PharmaSUG 2026 - Paper HT-453

Understanding Administrative Healthcare Datasets using SAS® programming tools.

Jay Iyengar, Data Systems Consultants LLC

ABSTRACT

Changes in the healthcare industry have highlighted the importance of healthcare data. The volume of healthcare data collected by healthcare institutions, such as providers and insurance companies is massive and growing exponentially. SAS® programmers need to understand the nuances and complexities of healthcare data structures to perform their responsibilities. There are various types and sources of administrative healthcare data, which include Healthcare Claims (Medicare, Commercial Insurance, & Pharmacy), Hospital Inpatient, and Hospital Outpatient. This training seminar will give attendees an overview and detailed explanation of the different types of healthcare data, and the SAS programming constructs to work with them. The workshop will engage attendees with a series of SAS exercises involving healthcare datasets.

INTRODUCTION

SAS® programmers in the healthcare industry should be familiar with the different types of administrative healthcare data, and data structures for each. The kinds of administrative healthcare data are varied and encompass healthcare claims, hospital admission data, hospital discharge data, and vital statistics data, such as birth records and death records. Hospital visit files are available from hospital associations, state governments, or from federal government agencies. Vital statistics data are only available from state governments, usually state departments of health. Claims files can be divided into Medicare Claims, Private Insurance Claims, and Pharmacy Claims. There are four types of Medicare coverage. Medicare claims could be Inpatient (Part A), Outpatient (Part B), Medicare Advantage plans (Part C), or Prescription Drug (Part D).

THE ORIGINS OF HEALTHCARE DATA

At some point in their lives, everyone visits a doctor for a regular checkup, or goes to a hospital emergency room for treatment. Anytime that an individual visits a doctor or a hospital for treatment, data is generated. Information concerning the visit is recorded, such as the medical diagnosis. When a patient checks in to be admitted at a hospital, the admitting diagnosis is collected and recorded. Other information including the date of admission is also recorded. Similarly, when an individual is seen by a physician additional information is collected, such as the principal diagnosis.

When an individual has x-rays taken, has blood drawn, or undergoes surgery, further information is recorded on the medical procedures the patient undergoes. Similar to hospital visits, each visit to a private physician generates a record. Physicians maintain medical charts which are patient health records. Some doctors keep electronic health records. In order to get reimbursed for the healthcare services that were provided, physicians and hospitals submit a bill known as a claim, to a patient's insurance company. Insurance companies process the claim and send payment for the services to the hospital or doctor. In healthcare industry terminology, physicians and hospital entities are known as healthcare providers. Insurance companies are known as healthcare payors.

The graphic in Figure 1 shows the entities involved in the healthcare delivery system.

Figure 1. Healthcare Delivery System.

PAYORS

In order to get treatment, everyone typically needs some form of health Insurance due to the high cost of healthcare. Many people get health insurance through their employers. Due to the Affordable Care Act (ACA), everyone is currently able to get health insurance through the Federal Government's Healthcare Marketplace with Obamacare. The federal government also provides health insurance for the elderly through Medicare, and health coverage for the poor through Medicaid. While Medicare is administered by the federal government, Medicaid is a program funded by the federal government. Medicaid is administered by state governments. There are other types of health insurance available. Healthcare coverage for the Military is called Tricare. Most health insurance is obtained through private insurance companies such as Blue Cross Blue Shield, Aetna, or United Health Group. The diagram in Figure 2 shows the payors in the healthcare system.

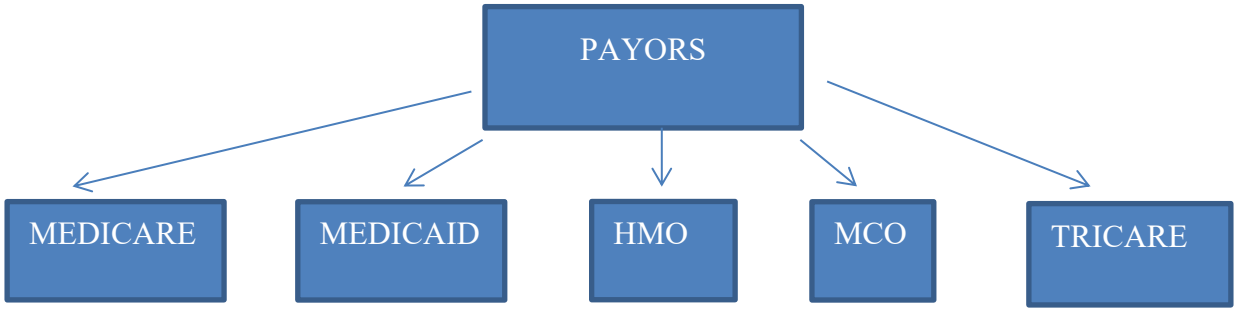


Figure 2. Healthcare Payors

HEALTHCARE CLAIM FORMS

Healthcare claims are submitted using claim forms, and there are separate claim forms for different healthcare providers. Currently, hospitals and other institutions submit claims using the CMS-1450 claim form, where CMS is the Center for Medicare and Medicaid Services. Originally the claim form used by institutions was the UB-92 form, which was developed by the Uniform Billing Committee. The types of institutions using this form include acute care hospitals, long-term care facilities, skilled nursing facilities, psychiatric institutions, and ambulatory care centers. More commonly, hospitals submit claims in electronic format, using the 837I claim format.

Physicians in private practice, and other private providers are known as healthcare professionals. Professionals submit claims using the CMS-1500 claim form. Originally, the claim form used by professionals was the HCFA-1500 form. The types of professionals using this form include private physicians, group physician practices, and clinics not associated with hospitals. Healthcare professionals can submit claims electronically using the 837P claim format.

The CMS-1500 and 837P forms were developed by the Accredited Standards Committee (ASC X 12) in conjunction with the American Medical Association (AMA). Pharmacies submit claims using an electronic form developed by the National Council of Prescription Drug Program (NCPDP). Although pharmacy claims are commonly submitted electronically, pharmacies have access to a universal paper claim form that was developed by the NCPDP.

THE STRUCTURE OF HEALTHCARE CLAIMS DATA FILES

Healthcare claims files are usually very large data files, containing millions of claims. It is not uncommon for the files to contain 100 million records for a single calendar year. In a single claims file, there usually are claims for thousands, and maybe even tens of thousands of patients. There are usually multiple records per patient in a single claims file, with each record pertaining to a separate claim. By definition, many patients in the file have multiple claims, with each claim corresponding to a distinct physician visit or hospital stay.

There are separate claims files for different providers. For example, hospitals and other institutions which provide healthcare services store claim records in facility claim files. Likewise, private physicians and practitioners save claim records in professional claim files. Pharmacies which submit claims for prescription drugs, store claims in pharmacy claim files.

Generally, claim files contain one record per claim. But this varies with the type of file. Claim records are usually split up into header records and detail records. The respective header and detail records provide different levels of information about the claim. The different records are stored separately in header files, and detail files. Each header record summarizes the services for the entire claim. Header files contain one record per claim. Detail records contain information pertaining to each service performed. Detail files contain multiple records per claim. However, if adjusted, a single claim may have several records. One record for the original claim, a second record for the claim adjustment, and another record for the final adjudicated claim. If a claim was denied, but then resubmitted by the provider, it will also have multiple records. Both header and detail files contain a claim identifier, which can be used to connect the files.

DIAGNOSIS CODES

Every record of an encounter with a healthcare professional or institution includes a recorded medical diagnosis in the form of a diagnosis code. Records for hospital inpatient stays, patient health records, electronic health and medical records, and healthcare claims all include diagnosis codes. There are different varieties of diagnosis codes, and different types of codes will appear on different kinds of claim records. Healthcare claim records may contain up to 8 or 12 distinct diagnosis codes. For hospital inpatient claim records, there are distinct types of diagnoses. For example, the diagnosis recorded when the patient is admitted is the admitting diagnosis. However, when the patient is seen by a physician, a primary or principal diagnosis is made and recorded. This is usually different from the initial diagnosis. The table in Figure 3 lists the different kinds of diagnosis codes.

Code	Description
ICD9	International Classification of Diseases – 9th revision
ICD10	International Classification of Diseases – 10th revision
DRG	Diagnosis Related Grouper
MDC	Major Diagnostic Category

Figure 3. Kinds of Diagnosis Codes.

ICD CODES

The most common type of diagnosis code is the ICD code. ICD codes were developed by The World Health Organization (WHO), and ICD stands for International Classification of Diseases. The Ninth Revision of the code, ICD9, was developed and published in 1975 and 1978, respectively. ICD9 codes were in use for many years. The Tenth Revision of the code, ICD10, was developed in 1994. However, implementation of the ICD10 code has been slow. In 1999, ICD10s were mandated for use in death certificates. ICD10s were eventually mandated for use in all systems starting in 2015, and officially replaced ICD9 codes.

Of all diagnosis codes, ICD codes have the highest level of specificity or granularity. ICD codes have the largest set of unique individual codes. There are as many as 15000 ICD9 codes, and 70000 ICD10 codes. Other diagnosis codes have been devised by grouping ICDs into a smaller set of categories. The ICD diagnosis code is a 3-digit code. These three digits define the disease or medical condition. Fourth or fifth digits can be added for extra levels of specificity. With the insertion of a decimal after the first 3 digits, ICD codes can contain a maximum of 6 digits. Figure 4 provides an example of an ICD9 code.

410.12 - Ischemic Heart Disease

Figure 4. ICD9 Diagnosis Code.

ICD-CM CODES

ICD-CM Codes are an adaptation of ICD codes which stands for International Classification of Diseases, Clinical Modification. ICD9-CM codes were developed by the National Center for Health Statistics (NCHS), to be used for hospital inpatient, outpatient, and physician office visits. ICD9-CM codes are a version of ICD9 codes which provides additional morbidity detail. ICD9-CM codes are updated annually and are maintained by NCHS as well as the Center for Medicare and Medicaid Services (CMS).

DRG CODES

DRG stands for Diagnosis Related Group. DRGs are diagnosis codes that are a higher level classification of ICD codes. DRG codes were developed by collapsing groups of ICD codes into smaller categories. DRGs are assigned by hospitals or facilities, and thus only appear on institutional records or claims. DRGs also use and incorporate procedure information. DRGs are used by the Medicare system to calculate payments and reimburse hospitals. The DRG code is a three-digit code. There are up to 511 distinct DRG codes. Figure 5 shows an example of a DRG code.

469 – Major Joint Replacement

Figure 5. DRG Code

PROCEDURE CODES

Every time a patient receives service from a healthcare professional, whether x-rays are taken, or lab tests or surgery is performed, a procedure code is generated. Just as with diagnoses, there are different types of procedure codes. The type of procedure code found on a specific record depends on the place of service, and healthcare provider. Different kinds of procedure codes will appear on different types of claim records. Procedure codes include ICD procedure codes (ICD-PCS), CPT codes, and HCPCS codes.

ICD9 PROCEDURE CODES

ICD procedure codes were developed by the World Health Organization (WHO) similar to ICD diagnosis codes. ICD procedure codes appear only on facility claims. They're primarily used for healthcare services provided in institutional settings. Any time professional charges are levied in a facility setting, these procedure codes are generated. Hospitals group ICD procedure codes with ICD diagnosis codes to calculate and assign DRG codes for billing purposes. The ICD procedure code is a two-digit code, followed by a decimal and two extra digits for additional specificity and detail. Figure 6 provides an example of an ICD9 procedure code.

35.10 - Open Heart Valvuloplasty

Figure 6. ICD9 Procedure Code.

ICD10-PCS CODES

ICD10-PCS codes are procedure codes which are based on ICD9 procedure codes but provide additional morbidity detail. Officially ICD10-PCS codes were developed to replace volume 3 of ICD9-CM procedure codes. ICD10-PCS codes were developed by CMS and 3M Health Information Systems in 1995 and released in 1998. Unlike other ICD codes, ICD10-PCS contain 7 digits and no decimals. There are over 72000 ICD10-PCS codes.

CPT CODES

CPT codes were developed by the American Medical Association. CPT stands for Current Procedural Terminology. CPT codes are used by physicians and other healthcare professionals for services provided in both physician offices and facility settings. Thus, CPT codes only appear on professional claim records or records where professional services were rendered. CPT is a five-digit code. There are three categories of these codes (CPT I, CPT II, CPT III). Most codes are CPT I. Figure 7 shows an example of a CPT code.

90847 - Family Therapy

Figure 7. CPT Procedure Code.

HCPCS CODES

Another type of procedure code is HCPCS codes. HCPCS stands for Healthcare Common Procedure Coding System. HCPCS codes are primarily used by the Medicare system, and other insurers, and often appear on Medicare claims. These codes were derived from CPT codes. HCPCS codes also have different levels. HCPCS level I is equivalent to CPT codes.

FACILITY CLAIMS

HEADER FILE

Facility claims header records include claim data for hospital visits, and other healthcare facilities. Header claim files have one record per claim. Present on the header record is the Claim Identifier (Claim ID) and Member ID which can be linked back to the Membership file. The header record includes several diagnosis codes, including DRG (Diagnosis Related Group), and ICD Codes. For the ICD codes, up to 8 diagnosis codes are recorded, including the principal diagnosis, and up to 7 secondary diagnoses. Other data elements unique to facility headers are BillType, Date of Admission, and Date of Discharge. Figure 8 below shows a typical sample of facility claims header records.

CLAIMID	MEMBERID	SEX	DOB	ADMITDT	DISCHGDT	DRG	AMTPAID	BILLTYPE	PDX	PPX
C042929171	M9440019833699700	F	9/2/1941	1/7/2011	1/7/2011	000	4287.54	332	V5481	
C042929554	M2743197531321900	M	9/10/1932	1/4/2011	1/19/2011	871	9708.92	111	0388	3722
C042931278	M4327445091704000	M	6/18/1942	1/2/2011	1/30/2011	853	79822.65	111	0389	4620
C042931758	M8582209243454100	M	10/27/1935	1/29/2011	2/4/2011	464	14576.23	111	99666	8006
C042931936	M8365920163413800	M	12/27/1936	1/28/2011	2/13/2011	207	36436.03	111	4821	9672

Figure 8. Facility Claims Header File

DETAIL FILE

Facility claim detail files contain multiple records per claim. They contain one record per service or procedure, designated as a claim line. Thus, each record in the detail file pertains to a specific service or procedure provided. Detail files can be linked to header files using Claim ID. Data elements unique to detail records include revenue codes and procedure codes. The revenue codes are specific to hospital cost centers. Radiology or Cardiology are examples of hospital cost centers. The detail records also include the Total Paid Charges for each service performed. Medicare claims records usually contain HCPCS procedure codes. Figure 9 shows a typical sample of facility claims detail records.

CLAIMID	CLAIMLINE	REVCODE	REVCODEDESC	PX	PXDESC	BEGINDOS	ENDDOS	UNITS	AMTPAID
C042929171	001	0023	HH PPS (HRG)	5CGKV	5CGKV	1/7/2011	1/7/2011	1	4626.67
C042929554	001	0120	ROOM-BOARD/SEMI			1/4/2011	1/19/2011	7	10476.87
C042929554	002	0210	CORONARY CARE			1/4/2011	1/19/2011	8	0
C042929554	003	0250	PHARMACY			1/4/2011	1/19/2011	999	0
C042929554	004	0258	IV SOLUTIONS			1/4/2011	1/19/2011	19	0

Figure 9. Facility Claims Detail File

SAS PROGRAMMING TOOLS

An essential initial step in working with administrative healthcare data files is to validate the data. A SAS programmer needs to examine the data to make sure it is accurate and correct. SAS has many useful tools to do this. Before anything else, a programmer should review the data set contents to discover how many variables are in it, how many observations, and which variables it contains. PROC CONTENTS is a standard construct for reviewing the descriptor portion of the data set. The code for PROC CONTENTS is displayed in Figure 10.

```
Proc Contents Data=FacilityHeader;
Proc Contents Data=FacilityDetail;
Run;
```

Figure 10. PROC CONTENTS

Appendix I contains the PROC CONTENTS report from the Facility Header data set. As you'll notice, PROC CONTENTS generates a report with high level information about the data set. The information contained in the report includes the number of observations, the number of variables, date it was created, date last modified, number of indexes, the size of the data set, and much more. This information is defined as the data set metadata. The report also includes a list of variables contained in the data set with each variables type, length, format, informat and label.

After reviewing the descriptor portion of the data set, the next step in data validation is to examine actual data values. Outputting frequency tables provides a way to examine data values for character variables. A frequency report provides the programmer with an exhaustive list of values for each selected variable, and summary record counts for each value. The report will detect and display any irregular values contained in the data set. A healthcare analyst can use the report to surmise whether record counts for variable values are reasonable. The code construct in SAS for generating frequency tables is PROC FREQ. The code for PROC FREQ is displayed in Figure 11.

```
Proc Freq Data=facilityheader;
    Tables TypeBill / List Missing;
    Format TypeBill $TOB.;
    Title 'Number of Claims by Place of Serivce';
Run;
```

Figure 11. PROC FREQ

Type of Bill (TOB) is a data element found on facility claim header records and is important to validate to verify the quality of the data set. TOB is a three-digit code. The first two digits indicate the place or setting of service. PROC FREQ output for Type of Bill is displayed in Figure 12.

TypeBill	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Hospital Inpatient	4208	63.24	4208	63.24
Hospital Outpatient	1621	24.36	5829	87.60
Skilled Nursing Facility – Inpatient	517	7.77	6346	95.37
Other	308	4.63	6654	100.00

Figure 12. Frequency Table – Type of Bill

As the output in Figure 12 shows, the highest claims volume (63.24%) was for hospital inpatient services. This is a reasonable and expected result since inpatient stays are a hospital's highest cost center. Claims for skilled nursing facilities had the lowest frequency, which follows logically since nursing facilities provide services for the elderly, which comprise a small part of the population.

PROC MEANS

Most healthcare data sets contain numeric variables, such as total charges. To validate these variables, it is important to obtain descriptive statistics, such as the mean, standard deviation, min and max values. These statistics give you the distribution of these variables, so you can evaluate whether the data is reasonable. Base SAS® has effective procedures, such as PROC MEANS, which produce descriptive statistics to examine numeric data. Figure 13 contains the code for PROC MEANS.

```
Proc Means Data=FacilityHeader N Mean Std Min Max;
    Var PaidAmt;
    Class Proc_Code;
    Title 'Total Charges by Procedure';
Run;
```

Figure 13. PROC MEANS

Total charges are the dollar amount a provider is billing for a given procedure or service. Charges above a certain amount for a given procedure would raise a red flag. They might be an indication of overbilling or healthcare fraud. Knowing the distribution of paid amounts by procedure provides valuable insights into the legitimacy of the data set. The PROC MEANS output for total charges is displayed in Figure 14.

Analysis Variable: AmtPaid Claim Paid Amount				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
6654	8400.58	9043.60	3003.55	175450.72

Figure 14. PROC MEANS Output

PROFESSIONAL CLAIMS

HEADER FILE

Professional claims header files includes one record per claim which summarizes information for that claim. It can be linked to the professional detail file using the claim identifier (Claim ID). Recorded on header claims are ICD10 diagnosis codes. The header record may contain up to 12 diagnosis codes. The principal diagnosis is not explicitly identified but is presumed to be the first diagnosis code. The header record includes a provider ID that identifies the physician who submitted the bill, which may be different than the physician who performed the service. Professional header claims also include patient demographic information, including gender and date of birth. Figure 15 provides a sample of professional claims header records.

CLAIMID	MEMBERID	SEX	DOB	DX1	DX2	CLAIMAMTPAID	PAIDDT	BILLINGPROVID
C020007703	M4770630862621200	M	6/19/1975	7394	7391	0	3/15/2011	P463460600
C020008363	M4765009932545200	M	11/13/1946	45340	12345	916.265358	3/10/2011	P183539100
C020013669	M1905076611277800	M	7/5/1957	2724	5715	141.50709	3/8/2011	P133506800
C020022324	M4745396702365100	F	7/10/1949	81342		639.951219	3/8/2011	P223461600
C020024380	M4696487421881300	F	3/14/1952	5712	78959	0	3/8/2011	P313457700

Figure 15. Professional Claims Header Records

DETAIL FILE

Similar to the facility detail file, professional claims detail files contain multiple records per claim, where each individual record pertains to a specific service or procedure performed. Unique to the detail record is place of service (POS), which indicates where the service was performed (i.e., doctor's office, hospital inpatient, emergency room). On the detail record are CPT procedure codes, which physicians commonly use to bill. Also present is NPI, the National Provider Identifier, which indicates the physician who performed the service, and provider taxonomy codes which categorize physician specialties. In Figure 16 is a sample of professional claim detail records.

CLAIMID	CLMLINE	BEGINDOS	ENDDOS	PNTRDX1	PNTRDX2	POS	PX	UNITS	LINEAMTPAID	SRVCPROVTAX
C020007703	1	1/19/2011	1/19/2011	7392	7840	11	98941	1	0	111N00000X
C020007703	2	1/23/2011	1/23/2011	7394		11	98941	1	0	111N00000X
C020008363	1	1/17/2011	1/17/2011	12345	45340	21	37620	1	572.67	2085R0202X
C020008363	2	1/17/2011	1/17/2011	45340		21	36011	1	282.87	2085R0202X
C020008363	3	1/17/2011	1/17/2011	12345		21	75940	1	47.88	2085R0204X

Figure 16. Professional Claims Detail Records

SAS PROGRAMMING TOOLS

In healthcare environments, after validating the data to ensure its quality and manipulating the data to make corrections to it, the next step is to analyze the data for regular reporting. Different players in the healthcare market have incentives to track this data. From the standpoint of a provider, the purpose of reporting may be to track costs or patterns of treatment which drive costs that impact a hospital's balance sheet. For example, certain medical procedures are very expensive, so analyzing the frequency of procedures at regular intervals is useful. From an insurance company standpoint, patients with chronic conditions need extensive care with large bills and reimbursements. Thus, it is worthwhile to track such patients that fit a certain risk profile.

Before analyzing claims files, it is important to determine the same claim hasn't been submitted twice, and thus the file you're using doesn't contain duplicate claims. PROC SORT is a good SAS construct to remove duplicate records. Use the NODUPKEY option to remove duplicate records based on a key variable. The SAS code for PROC SORT is in Figure 17.

```
Proc Sort Data=admhlth.professionalheader Out=PHeader Nodupkey;
   By ClaimID;
Run;
```

Figure 17. PROC SORT with NODUPKEY Option

Healthcare reporting and analytics are a set of project tasks which many SAS programmer/analysts must perform. SAS has many effective procedures to utilize in the production of periodic reports and analysis. Using PROC SQL, a programmer can produce detail or summary reports using intuitive coding structures. To produce summary reports that are grouped by categories, a programmer can reference summary functions in the SELECT statement with the GROUP BY clause to specify how the report is stratified. The code for PROC SQL is displayed in Figure 18.

```
Proc Sql;
Title1'Claims Volume by Provider Specialty';
Select SrvcProvTaxonDesc as Provider_Specialty,
       Count(ClaimID) as Claims_Volume,
       Sum(ClaimAmtPaid) as Claims_Payments Format=Dollar13.2
From Work.Header_Det
Group By SrvcProvTaxonDesc
Order By Claims_Volume Desc;
Quit;
Run;
```

Figure 18. PROC SQL with GROUP BY Clause

In this example, the variable SRVCPROVTAXON is used to group the report by Service Provider Taxonomy. Service Provider Taxonomy is a variable found on professional claim detail records which indicates a provider's specialty.

The PROC SQL code produces a report which contains counts of claims (claims volume), and claim payment amounts by provider specialty. The report is sorted in descending order by CLAIMS_VOLUME, using the ORDER BY clause, so provider specialties with the highest frequency appear at the top of the report. Using the report, a programmer/analyst can discern how many claims were submitted for visits to chiropractors, or the amount of payments made to general internists for patient care.

To adequately analyze healthcare data, it may be necessary to generate reports which have a more complex structure. PROC TABULATE provides such a structure. PROC TABULATE is capable of producing complex tables with multiple dimensions. In PROC TABULATE you have the advantage of using the CLASS statement to specify your grouping variables. Using CLASS, you can avoid using the BY statement, and a potentially time-consuming PROC SORT. In the TABLES statement you specify the dimensions of the table. You can specify up to three dimensions; row, column, and page. The code for PROC TABULATE is displayed in Figure 19.

```
Proc Tabulate Data=Header_Det_nodup;
   Class SrvcProvID DX1Desc;
   Var ClaimAmtPaid;
   Tables (SrvcProvID all='total')*(DX1Desc all='total'),
        N='Number of Claims'
        ClaimAmtPaid='Claim Payments';
   Title1'Claims Volume and Payments - Frequencies and percents';
Run;
```

Figure 19. PROC TABULATE

In the above example, the produced table is stratified by two variables; SRVCPROVID, and DX1DESC. That is, Service Provider ID, and Diagnosis Code Description, respectively. It is important to note that professional claims files include both a billing provider ID, the physician who billed for the service, and a service provider ID, the physician who rendered the service. The billing physician may be different from the servicing physician. For the sake of data analysis, it is important to use the service provider ID (SRVCPROVID), the physician that performed the service.

PROC TABULATE includes some helpful statistics to enhance reports. Using the PCTN keyword, you can compute percentages of different totals for each cell of the table. The totals pertain to the different row, column, or page dimensions. In brackets (<>), you specify the row, column, or page variable, and the denominator (keyword: ALL) used to calculate the percent. The output for PROC TABULATE is in Appendix II.

PHARMACY CLAIMS

Pharmacy claims are more straight-forward than facility or professional claim files. In pharmacy claim files, one record pertains to one prescription. This is the case regardless of whether it is a new prescription or a prescription refill. In these files, there is no claim identifier unlike the other claim files. However, the files contain several drug codes. On pharmacy claims is NDC, an 11-digit code that stands for National Drug Code. Also present on these records is AHFS, an 8-digit drug code which stands for American Hospital Formulary.

Pharmacy records also include GPI, the Generic Product Identifier, a 14-digit code which indicates the drug's therapeutic class. Other fields on the pharmacy record include the supply of the prescription or refill (QTY), and the date it was filled (FILLDATE). In Figure 20 is a sample of pharmacy claim records.

MEMBERID	DOB	NDC	LABEL	AHFS	PHARMACYNAME	QTY	FILLDATE
M0000074250011100	6/13/1933	00093550101	BUDEPRION TAB 100MG	28160492	Pharmacy C	30	4/4/2011
M0000074250011100	6/13/1933	00093550101	BUDEPRION TAB 100MG	28160492	Pharmacy C	30	5/4/2011
M0000074250011100	6/13/1933	00093550101	BUDEPRION TAB 100MG	28160492	Pharmacy C	30	6/3/2011
M0000074250011100	6/13/1933	00093550101	BUDEPRION TAB 100MG	28160492	Pharmacy C	30	7/3/2011
M0000074250011100	6/13/1933	00093550101	BUDEPRION TAB 100MG	28160492	Pharmacy C	30	8/2/2011

Figure 20. Pharmacy Claims File

SAS PROGRAMMING TOOLS

The Pharmacy Claims File contains the variable GPI14, the drug code defined above. The first two digits of the code indicate the class the drug belongs to, such as antidepressants. Since each record in the claims file is a prescription fill or refill, it is possible to create a report which contains the number of prescription refills grouped by drug class.

Before producing the report, it is necessary to manipulate the data a little. A SAS programmer needs to extract the two-digit drug code from the GPI variable. The SUBSTR function can be used to do this. The drug class descriptions are contained in a separate lookup table. The two-digit class code will be used to join the claims file to the lookup table and pull over the variable containing drug class descriptions.

To create a simple frequency report with the number of prescriptions by drug class, you should use PROC FREQ, which is the straightforward choice to produce such a report. To further enhance the report, you can display the counts in descending order, with the most frequent drug classes displayed at the top of the report. To add this feature, you code the ORDER=FREQ option on the PROC FREQ statement. The PROC FREQ code for the report is displayed in Figure 21 below.

```
Proc Freq Data=Pharmacy Order=FREQ;
  Tables GPI_DESC / Out=PH_FREQS List Missing;
  Title 'Number of Prescription Refills by Drug Class';
Run;
```

Figure 21. PROC FREQ code.

With some of the statistical and analytic procedures in SAS, users have the ability to create an output dataset containing the analysis results. PROC FREQ has this capability. Notice in Figure 16, that PROC FREQ uses the OUT= option on the TABLES statement to create an output SAS data set. A healthcare analyst may want to use this data set to produce additional analysis, or further refine the analysis.

For healthcare data analysis, it is valuable to create visual aids which illuminate the trends in your reports. SAS has effective tools for visual analytics. The Base SAS package includes ODS GRAPHICS procedures which produce a variety of graphs, charts, and plots. With PROC SGPLOT, you can create vertical and horizontal bar charts. The code for PROC SGPLOT is displayed in Figure 22.

```
Proc Sgplot Data=PH_FREQS;
   Vbar GPI_DESC / Response=Count barwidth=.5;
   Yaxis Label='Frequency Count' Min=0 Max=15000;
   Xaxis Label='Drug Class' Fitpolicy=Stagger;
   Title'Frequency of Prescription Refills by Drug Class';
Run;
```

Figure 22. PROC SGPLOT

With PROC SGPLOT you use the VBAR statement to create a vertical bar chart. You can modify the attributes of the chart axes using the XAXIS and YAXIS statements respectively. The output from PROC FREQ and PROC SGPLOT is provided in Appendix III. The reports clearly show that the drug class with the highest volume of prescription refills is anti-depressants.

MEMBERSHIP\ENROLLMENT FILES

In addition to utilization files, healthcare claims data files also include membership data. A set of claims files for an insurance plan usually includes a membership table, with information for each covered member. An insured member is considered the same as a patient from a provider's standpoint. A membership table can be linked to the claims file using a membership or subscriber identifier.

A member table may contain more than one record per member. Each record designates a specific enrollment period for a member. A member with multiple records in the membership table indicates a member with several enrollment periods. This may happen for several reasons. For instance, a new year starts a new enrollment period. Also, a member's coverage might be terminated due to switching employers, and the member re-enrolls in the same plan a few months after switching.

The membership table includes several key fields and data elements. The member identifier is the MemberID variable. The file includes dates of enrollment, including both starting and ending dates. The table also includes demographic information, such as date of birth and sex, which are occasionally found on claims files. In addition to these fields, a membership file usually includes benefit and eligibility information. A sample membership table is displayed in Figure 23.

MEMBERID	BEGINENROLL	ENDENROLL	DOB	SEX
M0000000860000100	1/1/2010	12/31/2011	5/14/1930	F
M0000008270000200	6/1/2009	12/31/2009	12/17/1928	F
M0000008270000200	1/1/2010	10/31/2010	12/17/1928	F
M0000013530000400	1/1/2007	12/31/2007	3/6/1937	M
M0000013530000400	1/1/2008	12/31/2008	3/6/1937	M

Figure 23. Membership File

SAS PROGRAMMING TOOLS

Working with a membership table allows a SAS programmer analyst to validate whether submitted claims pertains to a member or not. All submitted claims by a provider should be for healthcare services rendered to a patient who is a member of the health plan. If a claim was submitted for a patient who's not a member of the plan, it could be a case of health care fraud. Of course, if a claim is submitted for a patient who does not have active enrollment, the payor won't reimburse the provider. For these reasons it is important to connect claims files with the membership file.

For the purposes of analysis, it is worthwhile to use demographic variables to group data in a claims report. Demographic variables such as sex and date of birth are found on a membership table and can be obtained by joining the claims file with the membership file. A DATA STEP Merge can be used to join the two files. Then, the SAS programmer analyst can produce a report of claims volume and payments grouped by gender and age group

The DATA STEP Merge gives the programmer the ability to subset merge results into matches and non-matches. By using the IN= temporary variable, SAS output data sets can be constructed containing the matched and non-matched subsets. The IN= variable is assigned values which refer to one of the data sets being merged. Thereby IN= can flag records found in one data set or the other. To output the match results into SAS data sets the IN= variable is referenced in the subsetting IF statement.

For a DATA STEP Merge, the input data sets are required to be sorted on the BY variable first, which usually entails coding PROC SORT. Before merging with the claims file, it is good practice to remove duplicate records, if there are any, for members who have multiple enrollment periods. With PROC SORT you can use the NODUPKEY option to remove duplicate records based on a BY variable. In our example, separate SAS data sets containing claims from members, and claims from non-members respectively, are output from the match-merge. The code for the DATA STEP Merge is displayed in Figure 24.

```
Proc Sort Data=admhlth.members Out=Members Nodupkey;
     By MemberID;
Run;

Proc Sort Data=admhlth.facilityheader(Keep=MemberID DOB Sex) Out=FacHeader;
     By MemberID;
Run;

Data Claims_Members Claims_NonMembers;
  Merge FacHeader(In=B) Members(In=A);
    By MemberID;

   If A and B Then Output Claims_Members;
   Else If A and Not B Then Output Claims_NonMembers;
Run;
```

Figure 24. DATA STEP Merge – Membership File with Claims File

Now that demographic variables AGE and SEX are on a claim record, the analyst can develop a report which contains the frequency of claims, and claim paid amounts based on age group and gender. From the standpoint of an insurance company, it is worthwhile to track costs for specific age groups, and gender-specific age groups which constitute high-risk populations.

The REPORT Procedure is a versatile tool to produce reports with multiple grouping variables. With PROC REPORT, you use the DEFINE statement to set your grouping variables and analysis variables. You can select different analyses for each variable using statistic keywords. With the BREAK and RBREAK statements, you can compute subtotals for levels of group variables and report totals, adding further dimensionality. The PROC REPORT code is displayed in Figure 25, and the output from PROC REPORT is in Appendix IV.

```
Proc Report Data=ClaimsMembers Headline Headskip;
   Column Sex Age Count AmtPaid;
   Define Sex / Group 'Gender';
   Define Age / Group Format=AgeGrp. 'Age Group';
   Define Count / Analysis N 'Claim Volume';
   Define AmtPaid / Analysis Sum 'Claim Payments';

   Break After Sex / Summarize;
   RBreak After / Summarize;
Run;
```

Figure 25. PROC REPORT

DIAGNOSIS CODE LOOKUP FILE

Claims file records include diagnosis codes, but do not always include diagnosis descriptions. Diagnosis code descriptions are frequently contained in a separate file. This file contains a description for each unique diagnosis code. Many vendors and data organizations maintain diagnosis code lists, and provide them to users in a downloadable format, which can be incorporated into claims files. Diagnosis code lists are available for both ICD9/10 codes, and DRG codes. The ICD9 code list is quite extensive, since there are over 15000 unique ICD9 codes, with the codes being specific to the 4th or 5th digit.

The diagnosis code file usually just contains two fields; the code and the code description. The diagnosis code file serves as a lookup table and can be incorporated with claims by merging or joining by diagnosis code. A SAS programmer would use a DATA STEP Merge or PROC SQL Join to execute this. However, another valid Base SAS technique is to construct a format from the diagnosis code list. With this lookup method, you can simply apply the format to a diagnosis code variable on the claims file.

Using PROC FORMAT it is possible to generate a data driven format. To perform this, users reference the diagnosis code file as an input file to PROC FORMAT, instead of hard-coding lookup values in a VALUE statement. The input file is called a CNTLIN data set. PROC FORMAT generates the format using the variables in the CNTLIN data set. The CNTLIN data set must contain the required variables FMTNAME, START, and LABEL. The variable END is required If you are assigning a single label to a range of values. The code for PROC FORMAT with the CNTLIN option is displayed in Figure 26 below.

```
Proc Format Library=FORMATL CNTLIN=ICD9DX;
Run;
```

Figure 26. PROC FORMAT WITH CNTLIN OPTION.

In a CNTLIN data set a user might want to include the TYPE variable which indicates the format type; numeric or character. As shown in Figure 21, using the LIBRARY= option the format can be stored in a permanent FORMAT library. This allows the format to be used in a separate SAS program. To do this, SAS must search and find the format in a format library. The FMTSEARCH= global or system option must be used to find the format in the permanent format library.

Diagnosis code lists are valuable in healthcare reporting and analytics. From the standpoint of a hospital provider, admissions for specific diagnoses entail long hospital stays, with hospital inpatient stays being a significant cost area. From the standpoint of a payor, long hospital patient stays mean larger claim bills, and then larger reimbursements. Payors have an incentive to track specific diagnoses, such as chronic diseases, which increases the financial risk a patient poses to an insurance company.

Producing a report which displays the 10 most common patient diagnoses would be a priority task for a healthcare analyst to perform. The 10 most common diagnoses are referred to as the top 10 diagnoses. PROC FREQ is the primary SAS tool to produce frequencies. With the ORDER= option, the diagnoses can be output in descending order based on frequency. The top 10 frequencies can be saved in an output data set using the OUT= option. PROC FREQ code with the ORDER= option is displayed in Figure 27.

```
Proc Freq Data=FacHeader noprint order=freq;
  Tables PDX / List Missing Out=dxfreqs(obs=10);
Run;
```

Figure 27. PROC FREQ with ORDER= Option.

For hospital providers, performance measures rate the quality of care and service that a facility provides. For accreditation, hospitals are required to submit performance measures to national organizations. For hospital inpatient facilities, patient length of stay is a valid measure for provider reporting.

Length of Stay (LOS) is defined as the number of days between the admission date and discharge date. LOS can be computed from claims files. To enhance a report, Average Length of Stay by diagnosis can be computed. Average LOS can be computed from a DATA STEP using BY–GROUP processing. Figure 28 below shows the DATA STEP code to compute Average LOS by diagnosis.

```
Data FacH_Avg_Sum;
  Set FacilityH;

  By ICD9DX_Description;

  If First.ICD9DX_Description Then Do;
      Count=0;
      LOS_Sum=LOS;
  End;

  Count+1;
  LOS_Sum+LOS;

  If Last.ICD9DX_Description Then Do;
      AvgLOS=LOS_Sum/Count;
      Output;
  End;

  Keep ICD9DX_Description Count AVGLOS;
 Run;
```

Figure 28. DATA STEP with BY-GROUP processing.

With BY-Group processing, it is required to sort the data set on the BY variable using PROC SORT first. BY-Group processing creates two temporary variables; FIRST. and LAST. Within a DATA STEP, the FIRST. and LAST. variables can be used to flag the starting and ending records of a By Group, to perform computations at those points. Also, the SUM statement is used to aggregate records across By Groups. The code in Figure 28 produces a summary data set containing Counts of Claims and Average Length of Stay by Diagnosis. The report created from this data set is provided in Appendix V.

CONCLUSION

With all the distinct types and sources of data, administrative healthcare data structures are very complex and nuanced. The SAS System includes robust tools and constructs for tackling the tangled web of administrative healthcare data and reducing its complexity. For each type of task with healthcare data a SAS programmer or analyst needs to perform, from validation and quality control, to manipulation, to reporting and analytics, SAS provides a wealth of capable programming tools. This workshop was intended to give attendees some practical experience using these tools with healthcare data to provide exposure to what a programmer or analyst might encounter in a real-world setting.

REFERENCES

Marge Scerbo, Craig Dickstein, and Alan Wilson. *Healthcare Data and the SAS System*. Cary, NC: SAS Institute, 2001

Dickstein, Craig and Renu Gehring. 2014. *Administrative Healthcare Data: A Guide to its Origin, Content, and Application using SAS*. Cary, NC: SAS Institute Inc.

Wikipedia. “International Statistical Classification of Diseases and Related Health Problems”. Accessed August 22, 2018. https://en.wikipedia.org/wiki/International_Statistical_Classification_of_Diseases_and_Related_Health_Problems.

Wikipedia. “ICD10 Procedure Coding System”. Accessed August 22, 2018. https://en.wikipedia.org/wiki/ICD-10_Procedure_Coding_System.

ACKNOWLEGEMENTS

The author would like to thank Craig Brelage, PharmaSUG 2026 Operations Chair, Eunice Ndungu, PharmaSUG 2026 Academic Chair, Jyothi Ketavarapu and Scott Burroughs, Hands-on Training Section Co-Chairs, and the PharmaSUG 2026 Executive Committee and Conference Team for accepting my abstract and paper and for organizing the conference.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jay Iyengar
Data Systems Consultants LLC
Oak Brook, IL 60523
Email: datasyscon@gmail.com
Linkedin: https://www.linkedin.com/in/jisasprogconsult

Jay Iyengar is Director of Data Systems Consultants LLC. He is a SAS consultant, trainer, and SAS Certified Advanced Programmer. He’s been an invited speaker at several SAS user group conferences (WIILSU, WCSUG, SESUG) and has presented papers and training seminars at SAS Global Forum, Pharmaceutical SAS Users Group (PharmaSUG), and other regional and local SAS User Group conferences (MWSUG, NESUG, WUSS, MISUG). He was co-leader and organizer of the Chicago SAS Users Group (WCSUG) from 2015-19. He received his bachelor's degree from Syracuse University in Public Policy and Economics, and his master's degree from the American University.

TRADEMARK CITATION

APPENDIX I

Data Set Name	ADMHLTH.FACILITYHEADER	**r**	6654
Member Type	DATA	**Variables**	27
Engine	V9	**Indexes**	0
Created	10/21/2014 12:00:29	**Observation Length**	256
Last Modified	10/21/2014 12:00:29	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	27
First Data Page	1
Max Obs per Page	255
Obs in First Data Page	239
Number of Data Set Repairs	0
Filename	/folders/myfolders/SAS Data Sets/AdmHealthData/facilityheader.sas7bdat
Release Created	9.0401M1
Host Created	X64_7PRO
Inode Number	139102
Access Permission	rw-rw-r--
Owner Name	Sasdemo
File Size	2MB
File Size (bytes)	1835008

Alphabetic List of Variables and Attributes						
#	**Variable**	**Type**	**Len**	**Format**	**Informat**	**Label**
6	AdmitDt	Num	8	YYMMDD10.		Admission Date
9	AmtPaid	Num	8	12.2		Claim Paid Amount
11	BillType	Char	3	$3.	$3.	Type of Bill
1	ClaimID	Char	10	$10.		Claim Number
4	DOB	Num	8	YYMMDD10.	MMDDYY10.	Member Date of Birth
8	DRG	Char	5	$3.		Derived DRG
26	DRGDesc	Char	58			
7	DischgDt	Num	8	YYMMDD10.		Discharge Date
13	Dx_2	Char	5	$5.	$5.	Other Diagnosis 1
14	Dx_3	Char	5	$5.	$5.	Other Diagnosis 2
15	Dx_4	Char	5	$5.	$5.	Other Diagnosis 3
16	Dx_5	Char	5	$5.	$5.	Other Diagnosis 4
17	Dx_6	Char	5	$5.	$5.	Other Diagnosis 5
18	Dx_7	Char	5	$5.	$5.	Other Diagnosis 6
19	Dx_8	Char	5	$5.	$5.	Other Diagnosis 7
2	MemberID	Char	17	$17.		Member ID
27	PDXDesc	Char	35			
10	PaidDt	Num	8	YYMMDD10.		Claim Paid Date
12	Pdx	Char	5	$5.	$5.	Principal Diagnosis
20	Ppx	Char	5	$5.	$5.	Principal Surgical Procedure
5	ProviderID	Char	10	$10.		Servicing Provider
21	Px_2	Char	5	$5.	$5.	Other Surgical Procedure 1
22	Px_3	Char	5	$5.	$5.	Other Surgical Procedure 2
23	Px_4	Char	5	$5.	$5.	Other Surgical Procedure 3
24	Px_5	Char	5	$5.	$5.	Other Surgical Procedure 4
25	Px_6	Char	5	$5.	$5.	Other Surgical Procedure 5
3	Sex	Char	1	$1.	$1.	Member Sex

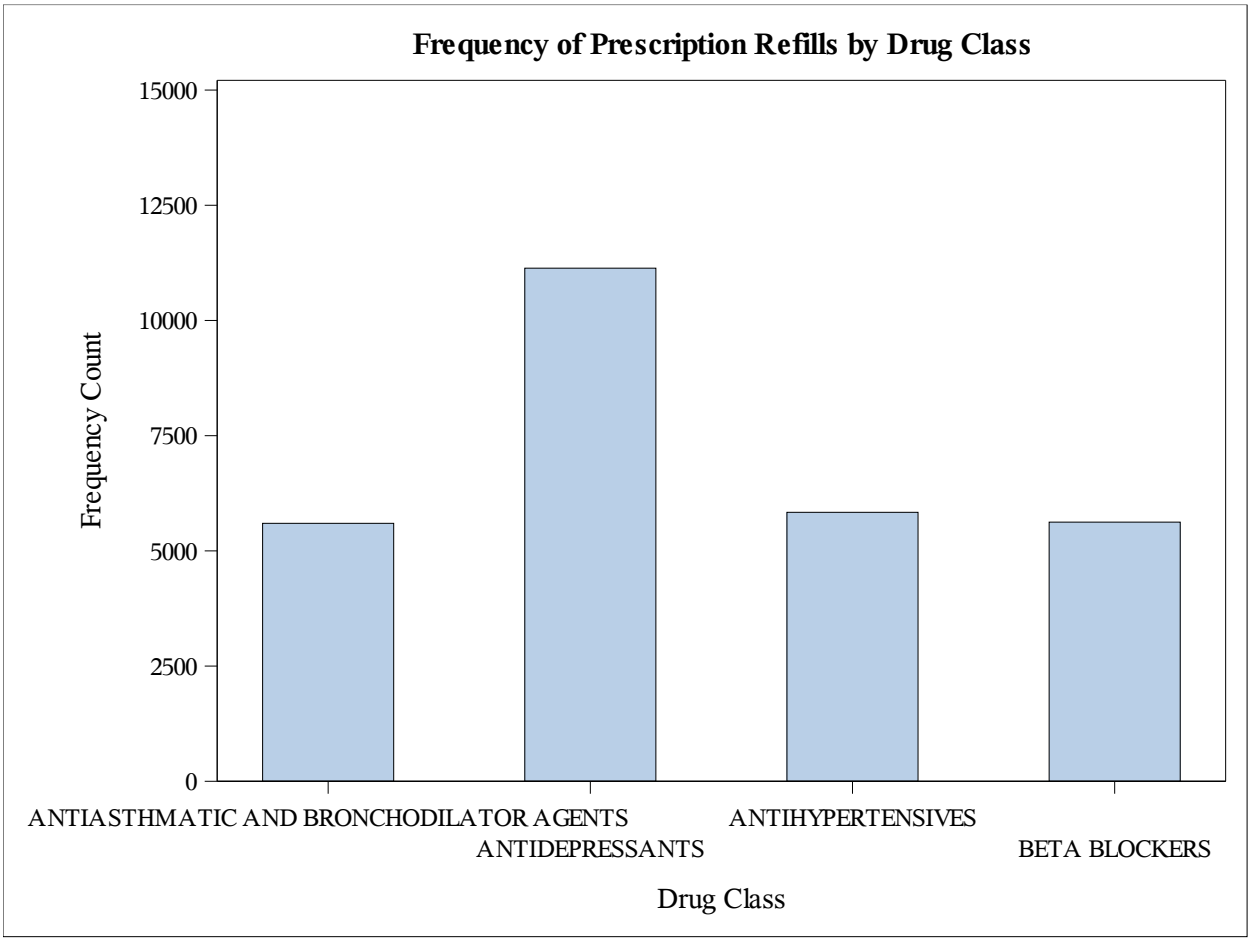
APPENDIX II

		Number of Claims	Claim Payments Sum
Provider	**Procedure**	1	$159.44
P623456834	**Procedure**		
	HOM HLTH AIDE/CNA PROV CARE HOM; HR	19	$7,515.54
	NRS CARE HOM; REGISTERED NURSE-HOUR	79	$21,462.37
	NURSING CARE THE HOME; LPN PER HOUR	10	$1,938.66
	OCCUPATIONAL THERAPY HOME PER DIEM	5	$2,174.04
	PHYSICAL THERAPY; HOME PER DIEM	9	$2,046.15
	Total	122	$35,136.76
P833477652	**Procedure**		
	NON-EMERG TRNSPRT; ENCOUNTER/TRIP	42	$1,822.09
	NONEMERG TRNSPRT: WHEELCHAIR VAN	99	$8,477.48
	NONEMERGENCY TRNSPRT; STRETCHER VAN	3	$619.78
	Total	144	$10,919.34

		Number of Claims	Claim Payments
			Sum
P843456976	Procedure		
	BREATHING CIRCUITS	1	$16.35
	CONTINUOUS AIRWAY PRESSURE DEVICE	1	$87.18
	DME MISCELLANEOUS	2	$35.68
	FCE MASK INTERFCE REPL FULL MASK EA	1	$0.00
	HUMDIFIR HEAT USED W/POS ARWAY PRSS	3	$344.60
	O2 CNTN GASEOUS 1 U = 1 CUBIC FOOT	65	$3,982.00
	O2 CONC 85%/>02 CONC PRSC FLW RATE	33	$4,401.10
	O2 CONTENTS LQD 1 U EQUALS 1 POUND	105	$5,607.84
	OXIMETER MSR BLD O2 LEVL NON-INVASV	1	$108.38
	PRTBLE GASEOUS O2 SYS RENTAL;	24	$1,251.03
	PRTBLE LIQUID O2 SYS RENTAL;	1	$28.31
	REGULATOR	1	$16.22
	RESP ASST DEVC BI-LEVL PRSS CAPABIL	5	$1,464.66
	RESP SUCTN PUMP HOME MODEL ELEC	1	$25.54
	STATION LIQUID O2 SYS RENTAL;	41	$5,784.84
	TRACHEOST/LARYNGECT TUBE CUFF PVC	2	$0.00
	VCV W/O PRESS SUPP INVASV INTRFACE	1	$0.00
	Total	288	$23,153.73

APPENDIX III

GPI_Desc	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ANTIDEPRESSANTS	11135	39.5	11135	39.5
ANTIHYPERTENSIVES	5837	20.7	16972	60.2
BETA BLOCKERS	5624	19.9	22596	80.1
ANTIASTHMATIC AND BRONCHODILATOR AGENTS	5598	19.9	28194	100.00



APPENDIX IV

Gender	Age Group	Claim Volume	Claim Payments
F	0-20	329	$2,864,724
	100+	3	$21,873
	21-40	850	$5,312,119
	41-60	835	$7,726,998
	61-80	1278	$10,347,868
	81-100	694	$4,881,086
F		3989	$31,154,669
M	0-20	311	$2,923,444
	21-40	245	$2,149,735
	41-60	659	$6,485,182
	61-80	1032	$10,036,820
	81-100	418	$3,147,592
M		2665	$24,742,773
		6654	$55,897,442

APPENDIX V

ICD9DX_Description	Count	Average Length of Stay
ATRIAL FIBRILLATION	69	4.8
CORONARY ATHEROSCLEROSIS NATIVE CORONARY ARTERY	137	2.8
ENCOUNTER FOR ANTINEOPLASTIC CHEMOTHERAPY	141	5.5
END STAGE RENAL DISEASE	73	25.5
LOC OSTEOARTHROS NOT SPEC PRIM/SEC LOWER LEG	85	3.1
OBSTRUCTIVE CHRONIC BRONCHITIS WITH EXACERBATION	84	3.5
OTHER SPECIFIED REHABILITATION PROCEDURE OTHER	157	16.5
PNEUMONIA, ORGANISM UNSPECIFIED	163	5.4
PREV C/S DELIV DELIV W/WO MENTION ANTPRTM COND	82	3.1
UNSPECIFIED SEPTICEMIA	96	7.6