PS2-45:

Conducting Multi-site Quality Assurance on Laboratory Results Data Incorporated into the Virtual Data Warehouse

Gwyn Saylor[1]; Christine Bredfeldt[2]; Melissa Butler[3]; Bruce Folck[4]; Paul Hitz[5]; Richard Krajenta[6]; Jessica Ogarek[7]; Lucas Ovans[8]; Roy Pardee[9]; Karen Riedlinger[1]; Mark Schmidt[10]; Andrew Sterrett[11]; Marsha Raebel[11]

[1]Kaiser Permanente Northwest; [2]Kaiser Permanente Mid-Atlantic, MAPMG; [3]Kaiser Permanente Southeast; [4]Kaiser Permanente Northern California; [5]Marshfield Clinic; [6]Henry Ford Health System; [7]Fallon Community Health Plan; [8]HealthPartners; [9]Group Health Cooperative; [10]Kaiser Permanente Hawaii; [11]Kaiser Permanente Colorado

Background/Aims: Assuring that complete and correct laboratory test results are brought into the VDW and associated with the correct Test_Type is an ongoing task. Many problems can occur. For example, a particular version of a test can incorrectly be left out when the data are extracted. Results can be duplicated, or results can be associated incorrectly with a Test_Type. Creating quality assurance checks to locate problems or validate data requires evaluating the same data record from several different angles. **Methods:** We conducted quality assurance on numerous laboratory test results that had been incorporated into VDW format across as many as 11 HMORN sites. Quality assurance programs were written to provide counts of test results across sites and over time. These programs also detailed result values, result units, patient location, result location, and other specific metrics. Descriptive statistics and graphical displays were used to facilitate assessment of across-site data variability and identify potential data quality issues. Sites were provided their site-specific results and results for all participating sites. **Results:** Masking site information, we will show examples of quality assurance checks and depict what type of problem each is designed to point out. Examples may include incorrectly mapped tests (e.g., "hemoglobin" test results where the result unit was in percent and was determined to actually be glycosylated hemoglobin results) and tests that require investigation by the site because the result unit is possible, but unusual (e.g., total cholesterol units of g/dl require verification of g/dl unit and conversion to mg/dl before use or verification that g/dl was entered or read incorrectly). Other examples can include the volume of test results now contained within the VDW (e.g., total cholesterol total n is over 26.6 million results). **Conclusions:** Laboratory result quality assurance requires assessment of problems as well as verifying expectations and standardizing data presentation. Multi-site comparisons, comparisons of multiple associated tests at one site, and in depth cross tabulation of single site data play a part in assuring accurate, complete laboratory data.

Keywords: Lab; Quality; Virtual Data Warehouse

doi:10.3121/cmr.2012.1100.ps2-45

PS2-46:

VDW Data Sources: Meyers Primary Care Institute

Jackie Fuller[1]; Jessica Ogarek[1]; Hassan Fouayzi[1]

[1]Fallon Community Health Plan

Background/Aims: The Virtual Data Warehouse (VDW) was created as a mechanism for producing comparable data across sites for purposes of proposing and conducting research. It is "virtual" in the sense that the data remain at the local sites; there is no multi-site physical database at a centralized data coordinating center. Local site programmers have mapped the data elements from their HMO's data systems into a VDW 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 Meyers Primary Care Institute to feed information into our local implementation of the VDW datasets. **Results:** The Meyers Primary Care Institute works with Fallon Community Health Plan (FCHP), a not-for-profit managed care plan in central Massachusetts closely aligned with Reliant Medical Group (RMG), a multi-specialty group practice. FCHP provides enrollment, demographics, pharmacy, and data on utilization; data are available for 1999-2010. RMG provides clinical data on their patients including vital signs and laboratory test results. Electronic text from RMG's EMR is pulled for specific studies. Tumor data are obtained from the tumor registries at our major hospitals on request; linked data have been obtained from the state tumor registry for patients who receive care from the RMG. The Meyers Primary Care Institute local implementation of the VDW contains medical details on 0.47 million FCHP enrollees. The VDW files encompass more than 22 million medical encounters, including more than 36 million diagnoses, and nearly 56 million procedures. We have more than 20 million pharmacy dispensing, nearly 1.6 million Vital Signs observations (2007-2009), and nearly 3 million lab results (1996-2009). **Conclusions:** The VDW provides an easily employable 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.ps2-46

PS2-47:

SAS Hash Objects are Memory Bound. So What?

William Tolbert[1]

[1]Group Health Cooperative

Background/Aims: If you are like most SAS programmers, you've heard amazing claims regarding SAS hash objects — lightning-fast lookups without the need to sort. You also know that hash objects are memory-bound, making them uselessfor your big data needs. This presentation outlines the benefits of hash objects when used with data step by-processing. GHRI uses hash objects for the VDW Utilization data build process. As you will see, hash objects are multi-purpose data containers that can dramatically reduce processing time, even when dealing with hundreds of millions of records. This talk will cover advanced features of the SAS hash object, including key summaries, check(), ref(), multikey objects, and sorting. **Methods:** Using hash objects with by-processing allows us to read our source data only once. This is a critical aspect of the current approach and the reason we achieve fantastic build times. The only requirement is that our source data be grouped. As each new by group is encountered, the hash objects are cleared. From 'first dot' to 'last dot', we load up our hash objects, putting procedures in one, diagnoses in another, and encounters in a third. On the last record of the by group, we out put our datasets, then begin the process again on the next group. **Results:** GHRI is now sourcing VDW Utilization data from our cost management data for 2004 to present. The input file is 130 gigabytes and has over 170 million records. We walk through this file only once, and output 5 datasets (utilization, diagnosis, procedure, invalid dx codes, and invalid px codes). The data are de-duped and sorted by consumer number and admit date. The process uses 26 hash objects and 16 hash iterator objects. The entire process is one data step. It runs in under 2 hours on a Windows 7 PC with 16 GB of ram. **Conclusions:** SAS hash objects are not just for lookups. They are flexible all-purpose data containers that allow programmers to exploit the inherent power of the data step.

Keywords: VDW; SAS; Virtual Data Warehouse

doi:10.3121/cmr.2012.1100.ps2-47

PS2-48:

De-duping the VDW: A Protocol for a Identifying Common Enrollees Across Sites Without Sending PHI

Roy Pardee[1]; Lucas Ovans[2]; Thomas Flottemesch[2]

[1]Group Health Cooperative; [2]HealthPartners

Background/Aims: Early on in VDW history, the sheer geographical distribution of HMORN member sites made it exceedingly unlikely that the same person would be found in more than one site's data. Thus our multi-site research has proceeded without much fear of double-counting anyone. But as the HMORN expands its membership to new organizations, this assumption becomes harder to justify. We now have 3 sites operating in Minnesota, for instance, and 2 each in Massachusetts, Northern California, Idaho, and Washington. **Methods:** This presentation will propose a protocol for creating a dataset at HMORN sites listing the local MRNs of people who