standardized file definitions. Content areas and data elements that are commonly required for research studies are identified, and data dictionaries are created for each content area, specifying a common format for each of the elements— variable name, label, description, code values, and value labels. Local site programmers have mapped the data elements from their HMO's data systems into this standardized set of variable definitions, names, and codes, as well as onto standardized SAS file formats. This common structure of the VDW files enables a SAS analyst at one site to write one program to extract and/or analyze data at all participating sites. **Methods:** This poster demonstrates the data in Harvard Pilgrim Health Care (HPHC) VDW files. **Results:** The HPHC VDW files contain details on 88 million pharmacy dispensings (2000-2010), nearly 89 million unique medical encounters (2000-2010), including .67 million hospitalizations, 69.6 million ambulatory visits, 183 million diagnoses, and 209 million procedures. Vital Signs and lab results are available for about 10% of HPHC enrollees who receive care at Harvard Vanguard Medical Associates, a local multi-specialty practice group and available per project basis. The VDW Enrollment, Demographic, and Census files are derived from internal HPHC historical membership files; the VDW Pharmacy and utilization files are derived from internal HPHC claims files; the VDW tumor data is derived from the Massachusetts Cancer Registry data; the VDW Death data is derived from Massachusetts death registry and HPHC internal data. **Conclusions:** The VDW at HPHC provides an easily employed unified central repository of data from all available source files. This resource enables the sharing of compatible data in multi-site studies, and also improves programming efficiency, accuracy, and completeness for local single site studies by expending resources to link these legacy systems only once.

Keywords: VDW; HPHC; Virtual Data Warehouse
doi:10.3121/cmr.2012.1100.ps2-38

PS2-39:

Key Challenges and Decisions in the Development of HMO Death Data

Jamila Gul[1]; Daniel Ng[1]; Wei Tao[1]; Daxin Zuo[1]

[1]Kaiser Permanente Northern California

Background: The mortality status of an HMO's members, whether a person is alive or dead, is critically important to health care research. The uses of mortality status include determination of causes of death, death rates, and for selection of study cohorts. Unlike other clinical activity and health statuses, mortality status cannot be determined solely with a health plan's internal data, even for well-defined populations, because health plan members often die outside of the plan's care delivery and claims payment systems. For such members, determination of mortality status and date of death requires linkage of health plan data to external sources of mortality data, such as state death certificate or federal Social Security data. This person-level statistical linkage is a complex, multi-step process that involves many decisions, assumptions, and choosing of priorities. Even when death data from internal systems is available, internal data sources can disagree, likewise requiring complex decisions to determine a member's "true" mortality status. **Methods:** This presentation highlights key challenges and decisions points in the development of death data logic at Kaiser Permanente Northern California (KPNC). **Results:** Key decisions in KPNC's death data logic include: handling internal membership data with conflicting personal identifiers, e.g., multiple Social Security Numbers associated with a single Medical Record Number; ranking, grading, and selecting from multiple matches between internal and external data; finding members with multiple MRNs; using utilization contact dates and membership activity dates to evaluate death records, judging reliability of internal health plan data, setting thresholds and criteria for confidence scoring, selecting between multiple possible dates of death, chart validation of final death data. **Conclusions:** The development of death data logic at KPNC is a complex process that requires intricate decisions, careful analysis, and a higher degree of discretionary judgment than is typical of building other data areas. The quality of final results depends upon validation and iterative improvements to the process.

Keywords: Death Data; HMORN; Virtual Data Warehouse
doi:10.3121/cmr.2012.1100.ps2-39

PS2-40:

Do Women Really Smoke and Drink During Pregnancy: A Detailed Look into Cleaning the Smoking and Alcohol Variables

Craig Hansen[1]; Marvin Adams[1]

[1]Kaiser Permanente Southeast

Background: In many epidemiologic studies smoking and alcohol are often examined as an exposure variable when focusing on health conditions associated with these lifestyle factors. Furthermore, smoking and alcohol have almost become standard confounding variables in epidemiologic studies when examining other exposures. We highlight the challenges associated with cleaning and analyzing the smoking and alcohol variables within various sources of administrative data when conducting pregnancy studies. **Methods:** We compared information on smoking and alcohol use among pregnant women from the VDW vitals table, social history table, and birth certificate data from Georgia State. Encounter data for smoking from the social history table was cleaned based the contact dates and quit dates falling within the pregnancy period and various combinations of 'yes', 'no', and 'quit' were examined based on these dates. Categories of smoking were created and compared to data in the VDW and on birth certificates. Similarly, we examined information on alcohol use during pregnancy from the social history table, including information reported in the 'alcohol comment' field. **Results:** (more detailed results will be available at the time of the conference):We found inconsistencies in how smoking during pregnancy was reported in birth certificate data versus the social history table. However, this inconsistency was only among smokers and former smokers, whereas information for non-smokers was consistent across the different sources. Within the social history table there were many pregnant women that had 'yes' for the 'drink alcohol' field, however many of these women had information in the 'alcohol comment' field stating that they did not drink during pregnancy. After creating levels of alcohol use during pregnancy we found inconsistencies with information on the birth certificate. **Conclusions:** While it is important to assess smoking and alcohol during pregnancy for the purposes of either a confounding variable or an exposure variable. We highlighted how these two variables (one from the VDW) are inconsistent across various data sources and both variables must be used with caution. NOTE: In order to present these data we feel a presentation is more suitable than a poster.

Keywords: Smoking; Pregnancy; Virtual Data Warehouse
doi:10.3121/cmr.2012.1100.ps2-40

PS2-41:

Mapping VDW Data Using SAS/GRAPH

Hongyuan Gao[1]

[1]Group Health Cooperative

Background: Most healthcare data have a spatial dimension. For investigators preparing a study, displaying the target population sizes or certain healthcare utilization in different geographic areas in a map will be very visual for investigators to know how the disease is spread and how much healthcare resources are used. For programmers who manage the database, presenting the geographic distribution of patient demographics, enrollment and utilization in a map could be the first step to check whether the data volume is normal and complete. When healthcare data are presented to audience, an informative map is worth a thousand words! However, map creation usually involves specialized software and is time-consuming. In SAS 9.2, creating highly-customized maps becomes surprisingly easy using GMAP procedure. This presentation/poster introduces several SAS macros that can transfer frequency tables of variables in VDW into colors and symbols on maps. **Methods:** Map creation in SAS/GRAPH involves two steps. The first step is preparing three datasets: a map data set which contains the information needed to draw map boundaries; a response data set which contains the information that will be displayed on the map; and an annotation dataset which customizes the map by adding text, symbols, contours, images and html links to the map. The second step is using the PROC GMAP procedure with the three datasets. SAS macros introduced here emphasize on the use of annotation. Annotation instructs SAS/GRAPH on how to customize graphs as if you were drawing with a pencil on a paper.