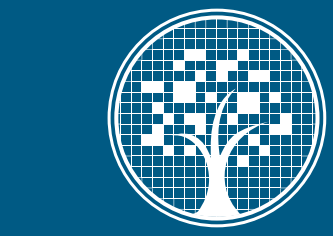

HCUP Methods Series

Final PNUM Linkage Variable Report of the State Databases Available Through the HCUP Central Distributor, 2003
Report # 2006-02

U.S. Department of Health and Human Services
Agency for Healthcare Research and Quality

Contact Information:
Healthcare Cost and Utilization Project (HCUP)
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
http://www.hcup-us.ahrq.gov
For Technical Assistance with HCUP Products:

Email: hcup@ahrq.gov

or

Phone: 1-866-290-HCUP

Recommended Citation: Busch J, Steiner C, Hickey K, Nordstrand N, Immel B, Whalen D. *Final PNUM Linkage Variable Report of the State Databases Available Through the HCUP Central Distributor, 2003.* HCUP Methods Series Report # 2006-02 Online. July 7, 2006. U.S. Agency for Healthcare Research and Quality. Available:http://www.hcup-us.ahrq.gov/reports/methods.jsp

Table of Contents

Index of Tables

INTRODUCTION

This document provides an evaluation of the PNUM_R data element in publicly-available HCUP data sets (i.e., those available through the HCUP Central Distributor) for the 2003 data year. States provide the PNUM data element to allow for constructing person-level analyses, and an encrypted version of this identifier is available on certain data sets available through the HCUP Central Distributor. Such an identifier is useful in conducting an array of health service research studies, including analysis of chronic conditions, readmission patterns, and resource utilization across settings. Ideally, this identifier is unique and specific to each individual over time and across settings. In practice, researchers have noted that PNUM may not always be an adequate identifier and that PNUM alone may not be useful in identifying individuals.

This activity is designed to assess the characteristics of the PNUM_R data element across states and settings. The specific objective is to examine the PNUM_R in light of three criteria essential for data quality (Statistics Canada, 2003). Brief descriptions of these criteria are presented below. Methods and results associated with each characteristic are described in sections that follow. Relevant aspects of data quality are:

- Completeness
- Accuracy
- Consistency.

Completeness

An obvious desirable characteristic for PNUM_R is that it is present (non-missing) for most or all of the records in each database. Completeness refers to the extent to which PNUM is present with each state's database(s). Not all states submit PNUM, and states that do submit PNUM may have missing values for some records. For this report, analyses were limited to those states that provided a version of PNUM through the HCUP Central Distributor in data year 2003. Specifically, the PNUM_R data element is used since this is an encrypted version of the PNUM data element supplied by the data source.

Accuracy

For non-missing values, a desirable attribute of PNUM_R is that it has the capacity to identify each person in a data set. Likewise, it is desirable if the same value of PNUM_R is used for all records that represent the same person. These characteristics describe ability of a data element to accurately capture what is designed to measure.

Consistency

Assuming PNUM_R values are present and the data element is well-constructed, a third desirable characteristic is the extent to which other data elements "agree" with PNUM_R when identifying individuals. For instance, if two observations match on PNUM_R then the two observations should represent the same person. Assuming other data elements are accurately coded, the two observations should have identical values for certain fields (e.g., DOB, FEMALE, AGE, RACE). Other fields (e.g., ZIP) should also agree provided the underlying characteristics do not change.

OVERVIEW OF METHODS

In this section the cross-cutting methods used in these analyses are described. Since various analyses used different inclusion/exclusion criteria, detailed descriptions of the analysis-specific methods are presented in the Completeness, Accuracy, and Consistency sections below.

Source Data

SID, SASD, and SEDD data were analyzed for each state that released a version of PNUM_R through the HCUP Central Distributor in data year 2003. Table 1 summarizes the availability of PNUM_R for states and databases included in this analysis. As noted in the table, relatively few states allow PNUM_R to be released through the central distributor. In addition to the 5 states that release a version of PNUM through the HCUP Central Distributor, 5 other HCUP Central Distributor states (Florida, Massachusetts, Nebraska, Virginia, and Wisconsin) include a version of PNUM on the intramural data files. Further evaluation of PNUM is available and AHRQ staff can be contacted to facilitate potential access to that data element for the states in the HCUP Central Distributor.

The scope of this report is limited to those states that include PNUM_R in the data year 2003 databases released through the HCUP Central Distributor. These data collectively represent approximately 3 million observations[1] from 5 states.

Table 1: Availability of PNUM_R by State and Data Type, 2003 Data Year

State	Inpatient (SID)		Ambulatory Surgery (SASD)		Emergency Department (SEDD)	
	In Central Distributor	PNUM_R Available	In Central Distributor	PNUM_R Available	In Central Distributor	PNUM_R Available
AZ	●	●				
CO	●		●			
FL	●		●			
IA	●					
KY	●		●			
MA	●				●	
MD	●		●		●	
ME	●		●		●	
MI	●					
NC	●	●	●	●		
NE	●		●		●	
NJ	●		●			
NV	●	●	●			
OR	●					
RI	●					
UT	●		●		●	●
WA	●	●				
WI	●		●			
WV	●					

[1] Observation: A record or row in the source data. In SID databases, each observation should represent a distinct inpatient stay.

Classification of Observations

Based on previous experience with PNUM_R-based analyses, each observation was classified as either "newborn," "maternity," or "all other." This sorting was done to identify possible sources of variation in terms of missing values (anecdotal evidence suggests that PNUM_R is often missing for newborns) and false matches (anecdotal evidence also suggests that newborns are sometimes assigned the same PNUM_R as their mother). For purposes of analyses, newborns were identified as any observation with an AGE of less than one year. Maternity discharges were classified using the NEOMAT data element, and any non-newborn, non-maternity observation was classified as "all other." The NEOMAT data element identifies discharges with neonatal and/or maternal diagnoses and procedures. After these initial classifications were assigned, the list of PNUM_Rs for maternity observations was compared to the list of PNUM_Rs for "all other" observations. Any value of PNUM_R that appeared in both the initial "maternity" list and the initial "all other" list was removed from the "all other" list.[2] This deletion was done to account for the possibility of maternity patients being admitted prior or subsequent to delivery. Thus, the resulting categories are mutually exclusive and exhaustive.

COMPLETENESS

The primary issue with respect to data completeness is the proportion of observations for which PNUM_R is non-missing. These analyses use all available records from each database. Tables 2, 3, and 4 summarize the proportion of observations with missing PNUM_R values for the SID, SASD, and SEDD, respectively. Appendix A includes a parallel set of tables that display the number (rather than proportion) of observations with blank/non-missing PNUM_Rs.

Results indicate that most observations contain a PNUM_R value, and that the proportion of observations with missing values is roughly equivalent across the three databases. Across all states represented here, PNUM_R is missing for approximately 20 percent of the observations in the SID, 39 percent of the observations in the SASD, and 21 percent of the observations in the SEDD. For both the SASD and SEDD states, the calculations are based on only one state, and the full research files may have a lower observed proportion of records with missing PNUM values.

Table 2: Proportion of Observations with Missing and Non-Missing Values for PNUM_R by Patient Type, Central Distributor SID 2003 Data

State	Category	Maternity (percent)	Newborns (percent)	All Others (percent)	All Patients (percent)
AZ	Observations with Non-Missing PNUM_Rs	92	62	97	91
	Observations with Blank/Missing PNUM_Rs	8	38	3	9
NC	Observations with Non-Missing PNUM_Rs	64	21	64	59
	Observations with Blank/Missing PNUM_Rs	36	79	36	41
NV	Observations with Non-Missing PNUM_Rs	94	54	97	90
	Observations with Blank/Missing PNUM_Rs	6	46	3	10
WA	Observations with Non-Missing PNUM_Rs	100	100	100	100
	Observations with Blank/Missing PNUM_Rs	0	0	0	0

[2] This step was ultimately not performed for Wisconsin data because the limited PNUM length resulted in almost all non-newborn observations being classified as maternity. These preliminary analyses resulted in approximately 450,000 maternity discharges, 75,000 newborns, and 125,000 "all other" discharges.

Table 3: Proportion of Observations with Missing and Non-Missing Values for PNUM_R by Patient Type, Central Distributor SASD 2003 Data

State	Category	Maternity (percent)	Newborns (percent)	All Others (percent)	All Patients (percent)
NC	Observations with Non-Missing PNUM_Rs	4	47	100	61
	Observations with Blank/Missing PNUM_Rs	96	53	0	39

Table 4: Proportion of Observations with Missing and Non-Missing Values for PNUM_R by Patient Type, Central Distributor SEDD 2003 Data

State	Category	Maternity (percent)	Newborns (percent)	All Others (percent)	All Patients (percent)
UT	Observations with Non-Missing PNUM_Rs	100	13	0	79
	Observations with Blank/Missing PNUM_Rs	0	87	100	21

ACCURACY

In terms of evaluating the accuracy of PNUM_R, the first issued addressed is the "capacity" of PNUM_R as it was provided by the data source. Implicit in the construction of PNUM_R is the concept that the data element is constructed in such a way as to allow unique identifiers for an entire population (e.g., all potential patients within a given state). By examining the number of characters used in each state's PNUM_R, it is possible to estimate the possible number of distinct values that can be represented by the PNUM_R data element. Analysis of the length of non-missing PNUM_R values for each state indicates that most states use PNUM_R that is at least 9 digits long, thus allowing for approximately one billion possible values.

Table 5: Nominal Length of PNUM_R Data Element by State, SID 2003 Data

State	Nominal PNUM_R Length	Approximate Number of Possible PNUM_R Values	Approximate Number of Inpatient Observations (Annual)
AZ	19	Several billion	650,000
NC	9	one billion	1,000,000
NV	12	one billion	250,000
UT	9-12	one billion	250,000
WA	12	one billion	600,000

Percent Duplicates

A second aspect of accuracy relates to the frequency with which each PNUM_R value occurs in a data set. Because some persons are readmitted in the course of a year, one would not expect to see each PNUM_R only once. Conversely, it is unlikely that a single person would be readmitted more than a dozen times per year (though some outliers with more than twelve visits in a year are possible). For these analyses, all non-missing observations were examined and measures were calculated as to whether each distinct PNUM_R appeared once or more than once in a data year. Observations linked to a PNUM_R that appears only once in that year are labeled "singletons," whereas observations linked to a PNUM_R that appeared two or more times have "recurring PNUM_Rs." For instance, suppose the PNUM_R 'A001A' appears only once in a data set and there are five observations with a PNUM_R value of 'A001B.' All five observations with a PNUM_R of 'A001B' would be flagged as having a recurring PNUM_R. The observation with a PNUM_R value of 'A001A' would be marked as a singleton. While the five 'recurring PNUM_R' observations may or may not represent the same person, these analyses provide some insight as to the efficacy of using PNUM_R to identify readmissions.

Results of these analyses are displayed in Tables 6, 7, and 8. Although the recurring PNUM_Rs do not necessarily represent readmissions, it is useful to compare the obtained values to benchmark readmission rates. For purposes of comparison, an all-population readmission rate of 5.5 percent was used. Based on this benchmark, one would expect to see the preponderance of observations in the "singleton" category. For example, with inpatient data one would expect to see approximately 5,500 observations with recurring PNUM_Rs for every 100,000 singletons.

Except in the case of the ED data, across all states and patient types, the number of observations linked to recurring PNUM_Rs is less than the number singleton observations. Within and across states, the proportion of singletons varied by patient type, with maternity discharges and newborn discharges representing a high percentage of singletons (73 percent and 72 percent, respectively) across all data types. For the "all other" population, the majority of observations (56 percent) were associated with singleton PNUM_R values, although the relative increase in observations associated with recurring PNUM_R values suggest this population is more likely to be re-admitted.

Table 6: Proportion of Observations Linked to Singleton versus Recurring PNUM_Rs, by Patient Type - Central Distributor SID 2003 Data

State	Category	Maternity (percent)		Newborns (percent)		All Others (percent)		Total (percent)	
		PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations
AZ	Singleton PNUM_Rs	92	70	76	70	79	60	81	62
	Recurring PNUM_Rs	8	30	24	30	21	40	19	38
NC	Singleton PNUM_Rs	91	78	89	49	73	50	76	54
	Recurring PNUM_Rs	9	22	11	51	27	50	24	46
NV	Singleton PNUM_Rs	93	85	91	74	77	56	81	62
	Recurring PNUM_Rs	7	15	9	26	23	44	19	38
WA	Singleton PNUM_Rs	93	85	91	83	75	53	81	62
	Recurring PNUM_Rs	7	15	9	17	25	47	19	38

Table 7: Proportion of Observations Linked to Singleton versus Recurring PNUM_Rs, by Patient Type - Central Distributor SASD 2003 Data

State	Category	Maternity (percent)		Newborns (percent)		All Others (percent)		All Patients (percent)	
		PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations
NC	Singleton PNUM_Rs	78	43	95	70	82	66	82	65
	Recurring PNUM_Rs	22	57	5	30	18	34	18	35

Table 8: Proportion of Observations Linked to Singleton versus Recurring PNUM_Rs, by Patient Type - Central Distributor SASD 2003 Data

State	Category	Maternity (percent)		Newborns (percent)		All Others (percent)		All Patients (percent)	
		PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations
UT	Singleton PNUM_Rs	59	28	69	36	73	46	73	45
	Recurring PNUM_Rs	41	72	31	64	27	54	27	55

Frequency with which PNUM_R Values Appear

Examination of the number of singleton PNUM_Rs led to an analysis of how frequently values of PNUM_R appeared. For these analyses, observations with missing PNUM_R values were excluded and list of distinct PNUM_R values was constructed for each database. The number of times each distinct value of PNUM_R appeared in the data set were then examined. Results for each database are included in Appendix A, with sample output below. Table 9 uses Nevada SID data to illustrate a typical pattern associated with these data. As shown in the table, a total of 91,982 observations are associated with various values of PNUM_R that appeared only once. Another 18,354 observations are linked to various other values of PNUM_R that appeared twice, and so on, culminating with one value of PNUM_R that appeared a total of 26 times. It is unlikely that this represents one person who was hospitalized 26 times in one year, and there are myriad possible causes for a given value of PNUM_R to appear numerous times in the same year. Such causes include submission of test data, clerical errors resulting in multiple discharge abstracts, and attempts to submit correcting or adjusted claims.

Table 9: Frequency of Occurrence for PNUM_R Values – Nevada Central Distributor SID 2003 Data

Frequency of Occurrence	Number of Observations
1	91,982
2	18,354
3	5,359
4	2,119
5	978
6	490
7	210
8	138
9	60
10	43
11	31
12	15
13	9
14	4
15	11
16	3
17	2
18	3
25	1
26	1

By comparison, Table 10 illustrates a somewhat unusual pattern: in this case a small set of PNUM_R values that appear very frequently, with one value occurring 7,636 times and seven other values occurring at least 100 times. In a practical sense, these PNUM_R values are effectively missing since they do not add any new information about each observation.

Table 10: Frequency of Occurrence for PNUM_R Values North Carolina Central Distributor SID 2003 Data

Frequency of Occurrence	Number of Observations
1	340,792
2	67,953
3	20,941
4	8,309
5	3,647
6	1,762
7	889
8	498
9	273
10	175
11	112
12	64
13	38
14	35
15	18
16	17
17	7
18	7
19	5
20	1
21	5
22	4
24	1
25	1
26	1
27	1
29	1
34	1
35	1
51	1
128	1
292	1
303	1
507	1
825	1
976	1
2,079	1
7,636	1

Number of Duplicate Records

Findings thus far illustrate that some values of PNUM_R are repeated in the database and some values appear quite frequently. A logical follow-on is to examine the databases for duplicate records by checking for observations that match on every data element. These “clone” records likely represent duplicate submissions from the facility or the data source.

Results of these analyses are presented in Tables 11, 12, and 13. Duplicate records do not appear to be a significant issue for most states, although a notable exception is North Carolina SASD. While the overall proportion of duplicate observations is still relatively low (less than one percent) for this state, the potential for more than 2,000 duplicates may impact certain types of analyses.

Table 11: Number of Duplicate Records by State and Patient Type – Central Distributor SID 2003 Data

State	Category	Maternity	Newborns	All Others	All Patients
AZ	Duplicate Records	2	10	12	24
	AZ Total Observations	100,162	100,609	462,745	663,516
NC	Duplicate Records	48	202	505	755
	NC Total Observations	129,529	127,959	811,226	1,068,714
NV	Duplicate Records	0	48	2	50
	NV Total Observations	34,993	35,450	170,344	240,787
WA	Duplicate Records	0	38	2	40
	WA Total Observations	83,463	84,828	420,742	589,033

Table 12: Number of Duplicate Records by State and Patient Type – Central Distributor SASD 2003 Data

State	Category	Maternity	Newborns	All Others	All Patients
NC	Duplicate Records	72	28	2,103	2,203
	NC Total Observations	24,911	8,823	1,181,961	1,215,695

Table 13: Number of Duplicate Records by State and Patient Type – Central Distributor SEDD 2003 Data

State	Category	Maternity	Newborns	All Others	All Patients
UT	Duplicate Records	0	0	4	4
	UT Total Observations	24,567	25,593	514,918	565,078

CONSISTENCY

The final set of analyses address the extent to which other data elements are consistent with PNUM_R. The underlying premise for these analyses is that if two or more records match on PNUM_R, these records should represent the same person. By extension, values for key demographic variables (e.g., FEMALE, DOB) for observations that match on PNUM_R should also match. In other words, if two observations match on PNUM_R, they should represent the same person, and both records should have the same value for age, sex, and other demographic data elements.

To calculate the consistency measures, observations that had non-missing PNUM_R values were selected and any duplicate records were excluded. Since consistency can only be evaluated using observations that match on PNUM_R, “singleton” observations were excluded from analysis. A final restriction was to exclude any value of PNUM_R that appeared more than 100 times in a data set. Since the databases released through the HCUP Central Distributor contain limited demographic information and would permit in-depth analyses, the results presented below reflect analyses conducted on the more detailed intramural database.

For the remaining observations the SAS MERGE function was used to identify sets or “clusters” of observations that match on PNUM_R. Within each cluster, the demographic values (i.e., FEMALE, AGE, DOB, ZIP, RACE) and MRN for each observation was compared to the corresponding values for all other observations in the cluster. Although not technically a demographic data element, MRN was also included in these analyses. This approach is analogous to conducting a series of pairwise comparisons for each set of records that match on PNUM_R. Each pairwise comparison results in either agreement or disagreement, and an overall agreement score is calculated for the cluster. Agreement scores are calculated for each

data element, and scores are averaged across all clusters. The result is an average level of agreement for observations that match on PNUM_R.

Results of these analyses for SID, SASD, and SEDD databases are presented in Tables 14, 15, and 16, respectively. By way of interpretation, these results describe the average agreement among records that match on PNUM_R. Thus, if two records from the Arizona SID have the same PNUM_R, there is a 93.2 percent chance they will have the same value for FEMALE, a 72.6 percent chance they will have the same value for AGE, and so on. Across all states, databases, and data elements, levels of agreement were generally high – typically in excess of 80 percent. The highest levels of agreement are obtained for gender, whereas low levels of agreement exist for AGE, a pattern that held across databases. Low levels of agreement are a function of either miscoding of data elements, false matches on PNUM_R, or missing values. Agreement on MRN is typically low, and could be the result of a number of factors, including admissions to different facilities or facilities assigning stay-specific MRNs. The RACE data was the most varied in terms of agreement, with some states approximating 75 percent agreement and others at approximately 40 percent.

Agreement on DOB is typically greater than agreement on AGE, presumably because some patients may have a birthday between service dates. In these instances, DOB would agree but age would not. In order to explore this possibility, a measure for Age ± one year was created. As the name implies, this measure treats any two AGE values that are within one year of each other as an agreement. This does result in higher levels of agreement than obtained for "strict" age comparisons, although in some case the level of agreement exceed that obtained for DOB, suggesting some agreement in the Age ± one year measure is due to false positives.

Table 14: Levels of Agreement for Demographic Fields on Observations with Matching PNUM_R Values – Central Distributor SID 2003 Data

State	Female	Age (strict)	Age (± 1 yr)	DOB	Year of Birth	Month of Birth	Day of Birth	ZIP Code	Race	MRN
AZ	93.2%	72.6%	86.9%	86.2%	86.7%	87.9%	87.0%	94.0%	88.8%	N/A
NC	98.0%	77.8%	95.7%	94.8%	95.4%	95.9%	95.3%	90.8%	66.7%	N/A
NV	100%	83.1%	100%	100%	100%	100%	100%	86.8%	N/A	N/A
WA	99.0%	84.1%	100%	100%	100%	100%	100%	91.9%	N/A	N/A

Table 15: Levels of Agreement for Demographic Fields on Observations with Matching PNUM_R Values – Central Distributor SASD 2003 Data

State	Female	Age (strict)	Age (± 1 yr)	DOB	Year of Birth	Month of Birth	Day of Birth	ZIP Code	Race	MRN
NC	98.3%	75.4%	96.5%	96.0%	96.3%	96.7%	96.3%	93.5%	N/A	N/A

Table 16: Levels of Agreement for Demographic Fields on Observations with Matching PNUM_R Values – Central Distributor SEDD 2003 Data

State	Female	Age (strict)	Age (± 1 yr)	DOB	Year of Birth	Month of Birth	Day of Birth	ZIP Code	Race	MRN
UT	99.5%	79.9%	99.2%	98.7%	99.1%	99.2%	99.0%	89.9%	36.8%	69.7%

SUMMARY

Taken together, these results illustrate the complexity associated with using and analyzing person-level identifiers, especially in the context of administrative health data. The purpose of a variable such as PNUM_R is to uniquely identify records that represent the same person. While

the results presented here do not allow us to state definitively whether any state's PNUM_R is effective, some general assessment as to the efficacy of PNUM_R can be made.

A first recommendation is that researchers wishing to use PNUM_R conduct relatively thorough exploratory analyses prior to using PNUM_R to link records for the same individuals. Examining the number of duplicate records, number of observations with the same PNUM_R, and proportion of observation missing PNUM_R are useful starting points.

In a research context, it would seem it is possible to use the PNUM_R data element to track individuals within a database, at least once certain conditions are met. The high levels of agreement obtained for some states and databases suggest that PNUM_R (either alone or in conjunction with other data elements) can be used to identify distinct individuals within a data set. It should be noted that these levels of agreement were achieved only after certain types of records were excluded from analysis.

Finally, the value of PNUM_R varies across a number of different contexts. Differences with respect to patient types (e.g., newborns, maternity) were observed, as were differences between states, and variation between databases (i.e., SID, SASD, and SEDD). These differences imply that PNUM_R is neither universally "good" nor universally "bad." Each state appears to have different issues surrounding their implementation of PNUM, which may present opportunities for future collaboration and development.

Reference

Statistics Canada (2003). Statistics Canada Quality Guidelines. Statistics Canada Catalogue no. 12-539-XIE.

Appendix A – Number of Observations with Missing and Non-Missing values for PNUM_R by Patient Type

Table 17: Number of Observations with Missing and Non-Missing values for PNUM_R by Patient Type, Central Distributor SID 2003 Data

State	Category	Maternity	Newborns	All Others	All Patients
AZ	Observations with Non-Missing PNUM_Rs	91,726	62,388	446,627	600,741
	Observations with Blank/Missing PNUM_Rs	8,436	38,221	16,118	62,775
	AZ Total	100,162	100,609	462,745	663,516
NC	Observations with Non-Missing PNUM_Rs	83,191	27,461	522,402	633,054
	Observations with Blank/Missing PNUM_Rs	46,338	100,498	288,824	435,660
	NC Total	129,529	127,959	811,226	1,068,714
NV	Observations with Non-Missing PNUM_Rs	32,977	19,147	165,611	217,735
	Observations with Blank/Missing PNUM_Rs	2,016	16,303	4,733	23,052
	NV Total	34,993	35,450	170,344	240,787
WA	Observations with Non-Missing PNUM_Rs	83,463	84,828	420,742	589,033
	Observations with Blank/Missing PNUM_Rs	0	0	0	0
	WA Total	83,463	84,828	420,742	589,033

Table 18: Number of Observations with Missing and Non-Missing values for PNUM_R by Patient Type, Central Distributor SASD 2003 Data

State	Category	Maternity	Newborns	All Others	All Patients
NC	Observations with Non-Missing PNUM_Rs	20,495	4,153	722,198	746,846
	Observations with Blank/Missing PNUM_Rs	464,179	4,670	0	468,849
	NC Total	484,674	8,823	722,198	1,215,695

Table 19: Number of Observations with Missing and Non-Missing values for PNUM_R by Patient Type, Central Distributor SEDD 2003 Data

State	Category	Maternity	Newborns	All Others	All Patients
UT	Observations with Non-Missing PNUM_Rs	441,700	3,441	0	445,141
	Observations with Blank/Missing PNUM_Rs	1,330	22,152	96,455	119,937
	UT Total	443,030	25,593	96,455	565,078

Appendix B – Number of Observations Linked to Singleton versus Recurring PNUM_Rs

Table 20: Number of Observations Linked to Singleton versus Recurring PNUM_Rs, by Patient Type - Central Distributor SID 2003 Data

State	Category	Maternity		Newborns		All Others		Total	
		PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations
AZ	Singleton PNUM_Rs	63,789	63,789	43,521	43,521	266,522	266,522	373,832	373,832
	Recurring PNUM_Rs	5,924	27,937	13,798	18,867	70,061	180,105	89,783	226,909
	AZ Total Non-Missing	69,713	91,726	57,319	62,388	336,583	446,627	463,615	600,741
NC	Singleton PNUM_Rs	65,109	65,109	13,398	13,398	262,285	262,285	340,792	340,792
	Recurring PNUM_Rs	6,379	18,082	1,626	14,063	96,771	260,117	104,776	292,262
	NC Total Non-Missing	71,488	83,191	15,024	27,461	359,056	522,402	445,568	633,054
NV	Singleton PNUM_Rs	28,104	28,104	14,180	14,180	91,982	91,982	134,266	134,266
	Recurring PNUM_Rs	2,210	4,873	1,354	4,967	27,831	73,629	31,395	83,469
	NV Total Non-Missing	30,314	32,977	15,534	19,147	119,813	165,611	165,661	217,735
WA	Singleton PNUM_Rs	70,915	70,915	70,118	70,118	222,580	222,580	363,613	363,613
	Recurring PNUM_Rs	5,562	12,548	6,541	14,710	74,410	198,162	86,513	225,420
	WA Total Non-Missing	76,477	83,463	76,659	84,828	296,990	420,742	450,126	589,033

Table 21: Number of Observations Linked to Singleton versus Recurring PNUM_Rs, by Patient Type - Central Distributor SASD 2003 Data

State	Category	Maternity		Newborns		All Others		Total	
		PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations
NC	Singleton PNUM_Rs	8,801	8,801	2,924	2,924	474,301	474,301	486,026	486,026
	Recurring PNUM_Rs	2,456	11,694	154	1,229	102,316	247,897	104,926	260,820
	NC Total non-missing	11,257	20,495	3,078	4,153	576,617	722,198	590,952	746,846

Table 22: Number of Observations Linked to Singleton versus Recurring PNUM_Rs, by Patient Type - Central Distributor SEDD 2003 Data

State	Category	Maternity		Newborns		All Others		Total	
		PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations	PNUM_Rs	Observations
UT	Singleton PNUM_Rs	6,396	6,396	1,248	1,248	191,828	191,828	199,472	199,472
	Recurring PNUM_Rs	4,469	16,841	558	2,193	69,404	226,635	74,431	245,669
	UT Total non-missing	10,865	23,237	1,806	3,441	261,232	418,463	273,903	445,141

Appendix C – Frequencies for Number of Occurrences for PNUM_R Values

SID

AZ : Unformatted Freq of PNUMCOUNT for ALL Records

pnumcount	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	373832	80.63	373832	80.63
2	65077	14.04	438909	94.67
3	15308	3.30	454217	97.97
4	5140	1.11	459357	99.08
5	2182	0.47	461539	99.55
6	954	0.21	462493	99.76
7	469	0.10	462962	99.86
8	284	0.06	463246	99.92
9	136	0.03	463382	99.95
10	74	0.02	463456	99.97
11	49	0.01	463505	99.98
12	42	0.01	463547	99.99
13	15	0.00	463562	99.99
14	7	0.00	463569	99.99
15	9	0.00	463578	99.99
16	4	0.00	463582	99.99
17	5	0.00	463587	99.99
18	5	0.00	463592	100.00
19	3	0.00	463595	100.00
21	3	0.00	463598	100.00
22	1	0.00	463599	100.00
23	1	0.00	463600	100.00
24	2	0.00	463602	100.00
25	1	0.00	463603	100.00
31	1	0.00	463604	100.00
39	2	0.00	463606	100.00
47	1	0.00	463607	100.00
51	1	0.00	463608	100.00
63	1	0.00	463609	100.00
73	1	0.00	463610	100.00
151	1	0.00	463611	100.00
482	1	0.00	463612	100.00
526	1	0.00	463613	100.00
1058	1	0.00	463614	100.00
1610	1	0.00	463615	100.00

NC : Unformatted Freq of PNUMCOUNT for ALL Records

pnumcount	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	340792	76.48	340792	76.48
2	67953	15.25	408745	91.74
3	20941	4.70	429686	96.44
4	8309	1.86	437995	98.30
5	3647	0.82	441642	99.12
6	1762	0.40	443404	99.51
7	889	0.20	444293	99.71
8	498	0.11	444791	99.83
9	273	0.06	445064	99.89
10	175	0.04	445239	99.93
11	112	0.03	445351	99.95
12	64	0.01	445415	99.97
13	38	0.01	445453	99.97
14	35	0.01	445488	99.98
15	18	0.00	445506	99.99
16	17	0.00	445523	99.99
17	7	0.00	445530	99.99
18	7	0.00	445537	99.99
19	5	0.00	445542	99.99
20	1	0.00	445543	99.99
21	5	0.00	445548	100.00
22	4	0.00	445552	100.00
24	1	0.00	445553	100.00
25	1	0.00	445554	100.00
26	1	0.00	445555	100.00
27	1	0.00	445556	100.00
29	1	0.00	445557	100.00
34	1	0.00	445558	100.00
35	1	0.00	445559	100.00
51	1	0.00	445560	100.00
128	1	0.00	445561	100.00
292	1	0.00	445562	100.00
303	1	0.00	445563	100.00

507	1	0.00	445564	100.00
825	1	0.00	445565	100.00
976	1	0.00	445566	100.00
2079	1	0.00	445567	100.00
7636	1	0.00	445568	100.00

NV : Unformatted Freq of PNUMCOUNT for ALL Records

pnumcount	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	134266	81.05	134266	81.05
2	20795	12.55	155061	93.60
3	5867	3.54	160928	97.14
4	2357	1.42	163285	98.57
5	1124	0.68	164409	99.24
6	577	0.35	164986	99.59
7	272	0.16	165258	99.76
8	176	0.11	165434	99.86
9	80	0.05	165514	99.91
10	56	0.03	165570	99.95
11	35	0.02	165605	99.97
12	21	0.01	165626	99.98
13	9	0.01	165635	99.98
14	5	0.00	165640	99.99
15	11	0.01	165651	99.99
16	3	0.00	165654	100.00
17	2	0.00	165656	100.00
18	3	0.00	165659	100.00
25	1	0.00	165660	100.00
26	1	0.00	165661	100.00

WA : Unformatted Freq of PNUMCOUNT for ALL Records

pnumcount	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	363613	80.78	363613	80.78
2	58513	13.00	422126	93.78
3	16485	3.66	438611	97.44
4	5968	1.33	444579	98.77
5	2581	0.57	447160	99.34
6	1361	0.30	448521	99.64
7	686	0.15	449207	99.80
8	341	0.08	449548	99.87
9	200	0.04	449748	99.92
10	119	0.03	449867	99.94
11	91	0.02	449958	99.96
12	63	0.01	450021	99.98
13	23	0.01	450044	99.98
14	13	0.00	450057	99.98
15	17	0.00	450074	99.99
16	15	0.00	450089	99.99
17	8	0.00	450097	99.99
18	5	0.00	450102	99.99
19	11	0.00	450113	100.00
20	3	0.00	450116	100.00
21	1	0.00	450117	100.00
22	4	0.00	450121	100.00
23	2	0.00	450123	100.00
25	1	0.00	450124	100.00
32	1	0.00	450125	100.00
36	1	0.00	450126	100.00

SASD

NC SASD: Unformatted Freq of PNUMCOUNT for ALL Records

pnumcount	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	486026	82.24	486026	82.24
2	77976	13.19	564002	95.44
3	17637	2.98	581639	98.42
4	5311	0.90	586950	99.32
5	2053	0.35	589003	99.67
6	961	0.16	589964	99.83
7	391	0.07	590355	99.90
8	238	0.04	590593	99.94
9	126	0.02	590719	99.96
10	67	0.01	590786	99.97
11	42	0.01	590828	99.98
12	32	0.01	590860	99.98

13	15	0.00	590875	99.99
14	18	0.00	590893	99.99
15	12	0.00	590905	99.99
16	7	0.00	590912	99.99
17	4	0.00	590916	99.99
18	6	0.00	590922	99.99
19	5	0.00	590927	100.00
20	2	0.00	590929	100.00
21	1	0.00	590930	100.00
22	2	0.00	590932	100.00
23	1	0.00	590933	100.00
24	1	0.00	590934	100.00
25	5	0.00	590939	100.00
27	1	0.00	590940	100.00
29	1	0.00	590941	100.00
31	1	0.00	590942	100.00
36	1	0.00	590943	100.00
46	1	0.00	590944	100.00
53	1	0.00	590945	100.00
71	1	0.00	590946	100.00
92	1	0.00	590947	100.00
156	1	0.00	590948	100.00
168	1	0.00	590949	100.00
325	1	0.00	590950	100.00
1176	1	0.00	590951	100.00
3894	1	0.00	590952	100.00

SEDD

UT SEDD: Unformatted Freq of PNUMCOUNT for ALL Records

pnumcount	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	199472	72.83	199472	72.83
2	43728	15.96	243200	88.79
3	14276	5.21	257476	94.00
4	6286	2.29	263762	96.30
5	3237	1.18	266999	97.48
6	1948	0.71	268947	98.19
7	1102	0.40	270049	98.59
8	788	0.29	270837	98.88
9	529	0.19	271366	99.07
10	417	0.15	271783	99.23
11	320	0.12	272103	99.34
12	268	0.10	272371	99.44
13	228	0.08	272599	99.52
14	179	0.07	272778	99.59
15	142	0.05	272920	99.64
16	135	0.05	273055	99.69
17	97	0.04	273152	99.73
18	77	0.03	273229	99.75
19	72	0.03	273301	99.78
20	68	0.02	273369	99.81
21	45	0.02	273414	99.82
22	61	0.02	273475	99.84
23	38	0.01	273513	99.86
24	34	0.01	273547	99.87
25	36	0.01	273583	99.88
26	27	0.01	273610	99.89
27	24	0.01	273634	99.90
28	29	0.01	273663	99.91
29	21	0.01	273684	99.92
30	19	0.01	273703	99.93
31	17	0.01	273720	99.93
32	20	0.01	273740	99.94
33	10	0.00	273750	99.94
34	8	0.00	273758	99.95
35	11	0.00	273769	99.95
36	10	0.00	273779	99.95
37	10	0.00	273789	99.96
38	7	0.00	273796	99.96
39	7	0.00	273803	99.96
40	7	0.00	273810	99.97
41	6	0.00	273816	99.97
42	7	0.00	273823	99.97
43	8	0.00	273831	99.97
44	4	0.00	273835	99.98
45	6	0.00	273841	99.98
46	5	0.00	273846	99.98
47	5	0.00	273851	99.98
48	1	0.00	273852	99.98

49	4	0.00	273856	99.98
50	2	0.00	273858	99.98
51	3	0.00	273861	99.98
52	3	0.00	273864	99.99
53	3	0.00	273867	99.99
54	2	0.00	273869	99.99
55	2	0.00	273871	99.99
56	3	0.00	273874	99.99
58	1	0.00	273875	99.99
59	2	0.00	273877	99.99
60	2	0.00	273879	99.99
61	1	0.00	273880	99.99
64	3	0.00	273883	99.99
67	1	0.00	273884	99.99
68	1	0.00	273885	99.99
69	1	0.00	273886	99.99
70	1	0.00	273887	99.99
71	3	0.00	273890	100.00
72	1	0.00	273891	100.00
74	1	0.00	273892	100.00
75	1	0.00	273893	100.00
77	1	0.00	273894	100.00
79	1	0.00	273895	100.00
83	1	0.00	273896	100.00
87	2	0.00	273898	100.00
98	1	0.00	273899	100.00
112	1	0.00	273900	100.00
115	1	0.00	273901	100.00
123	1	0.00	273902	100.00
128	1	0.00	273903	100.00