also appear in another site's data. Most notably—the protocol does not call for transmitting any unencrypted PHI. The core of the method involves the repeated use of a commutative cipher to encrypt identifiers, each site using its own key, and then comparing the results for overlaps. The end result of the protocol is that each site will know which of its MRNs represent people who also appear in other sites' data, and which site(s) those are. The presentation will describe the data flows, algorithms and processes necessary to create this data. **Results:** The protocol described can be run once to assess the actual overlap, and—if it is significant enough—periodically to create this data at the sites. Once created, the data could be made part of the VDW, allowing HMORN projects to easily assure that each person in each distributed cohort appeared in only one site's file.

Conclusions: Being able to say that we have actually investigated population overlap and either found it in fact to be negligible, or that we have a method for de-duplicating people is far preferable to hand-waving and bare assertions that any overlap must be insignificant. As the HMORN grows and VDW collaborators become more numerous, we should expect this question to take on more and more significance in the eyes of both our funders and the consumers of our research.

Keywords: VDW; Deduplicating; Virtual Data Warehouse
doi:10.3121/cmr.2012.1100.ps2-48

PS2-49:

The VDW Census File: Strengths, Issues, and Recommendations for the Future

Christine Bredfeldt¹; Julie Liu²; Roy Pardee³

¹Kaiser Permanente Mid-Atlantic, MAPMG; ²Kaiser Permanente Southern California; ³Group Health Cooperative

Background/Aims: The VDW Census file contains information about our members such as income and education. The information is estimated based on data collected by the decennial census for the geographic areas around our members' addresses (identified by geocodes). The VDW Census file was initially developed based on 2000 census data. Data from the 2010 census is now being released, and in many cases, the 2010 census data structure looks different than the 2000 census. For example, income and education were no longer collected in the 2010 census, but have been moved to a supplementary survey (American Community Survey, or ACS) collected over a different time period. **Methods:** The VDW Census workgroup reviewed the changes in the 2010 census data and compared the available information to the data in the existing VDW Census file. Many of the fields from the 'long-form' in the 2000 census were moved to the ACS. Where possible, the workgroup mapped the existing VDW Census fields to fields with similar, if not equivalent, data in the new census data structure. The workgroup also considered whether new data available from the ACS might be of interest to HMORN investigators. Where major changes in information availability were anticipated, the workgroup used online surveys to solicit feedback from a larger group of HMORN analysts and investigators. **Results:** Our primary goal was to identify changes to the specification that would allow updating the Census file to 2010 data. Our secondary goal was to identify that would allow sites to maintain multiple sets of census data to accommodate projects with different time periods. To accommodate multiple years of census data and reduce file sizes, we recommend splitting the VDW Census file into two files: one tracking members' geocoded addresses, and another containing the census data for all available geocodes. **Conclusions:** We found significant changes in the 2010 census data that require considerable changes to the VDW Census file. Once agreement has been reached on the new Census specification, a subcommittee will develop code to build a new VDW Census data file and recommendations on building a file containing members' geocodes over time.

Keywords: VDW; Census Data; Virtual Data Warehouse
doi:10.3121/cmr.2012.1100.ps2-49

PS2-50:

VDW Data Sources: Scott & White Healthcare

Debbie Godwin¹; Bhushan Mahajan¹; Dean Kjar¹; Anne Robertson¹; Teresa Ponder¹

¹Scott & White Health System

Background: The Virtual Data Warehouse (VDW) was created as a mechanism for producing comparable data across sites for purposes of proposing and conducting research. The database is "virtual" in the sense that the data remain at the local sites; there is no multi-site physical database at a centralized center. At the core of the VDW are a series of standardized file definitions. Content areas and data elements that are commonly required for research studies are identified, and data dictionaries are created for each of the content areas, 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 wide range of data sources used at Scott & White to feed information into our local implementation of the VDW datasets. **Results:** Scott & White local implementation of the VDW contains detailed medical information on Scott & White Health Plan (SWHP), EMR (Electronic Medical Records), Lab Services, Pharmacy Department, State Death Data, State Provider Data, CENSUS Bureau Data and Medical Staff Services Data from year 2001 to present. The VDW tumor data are loaded from the Scott & White Tumor Registry. VDW tables are built and maintained from these different sources within Scott & White. **Conclusions:** The VDW at Scott & White 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: Clinical Trial; Tracking Database; Virtual Data Warehouse
doi:10.3121/cmr.2012.1100.ps2-50

PS2-51:

Utilization Quality Assurance: Are We Better Yet?

Donald Bachman¹; Terry Field²; Christine Bredfeldt³; Mark Hornbrook¹; Alan Bauck¹; Heather Tavel⁴; Lucas Ovans⁵; Debbie Godwin⁶; Dean Kjar⁶

¹Kaiser Permanente Northwest; ²Fallon Community Health Plan; ³Kaiser Permanente Mid-Atlantic, MAPMG; ⁴Kaiser Permanente Colorado; ⁵HealthPartners; ⁶Scott & White Health System

Background: The HMORN Virtual Data Warehouse (VDW) Utilization files are used in almost every VDW research project for a range of purposes including selecting study populations, building disease registries, measuring health status, and evaluating resource use and appropriateness of care. Utilization data, including encounter, diagnosis and procedure data, comes from multiple data sources including legacy data, electronic health records, and claims. Because the data come from many sources and require complicated processes of standardization, the VDW tables can be very complex to build, potentially leading to inconsistencies across sites. Our objective was to assess, document, and improve the overall quality, availability, and completeness of the VDW Utilization data. To understand whether our QA approach was effective, we compared the current data quality to quality assurance data collected in 2009. **Methods:** The HMORN Utilization Work Group, together with KP CESR staff, developed quality assurance programs to build summary tables for participating sites. The Utilization Work Group then combined the summary tables from each site and graphically compared utilization rates, diagnosis capture, and other statistics across HMORN sites to provide a more complete picture of the variability across sites and identify potential outliers that may indicate data quality concerns. **Results:** Overall, we found that VDW Utilization data quality has improved considerably since 2009, as demonstrated by the reduction in variability across sites. In particular, rates of hospitalization, inpatient days, and doctor's office visits are considerably more consistent across time and sites. Residual differences likely reflect real-world variation in membership composition and standards of care. In addition, we identified areas with persistent variability that indicate a need for further exploration, such as rates of dialysis and out-of-office encounters. Finally, we also found between-site differences in the interpretation of the Utilization specification, such as the designation of principal and primary diagnoses. **Conclusions:** Identification and resolution of data quality problems through frequent use of the data, cross-site quality checks, QA programs that produce traffic light