consists of claims, encounter, and eligibility data obtained from the Medicaid agency in each State. This Medicaid data was combined with MH/AOD treatment data from the appropriate State agencies along with population demographics. The MH/AOD program administration is organized differently in each of the three States. In Oklahoma, one agency - the Oklahoma Department of Mental Health and Substance Abuse Services - administers both MH and AOD programs. Washington uses two separate agencies, the State Mental Health Agency and the Department of Alcohol and Substance Abuse, for mental health services and AOD services respectively. In contrast, Delaware divides responsibility between one agency that serves adults and another that serves children and adolescents. Both agencies administer MH/AOD services to their respective populations. Further discussion of source data can be found in Chapter Three, Four and Five.

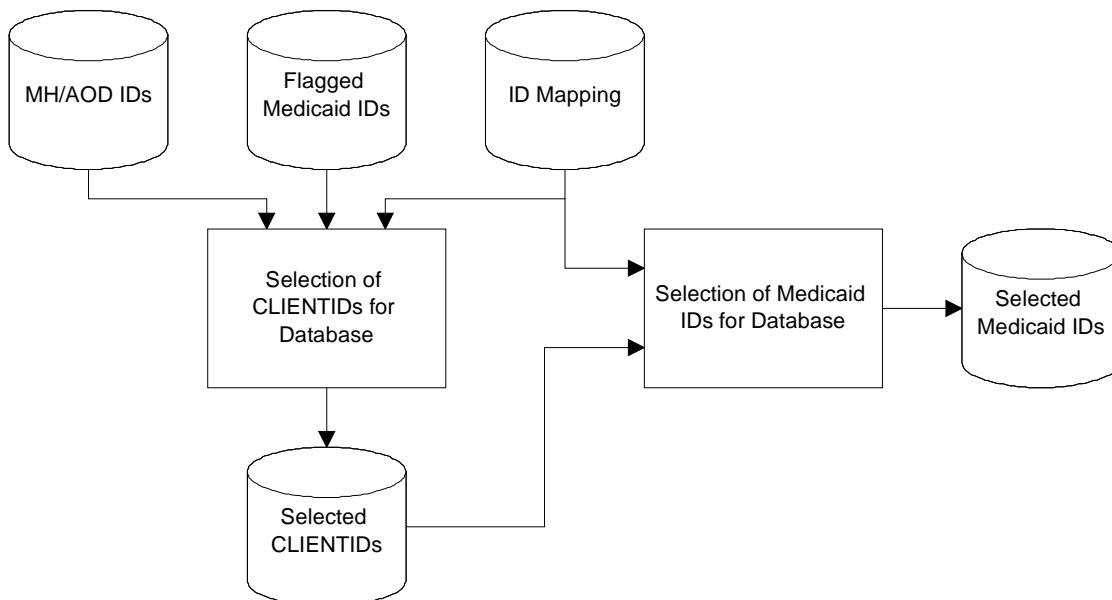
### POPULATION

The data base consists of all clients for whom at least one of the following conditions is true:

- They received at least one Medicaid service that was classified as an MH/AOD service.
- They received treatment, or were admitted into a treatment program, with one of the MH/AOD agencies.

In the discussion that follows, the SAS computer programs that were used to transform the data and build the IDB are referred to by number. The complete source code for these programs can be found in Volume 2 of this report. Program 400 identifies all the IDs for clients included in the data base according to the inclusion criteria in Appendix B. All available services for these clients, including routine medical, are then collected into the data base. Figure 1 depicts the data flow used in selecting the data base population.

**Figure 1 Selection of Data Base Population**



Recipients of Medicaid services are selected for inclusion in the data base if they received any services considered MH/AOD. This determination is made by examining:

- Provider type
- Type of service
- MH/AOD detailed revenue center code
- MH/AOD procedure codes (including State specific codes)
- MH/AOD diagnosis codes

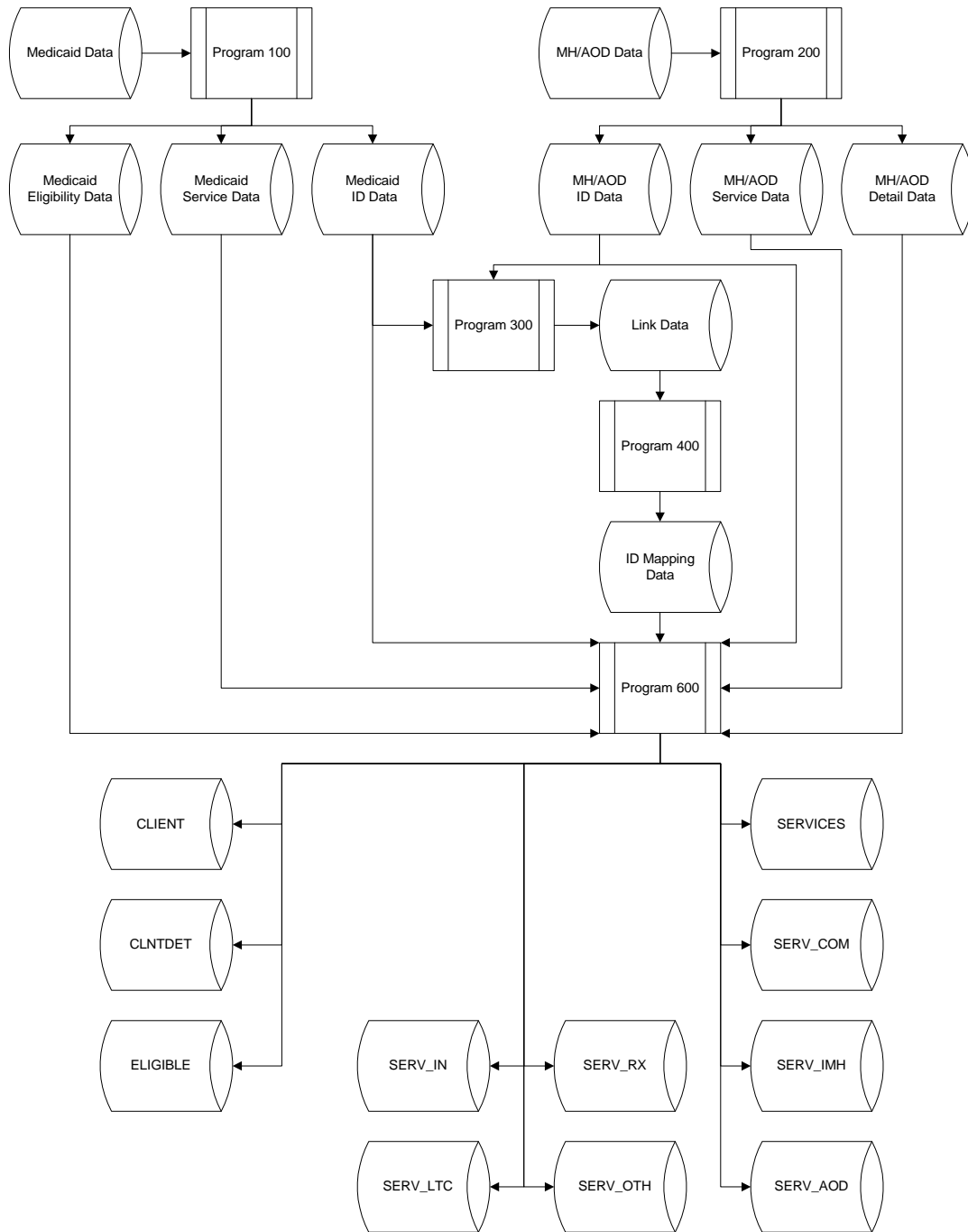
### **OVERVIEW OF PROCESSING STEPS**

Creation of the IDB proceeds in five separate, but related stages.

- Initial processing of Medicaid data, including SAS loads and variable recodes (Program 100).
- Initial processing of MH/AOD data, including SAS loads and variable recodes (Program 200).
- Linking of client IDs (Program 300).
- Selection of data base population and final mapping of client IDs (Program 400).
- Creation of the data base files (Program 600).

All programs are divided into discrete "steps", designed to perform specific tasks. The Users Guide provides additional detail. Figure 2 presents the flow of data graphically.

**Figure 2 Overview of IDB Processing Flow**



**STANDARDIZING VARIABLES AND COMMON CROSSWALKS**

In order to facilitate the linkage and analysis of records from disparate systems, identifying variables are standardized. For some variables, standardization consists of cosmetic changes such as changing all name variables to uppercase characters or removing hyphens from Social Security Numbers. Other variables require transformation of values, applied through the use of value crosswalks. A uniform set of crosswalks was implemented for common linking variables including the race and gender variables. This sometimes results in a loss of information. For example, data for some States includes separate variables

for race and ethnicity. In this case, processing combines information from the two variables into the IDB variable `race`. The loss of information is tolerated since the method brings us closer to our goal of a standardized, IDB. The data dictionary contains details of the uniform values for both race, gender and other identifying variables.

## LINKING

The goal of client linking is to combine person-level records from two or more separate data files. Linking implies that the files represent either the same population or overlapping populations, so that records from one source may be linked to records from another source. In this manner, we have linked client data for Medicaid eligibles with client data from MH/AOD programs. Records from two different files that have been linked are said to be “matched”. In the absence of a universally present deterministic identifier (such as Social Security Number), the most accurate and statistically reliable method of record linking is probabilistic record linking. This method matches records by comparing two records using multiple characteristics. With probabilistic record linking, all appropriate characteristics, such as name, date of birth, IDs, and other demographic information are used in determining matches. Similarity and dissimilarity between all characteristics of the two records are taken into account. No single agreement alone determines a match, nor does any non-agreement preclude a match. Matches are typically determined by a mix of characteristics in agreement and non-agreement. Probabilistic linking quantifies the importance of each comparison characteristic and weights each comparison accordingly by calculating the relative probabilities that a characteristic agrees for the matched and unmatched groups.

For instance, two records may disagree on last name and zip code, but agree on first name, gender, and date of birth. Linking routines weigh the relative importance of each comparison and compare the scores for this pair of records to those in the “matched” and “unmatched” populations. That they agree on gender is generally less informative than agreement on date of birth. If the overall pattern of agreement for this pair of records is more similar to the matched population than the unmatched population, the records are matched.

We have employed probabilistic linking as the technique for linking the Medicaid and MH/AOD client records. The specific methodology can be summarized as follows:

- The data files to be linked are concatenated to create a single data set.
- Factors, based on relative frequencies, are calculated from the concatenated data to scale weights developed in later steps.
- The concatenated data set is joined back onto itself (self-joined) to conceptually form the set of all possible pairs. In practice, to reduce the size of the resulting data set, only joined pairs with conditions suggesting a possible link are kept. This process is referred to as “blocking”. The constraints imposed to perform blocking are:
  - ◆ Social Security Numbers agree,
  - ◆ Phonetic last names agree,
  - ◆ Phonetic first names, date of birth, and gender agree, and
  - ◆ Phonetic names, date of birth, and gender agree.
- Comparisons of the joined data on identifying characteristics are made.
  - ◆ Exact (0 or 1) – comparisons for race, gender, and middle initial (name).
  - ◆ Relative comparisons (a continuous value on the range 0 to 1):

- First and last names – approximate string matching,
  - IDs – bigram,
  - Date of birth – relative difference in days, and
  - ZIP code – relative distance in miles between the centroids of two ZIP codes.
- Initial deterministic link to assess which client observations match. There are five sets of conditions that could result in a deterministic link, as follows.
  - ◆ There are exact matches on Social Security Number, Medicaid ID, date of birth, and gender.
  - ◆ There are exact matches on Social Security Number, date of birth, and gender, and one of the following is true:
    - There is a first name match of at least 80%,
    - There is a last name match of at least 90%, or
    - There is an exact match on middle initial.
  - ◆ There is a first name match of at least 80%, a last name match of at least 90%, an exact match on date of birth and gender, and one of the following is true:
    - There is an exact match on ZIP code, or
    - There is an exact match on race.
  - ◆ There is a first name match of at least 80%, a last name match of at least 90%, an exact match on date of birth, and one of the following is true:
    - There is a Social Security Number match of at least 90%, or
    - There is a Medicaid ID match of at least 90%.
  - ◆ There is a first name match of at least 80%, a last name match of at least 90%, an exact match on date of birth and middle initial.
- Comparison weights are calculated based upon the initial matches.
  - ◆ Weights: log of relative probabilities of characteristics for the matched and unmatched groups
  - ◆ Weights for both agreement and non-agreements
- Scores for each pair of joined observations are calculated on the joined data using the weights calculated in the prior step - adjusted with the scaling factors developed at the start of the process. Scores are calculated as the sum of the weight-comparison value products.
- An iterative process to refine weights and the matched population follows.
  - ◆ A minimum score for matched records based upon distribution of scores for matched and unmatched groups is calculated.
  - ◆ Scores are compared to the minimum, and a new matched/unmatched determination is made.
  - ◆ Weights are recalculated.
- An optional final step is to create a third population of uncertain matches which qualify for manual review.

## Confidentiality

Incorporated in the linking process is the assignment of a unique MEDSTAT client ID. Medicaid and MH/AOD agency IDs are mapped to this new client ID. A file with this mapping information is used in constructing the data base, but the ID file is not included with the data base. The file with the mapping information is stored separately from the data base and is the only method of converting from the MEDSTAT client ID to the original IDs. There are no formulas or functions that can convert the client ID back to the original source IDs.

## File type definitions

The IDBs consist of eleven (11) SAS data sets, connected through common client ID and service ID "numbers." The data base is divided into data sets with client information and data sets with service, or treatment, information. The client data sets are:

Client Core (CLIENT) – which contains general identifying variables such as race, gender, and birth year and month, and also contains client level MH/AOD flags. The data set primary key is client ID (CLIENTID); it is unique and there is one observation per CLIENTID. See the data dictionary for a complete list and description of variables.

Client Detail (CLNTDET) – this file contains detailed client variables from MH/AOD agency records describing client conditions at various points during treatment. The data set key is a combination of CLIENTID and date of update (UP\_DATE). There is one observation per key, and one or more observations per CLIENTID. The CLNTDET data set contains observations only for clients receiving treatment from the State MH/AOD agency. See the data dictionary for a complete list and description of variables.

Medicaid Eligibility (ELIGIBLE) – which contains information on Medicaid eligibility periods, including managed care information. The data set key is a combination of CLIENTID and beginning eligibility date (ELIG\_BEG). There is one observation per key, and one or more observations per CLIENTID. The ELIGIBLE data set contains observations only for clients participating in the Medicaid program. See the data dictionary for a complete list and description of variables.

Client data sets are connected through the CLIENTID variable. There is a one-to-many relationship between both:

- CLIENT and CLNTDET, and
- CLIENT and ELIGIBLE.

In other words, each client can potentially have several related records in the CLNTDET and ELIGIBLE files, indicating multiple periods of Medicaid eligibility and/or multiple episodes of involvement with State agencies.

Client data are connected to the service, or treatment, data through a one-to-many relationship between the client core and a service core data set (described below) through CLIENTID. The service data sets consist of a core service data set and seven service detail data sets representing distinct types of services. The service data sets are:

- Service Core (SERVICES) – which contains general service information describing the date and type, or category of service. All services/treatments have a corresponding observation in

the SERVICES data set. The data set key is a combination of CLIENTID and service ID (SERV\_ID). There is one observation per key, and one or more observations per CLIENTID. See the data dictionary for a complete list and description of variables.

- Medicaid Inpatient Service Detail (SERV\_IN) – which contains detailed information for Medicaid inpatient claims. The data set key is a combination of CLIENTID and SERV\_ID. There is one observation per key, and one or more observations per CLIENTID. See the data dictionary for a complete list and description of variables.
- Medicaid Long Term Care Service Detail (SERV\_LTC) – which contains detailed information for Medicaid long term care claims, including claims for inpatient psychological services. The data set key is a combination of CLIENTID and SERV\_ID. There is one observation per key, and one or more observations per CLIENTID. See the data dictionary for a complete list and description of variables.
- Medicaid Pharmacy Service Detail (SERV\_RX) – this file contains detailed information for Medicaid pharmacy claims, including, but not limited to prescription drugs. The data set key is a combination of CLIENTID and SERV\_ID. There is one observation per key, and one or more observations per CLIENTID. See the data dictionary for a complete list and description of variables.
- Medicaid Other/Outpatient Service Detail (SERV\_OTH) – which contains detailed information for Medicaid outpatient claims and all other Medicaid claims. The data set key is a combination of CLIENTID and SERV\_ID; there is one observation per key, and one or more observations per CLIENTID. See the data dictionary for a complete list and description of variables.
- Community Mental Health Treatment Detail (SERV\_COM) – which contains information for MH/AOD agency outpatient/community mental health treatments. The data set key is a combination of CLIENTID and SERV\_ID; there is one observation per key, and one or more observations per CLIENTID. See the data dictionary for a complete list and description of variables.
- Institutional Mental Health Treatment Detail (SERV\_IMH) – which contains information for MH/AOD agency institutional mental health treatments. The data set key is a combination of CLIENTID and SERV\_ID; there is one observation per key, and one or more observations per CLIENTID. See the data dictionary for a complete list and description of variables.
- Alcohol/Other Drug Treatment Detail (SERV\_AOD) – contains information for MH/AOD agency alcohol/other drug treatments, including detoxification treatments. The data set key is a combination of CLIENTID and SERV\_ID; there is one observation per key, and one or more observations per CLIENTID. See the data dictionary for a complete list and description of variables.

The service detail data sets are connected to the service core data set through CLIENTID and a service SERV\_ID in one-to-one relationships.

Observation counts, by State and type of file are shown in Table 1. A graphical representation of the eleven data sets in the IDB is shown in Figure 2.

**Table 1 Data Base Observation Counts by Data Set**

<b>Data Set</b>	<b>Delaware</b>	<b>Oklahoma</b>	<b>Washington</b>
CLIENT	27,594	157,502	123,180
CLNTDET	22,203	321,509	1,697,975
ELIGIBLE	27,305	643,485	100,429
SERVICES	1,739,494	6,023,613	10,938,661
SERV_IN	6,998	22,017	36,013
SERV_LTC	6,908	78,816	54,061
SERV_RX	342,122	955,477	2,250,408
SERV_OTH	1,004,461	2,722,921	5,785,222
SERV_COM	316,631	1,566,461	2,176,678
SERV_IMH	3,095	351,467	11,915
SERV_AOD	59,279	326,454	624,364



## **CHAPTER THREE: DELAWARE PROCESSING DOCUMENTATION**

Data processing for Delaware involved obtaining data from the State Medicaid agency, and obtaining MH/AOD data from the two agencies that administer children/youth programs and adult programs. We received and processed two Medicaid claims files, one Medicaid eligibility file, and twelve MH/AOD files from various regions and agencies. The twelve files consisted of data from the Division of Child and Youth Services and The Division of Alcoholism, Drug Abuse and Mental Health (DADMH). In order to reduce the burden on the State, MEDSTAT agreed to accept DADMH files provided by individual sources that included the following:

- Community mental health files (client and service),
- County mental health files for New Castle (client and service),
- County mental health files which combined data for Kent and Sussex counties (client and service),
- Psychiatric Hospital file, and
- AOD files (three fiscal years)

The nature and content of each of these files is discussed in the section on Delaware Mental Health/AOD Data, and processing details for each data source are described below. Because Medicaid FFS claims and encounter records were processed separately, these file types are described independently.

### **DELAWARE MEDICAID**

#### **Data Sources – Claims and Encounters**

Fee-for-service (FFS) claims and encounter records were received from Delaware's fiscal intermediary, Electronic Data Systems (EDS). Each file (i.e., FFS claims, encounters) was stored in a separate data base with different file layouts and formats. Both files consisted of variable length records with a single header record and repeating detail information. No additional data were requested from Delaware since the State had sent full date of service files for the prototype data base.

**Medicaid Claims**

The claims data base contained 16 record types with information specific to each claim type as shown Table 2.

**Table 2 Delaware Claim Counts by Claim Type**

<b>Claim Number</b>	<b>Fee-For-Service Claim Type</b>	<b>Number Of Records</b>
01	Drug	1,044,286
02	Adjusted Drug	3,781
03	Medical	1,017,069
04	Dental*	0
05	Screen	506,894
06	Professional, Crossover	168,430
07	Transportation	234,491
08	Vision	844
10	Professional, Adjusted	22,042
11	Inpatient	41,216
12	Nursing Home	948
13	Inpatient Crossover	3,518
14	Nursing Home Crossover	38,777
15	Inpatient Adjustment	9,030
16	Outpatient	148,651
17	Outpatient Crossover	19,489
19	Outpatient Adjustment	3,345

Claim types 09 and 18 were not used according to documentation received from Delaware.

Although there were no claim type 05 (Dental) on the file, there were a limited number of dental clinics and dental claims on the medical claim type. According to State contacts and EDS, Delaware has very few dental providers and the State is working to add more dentists to their providers. For 1996, children under the age of 18 received only limited coverage for dental services.

Our final file contained fewer than 2,000 claims for dental clinics and only 25 claims for individual dentists. Both State and EDS sources said they were not surprised at the small number of dental claims.

According to the State, all claims were fully adjusted. Their method of adjustment is to replace the original claim with a revised claim that contains the correct amount if an adjustment is necessary. We were advised to delete all voided claims and all denied claims.

FFS claims were transformed to uniform Medicaid service files using the crosswalks illustrated in Table 3.

**Table 3 Delaware Claim Counts by Claim Type and Provider Type**

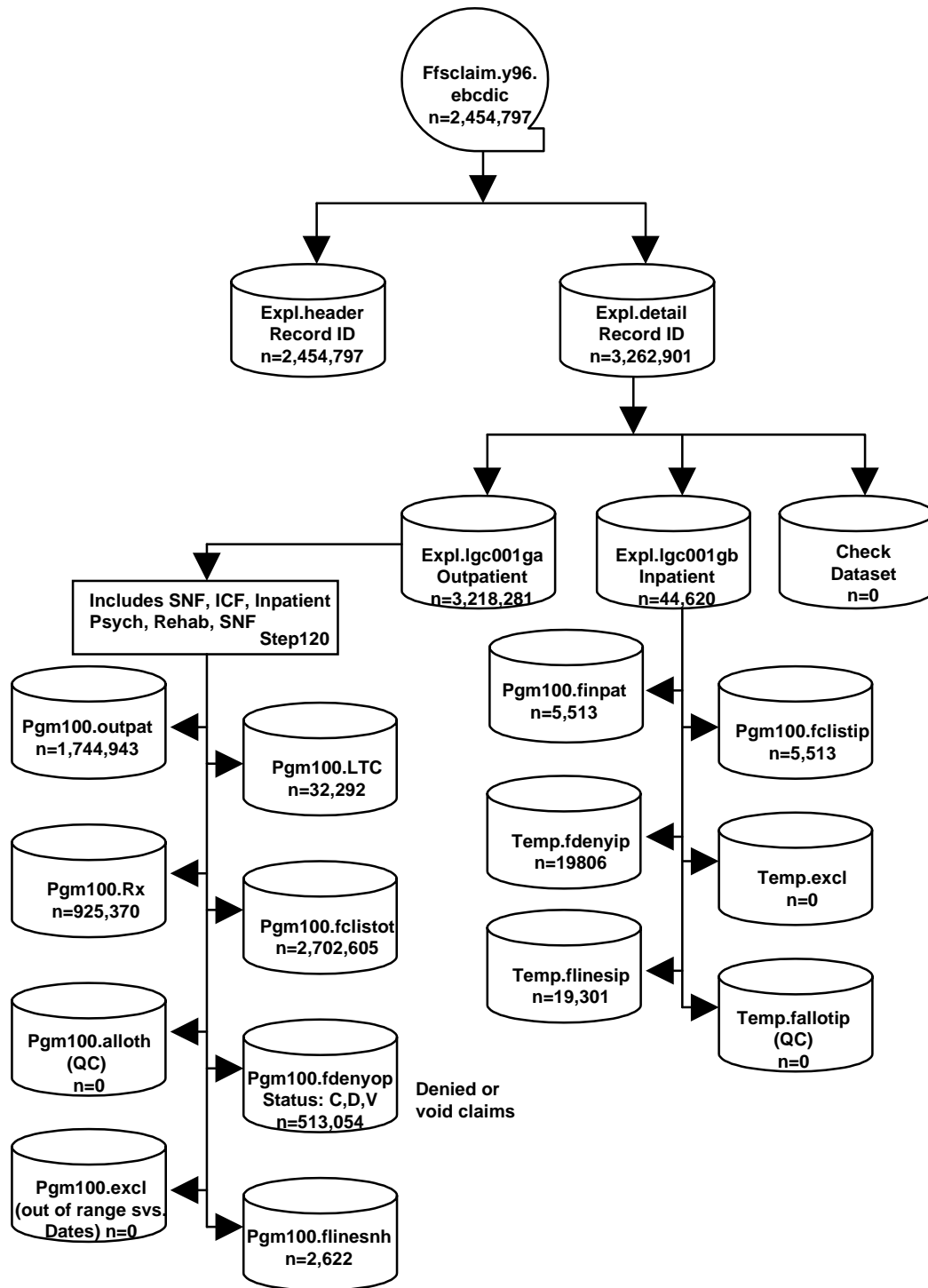
<b>File Type</b>	<b>Claim Type</b>	<b>Provider Category of Service</b>
Other Medicaid Services	03 - Medical Services	Any
	04 - Dental	Any
	05 - Screens	Any
	06 - Professional Crossover	Any
	07 - Transportation	Any
	08 - Vision	Any
	10 - Professional Adjustment	Any
	11 - Inpatient	Billing Provider = 06, Outpatient Hospital or Billing Provider = 26, Rehab
	13 - Inpatient Crossover	Billing Provider = 06, Outpatient Hospital or Billing Provider = 26, Rehab
	16 - Outpatient	Any except Billing Provider = 05, Inpatient Services
	17 - Outpatient Crossover	Any
19 - Adjustments	23 - Home Health 25 - Hospice 47 - Outpatient Crossover 48 - Outpatient Hospital 76 - Outpatient Clinical Lab	
Inpatient	11 - Inpatient	Any NEC
	13 - Inpatient Crossover	29 - Inpatient Crossover
	15 - Inpatient Adjustment*	Any NEC
	16 - Outpatient	Billing Provider = 05, Inpatient
Long Term Care	11 - Inpatient	15 - Inpatient Psych LT 21 or Billing Provider = 25, Inpatient Psych Services
	12 - Nursing Home	Any
	13 - Inpatient Crossover	15 - Inpatient Psych LT 21 or Billing Provider = 25, Inpatient Psych Services
	14 - Nursing Home Crossover	Any
Pharmacy	01 - Pharmacy Claim Record	20 - Family Planning Drug 50 - Pharmacy
	02 - Drug Adjustment	50 - Pharmacy

At the time of the prototype data base build there were a number of outstanding issues to be resolved with Delaware concerning crosswalks, use of provider type, payment fields and several areas where we felt additional information was needed from the State. Each crosswalk was reviewed by the State; minor adjustments were made and all outstanding issues were resolved.

Preprocessing of the data was relatively straightforward as shown in Figure 3. Additional steps were taken to examine specific types of claims. For clarity, these steps are not included in the preprocessing chart, and are not described below. There were three main steps taken to prepare claims before combining the files with encounter data:

- A total of 2,454,797 claims were loaded into SAS and written out with one line item or detail for each service. Data elements not needed to create the IDB were dropped from the file and initial frequencies and crosstabulations of all variables were performed to check for quality of the data and missing information.
- The SAS loaded claims were split into two files to facilitate processing records with similar characteristics at the same time. Inpatient records were kept in one file; all other records were kept in another.
- Uniform data elements were created, temporary datasets were created for quality control, and service files were saved for further processing. During this phase, a total of 532,860 line item records were deleted with a status indicator of void or denied.

Figure 3 Delaware Medicaid FFS Claims Processing



When preprocessing was complete, a total of 2,708,118 claims were written to the uniform Medicaid service file types in preparation for input to the final processing stream of the IDB. State specific claim types were distributed across the uniform file types as shown in Table 4.

**Table 4 Delaware Medicaid FFS Claim Counts by State Claim Type and IDB File Type**

Claim Number	Fee-for-Service Claim Type	Total	Inpatient	Rx	LTC	Other
01	Drug	924,518	0	924,518	0	0
02	Adjusted Drug	852	0	852	0	0
03	Medical	851,266	0	0	0	851,266
04	Dental*	0	0	0	0	0
05	Screen	439,289	0	0	0	439,289
06	Professional, Crossover	138,236	0	0	0	138,236
07	Transportation	206,308	0	0	0	206,308
08	Vision	536	0	0	0	536
10	Professional, Adjusted	11,912	0	0	0	11,912
11	Inpatient	3,019	2,793	0	182	44
12	Nursing Home	25,082	0	0	25,082	0
13	Inpatient Crossover	2,668	2,668	0	0	0
14	Nursing Home Crossover	541	0	0	541	0
15	Inpatient Adjustment	6,539	52	0	6,487	0
16	Outpatient	79,877	0	0	0	79,877
17	Outpatient Crossover	16,109	0	0	0	16,109
19	Outpatient Adjustment	1,366	0	0	0	1,366

***Application of Costs***

Delaware reports Medicaid payments on its FFS claims at the line item level for all claim types except institutional claims (nursing home, skilled nursing facilities), those reported on UB92 claim forms (hospital, home health) and drug claims (one line item per claim). For UB92 claims and drug claims, payments and billed amounts are reported at the header level.

***Exclusions***

During the course of creating the interim files, certain records were excluded from the service type files. For Medicaid FFS data, a total of 532,860 voided or denied line items were excluded on advice of the State.

A second type of exclusion was the detail line items from acute inpatient records. Early in the design process of the IDB the decision was made to aggregate and extract summary information for acute inpatient records rather than retain information at the line item level. These records containing line item detail are stored separately as an archived data set in the event they are needed at a future date.

The table below details the summary of exclusions:

**Table 5 Delaware Exclusions Types and Counts**

Type of Exclusion	Number
Total denied non-inpatient records	513,054
Total denied/void inpatient records	19,806
Total inpatient line items	19,301
Total excluded records	552,161

**Medicaid Encounter Records**

Delaware Encounter Data consisted of three record (pseudoclaim) types: medical, outpatient services, and inpatient services. There were no pharmacy or long term care services since Delaware does not provide pharmacy or LTC services under Medicaid managed care.

The number of records we received for each of three record types is as follows:

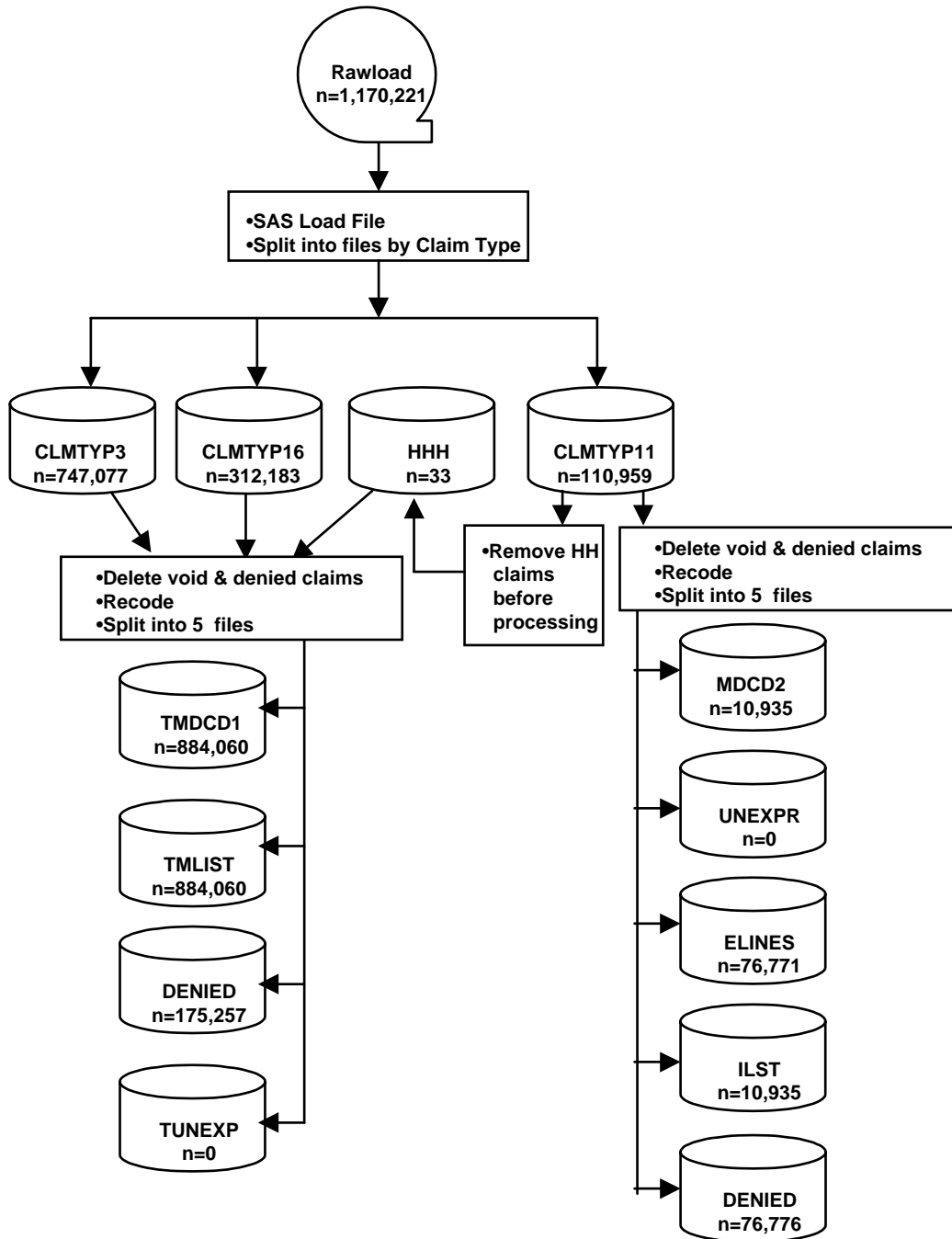
**Table 6 Delaware Medicaid Encounter Records Counts and Types**

<b>Record Type</b>	<b>Encounter Type</b>	<b>Number of Records</b>
03	Medical	747,077
11	Inpatient	110,959
16	Outpatient	312,183

Processing of the data was similar to and consistent with the processing stream for the Medicaid FFS claims data described previously.

- Data representing 1,170,221 encounters were loaded into SAS and all unnecessary data elements were dropped. We performed initial frequency distributions and crosstabulations to check data quality and investigate missing information.
- After sorting by claim type, data were stored into one of three files based on encounter type. Physician service encounters (claim type 3) following the HCFA 1500 format were saved to one file, whereas UB 92-format inpatient encounters (claim type 11) were saved to another file. Institutional outpatient encounters (claim type 16) were stored in a third file.
- For inpatient encounters, void and denied records were deleted and 33 home health (HH) records were removed from the inpatient file prior to recoding and processing. The 33 HH claims that were removed from this file were later added to and processed with the physician and outpatient records. The remaining inpatient encounters were then further divided into five files for quality control and further processing.
- For institutional outpatient encounters, void and denied records were deleted and the remaining claim types ('03' and '16') were merged with the 33 home health records removed from the inpatient data set. Data values were recoded into uniform format and output into four files for quality control and further processing.

**Figure 4 Delaware Medicaid Encounter Processing**



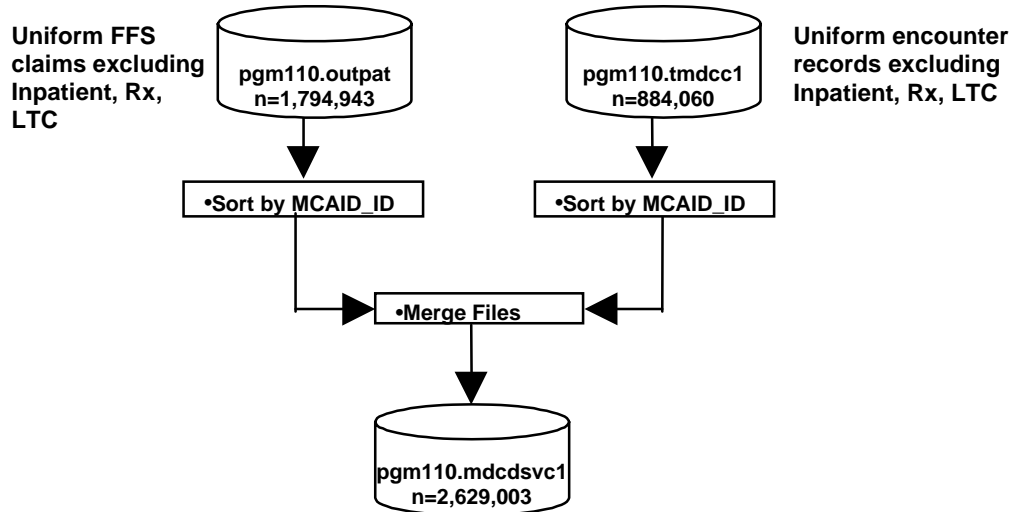
At the completion of interim processing for the FFS claims and encounter data the two datasets were combined for the final processing stream, as shown in the following charts.

***Merge of FFS Other Medicaid Claims with Encounter Other Medicaid Claims***

We sorted the FFS outpatient claims and the encounter outpatient claims by MCAID\_ID then merged the two files together as represented in Figure 5.



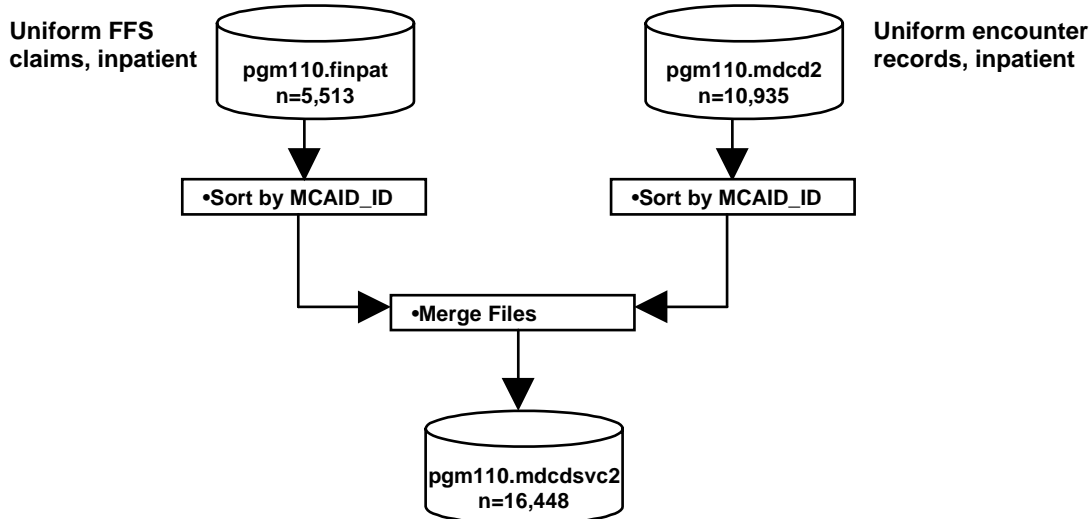
**Figure 5 Delaware Medicaid FFS and Encounter Processing: Other Medicaid Services**



**Merge of FFS and Encounter Inpatient Records**

We sorted the FFS inpatient claims and the encounter inpatient claims by MCAID\_ID then merged the two files together as shown in Figure 6.

**Figure 6 Delaware Medicaid FFS and Encounter Processing: Inpatient Services**



**Anomalies - Claims**

During the course of creating these interim files, several anomalies were observed in the FFS data. They are described below:

Procedure Code Data Element: Approximately fifteen percent of records on the Other Service File contain reimbursement information for monthly capitated services rather than procedure codes. We chose to retain the information, which actually indicates basis of eligibility and level of payment and set the procedure code indicator to '10', Other systems (State Specific). Possible alternatives involved setting the

field to blank or missing. The service category of ‘24’ capitated premium payments and uniform provider type of ‘321’ managed care provider unknown can be used to identify these records. As is common with capitated payments, there is no usable diagnostic information available.

Family Planning Services: A number of claims relating to family planning services were identified in the pharmacy file and the other services. However, based on State-specific category of service codes, no family planning services were identified on the inpatient file.

Difference in contents of FFS File and Encounter Data File: Because the FFS and encounter data contained substantially different data in different formats, it was necessary to use unique combinations of source values to create uniform data elements. Researchers will find it necessary to refer to the individual crosswalks for information regarding these combinations for key data elements.

Incoming source variables: In an attempt to retain State specific information used for key data elements, non-uniform data elements were created containing the State source information (ISERVCAT, IPROVTYP and IPROVSPC). For Delaware, this is somewhat problematic because the source data elements do not contain consistent information. For instance, in some cases ISERVCAT may refer to information from type of bill for an encounter record, or category of service for a FFS claim. The following table demonstrates the source data elements used to derive these key data elements:

**Table 7 Selected Delaware Crosswalks**

		Delaware Source Values		
Key Data Element	Data Element Containing Source Information	Fee For Service Claims	Encounter Records	
			Claim type=3	Claim type=11 or 16
Provider Type (PROVTYPE)	IPROVTYP	Provider Type Code	Provider Type Code	UB92 Type Bill
Service Category (SERVCAT)	ISERVCAT	Category of Service	Provider Type Code	UB92 Type Bill
Provider Specialty (PROVSPEC)	IPROVSPC	Provider Specialty Code	No data element available	

***Anomalies - Encounters***

In many cases, encounter records included missing data elements for key variables such as provider specialty and category of service. Researchers who do not wish to retain encounter records in analyses can identify these records using the FFS Indicator variable (FFS\_IND).

Inpatient Psychiatric Claims. A total of 47 inpatient psychiatric claims were transferred from the inpatient file to the “Other” (outpatient) service file. These were reported on HCFA 1500 claim forms and upon examination of key data elements these claims appear to be physician visits.

Payment Fields: The Medicaid payment fields contain the amount paid to the provider by the MCO or HMO rather than the actual amount paid by Medicaid. These payments can be identified by the Medicaid Payment Indicator (IND\_CAID=A, B OR C) The billing amount (AMT\_CHGD) is reflective of the FFS equivalent.

EPSDT services: Delaware assigned codes were used to identify EPSDT services. State-specific HCPC codes were used to define these services.

Provider IDs for Encounters: The provider identification number found in the header portion of the encounter record contained the MCO identification rather than the actual billing or performing provider ID. Where possible, the ID found in the detail portion of the record was used for billing and service information.

Adjustments: The State advised that the MCOs have their own system of adjusting encounter records. None of the encounter records submitted indicated any form of adjustment.

Assignment of Uniform Provider Type: Advice from the State indicated that provider type must be supplemented with additional information regarding services and type of bill submitted. See Appendix C for further information.

***Medicaid Eligibility***

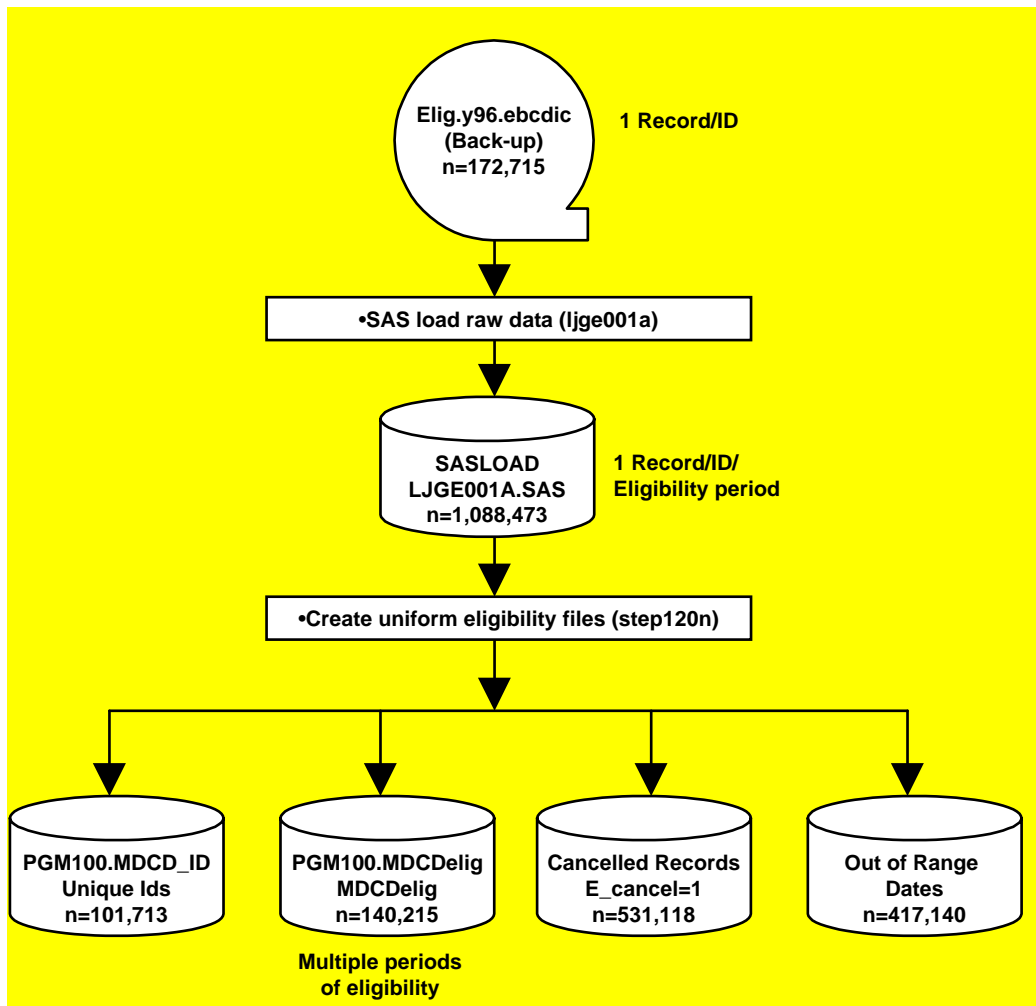
Delaware provided a historical file for the prototype data base and no additional files were requested for the final data base build. All eligibility records with dates outside the range of our data base were deleted from the file.

There was an outstanding question at the time of the prototype build with regard to records with recurring beginning dates of enrollment with multiple ending dates. The State provided us with information that any change in aid category will be reflected with a new beginning date. If the beginning dates have multiple end dates, we were instructed to use the most recent segment and delete anything that follows. A new algorithm was developed that retained correct enrollment periods.

**Table 8 Delaware Medicaid Eligibility Processing Summary**

<b>Description</b>	<b>Number</b>
Records received from Delaware	
One record per person	172,715
One record per enrollment period	1,088,473
Excluded records	
Out of range enrollment dates (not enrolled in 1996)	417,140
Enrollment segments that are no longer valid (only most recent kept)	531,118
Records on Interim File before final population selection	
One record per person	101,713
One record per enrollment period	140,215

**Figure 7 Delaware Eligibility Processing Overview**



**DELAWARE MENTAL HEALTH/AOD DATA**

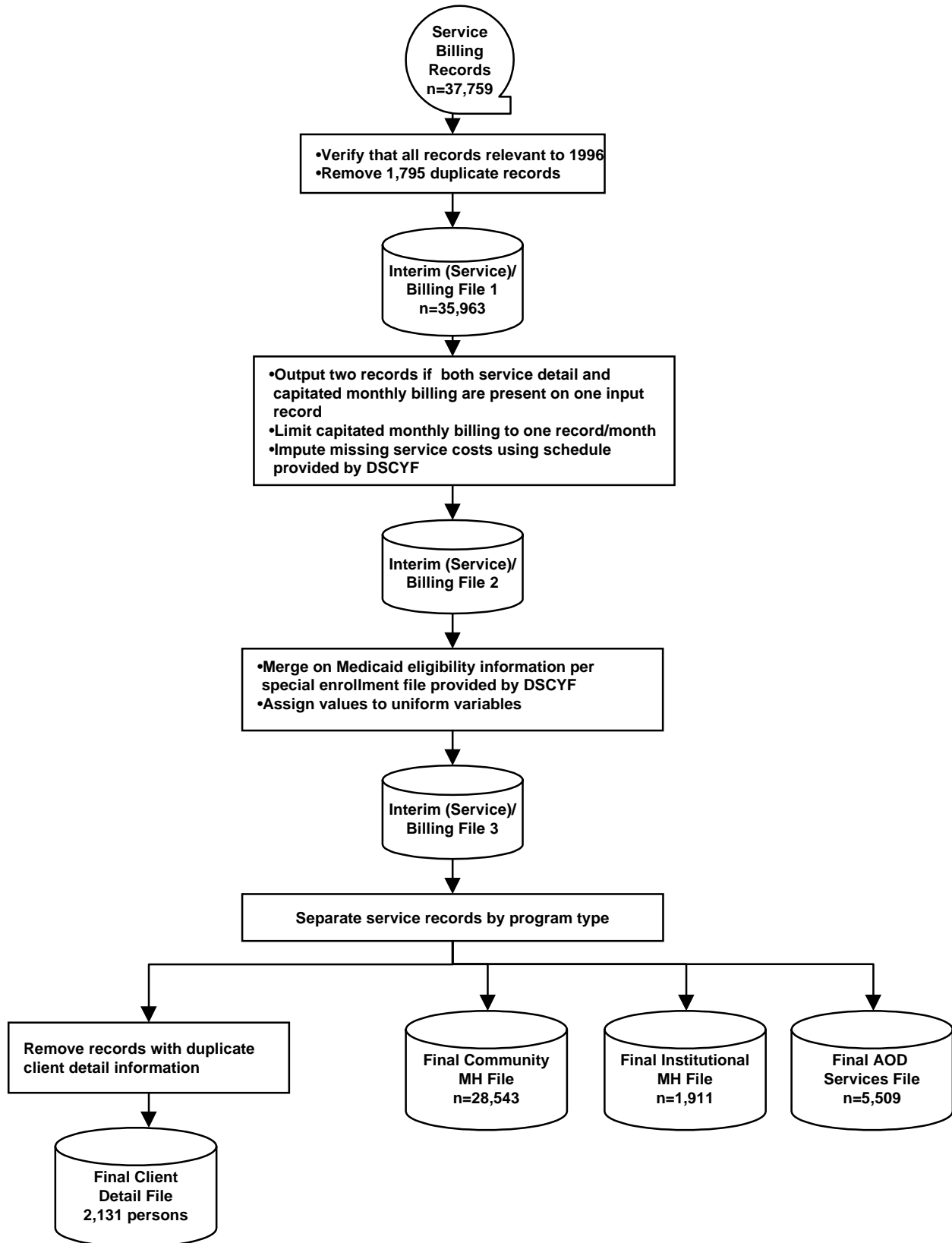
As mentioned previously, Delaware Mental Health/AOD data was submitted by several different agencies that addressed specific populations and service areas. Specific data sources and how they were combined are outlined below.

**Delaware Department of Services for Children, Youth and Their Families**

*Data Source and Processing*

The Department of Services for Children, Youth, and Their Families (DSCYF) created a special file for this project. The information on 1996 services came from the DSCYF Cost Recovery Unit and consisted of demographic, diagnostic, and high-level service data. The data base consisted of 37,758 records representing 2,131 persons served. An ancillary enrollment file was also supplied to allow identification of Medicaid eligibility. Approximately 73 percent of the youth were Medicaid eligible. Nearly 80% of the appropriate records were directed into the final community mental health file, roughly 5 percent were routed to the final institutional mental health file and approximately 15 percent were routed to the final AOD file, based on the program type value. A flowchart detailing the preprocessing of these data is shown in Figure 8.

**Figure 8 Delaware Youth Processing**



### ***Application of Costs***

Cost data was generally available on each record. Where costs were missing (n=1,044) an attempt was made to impute costs by service category, units of service, and provider ID through the use of a cost schedule provided by DSCYF. Such imputation was possible for 944 records (approximately 90% of the records missing costs).

### ***Exceptions***

There were 1,795 duplicate source records that were excluded. No records had to be excluded for being outside of calendar year 1996.

### ***Anomalies***

The available service codes were basically meant to be used for bundled reimbursement calculations and generally were not at the level of specific services that would be most meaningful for researchers. For example, Mental Health Outpatient Services (code=YY751) could include such diverse services as: client recreation, court appearance, emergency home visit, family consultation, individual session, substance abuse evaluation, parent training, and other services. Thus it was often not possible to map such high-level service categories into more specific services categorized in the IDB. Such service-specific detail is available from the DSCYF FACTS system for services after 1996 and should afford better service categorization if data are processed for subsequent years. Mappings of the 1996 DSCYF service codes to IDB program types and service categories can be found in Appendix C.

There were approximately 30 (3%) birth dates that were inconsistent with the youth program. These indicated an age of over 18 at the beginning of 1996. It is unknown if these represent data errors or special situations. In any event, they were left unchanged.

## **Delaware Adult Mental Health**

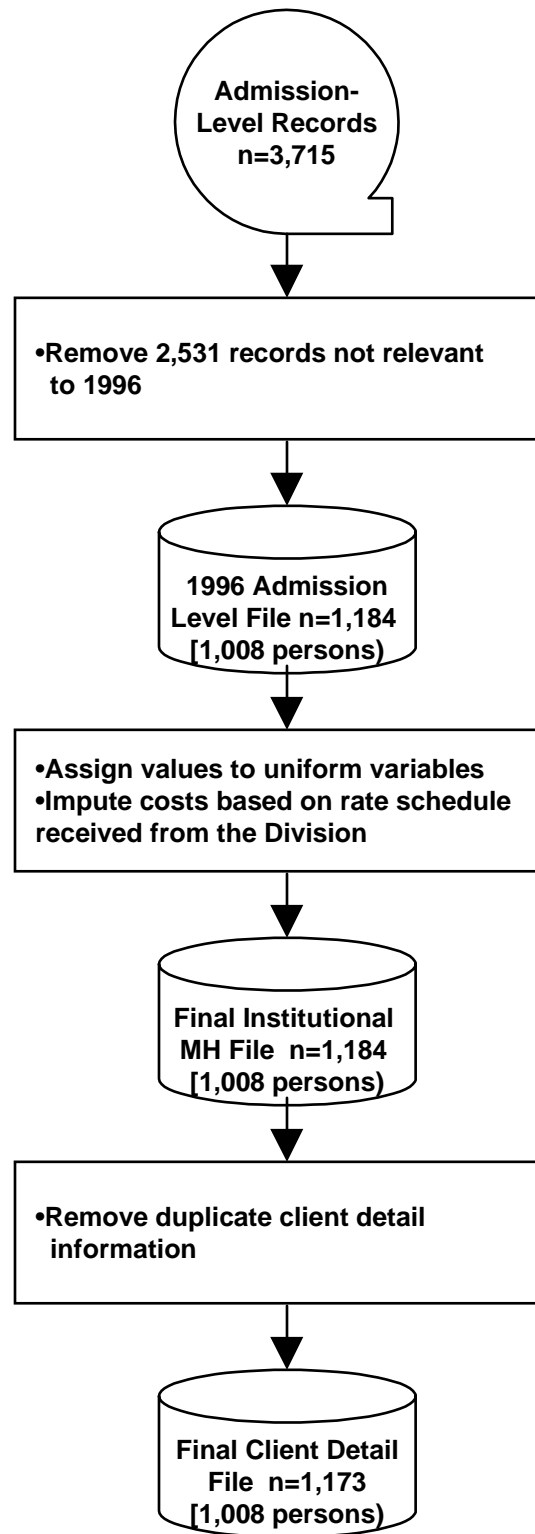
Data were obtained from four adult mental health programs: Delaware State Psychiatric Hospital, Community Treatment Team (CTT), Kent/Sussex Counties and New Castle County.

### ***Delaware State Psychiatric Hospital***

#### **Data Sources and Processing**

A file of 3,715 psychiatric hospital admissions was received. A flowchart detailing the preprocessing of these data is shown in Figure 9.

**Figure 9 Delaware State Psychiatric Hospital Processing**



Application of Costs

Costs information was not available on this file but costs were imputed on a *per diem* basis by ward and period within the year through the use of a cost schedule provided by Division of Alcoholism, Drug Abuse, and Mental Health (DADAMH).

Exceptions

There were 2,531 records excluded as not being relevant to 1996.

Anomalies

One record had an invalid ward code and therefore cost imputation was not possible.

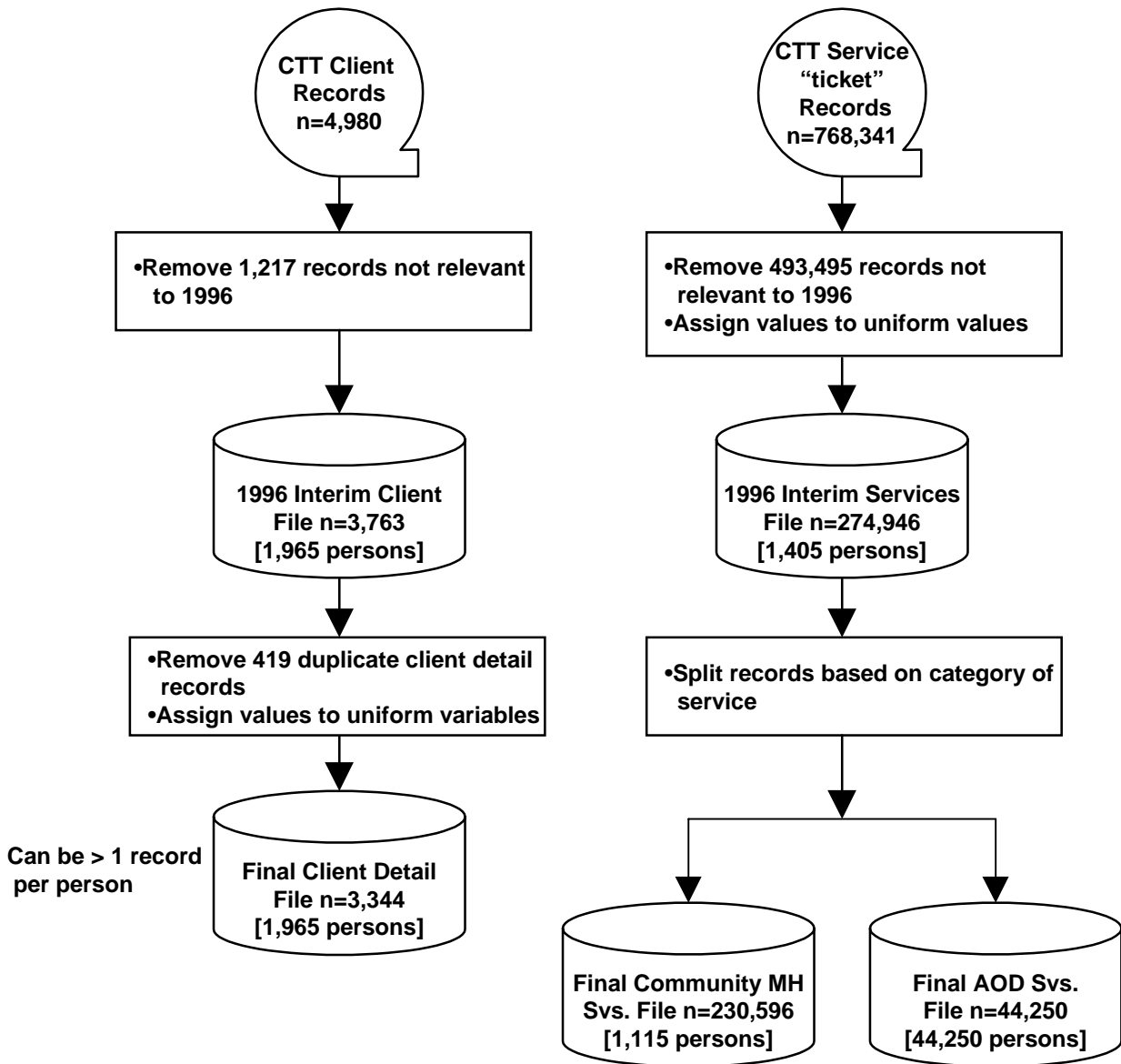
**Community Treatment Team**

***Data Sources and Processing***

CTT provided both a client file (4,980 records) and a service 'ticket' file (768,341 records). Most (84%) of the suitable service records were directed into the final community mental health file, but some (16%) were routed into the final AOD file based on category of service. A flowchart detailing the preprocessing of these data is shown in Figure 10.



**Figure 10 Delaware Community Treatment Team Processing**



### Application of Costs

Cost information was taken from the AMOUNT field on the service file. When this field contained a missing value (53% of the records), costs were imputed by number of 10-minute service increments represented on each record using the midpoint of a rate range on a schedule provided by DADAMH. Such imputation was possible for all records missing a value in the AMOUNT field.

### Exceptions

There were 1,217 records excluded from the client file as not being relevant to 1996. There were 493,495 records excluded from the service file for the same reason. Additionally, 419 of the remaining client records were excluded as duplicates.

## Anomalies

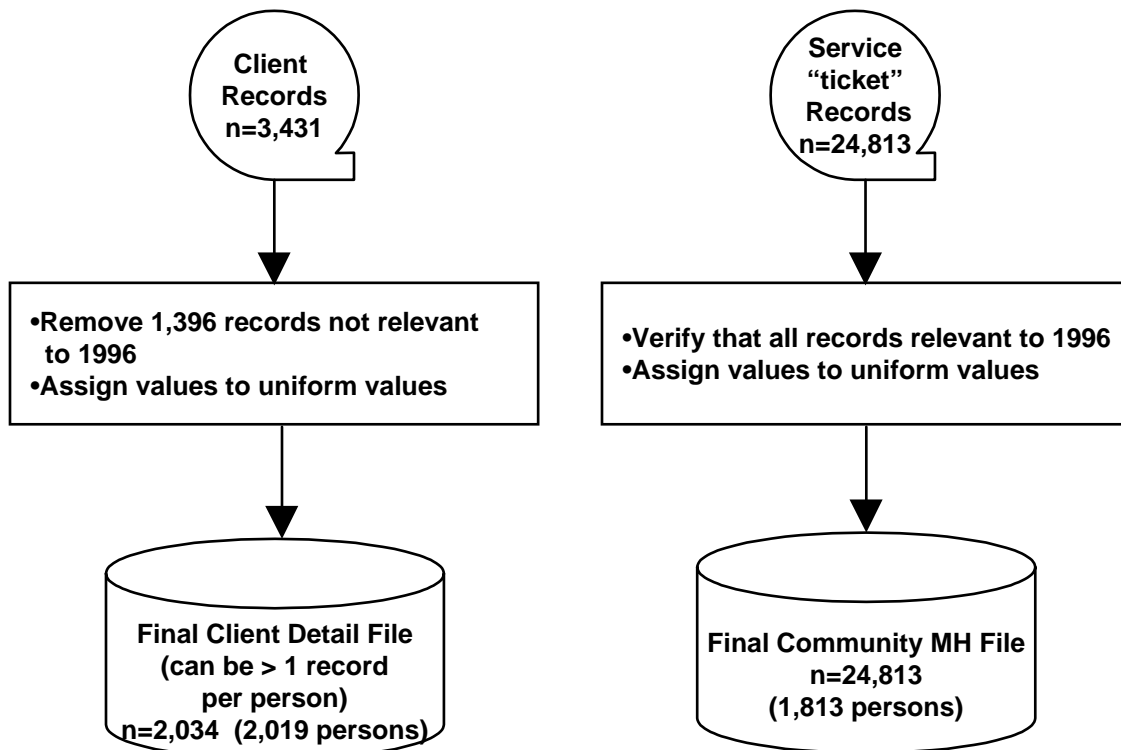
The client and service files were not explicitly matched. Client (demographic) records that did not have matching service records were kept and processed normally. Service records that did not have matching client records were assigned a dummy CLIENT\_ID (which links to a dummy final IDB client detail record) but otherwise were processed normally.

## Kent/Sussex Counties

### Data Sources and Processing

The Kent/Sussex Counties Program provided both a client file (3,431 records) and a service ‘ticket’ file (24,813 records). A flowchart detailing the preprocessing of these data is shown in Figure 11.

**Figure 11 Delaware Processing for Kent and Sussex County Data**



### Application of Costs

Cost information was not available on the service file and was imputed by number of 10-minute service increments represented on each record using the midpoint of a rate range on a schedule provided by DADAMH. Such imputation was possible for all records.

### Exceptions

There were 1,396 client records excluded as not being relevant to 1996. All service records were relevant to 1996 and were processed.

**Anomalies**

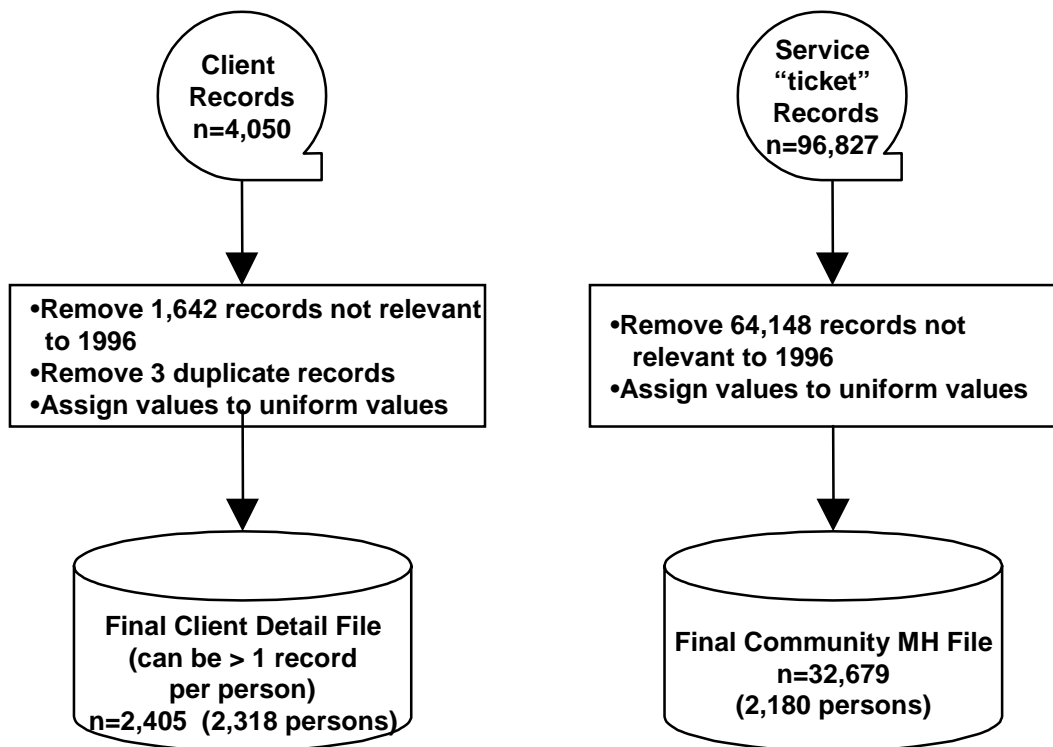
The client and service files were not explicitly matched. Client (demographic) records that did not have matching service records were kept and processed normally. Service records that did not have matching client records were assigned a dummy CLIENT\_ID (which links to a dummy final IDB client detail record) but otherwise were processed normally.

**New Castle County**

**Data Sources and Processing**

The New Castle County Program provided both a client file (4,050 records) and a service ‘ticket’ file (96,827 records). A flowchart detailing the preprocessing of these data is shown in Figure 12.

**Figure 12 Delaware New Castle County Data Processing**



**Application of Costs**

Some cost information was available on most (74%) of the service records. Cost was assigned from the AMOUNT field when it was not missing. Where AMOUNT information was missing, cost was assigned from the TPRICE field. Of the 8,344 records that completely lacked cost information, 1,832 had valid SERV\_QTY values which allowed cost imputation by number of 10-minute service increments using the midpoint of a rate range on a schedule provided by DADAMH. Costs were missing and could not be imputed on 6,512 (20%) of final service records.

***Exceptions***

There were 1,642 client records excluded as not being relevant to 1996. Three of the remaining client records were excluded as duplicates. There were 64,148 service records that were excluded as not being relevant to 1996.

***Anomalies***

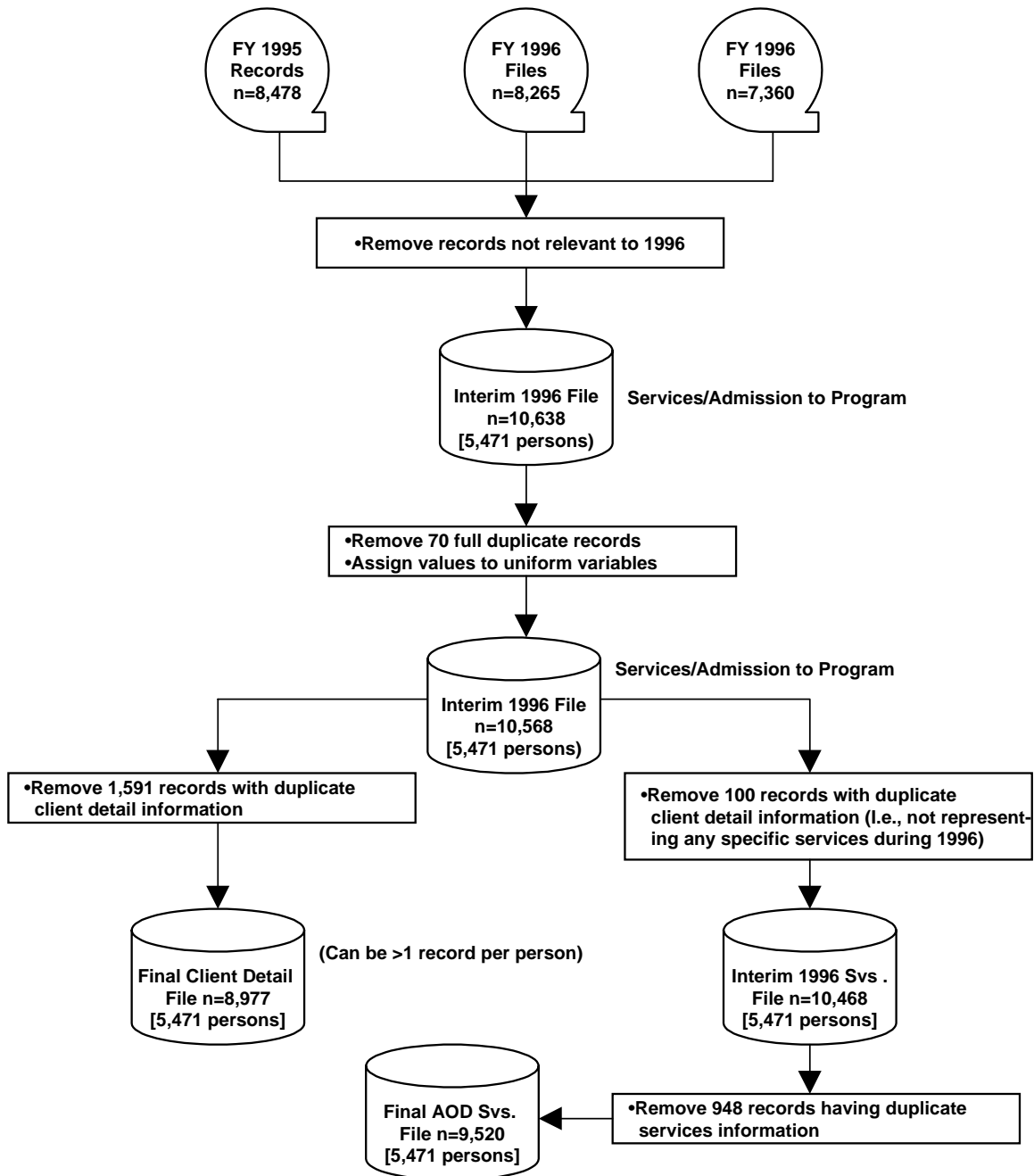
The client and service files were not explicitly matched. Client (demographic) records that did not have matching service records were kept and processed normally. Service records that did not have matching client records were assigned a dummy CLIENT\_ID (which links to a dummy final IDB client detail record) but otherwise were processed normally.

**Delaware Adult AOD**

***Data Source and Processing***

The adult AOD Program submitted three fiscal year files totaling 24,103 records on treatment episodes. These files contained client data and dates entering and leaving the program and service start and end dates, but little specific service information. A flowchart detailing the preprocessing of these data is shown in Figure 13.

**Figure 13 Delaware Adult AOD Process**



***Application of Costs***

Cost information was not available on the received data. Cost imputation based on specific services was not possible due to absence of suitable service information. Crude cost estimates were calculated, where possible, by apportioning the total dollars spent by a provider (or a group of providers) to the number of days that a record indicated a person’s association to that provider (i.e., start to end of a treatment episode). It should be noted that this number of person association-days may overestimate the number of actual service days.

For each provider with associated budgetary information (a partial schedule was obtained from DADAMH), it was possible to calculate the total number of association-days for all service records for that provider, and distribute the total spending proportionally to association-days represented on each record. In situations where multiple records for a person at one provider indicated overlapping dates, the association-days were reduced so that no particular date was counted more than once. This adjustment was not evenly distributed over all overlapping records. For example, if a person had a February 3 to February 15 record at a particular provider, and a February 10 to February 18 record at that same provider, 13 `association-days` were assigned to the first record, and the second record was adjusted to three `association-days` to remove the count of overlapping dates. Thus, cost estimates of records aggregated by person, or provider, are likely to be more realistic than record-specific cost estimates. Such cost estimation was possible for 62% of the records in the final service file.

### ***Exceptions***

There were 13,465 records excluded as not being relevant to 1996. Additionally, 70 records were excluded as duplicates. Some records had client or program information that was relevant to 1996 (e.g., a discharge date and discharge status from the program in 1996), but had service date ranges outside of calendar year 1996. Such records (n=948) were processed in the build of the final client detail file, but were excluded from the build of the final service file.

### ***Anomalies***

There were five cases where the value of the MCI (person identification number) was shared by more than one person. This was determined by finding greatly conflicting birthdates and/or gender and/or name values associated with particular MCI values. In these situations, the service records were processed normally, leaving the shared MCI in place, but in the final client detail file, all demographic and other variables (except STATE\_ID) were set to missing.

## CHAPTER FOUR: OKLAHOMA PROCESSING DOCUMENTATION

Data processing for Oklahoma involved obtaining data from the State Medicaid agency, as well as for the State MH/AOD agency. We received and processed four Medicaid claims files, one Medicaid eligibility file, and three MH/AOD files. Processing details for each data source are described below.

### OKLAHOMA MEDICAID

#### Data Sources – Claims/Encounters

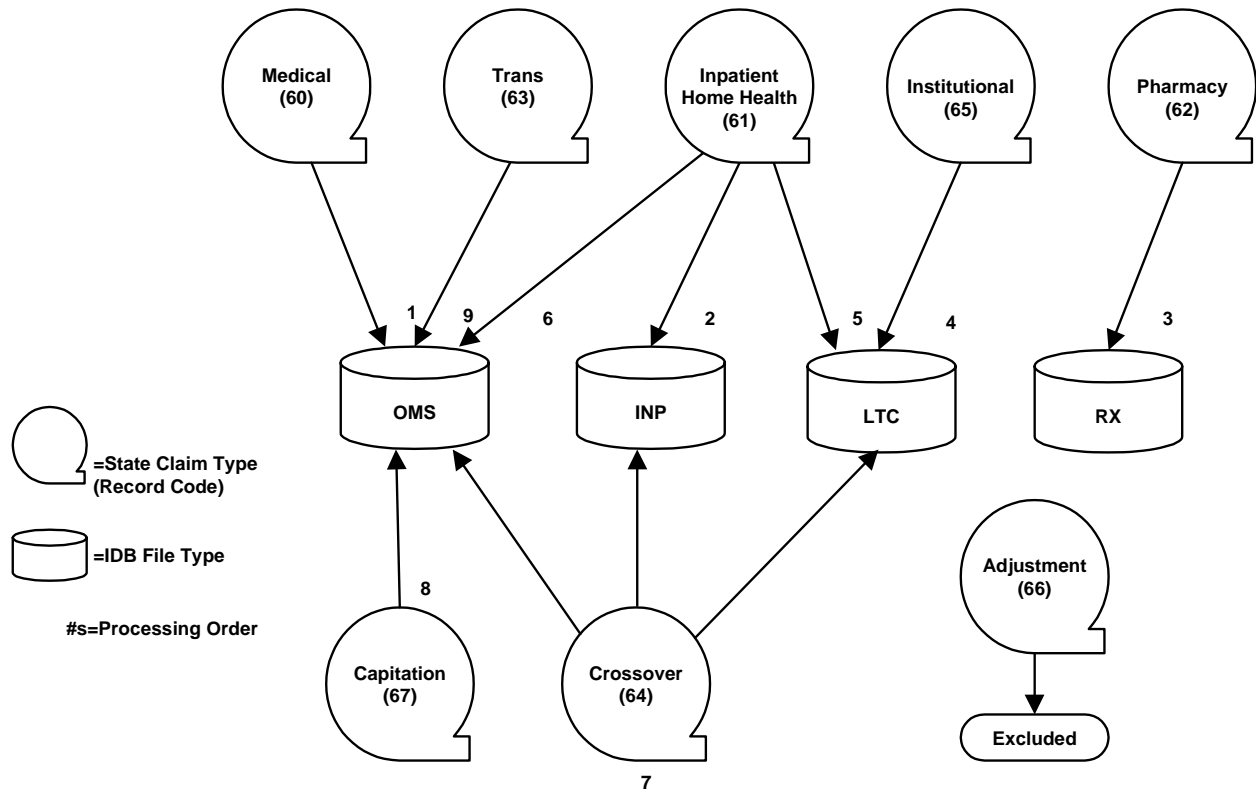
Data was received from Oklahoma Medicaid in three phases. The first acquisition included all claims/encounters paid during the first quarter of calendar year (CY) 1996 (3 monthly files). The second acquisition covered the remaining quarters of CY 1996. Given that the goal of the data base was to include all services occurring in calendar year 1996, a third set of data was obtained. This included all services occurring in calendar year 1996 but that were paid in 1997 and 1998. Approximately 140,000 claims/encounters that were paid in 1996 were also received as part of the third submission of data, but were excluded so as not to duplicate claims from the first two acquisitions. The first and second acquisitions of data also included denied claims. Because of the large volume of denials, the State eliminated them from the third acquisition.

The Oklahoma Medicaid claims and encounter data are stored within the same data base. The file consists of variable length records with different formats for each of the seven claim types:

- Medical
- Hospital/Home Health
- Pharmacy
- Transportation
- Crossover
- Institutional
- Adjustment
- Capitation

Because of the differences in the file types on the Oklahoma claims files, the claims were processed by file type for inclusion in the IDB. In most cases, all claims of one file type fed into one IDB file. However, in the case of crossovers and inpatient/home health record types, the claims were split based on a combination of "Provider Category of Service" and "Provider Type" as shown in Figure 14.

**Figure 14 Oklahoma Medicaid Processing**



**Application of Costs**

Oklahoma reports Medicaid payments on its fee-for-service claims at the line item level for all services except for services submitted on a UB92 (Hospital/Home Health). These claims only report costs at the header level. The payment for these claims is placed on the first line item and the remaining line items for the claim have zero-filled amount fields. The amount indicators are used to identify these services.

Encounter claims do not report Medicaid payments. The Medicaid payment for these services is carried on the premium payment. Oklahoma does not capture fee-for-service equivalents.



**Table 9 Oklahoma Medicaid Category of Service and Provider Type Crosswalk**

Oklahoma Record Code	IDB File Type	Oklahoma Provider Category of Service	Oklahoma Provider Type
61 Hospital/Home Health	Outpatient	15 – Outpatient 30 – Home Health 40 – Clinic	Any
	Inpatient	10 –Inpatient	Any except 26
		Missing – Unknown	Any
Long Term Care	10 –Inpatient	26 - Inpatient Psych. Services	
64 -Crossover	Outpatient	Any Except 10, 20, 25, 26	Any
	Inpatient	10 – Inpatient	Any except 26
	Long Term Care	20 – Skilled Nursing Services 25 – Intermediate Care*	Any
		26 – ICF/MR*	26 - Inpatient Psych. Services
	10 - Inpatient	26 - Inpatient Psych. Services	

\* According to the State, these values are no longer used. They were included in the crosswalk to catch any exception cases.

#### ***Data Sources – Eligibility***

The eligibility file used for the final IDB was the Oklahoma Eligibility History file. The data base was built using the version acquired in March, 1998 and included all Medicaid eligibility history as of the date the file was cut. Like the claims file, the eligibility file also consists of variable length records for each of eight file types:

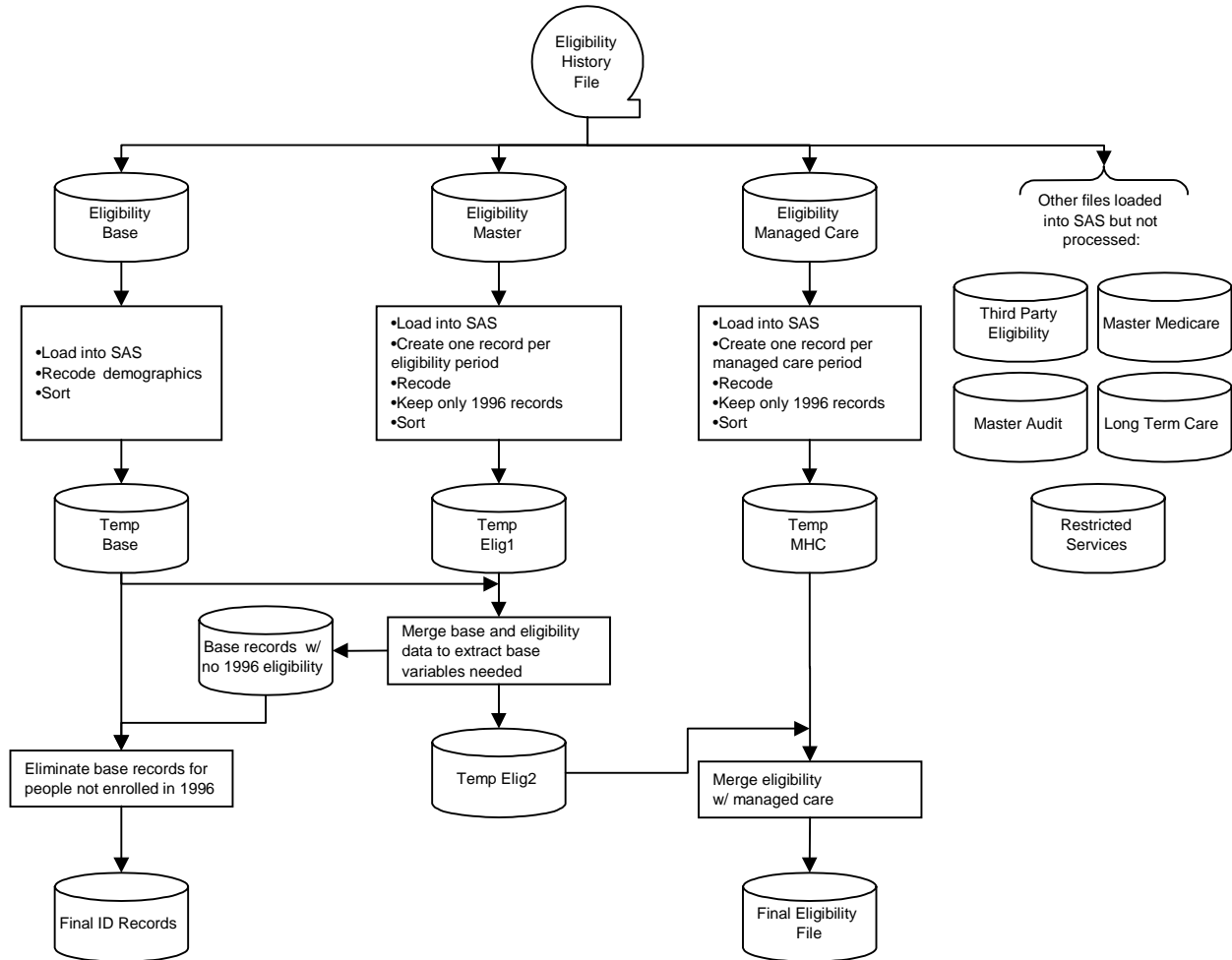
- Master Base
- Master Eligibility
- Third Party Liability
- Master Medicare Record
- Long Term Care
- Restricted Services
- Managed Health Care
- Master Audit

Only three of the above file types were needed to construct the IDB: "Master Base", "Master Eligibility" and "Managed Health Care". The "Master Base" record was used to derive the demographic information. The "Master Eligibility" record contained one segment for every monthly period of overall Medicaid enrollment. The Managed Health Care record also contains one segment for every monthly period of enrollment, but the information reported is enrollment in a managed health plan.

The eligibility processing was performed in one job after the initial SAS load. Information from the three record types (Base, Master and Managed Care) was combined by linking records based on person ID and, where necessary, time period. People were only retained in the IDB "client core" file if they had some period of enrollment within the calendar year. Enrollment records were combined with managed care records and, similarly, only retained if the managed care enrollment period occurred within 1996.

Oklahoma tracks Medicaid eligibility periods in monthly segments, so combining the information across the two sources was very straightforward. Figure 15 shows the general flow of the processing.

**Figure 15 Oklahoma Medicaid Eligibility Processing Stream**



**Counts and Summaries – Claims**

As mentioned previously, the Oklahoma claims data was received in three acquisitions. Tallies were kept by quarter for 1996 and 1997 date of payment data was combined. To minimize the size of the data to be processed denied claims were deleted at the time of the original load. Table 10 shows counts by quarter.

**Table 10 Oklahoma Medicaid Claims: Exclusions of Denied Claims**

Time Period Processed	Raw Input Records	Denied Claims Excluded	Raw Output Records
Quarter 1, 1996	3,475,030	934,761	2,540,269
Quarter 2, 1996	3,815,088	1,123,747	2,691,341
Quarter 3, 1996	3,366,410	869,873	2,496,537
Quarter 4, 1996	3,662,837	1,022,356	2,640,481
Quarter 1-3, 1997	1,538,343	0	1,538,343
Total	15,857,708	3,950,737	11,906,971

When the data was loaded into SAS, the claims were split based on their record type as defined by Oklahoma. In addition, one record was created for each line item of an input record, when multiple line items occurred within a claim. Most record types only contain one line item per claim. However, Medical and Hospital/Home Health record types did carry multiple line items per claim. Table 11 shows how the record counts (at the line item level) were distributed across record type and across quarter.

**Table 11 Oklahoma Medicaid Raw Input Claims by Record Code and Payment Quarter**

State Record Code	State File Type	Quarter 1, 1996	Quarter 2, 1996	Quarter 3, 1996	Quarter 4, 1996	Quarter 1-3, 1997	Total
60	Medical	889,953	878,383	827,216	835,845	839,055	4,270,452
61	Hospital/HH	511,689	470,864	441,627	366,728	484,658	2,275,566
62	Pharmacy	897,823	838,844	817,521	847,790	146,564	3,548,542
63	Transportation	7,288	10,273	10,210	7,987	7,951	43,709
64	Crossover	372,775	490,407	383,241	343,002	399,466	1,988,891
65	Institutional	64,929	64,715	62,601	63,335	27,624	283,204
66	Adjustment	45,370	117,674	64,211	90,884	0	318,139
67	Capitation	170,935	208,733	236,822	389,022	0	1,005,512
	Total	2,960,762	3,079,893	2,843,449	2,944,593	1,905,318	13,734,015

Each of these file types was processed separately for inclusion into the IDB. Based on our uniform file definitions, the claims were sorted into IDB uniform file types. The following table shows the counts of records obtained based on their final destination in the IDB.

**Table 12 Oklahoma Medicaid Raw Input Claim Counts by State File Type and IDB File Type**

State File Type	Total	Inpatient	RX	LTC	Other
Medical	4,270,452	0	0	0	4,270,452
Hospital/Home Health	2,275,566	630,605	0	176,478	1,468,483
Pharmacy	3,548,542	0	3,548,542	0	
Transportation	43,709	0	0	0	43,709
Crossover	1,988,891	28,653	0	14,613	1,945,625
Institutional	283,204	0	0	283,204	0
Adjustment	318,139	4,916	147,537	3,685	162,001
Capitation	1,005,512	0	0	0	1,005,512
Total	13,734,015	664,174	3,696,079	477,980	8,895,782

In keeping with guidelines for the structure and content of the IDB, certain exclusions were performed. These exclusions fell into three broad categories. The first category was for claims with service dates that occurred outside of the time period of interest (i.e., 1996). To be included in the data base, at least one day covered by the claim/encounter had to occur in calendar 1996. Also, in the last acquisition of data, there were claims that overlapped those submitted in earlier acquisitions. These claims were excluded using the criteria that they were part of the last acquisition (which should have only had dates of payment after 1996) but which had dates of payment in 1996.

The second exclusion was adjustment records (Record Type 66). According to advice obtained from Oklahoma, these are gross adjustments that could not be tied to an individual patient. These records were retained in a separate exclusion file in case they were needed at a later date for identifying total costs.

The last exclusion was the detail line items from acute inpatient claims. In keeping with standard IDB structural guidelines, aggregate summary information was extracted from inpatient claims and line item detail was excluded from the data base.

The following table shows a summary of the exclusions by category for each of the IDB file types.

**Table 13 Oklahoma Medicaid Claims Processing Exclusion Report**

<b>Type of Exclusion</b>	<b>Total</b>	<b>Inpatient*</b>	<b>RX</b>	<b>LTC</b>	<b>Other</b>
Out of Range Service Dates and Potential Duplicates	1,778,196	68,844	193,354	173,002	1,342,996
Gross Adjustments	318,139	4,916	147,537	3,685	162,001
Inpatient Detail Records	520,141	497,513	0	22,628	0
Total Observations Excluded	2,616,476	571,273	340,891	199,315	1,504,997

\* Observations in the Inpatient column include some records that were excluded at a later point in processing.

The following shows the number of Medicaid observations that fed into the final IDB generation for Oklahoma:

**Table 14 Oklahoma Medicaid Counts of Records Submitted for Linking Step**

<b>Record Type</b>	<b>Observations</b>
Inpatient	92,901
Rx	3,355,188
Long Term Care	278,665
Other	7,390,785
Total	11,117,539

***Counts and Summaries – Eligibility***

The eligibility data was loaded by file type as defined by Oklahoma. As described in the data sources section, there were only three record types that were retained in the final IDB. Aside from the exclusion by record type, the only other exclusions performed on enrollment data was to exclude enrollment periods outside of the calendar 1996 time window. The following table shows the record counts at each phase of the processing up to the point where the data fed into the linking process.

**Table 15 Oklahoma Medicaid Eligibility Input Record Summary**

<b>SAS Loaded Records, by Input File Type:</b>		
Master Base	(1 record per eligible)	908,476
Master Eligibility	(1 record per eligibility segment)	20,048,854
Managed Health Care	(1 record per eligible – split later in processing)	658,587
Third Party Liability	(not used in data base creation)	98,397
Master Medicare Record	(not used in data base creation)	282,294
Long Term Care	(not used in data base creation)	182,540
Restricted Services	(not used in data base creation)	144
Master Audit	(not used in data base creation)	908,476
<b>Observations Deleted:</b>		
Out of Range Eligibility Dates - Master Base		441,300
Out of Range Eligibility Dates - Master Eligibility		16,151,513
<b>Total Observations At Completion of Processing (prior to linking):</b>		
Client Level, Unique ID		467,176
Eligibility Level, per Period of Enrollment per Client		3,897,341
<b>Total Input Records Read:</b>		<b>3,947,390</b>

***Anomalies***

During processing of the data, certain characteristics were noted and flagged as either potential data anomalies or perhaps simply interesting findings. These characteristics fall into five major areas as follows:

- Coding Practices
- Pharmacy Topics
- Crossover Topics
- Inpatient Topics
- Remaining Topics

***Coding Practices***

**Distribution of Services:** On the Medical portion of the OMS file (without the Transportation, crossover, capitation and OPD claims/encounters), it was interesting to note that 37.3 percent were for physician services, 21.3 percent of the claims were for rehabilitation services, 7.3 percent were for EPSDT and 12.2 percent were for waiver services.

**Family Planning:** Family planning services can legitimately appear on the "Inpatient", "Pharmacy", and "Other Medical Services" files. It is advantageous for States to identify services that qualify as family planning wherever possible, as they receive a higher federal match on their payments for these services. Surprisingly, the number of family planning claims on the "Inpatient" file was very small (only 22 claims) and there were no claims identified as family planning on the "Pharmacy" file. It should also be noted that there were a few claims for family planning services performed by unexpected providers (i.e., mental health clinics).

Place of Service: The distribution of place of service values showed the expected percentages (e.g., "Office" - 15 percent, "Outpatient Hospital" - 23 percent.). However, there were two situations where the information was problematic. The first was the high number of claims classified as "Other Unlisted Facility" (i.e., 30 percent). This was due to two Oklahoma values that did not fit cleanly into the more specific categories: "Other Locations" and "Specialized Treatment Facility." There were two record types that did not carry a place of service value and none could be imputed for them (i.e., Capitation, and Crossover). Place of Service for these claims, was coded as "Not Available". The Outpatient Hospital and Transportation records also did not carry a Place of Service variable. However, in these cases, the correct value could be imputed because of a direct correlation to a uniform place of service value (i.e., Outpatient Hospital and Transportation).

Unknown Provider Types: There were a few situations on the "OMS" file where a uniform provider type value could not be determined. In addition to missing values, the following Oklahoma provider types were mapped to the "unknown" category:

- Room/Board (Provider type 31)
- Developmentally Disabled Service Division (DDSD) (Provider Types 44 and 62).
- Architectural Modification (Provider Type 48)

Inconsistencies in Provider Specialty Code Field: There were a small number of claims with an outpatient hospital provider type but reported physician specialties. No corrections were made to the data base to reconcile this inconsistency.

Definitional Issues Related to Category of Service. The Oklahoma data included two uniform categories of service that represented services not covered by Oklahoma. They are:

- LTC Psychiatric Services
- Religious Non-Medical Health Care Institutional Services

### Pharmacy Topics

Days Supply: Given that most prescriptions are written such that only a 30 day supply of a drug is dispensed at a time, it was surprising to see that over eight percent of the claims had a days supply of over 31 days (13.8 percent over 30 days). Of all pharmacy claims, 98.4 percent had days supply within two months with 99.9 percent under 120 days. There were some cases where the "Days Supply" exceeded 800 days with the largest outlier reporting 960 days.

Provider Type on the Pharmacy File: There were a few exceptional cases on the pharmacy file where the provider type "DME/Supply" rather than "Pharmacy." Upon examination of these claims, it was determined that they carried National Drug Codes (NDC) and were correctly placed in this file.

Outlier Payments and Units on Pharmacy File: While the mean payment for a drug claim in Oklahoma is about \$35 and 99 percent of the claims had payments under \$209, there were some large outliers. Some showed payments as high as 79K which are assumed to be expensive drugs such as AZT. These outliers also showed high quantity values in some cases. The maximum quantity on the "Pharmacy" file was 88,740 units.

### Crossover Topics

Pharmacy Providers on Crossover Claims: There were a few (556) claims with pharmacy provider types occurring on crossover claims. Examination of these claims determined that they contained CPT (Current

Procedural Terminology) codes representing DME (Durable Medical Equipment). Therefore, these claims were included in the "Other Medical Services" file rather than the "Pharmacy" File.

EPSDT Crossovers: There were approximately 18K crossover claims that represented EPSDT services. We are assuming these are services for disabled children covered under the Medicare-D program, but did not fully investigate them since the numbers were relatively small.

LTC Crossovers Provided by Inpatient Facilities: Six percent of the crossover LTC claims showed an Inpatient Hospital provider type. The service category reported on these claims was Nursing Facility Services. These claims represented slightly over \$723K in payments by Medicaid.

### Inpatient Topics

Inpatient Procedure Codes: The inpatient file can carry up to six procedure codes in Oklahoma. In most cases, the coding scheme used is ICD-9-CM. However, there are a few cases where there is a mixture of both ICD-9-CM and CPT-4. Procedure codes are only available on non-crossover claims. For the non-crossovers, the number of claims with no procedure code is higher than expected (45.5 percent).

Inpatient Payment Outliers: Over Ninety-nine percent of the claims in the inpatient claims had reasonable payment amounts (i.e., less than \$13K). Claims with payments over \$250K were examined and determined to be legitimate given the diagnostic information reported. Almost half of these cases involved neonatal claims. The remaining were various serious conditions from respiratory distress, severe burns, complications of diabetes, and so forth.

### Remaining Topics

Date of Payment Lag: Normally, it is assumed that the data base will be 95 to 98 percent complete if a six to nine month payment lag is used to capture all services for a calendar year. In Oklahoma, we received all services for 1996 paid up until October 1998 due to the time at which we requested our third and final acquisition of data. This resulted in a higher than expected percentage of claims in some file types. The percent of claims that were paid 10-22 months after the end of calendar year 1996 was 1.7 percent in the OMS file and 3.6 percent in the inpatient file. The percentages were less than one percent in LTC and Pharmacy.

Encounter Data. While every file type contained some encounter data, the amount contained in the LTC file and the Pharmacy file was negligible. Overall, the percent of encounters that fed into the linking process was only 3.7 percent of the entire file volume. 98 percent of encounters were in the OMS file and 1.2 percent were in the Inpatient file. Table 16 shows the number and percent of claims compared to encounters by file type. It also shows how the encounters were distributed across Oklahoma's native file types.

**Table 16 Oklahoma Medicaid FFS and Encounter Counts by State File Type and IDB File Type**

IDB File Type	Oklahoma File Type	Total FFS Claims	Total Encounter Claims	Total Claims	Encounter Percentages	
					Row Percent	Column Percent
Inpatient	Crossover	22,204	0	22,204	0.0%	0.0%
	Hospital/HH	65,782	4,915	70,697	7.0%	1.2%
	Total	87,986	4,915	92,901	5.3%	1.2%
LTC	Crossover	11,038	0	11,038	0.0%	0.0%
	Hospital/HH	13,452	22	13,474	0.2%	0.0%
	Institutional	254,153	0	254,153	0.0%	0.0%
	Total	278,643	22	278,665	0.0%	0.0%
OMS	Capitation	1,003,779	0	1,003,779	0.0%	0.0%
	Crossover	1,523,208	0	1,523,208	0.0%	0.0%
	Hospital/HH	1,127,495	74,233	1,201,728	6.2%	18.0%
	Medical	3,297,752	331,270	3,629,022	9.1%	80.1%
	Transportation	32,944	104	33,048	0.3%	0.0%
	Total	6,985,178	405,607	7,390,785	5.5%	98.1%
RX	Pharmacy	3,352,292	2,896	3,355,188	0.1%	0.7%
Grand Total		10,704,099	413,440	11,117,539	3.7%	100.0%

Capitation Payments: When comparing the total capitation payments found on the Oklahoma Medicaid files with the amounts reported on the 1996 HCFA 2082 report, the IDB numbers were significantly higher. The data base reports \$84 million whereas the HCFA 2082 reports slightly under \$61 million.

Diagnosis Codes. Some invalid diagnosis codes were identified in our initial runs. We performed an editing pass to identify these codes and mark them as invalid.

## OKLAHOMA MENTAL HEALTH

### Data Sources

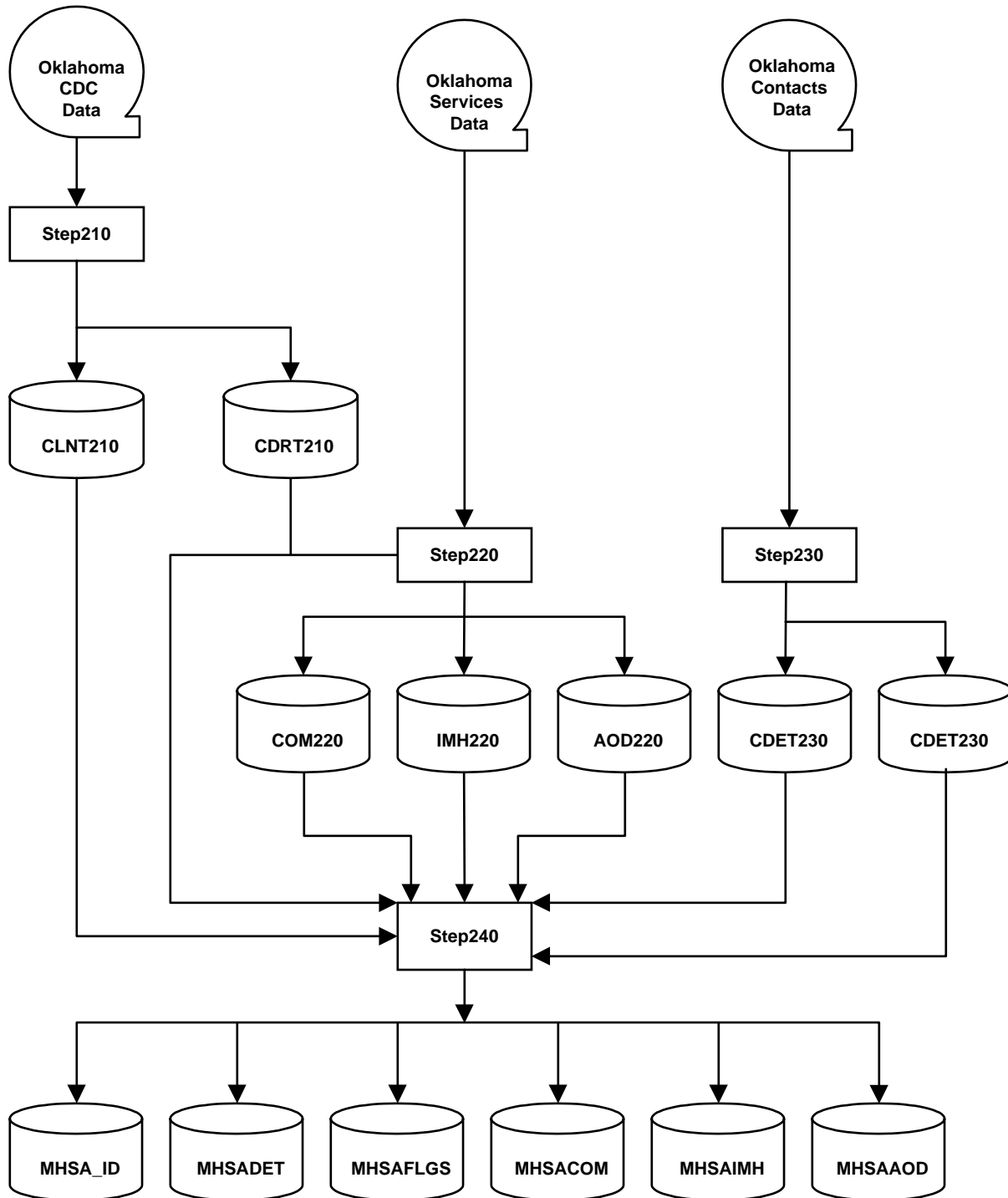
The files received from the Oklahoma Department of Mental Health and Substance Abuse Services (DMHSA) contained services provided during calendar year 1996. Four separate files were submitted to MEDSTAT for this project:

- Client Data Care (CDC) containing program of care and client demographic information;
- Services data;
- Contacts data containing contact services and limited demographic information; and
- Agency data containing information about service providers.

All data were supplied in the form of SAS data sets, compressed as PC Zip™ files. The PC Zip™ files were uncompressed and the SAS data sets uploaded to the DEC AlphaServer for processing. Only the CDC, Services, and Contacts data files were processed. DMHSA does not separately record Mental Health and AOD treatments. The services data contains a combination of both Mental Health and AOD information.



Figure 16 Determining the Output Data Set



**Application of Costs**

Cost estimates for mental health, alcohol, and other drug treatments were calculated on individual services. The estimates were based upon the quantity of service and cost factors supplied by DMHSA. These factors were furnished in two distinct formats: alcohol and other drug treatment factors were based on contracted rates for services, while mental health service factors were based on total department costs.

For the alcohol and other drug treatments, DMHSA provided a list of contracted rates (1996) for specific services as defined by the Oklahoma service codes. The Oklahoma service codes can be found on the IDB in MH/AOD service detail data sets (SERV\_COM, SERV\_IMH, and SERV\_AOD) as the variable ISERVICE. Cost estimates for mental health services, however, were derived from the 1996 DMHSA Cost Analysis Report. The report totaled costs and units across 35 categories of services. Cost-per-unit amounts were calculated from the totals, and the categories of service were mapped to the Oklahoma service type codes.

Together, the two sets of factors allowed for cost estimation for most services. Availability of cost data ranges from 67.0% for alcohol and other drug treatment data to 85.6% for institutional mental health services. Cost estimates are available for 71.3% of the community mental health services.

In some instances two cost factors could potentially be applied to a service code. Examples include case management and group counseling. In these cases, the cost factor used was determined by the client's program of treatment: the AOD contract rate cost factors were used where the treatment program was AOD, otherwise the mental health cost analysis factors were used.

## Counts and Summaries

The CDC data contain multiple observations per client. Separate observations exist for a client's admission to a program of treatment and for completion of the program. Additional observations can occur in the file for information updates and/or changes in a client's treatment program. The CDC data was used to populate three data sets:

- Client Core (CLIENT),
- Identifying Data (ID\_DATA), and
- Client Detail (CLNTDET).

A total of 256,339 observations were processed into 87,226 unique clients (CLIENT and ID\_DATA) and 256,339 Client Detail observations.

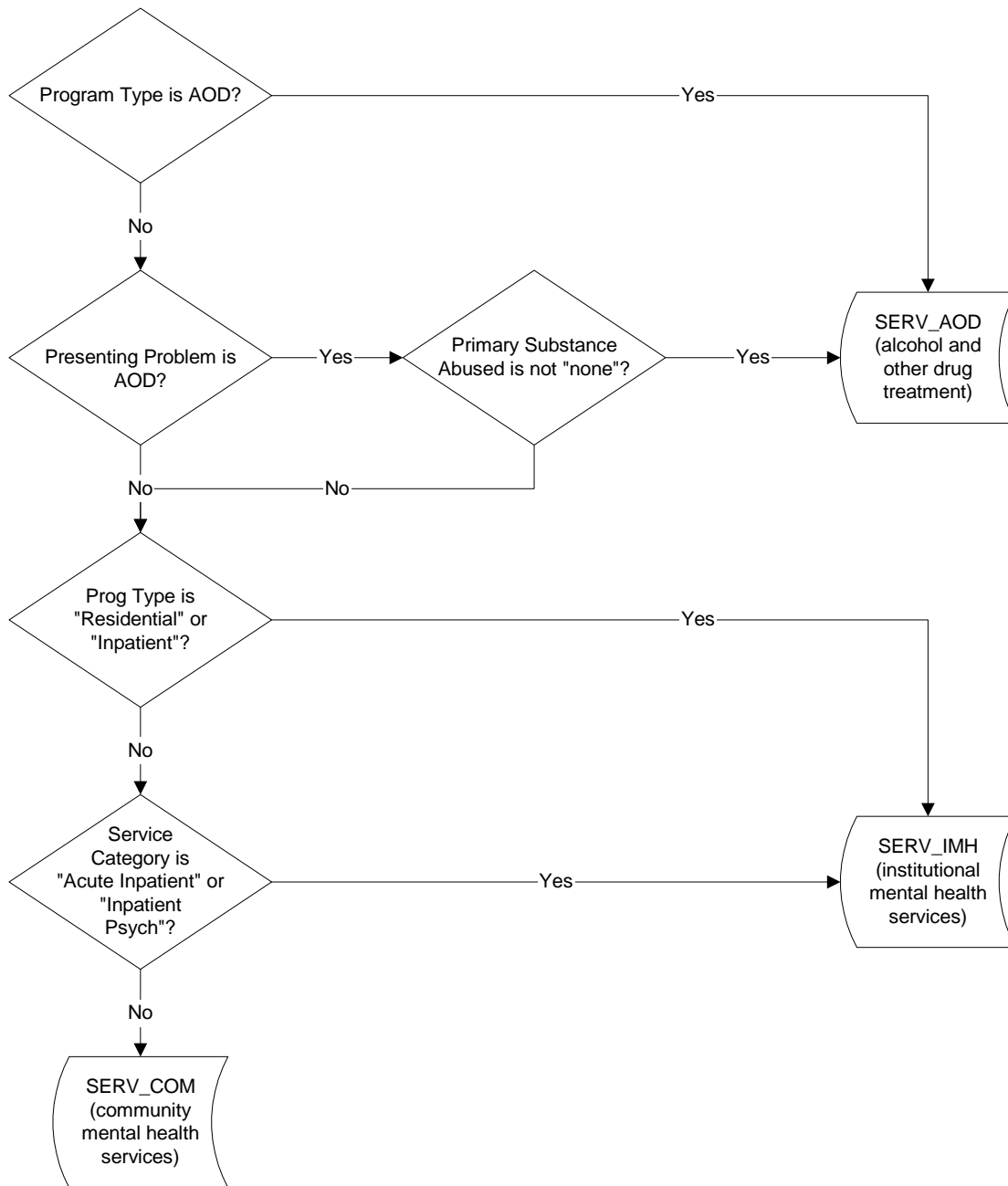
The Services data contain a single observation for each service. All service dates fell into the 1996 calendar year. The service data were combined with an extract from the Client Detail data to add information on the treatment program a client was undertaking. The service data were joined where client and provider IDs matched, associating the date of service with the most recent Client Detail date occurring prior to the date of service. The Services data contain 2,244,382 observations.

Service data were output to one of three data sets:

- Community Mental Health services (SERV\_COM),
- Institutional Mental Health services (SERV\_IMH), or
- Alcohol and Other Drug treatments (SERV\_AOD).

The output data set was determined from the category of service and the client's treatment program type, the primary presenting problem, and the primary substance abused. Figure 17 describes the algorithm used in determining the output data set.

**Figure 17 Determining the Output Data Set**



The Community MH data set contains 1,566,461 observations, the Institutional MH data set contains 351,467 observations and the AOD Treatment data set contains 320,454 observations.

**Contacts Data:** The Contacts data contain a single observation for each contact, with a limited amount of demographic information. The Oklahoma Contacts data contained a total of 65,170 observations. Processing of the Contacts data created 46,487 Client Core observations and 65,170 Client Detail observations.

The Client Core and Client Detail data from the CDC and contacts files were combined to create a total of 119,000 unique Client Core and 321,509 Client Detail observations. For linking, client data from the

treatment and contact data were combined into a separate ID data set. The ID data set contains information for 122,607 unique client numbers.

### ***Exceptions***

The Oklahoma DMHSA does not record diagnoses with services. The source for all diagnoses was the Client Data Core (CDC). On the Oklahoma portion of the IDB, all MH/AOD data is found on the client detail data set (CLNTDET). The "mental health treatment claim" and "alcohol/other drug treatment claim" flags were based upon uniform service category (SERV\_CAT) which was recoded from the Oklahoma service type code (ISERVICE). Likewise, there are no procedure codes on any Oklahoma DMHSA data. The "detoxification service" flag was based on the uniform service category (SERV\_CAT).

### ***Anomalies***

Client demographics are found on both the Client Data Core (CDC) and the Contacts data, although the Contacts data contains only a subset of the information available on the CDC. Client detail information is marked as unavailable for observations originating from the Contacts data.

Substance abuse information (substance, frequency, usual route of administration, and age at first use) is recoded independently of other information. No cross validation recodes were performed. This results in some unusual combinations such as non-zero frequency of use without any associated substance for abuse and smoking of alcohol (11 instances).

Treatment records are found on one of three service detail data sets based on a number of factors which include Program Type, Principal Presenting Problem, and Category of Service (see Figure 17 above). As a result, some unexpected services can be found on the service detail datasets, such as detoxification treatments in the Community MH data set (21 instances) and acute hospital inpatient services on the AOD data set (23 instances).

## **CHAPTER FIVE: WASHINGTON PROCESSING DOCUMENTATION**

### **OVERVIEW**

Data processing for Washington involved obtaining data from the State Medicaid agency and from the State MH/AOD agencies. MEDSTAT received and processed three Medicaid files and 16 MH/AOD files. Processing details for each data source are described below.

### **WASHINGTON MEDICAID DATA**

#### **Data Sources**

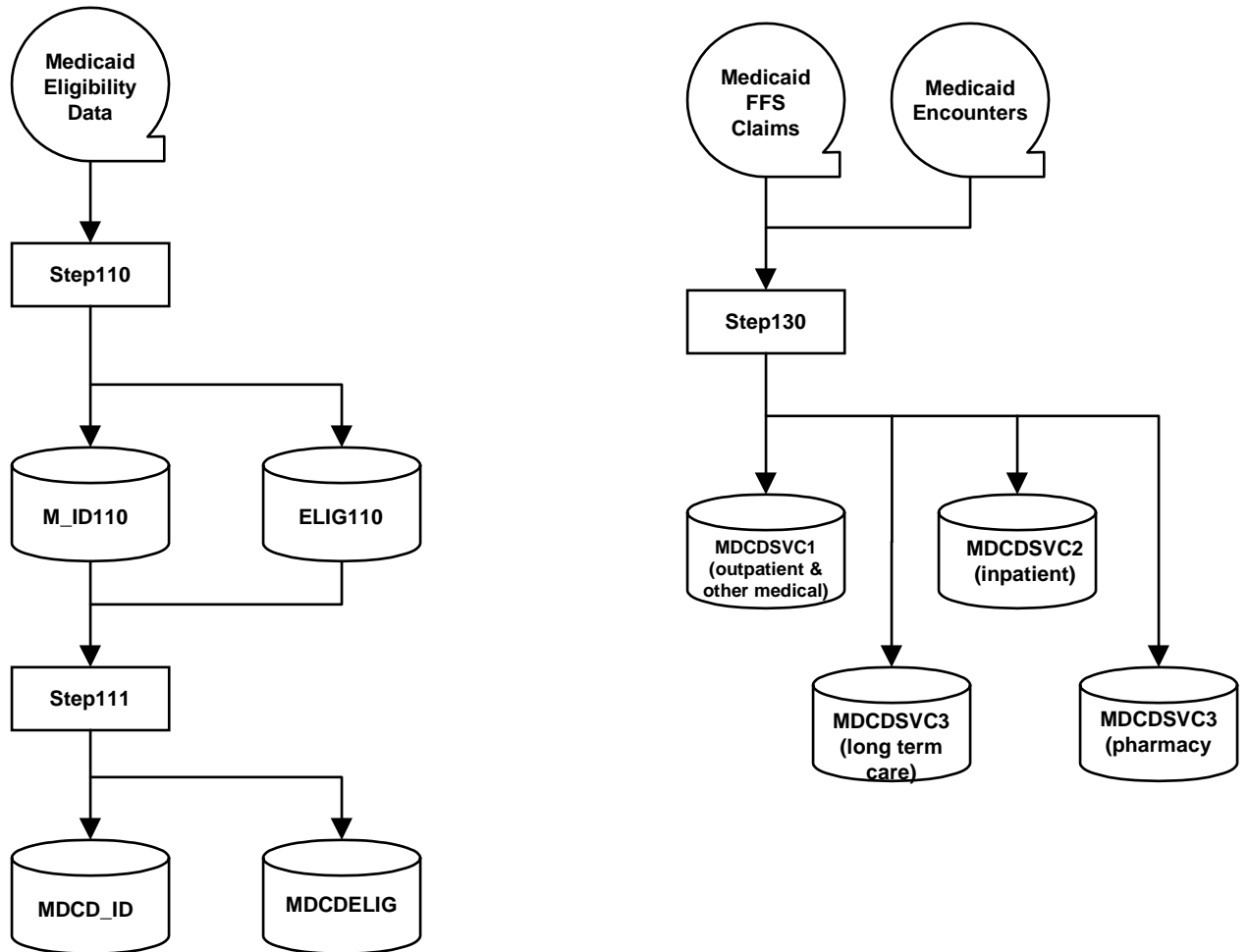
Washington Medicaid data were supplied in three files:

- Fee-for-service (FFS) claims
- Pseudo-claims for encounter data, and
- Eligibility

#### **Processing**

All files were submitted in EBCDIC format, and were loaded into SAS. As can be seen from the flowchart in Figure 18, the eligibility files we received were processed into the standard IDB eligibility file and into the Medicaid ID file for linking. Both the documentation and the file characteristics indicated that the FFS and encounter data followed the same file format, so these data were combined, and an indicator was added to the combined file to indicate the data source (FFS or encounter). Step 130 processed this combined information into the four types of IDB standard Medicaid claims.

**Figure 18 Washington Medicaid Processing**



A common practice with many Medicaid files is to record claim information (e.g. diagnoses, procedures, and line items) as repeating groups or sections within a claim record. Washington data, however, was provided in fixed length records. Because one record format was used for several types of claims, fields within some records were not used, resulting in values of blank or zero. For instance, the input variable `ADMISSION-DATE` was non-missing for all records, even though admission date does not apply in all cases (e.g., gross adjustments, pharmacy claims).

Washington Medicaid claim records contain a combination of summary and detail information. Summary information includes a patient identifier, summary diagnoses and procedure codes, and adjudicated charge totals for the claim. Detail information includes service dates, a line item revenue and procedure code, and detail adjudicated charges. For claims billed at the detail level, such as physician and pharmacy claims, each record represents a claim, and there is no difference between the detail claim charges and the total claim charges. Each of these records is processed as one observation using the detail charges. Inpatient claims typically include several line item charges which Washington records with one record per line item – repeating the summary information on each record. Inpatient claim data in the IDB was loaded from the first record of the claim with the following exceptions:

- Medicare Amount Paid (AMT\_CARE) claims records do not include a summary Medicare Amount Paid field, only a detail field. For inpatient claims, detail Medicare payment amounts from all line item records were summed to arrive at the AMT\_CARE amount.
- MH/AOD Revenue Center Flag (FREVCTR) – all line item records were scanned to determine the appropriate value for this flag.
- Procedure codes – Procedure codes can occur in two portions of the claim records: in a collection of six procedures on the header section, and in an additional procedure field that occurs on in each of the line item sections of the record. For inpatient claims, the collection of six procedures from the header was loaded into the data base, however, all line item procedure codes were examined during the initial processing for MH/AOD values. This may result in a client with an "YES" value for the MH/AOD Procedure Flag (FPROC), but with no MH/AOD procedures on any of that client's services.

Processed claims were output to one of four data sets corresponding to the four Medicaid service detail data sets in the IDB and augmented with the service core variables. Claims were assigned to one of the four files based on a combination of service category, claim form, and provider type, as shown below in Table 17.

**Table 17 Washington Claim Type Map**

<b>Condition</b>	<b>Output File</b>
Input service category of "inpatient psychiatric"	Long Term Care
Physician claim form Dental claim form EPSDT claim form Outpatient claim form Medicare Non-Institutional Crossover claim form Medical Vendor claim form Medicare Outpatient Crossover claim form	Outpatient/Other
Inpatient or Medicare Inpatient Crossover claim form and Input service category not one of: MHD Inpatient Psychiatric Hospital Involuntary Treatment Alcohol Detoxification	Inpatient
Inpatient Outlier claim form	Inpatient
Nursing Home claim form	Long Term Care
DRG or Inpatient or Medicare Inpatient Crossover claim form and Input service category is one of: MHD Inpatient Psychiatric Hospital Involuntary Treatment Alcohol Detoxification	Long Term Care
Practitioner or Outpatient Provider Type (PROVTYPE < 400)	Outpatient/Other
Hospital Provider Type (PROVTYPE 400 or 401)	Inpatient
Home Health Provider Type (PROVTYPE 600)	Outpatient/Other
Psychiatric Hospital or Long Term Care Provider Type (PROVTYPE between 402 and 601 inclusive, but excluding 600)	Long Term Care
Pharmacy Provider Type (PROVTYPE 700 or 702)	Pharmacy
All other claims	Outpatient/Other

The eligibility period data set contained one observation for each eligibility period for each person. Where appropriate, managed care periods were included in the eligibility data. Only one Managed Care period was included in each eligibility period. If more than one managed care period occurred during an eligibility period only the most recent was used. Where necessary, the begin and/or end dates of the Managed Care period were adjusted to prevent Managed Care dates outside the eligibility period.

As an example, consider an eligibility record with two eligibility periods (January through June, and July through December) and two Managed Care periods (February through March, and April through December). Processing would create two eligibility period observations from this record: one for the January through June period, and a second for the July through December period. The Managed Care dates for the first period would be April through June – the most recent overlapping Managed Care period (April-December), adjusted for the end of the eligibility period. The Managed Care dates for the second eligibility period would be July through December – the most recent overlapping Managed Care period (April-December), adjusted for the start of the eligibility period.

**Counts/Summaries**

Washington Medicaid FFS Claims and Encounters were supplied in one record format representing information from 14 unique types of claims, as follows:

**Table 18 Washington Claim Type Map**

<b>Claim Type</b>	<b>Count</b>
Inpatient Outlier	1,928 records
Drug Claim	7,022,525 records
Physician Claim	12,185,562 records
Dental Claim	2,157,966 records
EPSDT Claim	369,904 records
Outpatient Claim	1,808,471 records
Medicare Non-institutional Crossover Claim	831,266 records
Medical Vendor Claim	1,308,232 records
Gross Adjustment	10,592 records
DRG Claim	77,757 records
Inpatient Claim	14,466 records
Nursing Home Claim	211,025 records
Medicare Inpatient Crossover Claim	16,606 records
Medicare Outpatient Crossover Claim	140,678 records

As noted previously, the FFS and encounter data were received with an identical layout, so these data were combined and an indicator was added to indicate the source (FFS or encounter). This resulted in a total of 26,718,568 claims and pseudo-claims delivered in two files. Over 88% of these records were FFS claims (23,668,185 records), encounter records account for only 3,050,383 of the records. As described below, only the “header” or summary portion of inpatient claims was kept – all line item records were



discarded<sup>1</sup> in the initial processing run. The total number of inpatient line item records discarded was 561,590. This resulted in 26,030,142 records processed.

The output data sets contained a total of 26,030,142 observations broken down as follows:

- Inpatient: 72,192 observations
- Long Term Care: 220,497 observations
- Pharmacy: 6,730,956 observations
- Outpatient/Other: 19,006,497 observations.

## Exclusions

A total of 126,836 records were excluded from processing. Most of these were claims without a Medicaid ID number (116,244 records). The remainder of the excluded records (10,592) represents gross adjustments to health plans or other (non-individual) entities. In addition, there were 561,590 inpatient line item records that were processed as part of a previous inpatient claim (see above). No claims for services occurring before or after 1996 were found on the supplied data.

## Anomalies

There is a possibility that a small number of data values were set to “missing” values despite the presence of valid input data. All dates on the Washington files were recorded as the number of days since January 1, 1964, and many records had date values of zero. These values were converted to the missing value of “unknown” (.A in SAS). We realize that there is some possibility of incorrectly classifying people who were actually admitted (or received services) on January 1, 1964. However, the probability of this occurrence is so small that the error introduced into the data base as a result is negligible.

All but two sets of data transformations are documented in the data dictionary. Each of the undocumented transforms deals with dates, which appear in one of two formats. As previously noted, various service dates are coded as the number of days since January 1, 1964. These were converted to SAS dates and formatted using the appropriate date formats listed in the data dictionary. The second transformation involved recipient date of birth, which was source coded in YYYYMMDD format. A “1” was appended to the year and the result was converted to a SAS date (i.e. "9760704" became "19760704" prior to SAS conversion).

The majority of the information needed to create the final data base was available. Missing data was encountered in the following fields:

- Covered Days,
- Leave Days (long term care),
- Date Prescribed (pharmacy claims),
- Level of Care,
- Diagnoses 7, 8, and 9,
- Procedures 7, 8, and 9,

---

<sup>1</sup> Inpatient line item records were examined for MH/AOD indicators in the procedure and revenue code fields. In addition, the Medicare amount value for inpatient claims was totaled from all line item Medicare amount fields for the claim. Line item records were discarded after these operations were performed.

- Procedure Dates,
- Procedure Modifiers,
- Medicare Coinsurance Amount
- Medicare Deductible Amount.

The AMT\_PAID filed was missing for all encounter records. As a result, costs could not be assigned for these records.

The Washington Medicaid eligibility file contained information on a total of 380,078 persons and included a significant amount of information on each person. Eligibility records contained "header" information with name and demographics fields, up to 24 segments of eligibility information, and up to 12 segments of managed care information. The data were processed to create a SAS data set with one observation for each eligibility period that included some part of 1996. This data set contained a total of 6,586,762 observations on 380,078 Medicaid IDs. During the final data base build this count was reduced to 100,429 observations for 41,566 client IDs. As with the claims and encounter data, all dates were recorded as the number of days since January 1, 1964, and a conversion was required to produce SAS dates. Values of zero were recoded to "unknown" (.A). All records with a Medicaid ID were processed.

Many Medicaid services received from the State contain unusable Medicaid ID numbers. A total of 740,014 different erroneous IDs were found. These IDs could not be matched with eligibility data, and no information about the recipient could be determined. Some of these claims are MH/AOD related<sup>2</sup>, however, and were included on the data base. These claims are distributed across four of the Medicaid service detail data sets:

- Inpatient: 35,981 claims
- Long Term Care: 54,046 claims
- Pharmacy: 2,249,129 claims
- Outpatient/Other: 5,781,536 claims

During processing of the data, certain characteristics were observed that were flagged as either potential data anomalies or simply interesting findings. They are as follows:

- There seems to be a slight shortfall of claims occurring during the last two months of 1996.
- Washington Medicaid data contains neither Medicare coinsurance nor Medicare deductible dollar amounts.
- No inpatient encounter claims were found.
- Most (83%) of the Long Term Care claims do not contain a diagnosis code.
- Long Term Care claims do not contain level of care information.
- The most frequently found diagnosis code on the outpatient and other medical services data is V990, which is not a valid ICD-9 code.

## **MENTAL HEALTH DATA**

The Mental Health Division of the Washington Department of Social and Health Services, delivered two sets of data files for use in the construction of the IDB. The first set of data covered State inpatient

---

<sup>2</sup> As determined by a MH/AOD diagnosis, procedure, service type, provider type, or revenue code.

psychiatric hospitals. The second collection of data covered outpatient and community inpatient mental health services. Since the issues reported here are specific to the data source, this section of Chapter 5 is divided into the following three subsections:

- State Inpatient Psychiatric Hospitals
- Community Inpatient Hospitals
- Outpatient Mental Health Services.

## **State Inpatient Psychiatric Hospitals**

### *Data Sources and Processing*

The first set of data we received covered State inpatient psychiatric hospitals and was comprised of the following five data base table extracts:

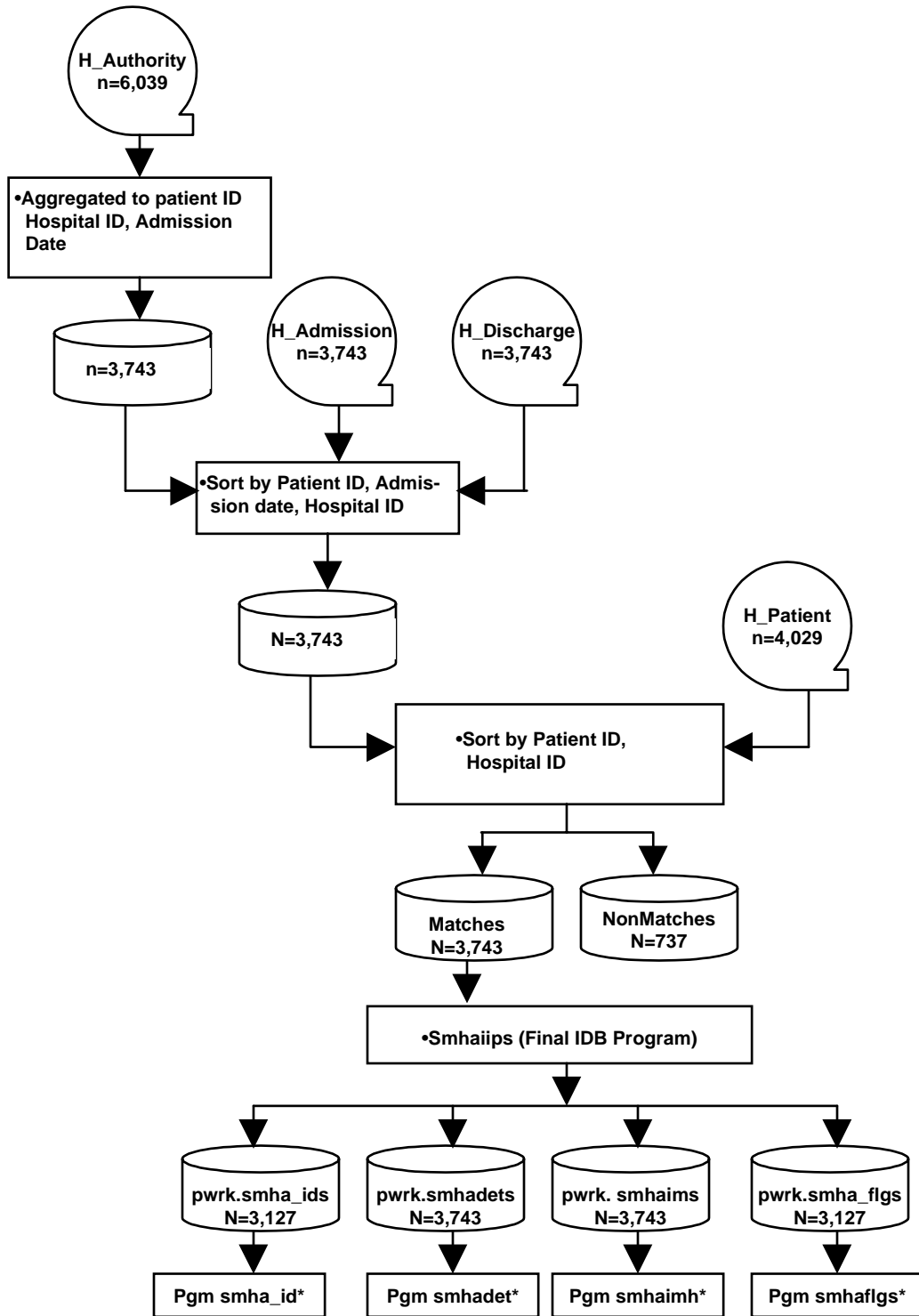
- H\_Patient: patient-level data (N = 4,029)
- H\_Admission: data on individual admissions (N = 3,743)
- H\_Discharge: data on individual discharges (N = 3,743)
- H\_Authority: This table contains a collection of information relevant to the legal authority for WA State Hospital patients.<sup>3</sup> There were a total of 6,039 records in this table, which was reduced to 3,743 records when aggregated to the Patient ID, Hospital ID, Admission Date level.
- H\_Movement: all movement records for the admissions identified in H\_Admission (N = 29,736)

The IDB build from SMHA State inpatient psychiatric hospital data includes information from the first four source files listed above. A flowchart of the major data merging steps undertaken during the IDB build is presented in Figure 19. The flowchart shows the source file and number of observations in the file, the sort key used to perform the merging of data, when applicable, and notes about the merge results, if applicable. These data represented 3,743 hospital stays for 3,127 clients.

---

<sup>3</sup> This table contains data elements that provide information about: (a) under what legal authority a patient was committed (e.g., criminal insanity, voluntary commitment, commitment for sexual psychopathy, etc.); (b) the begin/end dates of this authority; (c) the reason for commitment (e.g., gravely disabled - health and safety needs not met; serious harm - risk to self/others, etc.); and (d) hospital ID number.

**Figure 19 Washington State Inpatient Psychiatric Hospital Data Processing**



### ***Application of Costs***

The Department of Social and Health Services provided documentation that listed the daily rates for each State inpatient psychiatric hospital. As a result, the algorithm to assign total costs for each 1996 stay simply computes the number of days a patient was hospitalized during 1996 (non-1996 days are excluded), and multiplies the result by the daily rate for the corresponding facility. The day of admission is excluded from the algorithm per Washington State policy. In the rare case when the admission and discharge dates for a hospital stay are equal to one another, the algorithm assigned a "1" to the number of days hospitalized data element.

### ***Exclusions***

The Department of Social and Health Services prepared the data files using the criteria established for the project. They provided data for clients who were admitted during calendar year 1996 or who were admitted before January 1, 1996 but were still a patient on that date. Accordingly, no State inpatient psychiatric hospital utilization data delivered to MEDSTAT were excluded due to admission criteria being out of range of the specifications. Review of the State Inpatient Psychiatric Hospital flowchart, however, shows that 737 patient-level records, from file H\_Patient, were excluded from the IDB build because they did not have a corresponding hospital utilization record.

## **Community Inpatient Hospitals**

### ***Data Sources and Processing***

MEDSTAT received the following three community inpatient data files:

- CHD.DAT: community inpatient hospital recipient demographic information (N = 40,945)
- CHPROV.DAT: community inpatient hospital provider data (N = 40,945)
- CHSVC.DAT: community inpatient hospital recipient service data (N = 40,945).

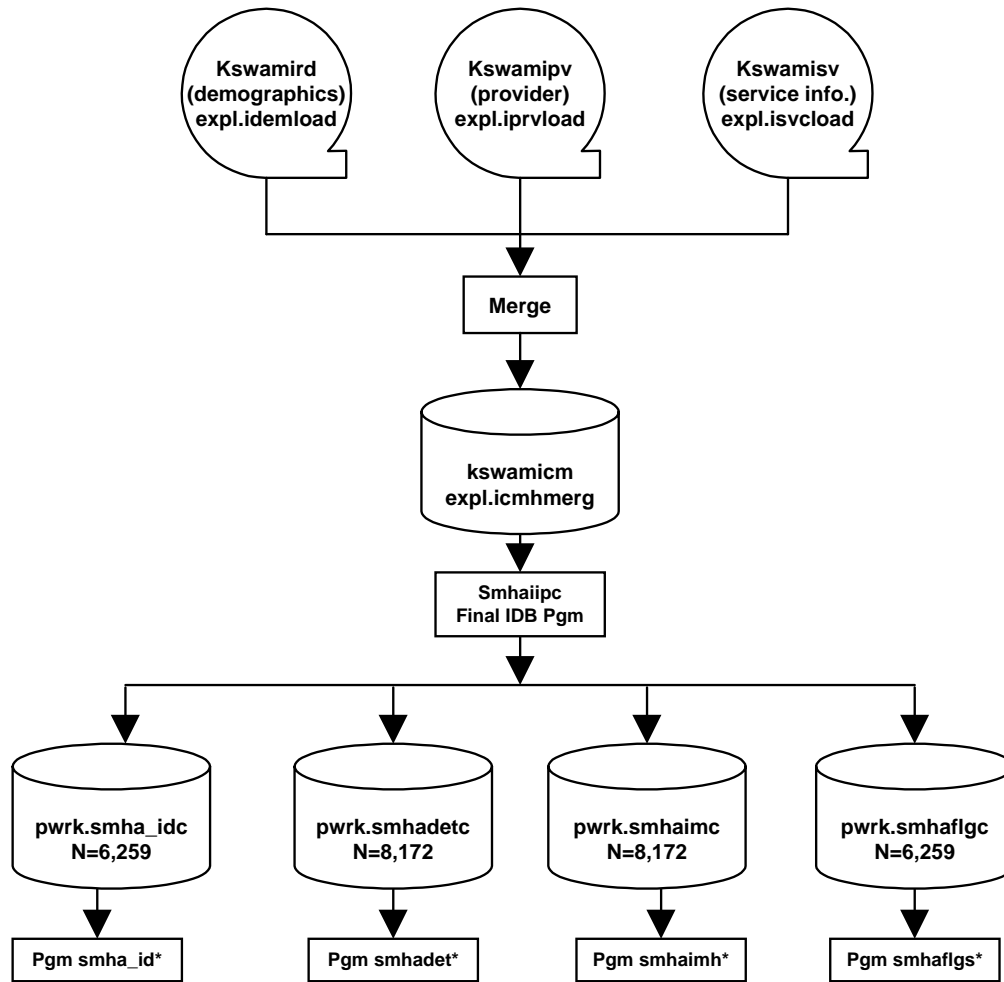
The IDB build from the community inpatient hospital data utilizes information from all three files listed above. Each of the files were sorted and merged by client PIC number and admission date. There were no mismatches of data. These data represented 8,172 hospital stays for 6,259 clients.

Four of the component files of the IDB are derived from community inpatient hospital data. They are:

- client identifying data,
- client detail data,
- institutional mental health services,
- client treatment/service diagnosis flags.

Figure 20 below shows the processing steps performed in the initial transformation of the Washington community inpatient hospital data.

**Figure 20 Washington Community Inpatient Hospital Processing**



\*Please refer to flowchart entitled: "SMHA File Interleave Programs"

***Application of Costs***

The Washington community inpatient hospital data contains financial information fields. Consequently, the process used to assign total costs to each 1996 stay is to simply copy the value from the "Remit Amount" field.

***Exclusions***

The Department of Social and Health Services built the data files using the criteria established for the project (i.e., they provided data for clients who were admitted between January 1, 1996 and December 31, 1996). Accordingly, no data delivered to MEDSTAT were excluded due to admission criteria being out of range of the specifications. However, it is possible that there were admissions for patients who were admitted before January 1, 1996 but were still a patient on that date. A frequency distribution of admission year by discharge year failed to show any hospital stays that met this criterion. It is likely that the IDB is missing a small percentage of such stays.

## **Outpatient Mental Health Services**

### ***Data Sources and Processing***

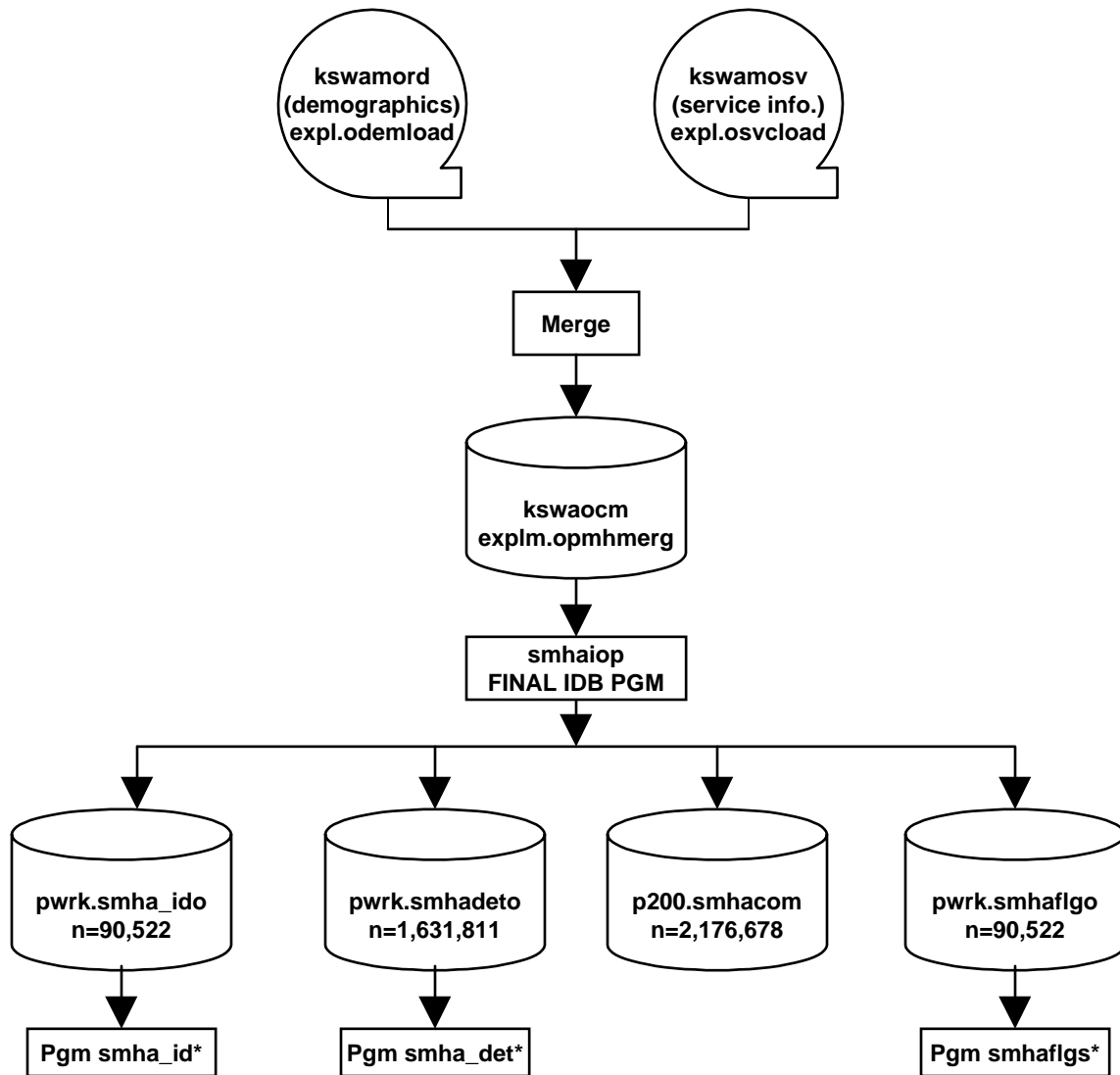
Washington outpatient mental health data were provided in two data files:

- D.DAT: outpatient mental health client demographic data (N = 97,608; N = 90,521 when duplicate Client IDs were eliminated)
- S.DAT: outpatient mental health service data (N = 2,176,678).

The IDB build from SMHA outpatient mental health data includes information from both source files listed above. MEDSTAT performed a simple merge by Client ID of these two outpatient files. Sixteen outpatient service records did not have a corresponding demographic record. Currently, these 16 records are included in the IDB, although the ability to link these data was severely restricted since there is no associated demographic data.

Four of the component files of the IDB are derived from community outpatient data. They are: client identifying data, client detail data, institutional mental health services, client treatment/service diagnosis flags. As Figure 21 shows, the merged file was used to produce the appropriate set of IDB data elements and number of records required for each component.

**Figure 21 Washington SMHSA Community Outpatient Programs**



***Application of Costs***

The Department of Social and Health Services provided a schedule of maximum allowances for community outpatient mental health services. Each outpatient treatment modality had a billing time associated with it. In addition, two sets of rates and procedure codes were assigned to each treatment modality. For example, Individual Treatment Services are billed per quarter hour. Procedure code 0276M is used to estimate costs for “In Facility Individual Treatment Services” at a rate of \$17.69 per quarter hour; while procedure code 0277M is used to estimate costs for “Out Facility Individual Treatment Services” at a rate of \$25.75 per quarter hour. Unless otherwise noted, the algorithms for total estimated 1996 costs for outpatient mental health services included the following steps:

1. Inspect procedure code value to determine proper rate to assign.
2. Inspect service minutes field and adjust it, if necessary, to quarter hour or hourly billing unit.
3. Multiply the outcome of Step 2 by the appropriate rate for each procedure code.



There are three procedure codes for which appropriate cost estimates could not be performed. These are “per person” services, “per event” services and a catch-all procedure code. Although there is rate information associated with these services, the comparable billing-unit information from the original individual outpatient treatment records was lost when the data were aggregated to monthly records. Consequently, it is impossible to discern the proper unit values to feed into the cost algorithm. The services for which cost information will be unavailable are:

- Medical Management – Individual (per person)
- Medical Management – Group (per event)
- Other Direct Services (catch-all).

Further, cost information is also unavailable for those outpatient service records with zero service minutes.

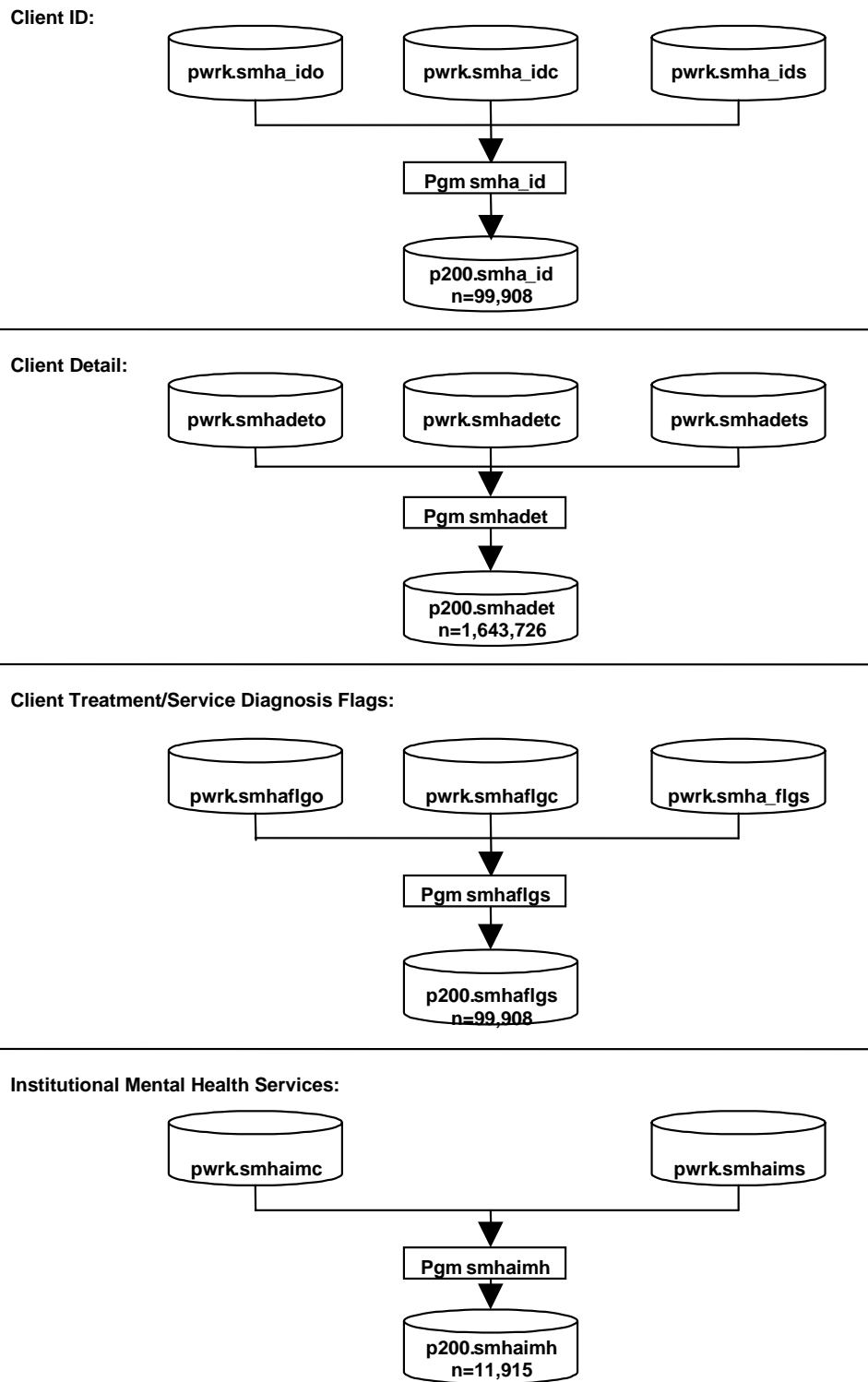
### ***Exclusions***

The Department of Social and Health Services provided the data files to be used for the IDB using the criteria established for the project (i.e., they provided data for clients who were served between January 1, 1996 and December 31, 1996). Accordingly, no data delivered to MEDSTAT were excluded due to service date criteria being out of range of the specifications.

### **Combining the Mental Health Data**

Figure 22 below shows how the output files from the outpatient mental health services data processing were interleaved with the corresponding output files from the State inpatient psychiatric hospital data processing and the community inpatient mental health data processing. The four interleave steps displayed in Figure 22 produce the final IDB component files that embody Washington’s mental health program data.

**Figure 22 Washington Mental Health File Interleave**



**DEPARTMENT OF ALCOHOL AND SUBSTANCE ABUSE (DASA)**

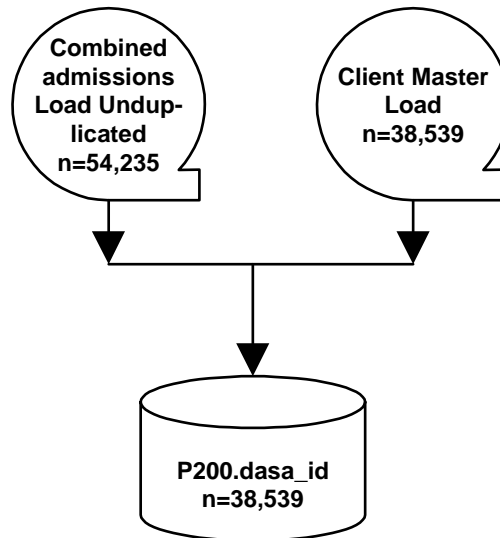
**Data Sources and Processing**

The Washington Department of Alcohol and Substance Abuse provided the following seven data files to be used in the construction of the IDB:

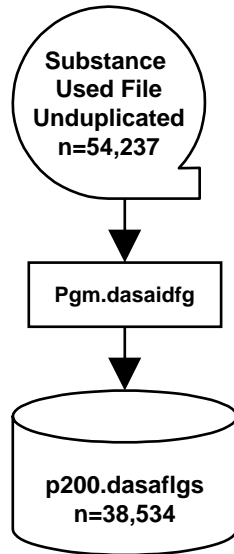
- Combined Admission, Admission Treatment Milestone, and Discharge (N = 54,245)
- Client Master (N = 38,539)
- Substance Used (N = 160,576; N = 54,237 when aggregated to client ID, admission date, provider number level)
- Service Funding (N = 59,385)
- Treatment Activity (N = 596,739)
- Facility (N = 228)
- Facility Services (N = 2,969).

The IDB includes information from the first five source files listed above. Flowcharts of the major data merging steps undertaken to construct each component file of the IDB build are included in this section (Figure 23 through Figure 26 below). These flowcharts show the source files and the number of observations they contain, the sort key used to perform the merging of data (when applicable) and the output files and corresponding number of observations.

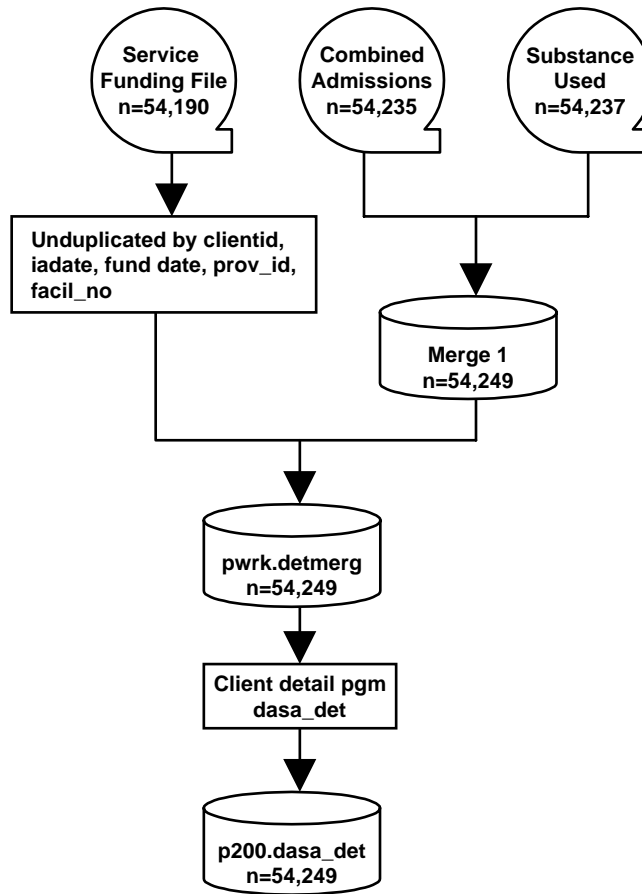
**Figure 23 Washington DASA Client ID File**

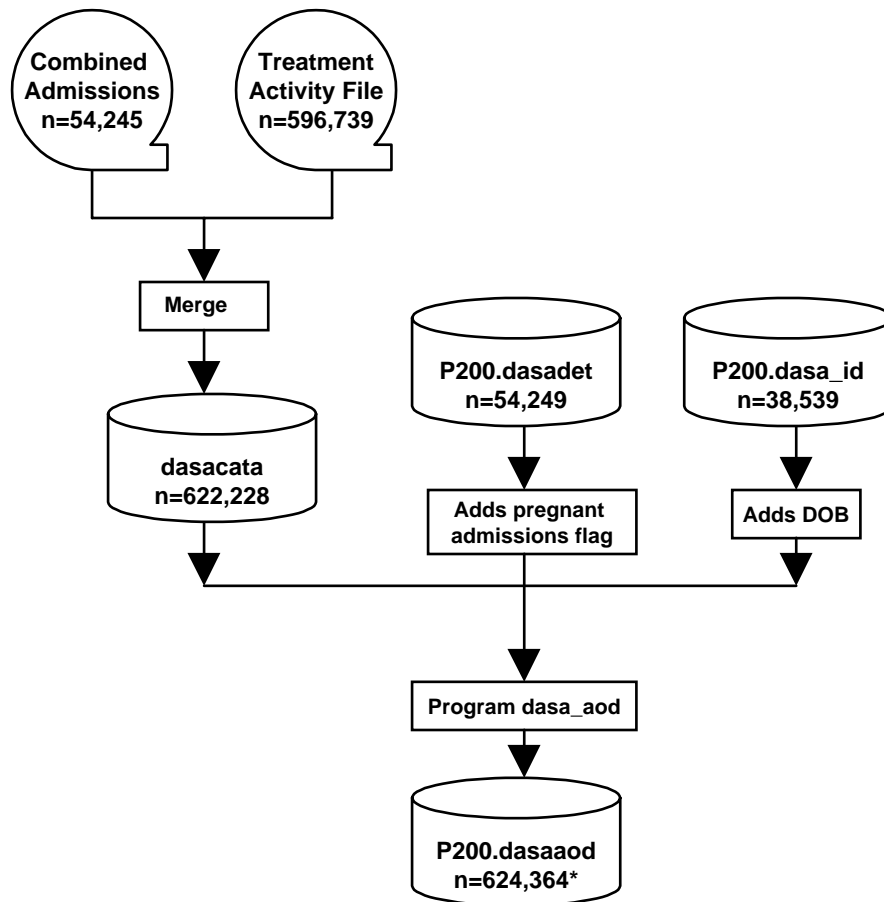


**Figure 24 Washington DASA Treatment/Service Diagnosis Flags**



**Figure 25 Washington DASA Client Detail File**



**Figure 26 Washington DASA AOD Treatment Program**

\*The additional 2,136 AOD treatment records represent bundled methadone maintenance records generated by The MEDSTAT Group to capture methadone maintenance costs.

## Application of Costs

DASA provides a comprehensive and varied set of services to its clients. These services range from ambulatory and outpatient services (e.g., a variety of different types of counseling services, acupuncture, methadone maintenance, etc.) to varied residential and institutional services. In addition, DASA administers special programs to serve high-risk populations such as pregnant and parenting women.

Each of these different services required the development of unique cost algorithms in order to accurately estimate the 1996 costs associated with each type of service. This section begins with a description of the general principles used in the DASA cost algorithms, followed by more detailed information regarding the unique cost algorithms developed for DASA data.

The following general principles were used in the development of DASA cost algorithms:

- All algorithms calculate costs for 1996 services only. Consequently, if a client was admitted to a residential facility prior to January 1, 1996 the cost algorithm computes the number of 1996 treatment days starting with January 1, 1996. Similarly, if a client enrolled in the

methadone maintenance program prior to January 1, 1996 the cost algorithm only counts treatment days starting with January 1, 1996. The same logic is applied to those services that continue forward into 1997.

- For residential treatment services with four or more hours of service on either the day of admission or discharge, the treatment period was set equal to a full day and the daily rate was applied.
- The cost was set to missing for residential treatment services with missing or blank discharge dates. For these records, the cost has a value of “unknown” assigned to it.
- Any DASA records containing a treatment modality of transitional housing were assigned a cost value of “unknown”.
- Any DASA records with a treatment activity type of “case management” were assigned a cost value of “unknown”.
- Any outpatient or intensive outpatient treatment modality records from the combined admissions file that do not have corresponding outpatient treatment activity records were assigned a cost value of “unknown”. This situation occurred for DASA clients who were admitted to an outpatient treatment program but never received services. This situation may also occur for DASA clients that enroll in an outpatient treatment program near the end of 1996, but do not receive services until early 1997.

Where appropriate, MEDSTAT modified the general routines to accommodate specific scenarios and situations. These are described below:

- Youth Detox/Crisis Stabilization: DASA provided an average weighted daily rate of \$172.09 for services provided by contracted youth detox/stabilization facilities. MEDSTAT created a “youth\_dx” data element that was set to true for clients residing in any of the seven facilities that appeared on the list provided by DASA. If a client had a treatment modality equal to “DX” (detox) and the youth\_dx data element was set to true, then the total estimated cost for 1996 was set equal to the number of 1996 days of treatment multiplied by \$172.09.
- Youth Residential Services: DASA provided a weighted average daily rate of \$95.09 for services provided by youth residential facilities. DASA instructed MEDSTAT that clients under 18 years of age would be considered “youth.” Accordingly, MEDSTAT computed client age using either the beginning date of service or January 1, 1996 for those youth clients whose treatment started just prior to the beginning of 1996. The cost algorithm determines whether the treatment is residential or not, using the treatment modality data element. If the treatment modality designates residential care and the client was identified as “youth”, then the total estimated cost for 1996 was set equal to the number of 1996 days of treatment multiplied by \$95.09.
- Pregnant and Parenting Women (PPW) Residential Services: DASA provided an average weighted daily rate of \$97.61 for Pregnant and Parenting Women residential services. MEDSTAT inspected the values from the data element FPREG (Pregnant on Admission Date) to determine whether the client was a pregnant or parenting woman. If a PPW client had a treatment modality reflecting residential or long-term residential services, then the total estimated costs for 1996 was set equal to the number of 1996 days of treatment multiplied by \$97.61.

Adult Residential or Detox Services: DASA provided a rate sheet that displayed rates MEDSTAT should apply to the following set of residential or detox services that are rendered to adult DASA clients:

- Intensive inpatient
- Differential diagnosis
- Long-term residential
- Detox
- Recovery house
- Extended care
- MICA residential

DASA instructed MEDSTAT that clients aged 18 years or older should be classified as “adults.” MEDSTAT computed client age using either the beginning date of service or January 1, 1996 for those adult clients whose treatment started just prior to the beginning of 1996. The treatment modality data value was inspected next to determine which DASA rate should be used in the cost algorithm. If the client was classified as an adult and the treatment modality data element had a value reflecting residential or detox services, then the total estimated cost for 1996 was set equal to the number of 1996 days of treatment multiplied by the daily rate corresponding to the DASA service.

Methadone Maintenance: DASA clients enrolled in the methadone maintenance program are entitled to a set of services that include:

- Counseling services (e.g., individual, group, conjoint, family, etc.)
- Urinalysis
- Acupuncture
- Methadone dose change.

These services are bundled into a daily rate. DASA provided the daily rate of \$9.15 for methadone maintenance provided from January 1, 1996 through June 30, 1996. A daily rate of \$9.33 was provided for methadone maintenance treatments provided from July 1, 1996 through December 31, 1996.

The Alcohol and Other Drug (AOD) Treatments component of the IDB includes treatment level data. Methadone maintenance is a program rather than a treatment. However, in order to represent the total bundled costs for methadone maintenance in the AOD IDB component, MEDSTAT generated a bundled methadone maintenance treatment record to capture the total costs. The methadone maintenance cost algorithm inspects the program admission and discharge dates to determine the total number of 1996 days of treatment. The following 3 scenarios could occur for clients enrolled in DASA’s methadone maintenance program:

- All 1996 days of treatment occur in the first half of the year, between January 1 and June 30, 1996.
- All 1996 days of treatment occur in the second half of the year, between July 1 and December 31, 1996.
- The 1996 days of treatment span both halves of 1996.

MEDSTAT computed 2 cost data elements, then summed them to arrive at the total estimated 1996 costs. The COST1 data element calculated costs using the daily rate and number of treatment days that fall in the first half of 1996. The COST2 data element calculated costs using the daily rate and number of

treatment days that fall in the second half of 1996. The COST1 and COST2 algorithms simply multiplied the number of treatment days during the corresponding half-year period by the appropriate daily rate.

DASA clients enrolled in methadone maintenance may require use of childcare services. Estimated costs for 1996 childcare services are calculated using either a \$4.50 hourly rate or a \$28.00 daily rate. MEDSTAT inspected the number of hours in the quantity of service field to determine total childcare costs. Partial hours of childcare were rounded up to the next largest whole hour. For quantities in excess of 24 hours, MEDSTAT calculated costs using the daily rate first, and then multiplied any residual hours by the hourly rate to determine the grand total estimated costs.

Outpatient Treatment Services: DASA provided a rate sheet that displayed rates MEDSTAT should apply to the following set of outpatient services that are rendered to DASA clients:

- Individual, family or conjoint counseling (rates were provided for both brief or full visits)
- Group counseling
- Child care services
- Acupuncture (for clients that are not enrolled in the methadone maintenance program).

For each of the above outpatient services, there was a rate for treatments provided between January 1 through June 30 and another rate for treatments provided between July 1 and December 31. MEDSTAT used the DASA treatment activity type and event duration data elements to determine the total estimated costs for these 1996 outpatient services. MEDSTAT assigned an “unknown” cost to those outpatient treatment records with a service quantity of zero.

For individual, family, or conjoint counseling, MEDSTAT had to first determine whether the visit was a brief visit or a full visit. DASA defines a brief counseling visit as having duration of 1 through 30 minutes. A full counseling visit is from 31 through 60 minutes. MEDSTAT determined the status of the visit (i.e., brief or full visit), then simply assigned the appropriate visit rate to the treatment record. There were a small number of counseling treatment records that had an event duration in excess of 60 minutes. For these counseling treatment records, we rounded the event duration to the next highest whole hour, then multiplied that result by the proper full visit rate, depending on whether the visit occurred during the first half or second half of 1996.

For acupuncture services provided to DASA clients not enrolled in the methadone maintenance program, we assigned the cost by examining the treatment date and determining whether the acupuncture occurred during the first half or second half of 1996.

Estimated costs for 1996 childcare services were calculated using either a \$4.50 hourly rate or a \$28.00 daily rate. MEDSTAT inspected the number of hours in the quantity of service field to determine total childcare costs. Partial hours of childcare were rounded to the next highest whole hour. For quantities in excess of 24 hours, MEDSTAT calculated costs using the daily rate first, then multiplied any residual hours by the hourly rate to determine the grand total estimated costs.

Costs for group counseling services are accumulated for each quarter hour of counseling. In order to calculate the total estimated costs for each 1996 group counseling treatment record, MEDSTAT determined how many quarter hours were represented by the hourly event duration. That value was multiplied by the appropriate group counseling rate, depending on whether the counseling occurred during the first half or second half of 1996.



## Exclusions

The Department of Social and Health Services built the data files for the IDB using the criteria established for the project (i.e., they provided data for clients who were served between January 1, 1996 and December 31, 1996<sup>4</sup>). Accordingly, no data delivered to MEDSTAT were excluded due to admission criteria being out of range of the specifications. However, DASA assessment records were not included in the data delivered to MEDSTAT. These assessment records will be included in the 1997 data delivery.

---

<sup>4</sup> Exclusion Criteria: Admissions prior to 01/01/1996 were excluded using the following criteria: Cases with no activity within 60 days of 01/01/1996 for GC, IO, MT, OP, MO; 15 days DX; 30 days DD; 40 days II (Y-Contract 100); 90 days RH, TH; 195 days EC, LT. This information appeared in the data delivery letter to MEDSTAT.



## CHAPTER SIX: PERSON SUMMARY FILE

One of the known limitations of the original data base specifications and prototype data base is the requirement that several files be joined or merged in order to conduct some of the more common preliminary investigations and analyses. To facilitate these efforts, a person summary file (PSF) has been constructed and included as part of the final IDB. This file provides aggregate utilization (counts of services or days), expenditures, and Medicaid enrollment information (including number of months of enrollment) at the client level. To create the file, a series of five processing steps were utilized to abstract the information from each of the eleven files contained in the data base (see Figure 27).

To ensure confidentiality, client zip code information is converted to an urban/rural FIPS code value obtained from the U.S. Department of Agriculture web site.<sup>5</sup> These rural-urban continuum codes are part of a classification scheme for all U.S. counties by degree of urbanization and proximity to a metropolitan area. A description of the continuum and codes used for the three States can be found in Appendix D.

### DEVELOPMENT

The person summary file is based on a fairly simple design and consists of data elements that summarize a person's service history in terms of total payment (and cost) amounts, total number of services received and a limited number of data elements describing characteristics of the client. Payment amounts, costs and counts of services were derived using the uniform services categories by creating "buckets" for each month and combination of service and allocating amounts and service counts to these buckets. In the course of creating these files, several issues came to our attention that resulted in a more complex approach to aggregation of services:

- The need to differentiate between MH/AOD and other types of medical costs/services such as lab work, preventive services and other non-MH/AOD service.
- The need to differentiate between Medicaid MH/AOD costs/services and MH/AOD State agency costs/services.
- The need to differentiate between counts of services for Detox provided by any payer and counts of services for any other service including MH/AOD.
- The need to differentiate between Medicaid FFS payment amounts and Medicaid managed care payment amounts. Since dollar amounts present in encounter data records are representative of the amount paid by the HMO to the provider rather than actual Medicaid amounts, only FFS Medicaid payment amounts are included.
- In contrast to Medicaid FFS payment amounts, number of services provided by Managed Care organizations needs to be included but counted separately from FFS services. This allows rates to be accurately calculated using the FFS counts.

In order to capture all possible combinations, a total of 912 cells or 'buckets' are created in STEPPSB2 (see Figure 27) of the processing stream for the PSF. Subsequent processing summarizes the total counts of services, dollars, and costs into data base variables and drops the 912 temporary buckets from the final PSF.

It is important to note that if a client has multiple records in the Client Detail File, PSF data elements such as county information, marital status and legal status are taken from the most recent, non-missing

---

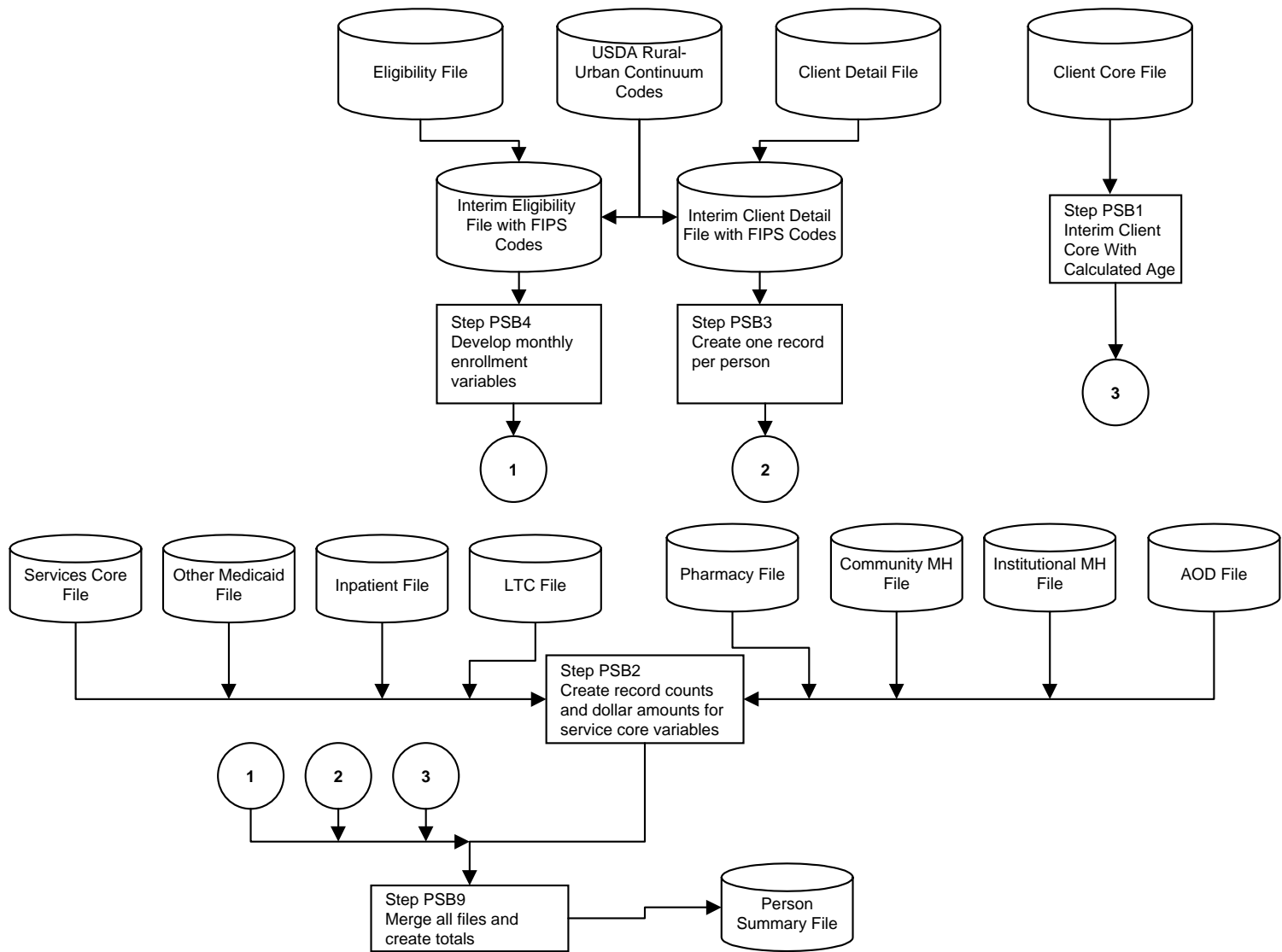
<sup>5</sup>This information can be found at <http://www.econ.ag.gov/briefing/rural/data/code93.txt>.

information. The exception to this rule is information regarding any prior AOD history information, which is retained in the PSF from the first record indicating evidence of the treatment or problem.

Eligibility information is available at the month level, as is managed care information and county information when provided by the States. A matrix of data element availability by State can be found in Appendix E. Detailed counts of services and dollar amounts for each State can be found in Appendix F.

In addition to commonly used service information, we include Severe Mental Illness (SMI) and Severely Emotionally Disturbed (SED) indicators on the file even though Oklahoma was the only State to incorporate these data elements. Future versions of the data base may include an algorithm for deriving these indicators that could be applied to clinical values across all States.

**Figure 27 Overview of Processing for Person Summary File**



## **CHAPTER SEVEN: POTENTIAL LIMITATIONS AND TENTATIVE PLANS FOR IMPROVEMENT**

As with any research project of this scope, decisions were made regarding methodological and other issues that may affect the types of analyses that can be conducted, and the generalizability of the results. This section describes the known limitations of the data base, and preliminary strategies for mitigating the impact of these limitations.

### **OVERVIEW**

The IDB has been designed to a resource allowing analysis and reporting at the individual source file level, across the files comprising the IDB, and at an aggregate personal summary level file. To this end, data from various sources within three States were transformed and standardized into a common format while preserving as much of the original source data detail as possible. A decision was made to preserve source data (e.g., claims and eligibility information) at the most detailed level possible. This allows for a wide range of potential views and analyses of the resulting data base. An implication of this choice is that some of the more common types of analyses will require preliminary manipulation of the data structures (e.g., merging or joining data sets) prior to any analytical programming.

As an example, information regarding outpatient claims is stored in one IDB file whereas inpatient claims and demographic data are stored in two other files. In order to obtain demographic information about persons with particular inpatient and outpatient characteristics, data from all 3 files must be merged prior to analysis. Although at first glance this may seem inefficient, the approach when all factors are considered is actually very desirable. There are severe drawbacks to creating either a single, very large file or to creating a series of files that include all possible combinations of possible merges (e.g., one summary for inpatient and outpatient, another for inpatient, outpatient and pharmacy, etc.)

Some research objectives allow working with data at an aggregate level as a means of expediting analysis. In order to facilitate such analyses, we are delivering an IDB person summary file (PSF). The PSF consists of a set of commonly accessed data elements summarized by person. The person summary record contains all basic demographic information such as age, sex and race and aggregate enrollment information in the form of number of months of enrollment and the eligibility group in which the enrollee was enrolled for the longest period during the year. In addition, it contains annual utilization information such as number of outpatient visits, inpatient discharges, inpatient days and total Medicaid expenditures by service category. See Chapter Six for a more detailed description of the Person Summary File.

This chapter includes a discussion of data base limitations and a description of recommended exclusions for research purposes. We strongly recommend that potential users of the IDB read this chapter in order to acquire a basic familiarity with the extent of data differences across States and record types. An understanding of these subjects is critical for proper use of the data base. A summary of some specific limitations follows.

### **SPECIFIC LIMITATIONS WITH RESPECT TO DATA AND DATA BASE DESIGN**

#### **Costs for Mental Health and Alcohol and Other Drugs**

Each State provided estimated cost rates for various levels of non-Medicaid MH/AOD services. If cost detail was not available on source records, and in a significant percentage of cases it was not, then the State rate estimates were used to impute costs. We consider these rates to be estimates rather than true costs incurred.

### **Overlapping Services between Medicaid and State MH/AOD Agencies**

In many instances, services provided by the State MH/AOD agencies overlap with services provided by Medicaid. Dollar amounts for these services may also be (legitimately) reported on both State agency files and Medicaid files. Therefore Medicaid dollars cannot be added to State Agency costs to derive a total cost for MH/AOD spending. Adding the two costs together amounts to double counting and may artificially inflate the cost estimates. We recommend that the dollars reported by MH/AOD be used to make general estimates of higher or lower costs for services.

### **Medicaid Encounter Records**

For many States, the collection of encounter data was still in its infancy in 1996, and many States have been unable to collect valid and complete information. For this reason researchers may choose to exclude encounter records during analysis. We have included encounter records on the relevant service files because for some analyses they are crucial. Researchers should proceed with the understanding that many encounter record data elements are missing or lack specificity and should conduct their own data validity assessments specific to their particular research goals.

### **Payment Amounts on Encounter Records**

Medicaid payment amounts are included on encounter records if a dollar amount was provided by the State. However these amounts generally represent payments made to the provider by the HMO, not the actual amount Medicaid paid for the service. Therefore they should not be included as part of aggregate summarization of Medicaid costs.

### **Medicaid Crossover Recipients**

The quality of data on Medicaid crossover claims (those clients enrolled in both Medicaid and Medicare and have at least one claim paid by both payers) is often incomplete. Researchers will want to assess data completeness for these individuals and may want to exclude them from studies.

### **‘F’ Flags Set for Mental Health and Alcohol and Other Drug Use**

These flags (e.g., FALCABSE, FSCHIZO) are carried on the client core file and represent whether a client has had a mental health or substance abuse diagnosis at any time during the year. A record is flagged if the client has a relevant primary or secondary diagnosis. For studies limited to people being treated for a MH/AOD condition based only on primary diagnosis, the variables AOD\_CLM and MH\_CLM from the service core file should be used.

### **Inconsistencies on Files**

There are several instances where conflicting information occurs. For example, there are occurrences of clients with gender equal to ‘male’ that also have the flag set for a pregnant condition. Although every attempt was made to check for this type of inconsistency and validate our processing efforts, we did not go so far as to change values obtained from our data sources and these values remain on the files. See individual State sections for further discussion.

### **Records in the Client Core File with No Corresponding Service Records**

Some State agencies provided files containing records for clients that did not receive services in 1996. We have chosen to include these persons in the data base since we presume they have made some form of contact with the MH or AOD agency. Medicaid enrollees with no services were not included in the client core file.

### **Service Records with No Corresponding Client Record**

A small proportion of State agency service records did not have a corresponding client record. The same is true for Medicaid service records. These records are retained on the service files but are missing demographic information. They will probably prove unusable for research purposes but were included in the IDB in the interest of completeness.

### **Medicaid Inpatient Psychiatric/Long Term Care**

A decision was made to place inpatient psychiatric cases in the LTC file. This approach is similar to the one used in Medicaid 2082 reporting systems. The Medicaid Long Term Care File (LTC) includes multiple records per user of long term care services based on admission, discharge and interim assessment dates. This file contains inpatient psychiatric care, skilled nursing, and intermediate care facilities.

### **Assignment of MH/AOD records**

As we worked with the data, we found that services were reported somewhat differently in the mental health community as opposed to the AOD community. A typical Institutional Mental Health File consists of multiple records per user based on admission into a Mental Health Institution. By contrast, a typical AOD Alcohol and Drug Service File includes multiple records per user based on utilization of AOD services. To report the services for each, the non-institutional MH/AOD and the institutional MH/AOD files have been replaced with the following three files:

- Institutional Mental Health File
- Alcohol and Drug Service File
- Community Mental Health File

Detox services provided by Medicaid often had no primary DX of MH/AOD. A decision was made that these dollars would go to Medicaid medical services, however, detox services would be counted. A general hierarchy for assigning costs to Medicaid services was determined to be as follows:

- If primary DX=AOD, the record goes into the AOD group (Detox or non-Detox)
- If primary DX=MH with Detox (count services), the record goes into the MH group
- If primary DX=MH with any other service, the record goes into the MH group
- If no primary DX of AOD/MH, the record goes into the medical group (we count Detox)
- For those included in the file strictly on the basis of procedures (and not diagnoses), we will allocate to MH/AOD if possible and count Detox services.

### **Medicaid Inpatient Line Items (Missing UB92 codes)**

Since data were submitted at the line item level and no inpatient line item file was created, line item data were rolled up to create discharge level summaries. Information from all line items is summarized into a single record for each inpatient stay. The current data base structure consolidates all inpatient charges for a claim to a specific single amount. Adding line item data would break down charges by specific services based on UB-92 revenue codes. Adding the line-item data would not require additional source data — the necessary data is included with the source Medicaid inpatient claims files, but it would significantly increase the size and complexity of the data base, which is why it has presently been deferred.



## **DATA BASE DESIGN RECOMMENDATIONS AND ENHANCEMENTS**

The current version of the IDB presents calendar year 1996 data and is now complete. However, the direction of subsequent IDB development is an open issue. Potential directions include extending the years of data for the three original States and/or adding additional cohorts to the data (i.e., social services, criminal justice, or school data). Participants at the Technical Expert Panel (TEP) meetings, particularly the November 1998 meeting, discussed these options and all were deemed valuable. These discussions quickly focussed on questions of treatment outcomes and the need for longitudinal data. A consensus emerged to extend the data base to additional years to create a longitudinal data base, rather than forming on a broader data base.

The recommendations that follow were submitted to SAMHSA on September 3, 1999 in the contract deliverable for Task 29 entitled "Proposal for Updating the Integrated Data Base and Feasibility Study (Contract # 270-96-0007).

### **Adding Data for Subsequent Years**

The addition of a second year of data to the data base will create the foundation of a longitudinal data base and provide the ability to analyze outcomes, identify new MH/AOD clients, and determine Medicaid enrollment patterns.

We recommend creating separate files for the 1997 data, rather than appending new data to the 1996 files. Appending to the 1996 files would create files that are too large and cumbersome to be of practical value. It would also necessitate the addition of identical data elements for 1997 to the 1996 data base, which in turn would greatly increase the amount of documentation needed and double the data storage requirements.

### **New or Upgraded Source Data**

The 1997 Delaware Child and Youth MH/AOD data for 1997 will be derived from Delaware's Family and Child Tracking System (FACTS) data base. This data base contains more information than was available with the 1996 data and technically represents a replacement data source rather than an additional source. Use of the FACTS data base will enhance the IDB by providing more detailed information on children's services sponsored by the agency.

Data provided by the Washington Department of Alcohol and Substance Abuse (DASA) for 1996 did not include client assessment data. This information was included with the MH/AOD data provided by both Oklahoma and Delaware and can be found in the IDB for both of those States. Assessment data is available from DASA as a separate file and Washington officials have expressed an interest in adding these data.

### **Additional Variables**

A minimal number of new variables are proposed for future versions of the IDB. These include Federal Medicaid eligibility categories, Maintenance Assistance Status (MAS), Basis of Eligibility (BOE) and case identification number on the Medicaid eligibility file, and provider ID code on the Medicaid service files. Of particular interest are DSM-IV DX codes from all MH sources and additional cost information. Also recommended are addition of standardized flags, such as severe mental illness (SMI) and severely emotionally disturbed (SED). ISMI and ISED are already used in Oklahoma. Planned changes for MH/AOD data include additional consistency checks for the AOD variables DRUG<sub>n</sub>, FREQ<sub>n</sub>, ROUTE<sub>n</sub>, and DAGE<sub>n</sub>. These apply to substance, frequency, route and age at first use, respectively.

### **Provider Files**

All services in the data base are associated with a provider identifier (SERV\_ID or BILL\_ID). This variable contains unmodified source values provided by the participating agencies. No other processing of provider information was done with the 1996 data, although provider information was received from all States and all agencies. In order to link services recorded on both Medicaid and MH/AOD files, it is necessary to link the providers by identifying information. For this reason, we propose the creation of a provider crosswalk table linking Medicaid and MH/AOD identifiers. Such a crosswalk table will require processing the provider data from each source and linking providers with both automated and manual methods. This is a significant amount of work. Our intent is to limit this crosswalk to a minimal amount of critical information. We may be able to apply the same general probabilistic linking techniques that we use for clients (albeit with different parameters) to the task of linking providers as a means of expediting the process.

For 1996, a number of observations were associated with Part H providers. It is not clear if these represent waiver services or something else. Additional documentation for these providers was also requested, and will be necessary in implementing crosswalks for future versions of the IDB.

## **CHAPTER EIGHT: DISCUSSION OF RESOURCE REQUIREMENTS AND UTILIZATION**

This section addresses the potential human and technical resources associated with implementing a variant of this IDB for a single State. It is important to note that this is not MEDSTAT's implementation plan. Our implementation plan will be outlined in a future deliverable for Task 26. Rather, this chapter is included as a guide to an organization that is considering implementation of an IDB project and needs to derive estimates of time requirements, staff requirements, levels of effort and other required resources.

In addition to the intrinsic requirement of the source data itself, there are three chief components to implementing the data base: staff resources, software, and hardware.

### **STAFF RESOURCES**

This section describes the approximate staff time necessary to create the data base for any State that may choose to do so. We assume that modifying and using the computer programs developed by MEDSTAT can minimize the level of effort. The level of effort will vary depending upon:

- the complexity of State source data and the level of familiarity users and designers have with the various data sources,
- the skill level of staff,
- the desired level of manual review in the linking process, and
- the attention paid to data quality and problem resolution.

General estimates for implementing a State-specific IDB are provided below. In preparing these estimates, we have assumed that the format of the source data is essentially the same as provided to MEDSTAT under the current contract and that States do not wish or do not need to significantly modify the programs and routines we have developed. Although some of the steps listed below may not be necessary for a given State, the steps reflect what we consider to be the most efficient order in which to perform the processing. Although it may be stating the obvious, it is worth noting that the entity building an IDB must obtain intra- and inter-agency data sharing agreements prior to embarking on such a project. It is our experience that these agreements can be difficult and time consuming to obtain. The estimates below reflect person-days, not elapsed time.

Level of Effort Estimates (per State):

- review and update crosswalks for data- and policy-related changes to key data elements (e.g., new provider type values): 5 to 10 days per data source
- preparation, loading, processing, and reviewing the Medicaid data: 15 to 30 days per State.
- preparation, loading, processing, and reviewing the MH/AOD data: 10 to 40 days.
- linking the Medicaid and MH/AOD clients:
  - ◆ with manual reviews: Eight days
  - ◆ without manual reviews: Four days
- selection of data base population: One day
- creating the data base files: Two days

- data quality review: Five days

At a minimum, implementation will require one mid-level SAS programmer, and potentially one programmer per data source. This person can load appropriate data files and perform the linking using the SAS routines developed by MEDSTAT. We strongly recommend one or more health care analysts be used as consultants for selecting the data base population and resolving data issues as they may arise.

## **SOFTWARE**

The chief application software component for this effort is SAS, which is widely used at the State level. Source code for all programs (including those for linking) is included as part of this report. SAS code is highly portable across different platforms and backward compatible across versions. We are highly confident that the code developed under this contract can be easily adapted and executed using versions of SAS at or above version 6.12. Although SAS is required to run the programs, the actual source code can be viewed in any text editor including Emacs, WinEdit, Notepad or BBEdit (or even Microsoft Word or Corel WordPerfect).

While SAS is a powerful and widely used language, we realize not all States will necessarily implement the IDB in SAS. To accommodate this, we have included comments to indicate the purpose or function of each code block. These comments could allow skilled programmers to convert SAS code to another programming language such as C/C++, Perl, or Pascal, or with more effort to a data base programming language such as Oracle PL/SQL.

We have also elected to use SAS as the native data storage format, and likewise foresee that not all States may wish to use SAS files to store and analyze data. Other analytic platforms such as SPSS are clearly viable, as are scalable relational data bases such as Sybase Adaptive Server, Oracle Server, Microsoft SQL Server or IBM's DB2.

Should States wish to implement a version of the online data dictionary similar to the one we have developed, they will also need Microsoft Access 97 or other data base. Table structures for the data dictionary are described in Chapter 9, and could be implemented using any standard relational data base.

## **HARDWARE**

States will need to evaluate whether a project of this size and scope can be accomplished using existing hardware resources or whether dedicated resources need to be acquired. A very significant advantage of the SAS approach we have taken is that the same source code can be executed on a wide range of hardware, from PC class computers through mid-range UNIX machines to very large IBM mainframes.

Clearly the scale of the hardware resources needs to match the volume of the data that will be manipulated. Our data base construction efforts on data supplied by Delaware, Oklahoma and Washington ran comfortably on a midrange UNIX machine with 120 GB of disk and 1 GB of RAM. Our efforts proceeded one State at a time and the largest jobs were so resource intensive that we could not run other jobs simultaneously. The most processor-intensive task was linking, which in some cases required several hours on a mid-range computer. States with a very small amount of data could conceivably perform data base construction on a large PC class machine with 256 MB of RAM and 20 to 40 (or more) GB of disk. The largest States, such as California, would potentially require a large-scale UNIX server or a mainframe class machine with an appropriately scaled amount of disk.

A second issue relates to the format of the source data. Media-specific hardware may be required to accommodate the range of input data formats. In the course of performing this contract, we received data on diskette, CD-ROM, and magnetic tape. In the future, data may also be received in other magnetic

formats such as DAT or DLT tape, other optical formats such as DVD and in electronic formats via a LAN/WAN or over the Internet through FTP or HTTP protocols.

### **OPPORTUNITIES FOR EFFICIENCIES**

Based on our experience and work with the three States, we anticipate the greatest opportunity for efficiency lies in the construction of crosswalks from source values to uniform values. While we anticipate that programmers within each agency will be familiar with their data, they are likely to be less well versed in the data standards used by other agencies. A significant amount of MEDSTAT's effort in creating the prototype involved the development of crosswalks and mappings to the uniform values. This is much more an analyst level task rather than a programming task, although both kinds of effort are ultimately required.

We recommend that States review the use of a joint task team prior to beginning work as a means of facilitating implementation of the data base. The rationale for this recommendation is that many of the issues the team are likely to face involve intra- and inter-agency data and definitions. Individuals within the various State agencies will have the in-depth knowledge of the source data that is crucial, which will aid in resolving crosswalk and data issues during data processing and review.



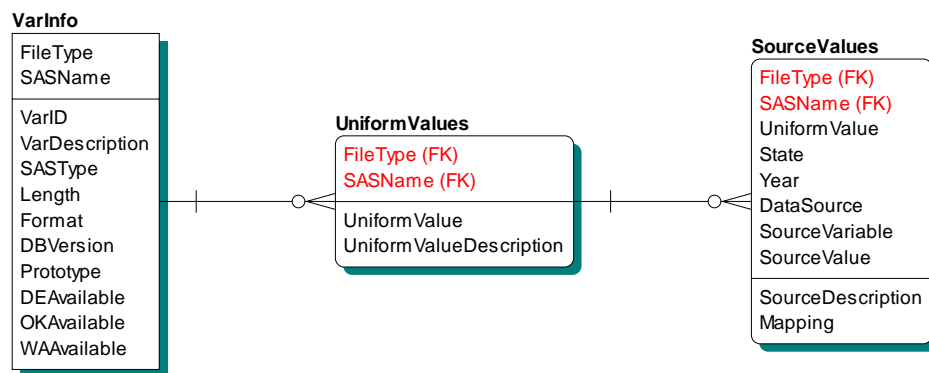
## CHAPTER NINE: DESCRIPTION OF DOCUMENTATION

### OVERVIEW

We had two primary goals in developing the data dictionary and documentation system: (a) to document crosswalks from source data to standardized data elements and uniform values, and (b) to make an easily updated documentation system available to a broad audience. To accomplish these two distinct goals, we developed one system for storing the metadata that are used to create the data dictionary and another for displaying the data dictionary.

The first component of the data dictionary consists of a simple Microsoft Access data base that stores the contents of the data dictionary. This data base includes three main tables, which store information about the 1) data elements, 2) uniform values, and 3) source data. A data entry component allows source values to be entered into the data base in a multi-user environment.

**Figure 28 Data Base Schema for Documentation Data Base**



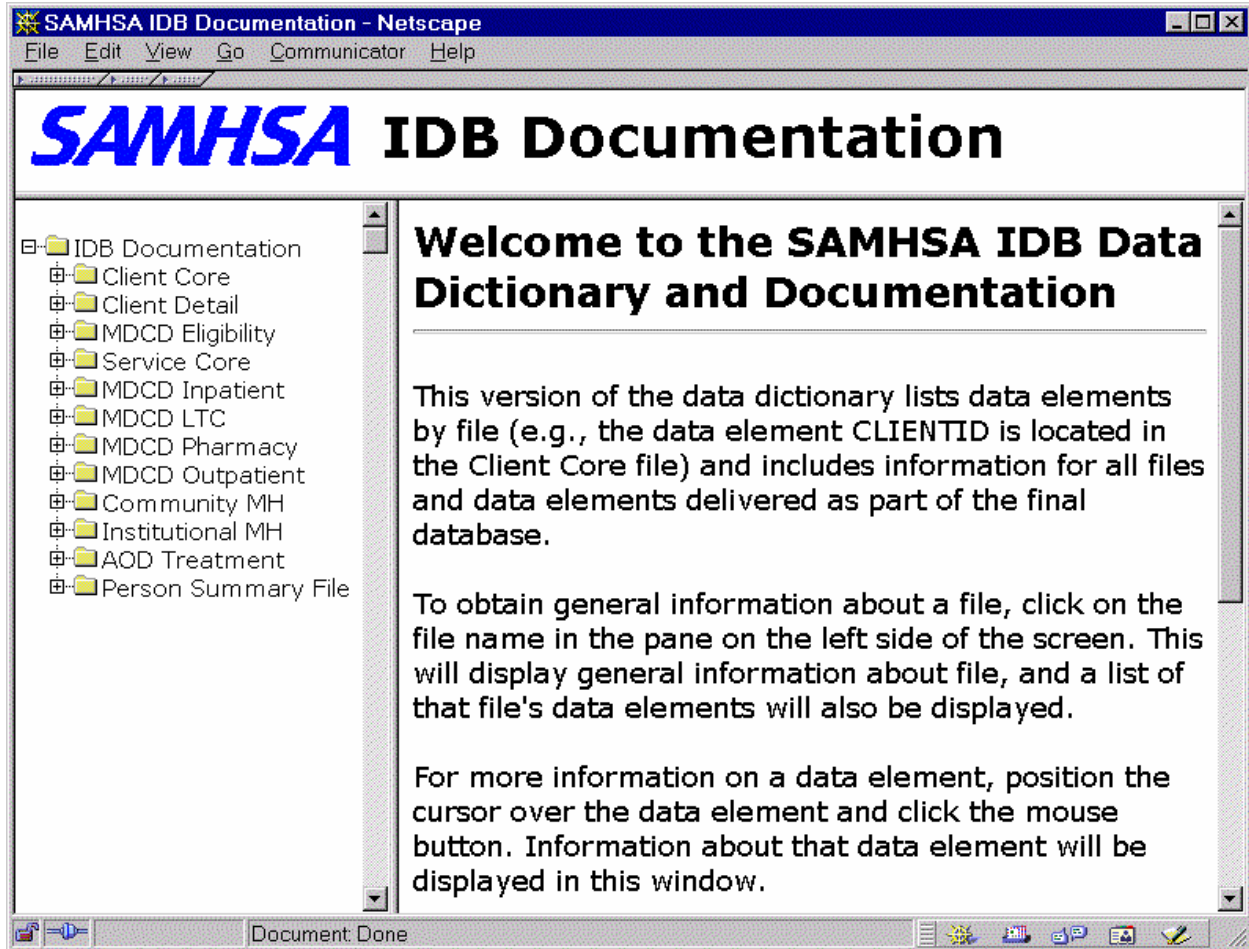
Complex reports can be created and printed for some or all of the data elements. Although it is powerful, there are several important limitations to this system. First, the data base uses Microsoft Access 97 (or the runtime version of Access 97). Consequently, some effort would be required to convert the system to operate in other environments (e.g., Oracle, Sybase). Second, the most comprehensive report is more than 300 pages long. Such reports cannot be distributed or updated efficiently. Moreover, using paper as the primary mechanism for distribution would be inefficient, given the potential for change in the source data systems.

In our review of alternative distribution methods, we found that the best alternative was to distribute a collection of hypertext markup language (HTML) files. Such files can be opened in any web browser, and they can be placed either on a web server or on a standalone PC. Since web browsers are available for almost every platform and operating system, the data dictionary can be viewed on virtually any computer. The files that comprise the data dictionary currently fit on a single CD, and the entire data dictionary can be updated by copying or downloading newer files.

HTML files are created by exporting data from the Access data base. As many as four files are created for each data element (one "header" file and one file for each State with available source data). The system automatically generates hyperlinks from the main page for each data element to the various State-specific pages. The HTML version of the data dictionary also displays the contents for files and data elements in a familiar tree-like user interface through a JavaScript component. This collection of files can be distributed independently or integrated into a web site.

The following section describes the data base documentation in detail. Also, it introduces the methods used to enter information at the uniform data element level and State Specific level. We encourage the reader to refer to the electronic file for a more complete view of the documentation.

Figure 29 IDB Documentation Main Screen



## STRUCTURE

The data base documentation is to allow researchers to access information on three levels: the variable level, the uniform level, and the State-specific level.

### *Variable Level*

For each data element listed in the data dictionary, the following items are displayed, along with a list of uniform values for that data element:

*SAS Name.* The SAS Name is a data element name that conforms to SAS naming conventions (no longer than eight characters). This is the data element name within the SAS files that make up the data base.

*Variable ID.* A MEDSTAT-assigned ID number. This is an arbitrary but unique identifier for each variable.



*Variable Description.* Description of each data element. For example, the variable description for the data element `PROVTYPE` would be “Provider Type.”

*Type.* Refers to how the data element is stored in SAS. The data type is either character or numeric.

*Length.* Amount of storage required for each variable. For a character variable it is the length of the character string. For a numeric variable it is the number of bytes used for storage.

*Format.* Provides a transformation of data values. For example, the dollar sign (\$) formats numbers with commas and dollar amounts with dollar signs. The ‘YES/NO’ format translates the number ‘1’ to ‘yes’ and the number ‘2’ to ‘no’. We developed format libraries for most of the data elements in the data dictionary.

*Comment.* Provides further information about a variable that might be of interest to researchers.

### ***Uniform Level***

A SAS name is assigned to each uniform variable. Each uniform variable is the result of transforming similar values from all three States into a common set of uniform values. Where possible, uniform values are consistent with commonly used data standards. For example, Figure 32 shows the uniform-level information for the data element `PROVTYP`. To see more detailed information for any one of the uniform values listed, the user can click the State link located directly above the table. Users can view data at the State specific level where one or more State values may be mapped to a single uniform value.

Figure 30 Uniform Level Documentation for PROVTYPE

**SAMHSA IDB Documentation**

**SAS Name:** PROVTYPE  
**Variable ID:** 6043  
**Variable Description:** Provider type, uniform value  
**Type:** Char  
**Length:** 3  
**Format:** \$PROVTYPE.  
**Comment:**

[DE Source Values](#)

Uniform Value	Description	Comment
215	Case Manager/Social Worker	
100	Medical Doctor or Osteopath	
200	Unspecified and All Other Practitioners	
201	Audiologist and Speech Pathologist	

**State Specific Level**

State-specific documentation provides detailed information on source data crosswalks from all three States. For each State, source values are described and mapped to their uniform values. The State-specific documentation contains uniform values along with brief descriptions, data years that apply, data sources, source variable names, source variable values and their descriptions, a section on variable crosswalks or “mapping”, and a column for comments.

Figure 31 Documentation Showing Source Values

The screenshot shows a Netscape browser window titled "SAMHSA IDB Documentation". The main page displays the "SAMHSA IDB Documentation" logo and a tree view of variables on the left. The variable "PROVTYPE" is selected, and its details are shown in the main content area:

**SAS Name:** PROVTYPE  
**Variable ID:** 6043

A secondary window titled "DE Source Data - Provider type, uniform value - Netscape" is open, displaying "Delaware source data" for the same variable. It includes a table with the following data:

Uniform Value	Uniform Value Description	Year	Data Source	Source Variab
100	Medical Doctor or Osteopath	1996	DE - Medicaid Claims	PROV-TYPE;REG-CC
200	Practitioners			
201	Audiologist and Speech Pathologist			

The browser's address bar shows "Document: Done" and the status bar at the bottom displays various system icons.

*Data Sources.* The data bases on the State-specific level contain documentation for the following data sources:

**Table 19 Data Source Listing by State**

State	Data Source
Delaware	Medicaid Claims
	Fee for Service
	Encounters
	Inpatient
	Medicaid Eligibility
	Adult Mental Health/Alcohol and Other Drug
	Kent County
	Sussex County
	Newcastle County
	Community Mental Health
Institutional Mental Health	
AOD	
Client Ticket file	
Youth and Child	
Oklahoma	Medicaid Claims
	Medicaid Eligibility
	Mental Health/Alcohol and Other Drug (MH/AOD)
Washington	Medicaid Claims
	Medicaid Eligibility
	DASA
	MH State Hospital
	MHD Community Ambulatory
	MHD Community Inpatient

*Source Variable Names.* The Source Variable name listed in the documentation (shown in Figure 32) is the original variable name found in documentation provided by the State. This provides a link to the original data for further investigation by researchers and States. In some cases, several source variables are used to document a single uniform value. For example, in Figure 32, the source variable PROV-TYPE is the primary variable used to define the uniform data element PROVTYPE. The other two source variables, PROV-CAT-OF-SVS-CODE and PROV-SPEC-CODE, are source variable names for ‘Category of Service’ and ‘Physician Specialty’, both of which also help define PROVTYPE.

**Figure 32 Source Value Map for PROVTYPE**

Uniform Value	Uniform Value Description	Year	Data Source	Source Variable	Source Value	Description
100	Medical Doctor or Osteopath	1996	WA - Medicaid Claims	PROV-TYPE;PROV-CAT-OF-SVS-CODE AND PROV-SPEC-CODE	35 OR 37;20 AND 70	Maternity Support Services;Physician General Practice AND Clinic
100	Medical Doctor or Osteopath	1996	WA - Medicaid Claims	PROV -TYPE;PROV-CAT-OF-SVS-CODE	59;20 OR 20	Hospital Type III;Physician Emergency Room OR Physician/General Pract

In some cases however, a particular data element has no source variable name. For these, 'N/A' is frequently substituted and a comment is added to direct the researcher to additional documentation. For example, in Figure 33 *SERVFILE* (Delaware Youth) has no source variable name associated with it. However, the source variable *PROGTYP* best represents the 'source' for *SERVFILE*, so it was used as an alternate variable name. Note that the comment sections for Delaware Youth direct the researcher to documentation for *PROGTYP* to complete the crosswalk for *SERVFILE*. Comments may also be included to further clarify the definition.

The source value column may contain a number, a list of numbers, or a list of codes provided by the State.

**Figure 33 Source Value Map for *SERVFILE***

Year	Data Source	Source Variable	Source Value	Description	Mapping	Comment
1996	DE - Youth	PROGTYP	02	AOD Treatments	If PROGTYP=02, Then <i>SERVFILE</i> =AOD	See Mapping and Documentation for PROGTYP
1996	DE - Adult MH/AOD	N/A	N/A	Community Based MH Treatments (K)	<i>SERVFILE</i> ='COM'	
1996	DE - Youth	PROGTYP	Any Other Value	Community based MH Treatments	If PROGTYP=Any Other Value, Then <i>SERVFILE</i> =COM	See Mapping and Documentation for PROGTYP

*Mapping.* Mapping allows users to see the correspondence between values and variables. This and to understand how uniform values are assigned. This is accomplished through one of two methods:

- Direct Mapping
- Integrated Mapping

An example of direct mapping is:

If *RECIP-RACE-CODE*='W' then 01=W

In this example, if the uniform variable *RACE* represents 'White/Caucasian' and has uniform value '01', then all source values that represent 'White/Caucasian' are mapped to '01'. For example, if Oklahoma Medicaid Eligibility assigns the value 'W' for its 'White/Caucasian' category, then for Oklahoma Medicaid Eligibility, 01='W'. Direct mapping is characterized by a one-to-one correspondence between source variable names and uniform values.

In contrast, integrated mapping involves more than one variable, as in the following example:

If *PROC\_IND*=01 AND (00100 <= *PROC\_CD* <=01999), Then *SERV\_UNT*=06.

For *SERV\_UNT* to be set equal to '06', both *PROC\_IND* must equal '01' and *PROC\_CD* must have values between '00100' and '01999'. What makes integrated mapping complicated is that there may be one or more variables that define *PROC\_IND*=01, and perhaps still another set of variables that define *PROC\_CD*. For this one statement as many as six variables could be involved.

For Provider Type, Physician Specialty, and Category of Service, the correspondence between values and variables is clearly established by the source values. Consequently, the mapping column is left blank to avoid unnecessary repetition.

*Symbol Descriptions.* Throughout the documentation on the State specific level, various symbols consistently represent documentation “rules.” The most frequently used symbols are listed below, along with their interpretation and examples of how they are used in programs. The names provided in the program example may differ from names provided in the documentation because only source variable names are used at the documentation level.

**Table 20 Documentation Conventions - ; (Semicolon)**

Symbol	Description
;	(Semicolon) Separates two individual statements that work in conjunction with one another, but do not necessarily have to be used together to make a statement true. Each statement may be used individually however, when put together, they provide a more complete understanding of how a value is obtained. The semicolon is mainly used in documenting Provider Type, Provider Specialty, and Service Category and represents a type of “crosswalk” map where the three are interrelated. Example: If IPROVTYP='01' Then Do; If ISERVCAT='35' Then PROVTYPE=100

**Figure 34 Sample Map for PROVTYPE Value of 100**

Uniform Value	Uniform Value Description	Year	Data Source	Source Variable	Source Value	Description
100	Medical Doctor or Osteopath	1996	OK - Medicaid Claims	PROV-TYPE;PROV-COS	01;35	Hospital;Physician Services

As seen in Figure 34, the source values 01 (for PROV-TYPE) and 35 (for PROV-COS) are used to define Provider Type uniform value 100, which indicates a Medical Doctor or Osteopath. PROV-COS, the Oklahoma Medicaid Claims source variable name for Provider Category of Service, may also be used to define Category of Service in separate documentation.

**Table 21 Documentation Conventions - OR**

Symbol	Description
OR	Used to indicate a range of source values that may be applied to any given uniform value. Example: If ('IR', 'RD', 'RL', 'RS'); Then SERVLOC=1

**Figure 35 Service Location Source Values**

Uniform Value	Uniform Value Description	Year	Data Source	Source Variable	Source Value	Description	Mapping	Comme
1	Provider's facility	1996	DE - Adult MH/AOD (AOD)	TXUNITID	'IR' OR 'RD' OR 'RL' OR 'RS'	Rehab/Residential in Hospital OR Detoxification Not In Hospital OR Rehab/Residential Long Term OR Rehab/Residential Short Term	1='IR' OR 'RD' OR 'RL' OR 'RS'	

As shown in Figure 35, for Service Location uniform value 1 (uniform description 'Provider's facility'), four different source values apply: 'IR', 'RD', 'RL', or 'RS.' This is another example of direct mapping.

**Table 22 Documentation Conventions - AND**

Symbol	Description
AND	(all caps) Combines two separate statements that are true only when used together.
	Example: If ('3') ; Then Place of Service=11

**Figure 36 Sample Service Location Map**

Uniform Value	Uniform Value Description	Year	Data Source	Source Variable	Source Value	Description	Mapping
11	Office	1996	OK - Medicaid Claims	Place of Service	3	Doctor's Office	If Place of Service=3 AND Recore Code=60; then 11=3

In the example shown in Figure 36, Oklahoma Medicaid Claims assign a value of '3' for 'Doctors office' under the category 'Place of Service'. Records with a source value '3' and a Record Code of 60 are mapped to the uniform value '11.'

For Oklahoma, the source data had different formats depending on the record types that map to the four Medicaid files (Other Medicaid, Long Term Care, Inpatient, and Pharmacy). Therefore, record codes that differentiate the record types are sometimes employed as in this example. Since a record code is not a source variable name, this also is an example of direct mapping.

**Table 23 Documentation Conventions – N/A**

Symbol	Description
N/A	Not Available or Not Applicable

**Table 24 Documentation Conventions – ^ (Carat)**

Symbol	Description
^	Not equal to, or not within given values.
	Example: If ^(49, 27, 38, 42) ; Then SERVCAT=62