Methods to Prepare Hospital Discharge Data

By

Linda Remy, MSW, PhD
Ted Clay, MS
Geraldine Oliva, MD, MPH

Family Health Outcomes Project
University of California, San Francisco
3333 California Street, Suite 335
San Francisco, CA 94118
Phone: 415-476-5283
Fax: 415-502-0848
Email: Linda Remy: lremy@well.com
Email: Gerry Oliva: OlivaG@fcm.ucsf.edu

June 2004

TABLE OF CONTENTS

TABLE OF FIGURES

METHODS TO PREPARE HOSPITAL DISCHARGE DATA

All California hospitals except certain state or federal facilities are required to submit patient discharge data (PDD) summarizing the course of care for each discharged patient to the Office of Statewide Health Planning and Development (OSHPD). These large datafiles contain arrays of diagnoses and procedures, as well as other information to describe the patient, geographic characteristics, and the clinical course of care for every patient discharged in a given year.

OSHPD distributes the PDD to qualified researchers such as the Family Health Outcomes Project (FHOP). The FHOP human subjects protocols permit us to have the confidential PDD, for all discharges and ages, from 1983 forward. Currently we have processed all years through 2000 and are about to start with the 2001 and 2002 files. This document presents an overview of the methods we developed to create the core files we use as the source for the different PDD-based research and data products that FHOP distributes.

CONFIDENTIAL MASTER FILES

The confidential PDD includes the following data elements: Social Security Number (SSN), dates (birth, admission, discharge, procedures) and 5-digit ZIP of patient residence. We were permitted to obtain the confidential PDD with these confidential elements because some of our work involves linking PDD records either within the file in a given year or over multiple years, or to other data such as the Vital Statistics mortality files which we access using an encrypted SSN we create. To protect confidentiality, work with these files is done on stand-alone work stations. Access to the files is restricted to two key members of the FHOP research team.

Files with the SSN never reside on the work stations. We developed an algorithm to create an encrypted SSN (SSNC) that is applied as the raw PDD are read into SAS. The encryption method uses a random number process to reassign digits, while maintaining the ability to make soft-linkages to correct for data entry errors. The algorithm includes a second routine to create a second identifier (SSNCN) based on SSNC if available or other data if SSNC is not available. With exact dates of birth, admission and discharge, admission source, and disposition, and the patient's ZIP code of residence, SSNCN allows us to "soft-match" likely transfers and readmissions. Once records are linked for a given study, a unique ID is created and the SSNCN is dropped from the analytic files. This method is described elsewhere.[1] SSNs were not available before 1990. Thus linkage is more reliable for certain years and conditions than for others. We subsequently pursued extending the linkage methodology to children and young adults, and added extensive reliability checks to the procedure, given the increased uncertainty.[2]

The first sequence of programs reads the data, encrypts the SSN, creates SSNCN, does minimal edits, and creates some variables. Details vary from year to year depending on the structure of the incoming data. We check results with various diagnostic listings and descriptive statistics. For example, we may identify unformatted values for which we need to update format libraries. We also may identify data errors that need to be addressed in subsequent programs.

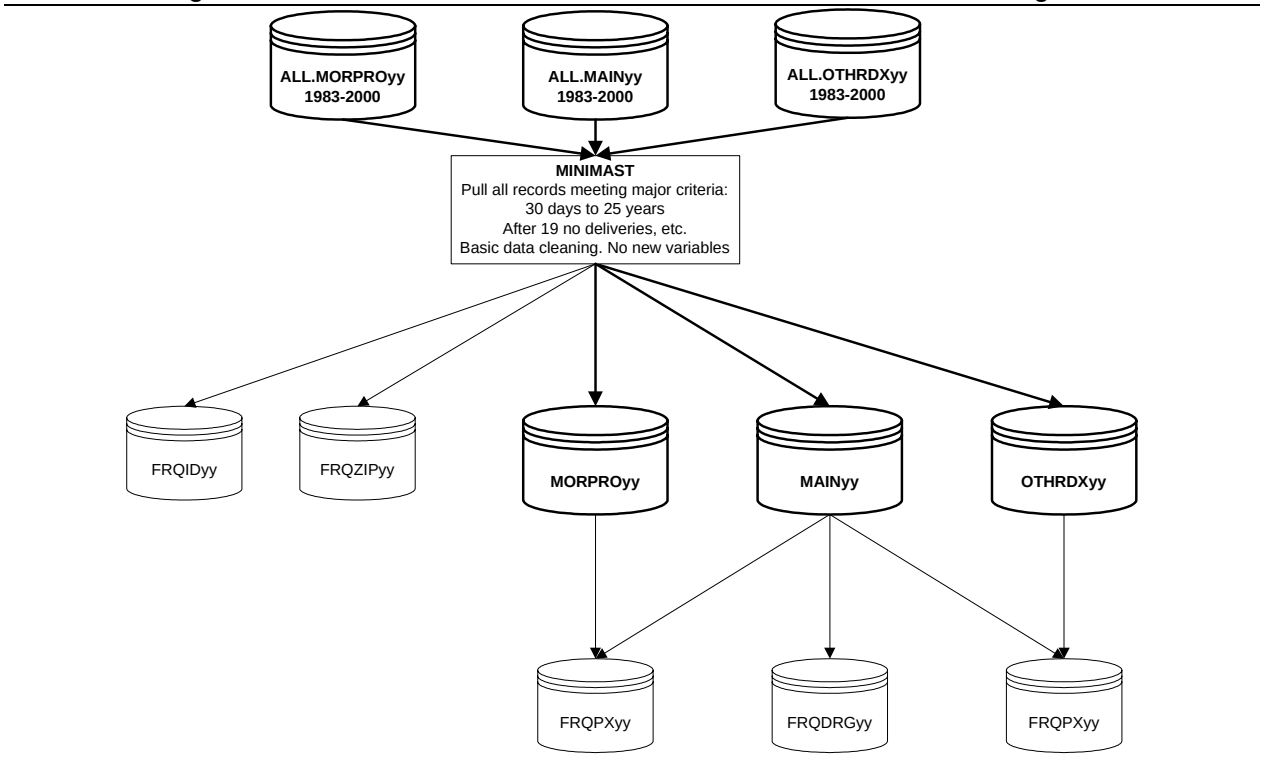
The very large PDD file is output into a main file of core data elements (MAINyy), with extra diagnoses (OTHRDXyy) and procedures (with their associated dates) (MORPROyy) removed to smaller files to reduce computing storage and overhead needs. The three files can be linked as needed using a unique discharge record identifier (YR_OBS) and SSNCN to link cases for the same person. The program also outputs a small file of records with date errors (DTERRSyy), and a small file with other variables we do not want (XVARSyy).

In addition, this series of macros outputs two summary files (FRQIDyy, FRQZIPyy) counting all admissions by geographic characteristics such as Health Service Area (HSA), Health Facilities Planning Area (HFPA), county, hospital, type of care and admission source, and ZIP. The use of these files is described in a subsequent section.

When we are confident the master files read in correctly and have identified any format or data quality issues, we run another macro (MINIMAST) to subset the data to our primary population of interest: California resident children age 28 days to young adults through age 24 at admission. The macro examines records within the range of California ZIPs or who reside in California counties and have known sex, date of birth, admission date, and discharge date. It then calculates age at admission and age at discharge, selecting records in our target age window. Because some patients may be admitted at 24 to one hospital, transferred and discharged from another hospital at age 25, we pull records through age 25. In the case of deliveries, we limit the upper age to 19. In the case of neonates, to be sure we are not getting newborns still in the hospital at 28 days, we exclude records within DRG ranges 385-391 and 760-779. Again, the macro outputs three individual-level files, a date errors file, an extra variables file, and two summary files. We use the same file names because the output data are stored in different directories.

Figure 1 is a summary of steps involved in going from the master files with all discharges to the mini-masters. Datasets are designaed with the drum shapes. The arrrows indicate the direction of the data flow. The rectangle names the program and summarizes its major tasks. Output files are below the program rectangle.

Figure 1: Create Mini-Masters and Related Files from All Discharges



GEOGRAPHIC CLASSIFICATIONS

To carry out various studies, we have to track changes over time in geographic identifiers. Although higher-level boundaries such as county are relatively stable, areas encompassing other types of geographic identifiers such as cities or ZIPs can change over time. The files we used contain a variety of geographic identifiers. For example, depending on the year of the data, the PDD includes two sets of geographic variables for the patient and the hospital: ZIP, county, Health Service Area (HSA), and Health Facilities Planning Area (HFPA). Before 1990, OSHPD identified the ZIP and county of the hospital discharging the patient and the patient ZIP of residence, but not the patient county of residence, HSA or HFPA.

For health planning purposes, California is divided into 139 Health Facility Planning Areas (HFPA).[3] The HFPA are subsumed under 14 Health Service Areas (HSA). Each HSA includes 1 to 6 of California's 58 counties within its borders. Los Angeles (LA) County is assigned to one HSA and represents about one-third of all hospital discharges annually. That County has divided itself into eight Service Provider Areas (SPA), which it uses for county planning.

The LA County Department of Public Health provided a file identifying ZIPs associated with each SPA. In some county-level analyses, we treat SPAs as "counties", increasing the number from 58 to 65.

We created and maintain a master file of every California ZIP ever reported in the PDD, vital statistics (birth, mortality), the United States Postal Service, the 1990 and 2000 census, and data from commercial vendors, together with the associated county, HSA, HFPA, and SPA. Using this master file, we impute values for missing geographic identifiers. In a separate report, we described the method by which we handle these geographic variables longitudinally.[2]

This file does not have addresses or latitude and longitude indicators, so more precise geocoding is not possible.

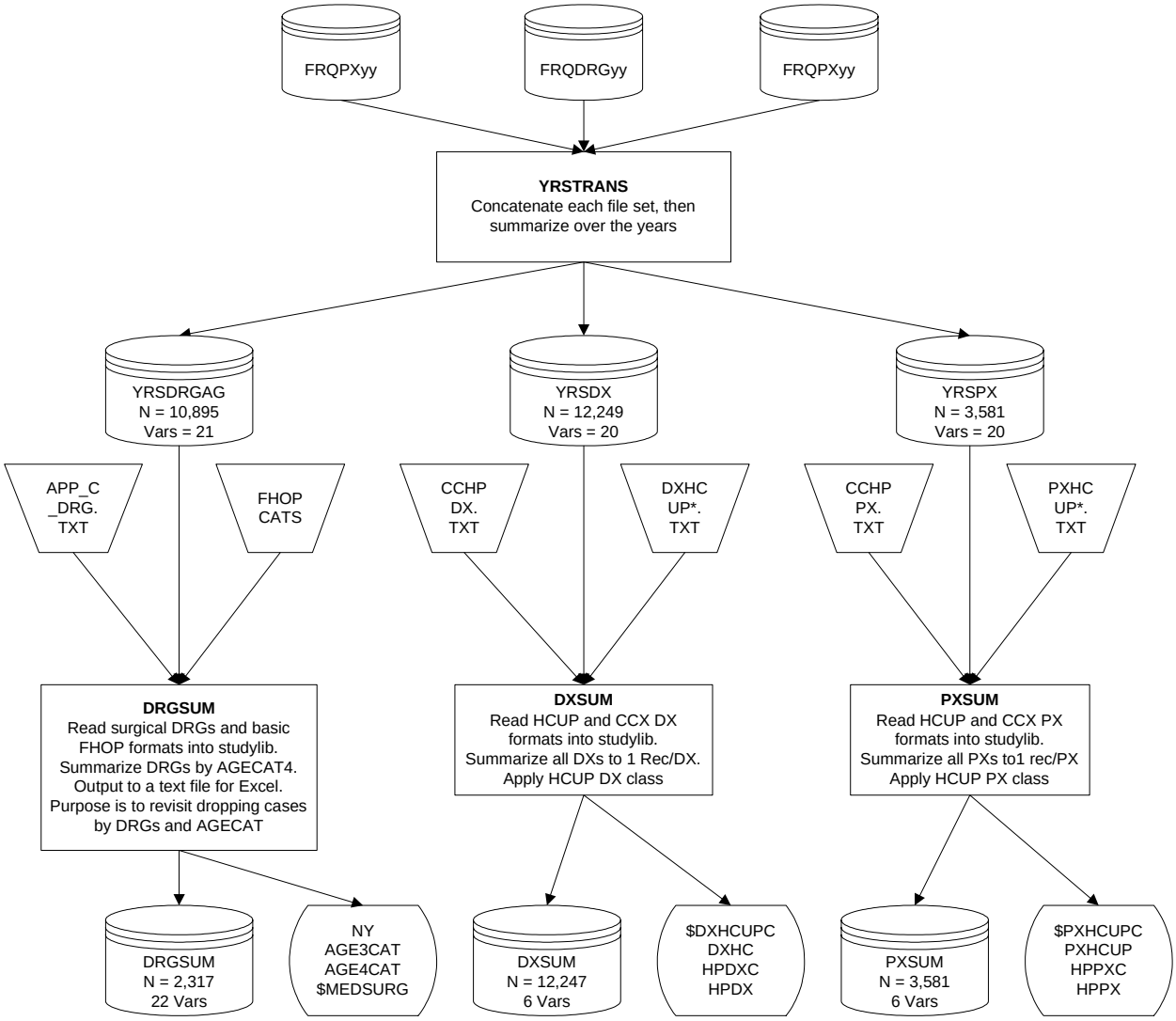
CLINICAL CLASSIFICATION

In the PDD, patient clinical characteristics are classified using the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) as to its principal and secondary diagnoses, the Diagnostic Related Group (DRG), the Major Diagnostic category (MDC), and after 1990, E-codes which classify environmental events, circumstances, and conditions as to the cause of injury, poisoning, and other adverse effects.[4]

Depending on the year, the PDD contains an array of up to 24 diagnoses and 20 procedures, and classifies the patient into a DRG and MDC. Each year we summarize the clinical classification variables found in the MINIMAST files. These are added to summary master files for all clinical characteristics found for our population since 1983.

Figure 2 is a graphic representation of the program flow to summarize the clinical classification variables.

Figure 2: Summarize Clinical Variables



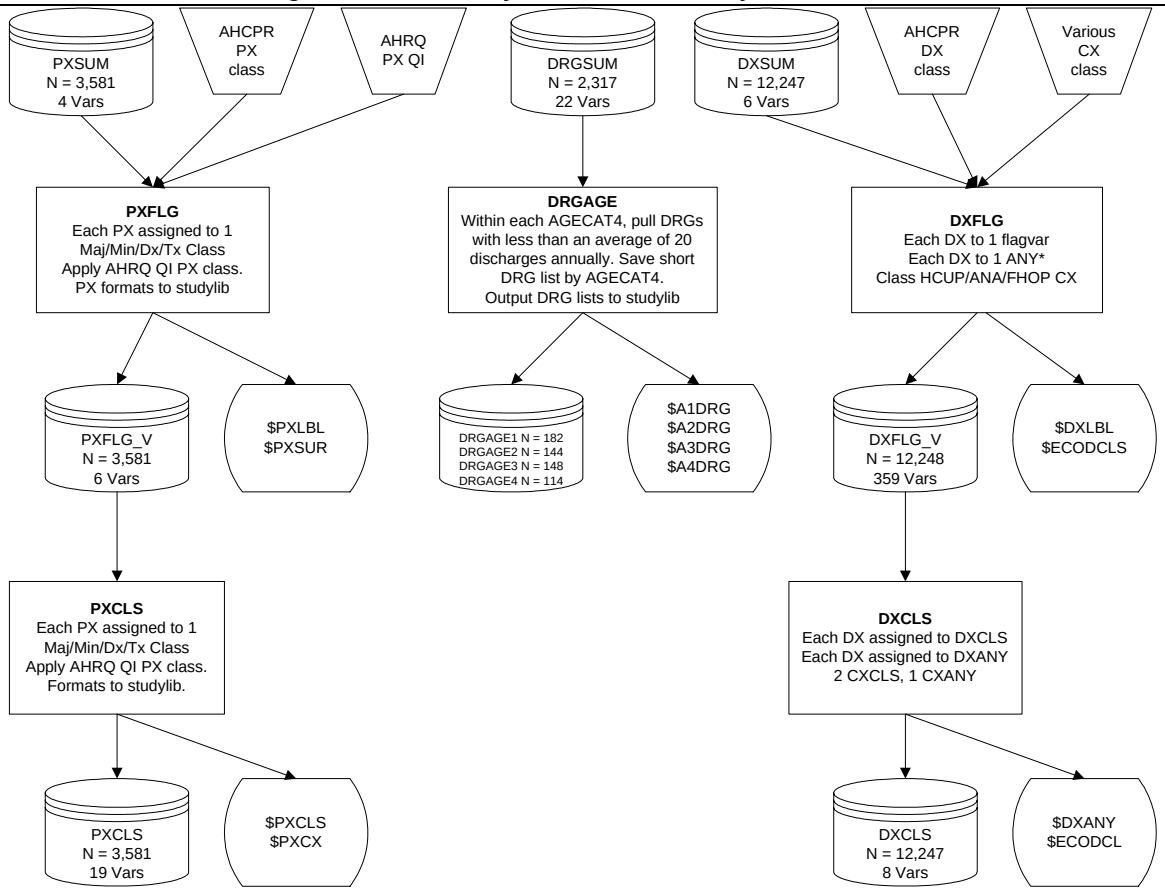
A series of programs classify these into clinically meaningful categories based on software from the Health Care Utilization Project (HCUP), the Agency for Health Research and Quality (AHRQ), and national researchers.[5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] The core software is based upon the Federal Clinical Classification Software (CCS) valid for the time period January 1980 through September 2002. CCS aggregates individual ICD-9-CM codes into broad diagnosis and procedure groups for statistical analysis and reporting purposes. It incorporates external injury mechanisms (E-code) classifications developed by the Centers for Disease Control.[17] We also obtain annual updates of the diagnosis, procedure, and DRG formats from AHRQ.

In addition, we maintain a file of DRG weights, which represent average resources required to care for cases in a particular DRG relative to the average resources used to treat cases in all DRGS and as such reflect increased risk.[18] OSHPD provided a file of historical DRG weights for discontinued DRGs, which we then updated from the HCFA website. Also related to this is a file

of longitudinal hospital-specific case mix weights based on the average DRG weight. We use the most currently available DRG weight.

The end product is a series of Excel files we use to create SAS formats. Currently our clinical classification files are current for codes valid from 1980 through 2002.

Figure 3: Classify Clinical Summary Variables



BRIDGE DEFINITIONS OF CATEGORICAL VARIABLES

Since the 1983 PDD was released, some variables have been redefined. We have implemented decision rules to permit longitudinal analyses of these variables. Age is categorized depending on the needs of a given study. The following recodes are accomplished as part of a macro included in the major classification macro program.

Race/Ethnicity. OSHPD redefined this variable in 1995 by separating race and Hispanic ethnicity. The federal government subsequently issued bridging guidelines for recoding race/ethnicity.[19] These guidelines recommend that longitudinal investigations use their recommended groupings until sufficient years are available to permit more detailed analyses of the complexities of race and ethnicity.

California requires researchers to use the Department of Finance (DOF) population estimates. In calculating population, DOF puts all "other race/ethnicity" but American Indian into "White". Since we use their population estimates, we follow the same rule.

Following these guidelines, we developed macros to reclassify race/ethnicity for longitudinal analysis and reporting purposes. The result is five race/ethnic groups: White, Black, Hispanic All-Race, Asian, American Indian. In calculating total population rates we include all cases, but we do not calculate rates for American Indians because their numbers are small and we believe unreliable because of definitional issues.

Expected Source of Payment. The PDD variable summarizing expected source of payment, or insurance status, underwent several changes between 1983 and 2000, as managed care (MC) took hold. We are not able to assess the impact of MC longitudinally, since MediCal MC was not available on the PDD until the late 1990s.

To bridge the variations in definitions over time, we summarize expected source of payment into the following categories: MediCal/Medicare (a few children annually have this latter coverage), HMO/PHP (managed care, private sector), uninsured, and employment-related coverage (fee-for-service (FFS), CHAMPUS, workers compensation, and other). In analysis we often group the data into two categories: Private (HMO/PHP, other employment-related) and Public (MediCal/Medicare/uninsured). The grouping is the only way we can realistically track longitudinal changes, because FFS coverage virtually vanished, paying for about 5% of admissions for our population in 2000.

Admission Source. This variable was restructured in 1995, breaking it in two parts, source of admission and type of admission. To permit longitudinal analysis we developed a macro incorporating decision rules to identify four admission sources: routine, emergency room, transfer from other hospital, and newborn. We discard all newborn records in selecting records for our MINIMAST files.

Disposition of Patient. We classify this variable into three categories: Return home, transfer to another facility, and death. We assign leaving against medical advice as a return home. Because deaths occur so infrequently in this population, we often create a variable Non-routine Disposition, which has a value of 1 if the patient transferred to another facility or died, and a value of 0 otherwise.

CLASSIFY OBSERVATIONS

The program to classify the observations is complex. It converts several files created earlier into temporary formats, includes macros written to longitudinally bridge changes in certain categorical variable definitions, and imputes geographic data where needed.

Basic Descriptors. Age categories, race/ethnicity, admission source, disposition, and payor are calculated using the macros described earlier.

Procedures. The number of procedures on the record are counted. Every procedure is flagged as to whether it is surgical (PXS) and then assigned into the two CCS categories. Next the principal procedure is classified into one of the following four categories: major diagnostic (PXMAD), minor diagnostic (PXMID), major therapeutic (PXMAT), or minor therapeutic (PXMIT). Then we look again across all procedures to see if any of these four procedure

classes are present in any position. Depending on the needs of a given study, other procedures can be grouped into these major categories using the formats developed.

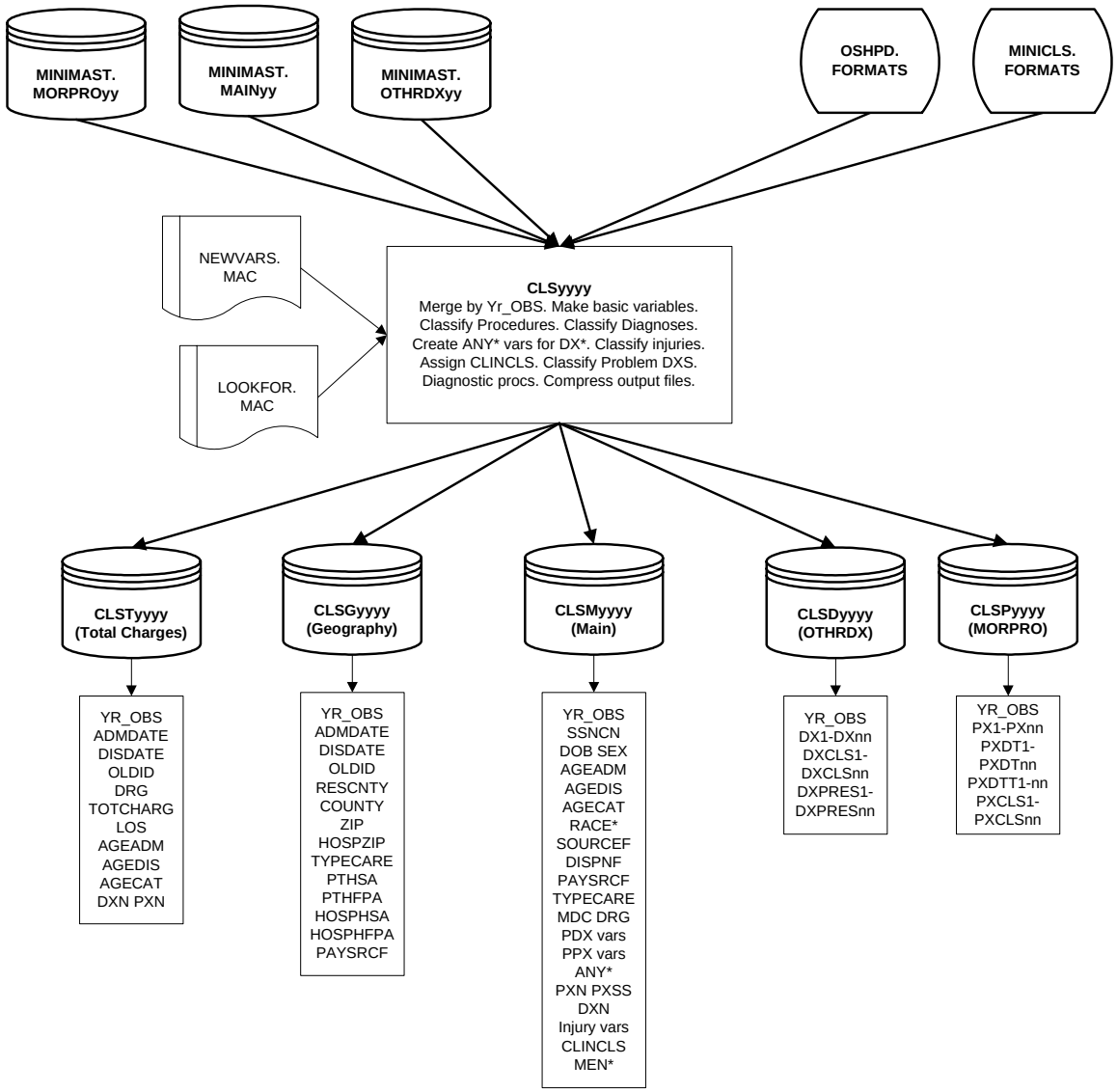
Diagnoses. The number of diagnoses on the record are counted. Every diagnosis is classified into the CSS categories (DXCLS) and then into larger classifications (DXANY), retaining the original ICD-9 codes. Secondary diagnoses are searched to flag any records that fit into certain diagnostic categories of interest. Injuries are identified based on the principal diagnosis and the principal E-Code. Then records are assigned into five mutually exclusive areas of importance (CLINCLS) based on the principal diagnosis, in the following order: injury, ambulatory care sensitive (ACS), pregnancy, mental illness (not injury related), medical, and surgical. Finally, we look across the array of diagnoses for any record indicating a mental illness.

County of Residence (PTCTY HOSCTY). Before 1990, OSHPD identified the ZIP and county of the hospital discharging the patient and the patient ZIP of residence, but not the patient county of residence. We used our ZIP master file to assign county of residence. For Los Angeles County, this includes assignment into SPAs.

Health Service Area (PTHSA HOSHSA). Following a legislative sea change in the late 1980's, California has not used HSAs for planning but they still exist legislatively and OSHPD still reports them. This permits us to examine the regional impact of "free market" healthcare restructuring on California's children needing hospital care. This variable is available in most files OSHPD releases.

Figure 4 summarizes the major steps in the program to classify the individual-level PDD and write out the results into specific datasets.

Figure 4: Classify Observations and Store Data in Specific Subsets



CARE QUALITY INDICATORS.

This sequence of programs macro evaluates every PDD record for an indication of care quality indicators (CQI), using materials provided by other researchers.

Adverse events by definition. As an early attempt to measure adverse outcomes, the ICD-9-CM included a series of diagnosis and procedure codes that reflect adverse outcomes no matter their position in the array of diagnoses or procedures. These can be used to evaluate performance depending on their position and sequencing of events. For example, one hospital may do knee surgery on a child who was discharged routinely, developed complications the next day, resulting in readmission to a second hospital whose principal procedure involved

correcting the previous hospital's error and resulted in an ELOS. We may not see the adverse event on the first hospital record, but would see it and the reparative surgery on the second.

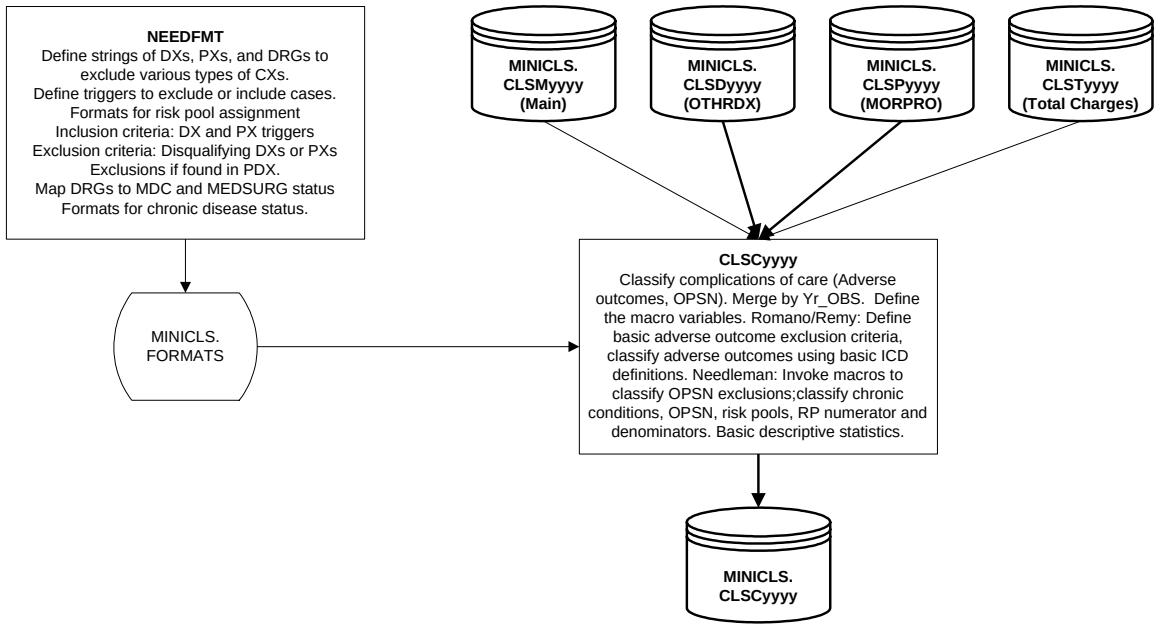
AHRQ Quality Indicators. A more recent initiative is the development of the Agency for Health Research and Quality (AHRQ) software for quality indicators. This evaluates events by clinical condition that are indicators of poor quality care.[6] [7] [8] Most events are for the adult population, but we did use those defined for children.

Outcomes Potentially Sensitive to Nursing (OPSN). Needleman very generously provided the software his group developed to classify cases into OPSN.[15]

Chronic Conditions. Needleman provided his code to identify the following chronic conditions: cancer, HIV/AIDS, pulmonary, coronary artery disease, congestive heart failure, severe chronic liver disease, diabetes, renal, nutritional, dementia, and functional, based on the work of Iezzoni and associates.[20] Relatively few children had chronic conditions. We created a summary variable counting the number of these conditions (CHRONN) and also made an indicator as to whether the child had none or 1 or more chronic conditions (ANYCHRON).

Risk Pool (POOL). Needleman also provided his code to assign cases to four mutually exclusive risk pools based on the DRG within which to assess Outcomes Potentially Sensitive to Nursing (OPSN). His risk pools are medical, major surgery, minor surgery, and all other (typically high-risk) cases that were outside his study design for various reasons.

Figure 5: Classify Care Quality Indicators



RESOURCE UTILIZATION

Length of Stay. Between 1983 and 2000, about 5% of discharges had a LOS of zero days; that is, the child was admitted and discharged on the same day. All records admitted and discharged

on the same day were changed to a LOS of 1 day to more accurately reflect the family and social burden of admitting and discharging a sick child. It also enables us to do log transformations or other transformations that are difficult with a value of 0. Some children are admitted to long term facilities for periods of many years. Because OSHPD coding rules require charges to be reported for the year, we truncate the LOS upper range at 365. Both recodes at the lower and upper end of the distribution makes it possible to calculate a (relatively) realistic charge per day. Other than these basic recodes, LOS is classified according to the needs of a given study.

Total Charges. About 9.3% of records were missing charges since 1983. This ranged from 10.9% in 1983 to 7.2% in 2000. Charges are missing non-randomly, because OSHPD does not require Kaiser and children's hospitals to report this. However, charges are reported when non-Kaiser members receive care in Kaiser facilities or Kaiser members receive care in non-Kaiser facilities.

In general, charges reported here are higher that actual reimbursements, and do not provide a clear picture of the cost of hospital care. However, charge information provides a sense of the relative cost of care, thus allowing comparison between groups of cases.

To better estimate the total economic burden of early childhood hospitalization, we impute charges for records lacking them, using charges converted to 2002 dollars to control for inflation. Using records with a charge, we create a file containing the original charge and the charge converted to 2002 dollars (CHG2002), from the monthly Medical Care Consumer Price Index.[21] This is converted to dollars per day, trimmed at the 99th percentile, and converted to logged dollars per day. Within DRG and year, for DRGs with 30 or more cases, we obtain a predicted log charge per day, controlling for the child's age, length of stay, number of diagnoses and number of procedures. If the DRG had fewer than 30 cases a year, we combine years within 3-year rolling averages. If the DRG still had too few cases, we did the regression over all years. The predicted logged dollar per day was converted back into 2002 dollars.

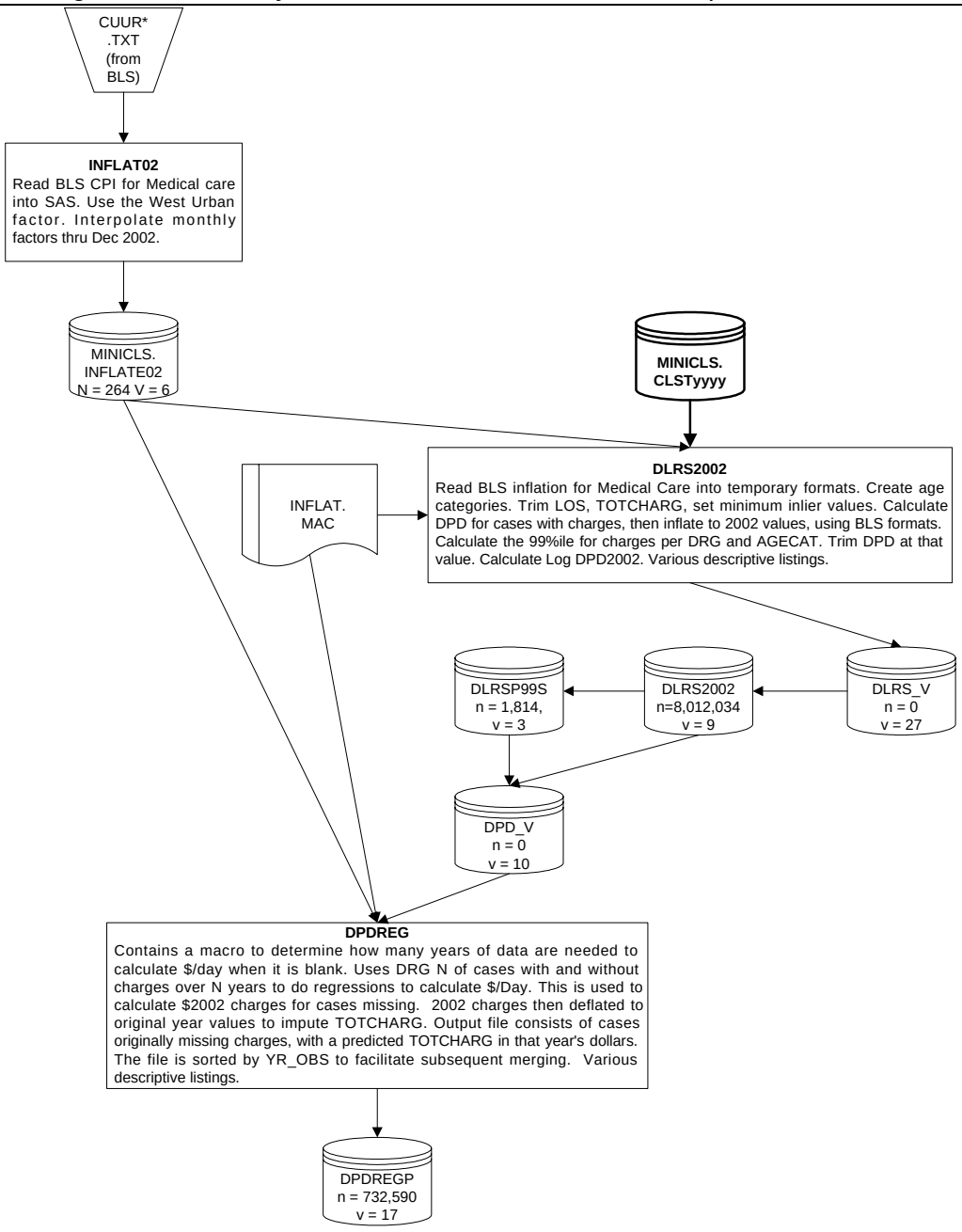
We impute charges using the predicted value based on the child's age, number of diagnoses and number of procedures, DRG, and year of discharge . Then we convert the dollars back to their original value within each year. We merge records for cases missing a charge back into the master file and convert all charges to 2002 dollars. At the end of the process, we have the original charge, a total charge based on the original or imputed value, the inflated charge in 2002 dollars, and a variable flagging cases that had been imputed. These variables are transformed or further categorized based on the needs of a given study.

Over all the years, we imputed charges on 7% of Medi-Cal cases, 2% on Private/Other, 4% on uninsured, and 39% of HMO/PHP. Average charges on imputed cases were $3,422 less than average charges for cases that had not been imputed.

We examined whether the imputation affected results in two ways. We made a dummy variable flagging a record as having an imputed charge. Then we checked the bivariate correlation of the dummy variable to the CHG2002 in each year. The correlation was -.022 in 1983 and -0.016 in 2000, both non-significant. We also added the dummy variable as the last variable in multivariate models involving charges, and it was statistically non-significant.

Figure 6 summarizes the sequence of programs to impute total charges. This sequence needs to be rerun as each year of data are added. Depending on the needs of a given study, we use the imputed charges as reported or the charges converted to current dollars.

Figure 6: Classify Observations and Store Data in Specific Subsets



NEXT STEPS

We treat these files as the "well" from which we draw data for particular analyses or public health products that FHOP distributes. The specific records we seek, variables we use, and methods followed to analyze or summarize the data are specific to each project. This intent of this document to summarize in one location the method by which the basic files were created. It should be read as the anchor for understanding the source files for our various products.

ENDNOTES

[1] Romano P, Luft HS, Remy LL (Dec 1993). Annual Report of the California Hospital Outcomes Project. Volume Two: Technical Appendix. California Health and Welfare Agency, Office of Statewide Health Planning and Development.

[2] Remy L, Clay T, Oliva G. California Child and Youth Injury Hot Spots Project 1995-1997, Volume Three: Technical Guide, Sacramento, CA: California Department of Health Services, Maternal and Child Health Branch, August 2000.

[3] Health Facility Planning Areas, January 1983. P-800-2, Revised 07/18/83, OSHPD, pursuant to Section 90811, Subdivision 7, Title 22, of the California Administrative Code.

[4] Public Health Service and Health Care Financing Administration. International classification of diseases, 9th revision, clinical modification. Vols. 1, 2, and 3; fifth edition. Washington, DC: Public Health Service; 1994. DHHS Publication No. (PHS) 94 1260. Originally developed by the World Health Organization, the ICD-9-CM is used to classify patient morbidity and mortality.

[5] Clinical Classifications Software (ICD-9-CM) Summary and Download. Summary and Downloading Information. Agency for Health Care Policy and Research, Rockville, MD. http://www.ahrq.gov/data/hcup/ccs.htm. Accessed 19 May 2002.

[6] Prevention Quality Indicators, Version 2.1. January 2003. Agency for Healthcare Research and Quality, Rockville, MD.http://www.qualityindicators.ahrq.gov/data/hcup/prevqi.htm

[7] Inpatient Quality Indicators, Version 2.1. March 2003. Agency for Healthcare Research and Quality, Rockville, MD. http://www.qualityindicators.ahrq.gov/data/hcup/inpatqi.htm

[8] Patient Safety Indicators. March 2003. Agency for Healthcare Research and Quality, Rockville, MD. http://www.qualityindicators.ahrq.gov/data/hcup/psi.htm

[9] AHRQ Quality Indicators—Guide to Prevention Quality Indicators: Hospital Admission for Ambulatory Care Sensitive Conditions. Rockville, MD: Agency for Healthcare Research and Quality, 2001. AHRQ Pub. No. 02-R0203.

[10] AHRQ Quality Indicators -- Prevention Quality Indicators: Software Documentation, Version 2.1 – SAS. Rockville, MD: Agency for Healthcare Research and Quality, 2001. AHRQ Pub. No. 02-R0202.

[11] Billings J, Zeitel L, Lukomnik J, Carey TS, Blank AE, Newman L. (1993). Impact of socioeconomic status on hospital use in New York City. Health Affairs 1993 Spring; 12(1): 162-73.

[12] Backus et al. Effect of managed care on preventable hospitalization rates in California. Medical Care. 40(4), 315-324. 2002.

[13] Agency for Health Care Policy and Research. (1988). Healthcare Cost and Utilization Project. Outcome, utilization, and access measures for quality improvement. AHCPR Pub. No. 98-0035.

[14] American Nurses Association. (1997). Implementing Nursing's Report Card. A Study of RN Staffing, Length of Stay and Patient Outcomes. Washington, DC: American Nurses Publishing.

[15] Needleman J, Buerhaus PI, Mattke S, Stewart M, Zelevinsky K. Nurse staffing and patient outcomes in hospitals. Final Report for Health Resources Services Administration Contract No. 230-99-0021. February 28, 2001.

[16] Kovner C, Gergen P. Nurse staffing levels and adverse events following surgery in U. S. hospitals. Journal of Nursing Scholarship. 1999, 30:315-21.

[17] Centers for Disease Control and Prevention. Recommended framework for presenting injury mortality data. Morbidity and Mortality Weekly Report, 1997, Aug 29; 46(no. RR14): 1-30.

[18] Get citation

[19] Provisional Guidance on the Implementation of the 1997 Standards for Federal Data on Race and Ethnicity. Executive Office of the President, Office of Management and Budget, Washington, D.C. 20503. December 15, 2000

[20] Iezzoni LI, Foley SM, Heeren T, et al. (1992). A method for screening the quality of hospital care using administrative data: preliminary validation results. Quality Review Bulletin, 18,361.

[21] Bureau of Labor Statistics Data, Consumer Price Index-All Urban Consumers, West Size A. Downloaded from http://146.142.4.24/cgi-bin/dsrv.).