Virtual Data Warehouse

C-A5-01:

Clarity and Research

Jamila Gul, BS¹; Mark Schmidt, BS²; Lauren Perkins, BS³

¹Kaiser Northern California; ²Kaiser, Hawaii; ³Kaiser Permanente

Clarity is the reporting database for Epic data. Clarity was created to extract data from the Epic production server and store it in a relational database and a dedicated reporting server. Clarity can reside on variance platforms such as Teradata or Oracle. Most tables in Clarity are updated nightly by a feed from Chronicles. There are also weekly and monthly updates for the more static tables. Queries and reports generated from Clarity can be very comprehensive and provide complete information at the individual data point level.

Keywords: Virtual Data Warehouse, Data extraction, Data storage
doi:10.3121/cmr.2010.943.c-a5-01

C-A5-03:

Virtual Data Warehouse (VDW) Architecture and Implementation

Debbie Godwin, BS¹; Monica Silcott¹; Daniel Ng, BS, BSE, MBA²

¹Scott & White; ²Kaiser Permanente Northern California

Background and Aim: VDW system architecture varies from one institution to another. This provides an opportunity for the development of best practice in system architecture design. The word architecture describes the underlying structure of networks, computer hardware and software. Well-designed systems-level architectures are critical to the success of large-scale projects. The explosion of technological opportunities and customer demands has driven up the size, complexity, cost, and investment risks of projects. Without solid architectures, these projects lack the firm foundation and robust structure on which to build. Solid systems architecture and design is a required phase following data gathering and analysis which should precede any design prior to any implementation. With functional requirements established, the system architect can design a system to perform in an optimized environment, ultimately producing robust data and customer satisfaction. Architecture choices Hardware platform DBMS –Relational Database Management System (RDBMS)Query/analysis tools Maintenance/service strategy. For bioinformatics research, the most important qualities of system architecture are: extensibility for new functions/features; interoperability for cross-site collaboration; security due to regulatory and ethical requirements; serviceability/maintainability to support independence from enterprise/business computing; and performance for analysis of large datasets. Aim of this study was to determine best practice architecture choices resulting in performance improvements per institutions data volume. **Methods:** Performance benchmarking analysis was gathered from KPNC and S&W comparing the effect on VDW performance of different design choices and implementations of database software and hardware. Our performance measurement focused on the delivery speed of data return. Measurements were made on various data areas at both peak and off-peak times. **Results:** KPNC has measured via benchmarking on VDW tables that Oracle DBMS can increase performance compared to SAS datasets. KPNC was able to measure separately the performance improvement from migration to a more powerful hardware platform. **Conclusions:** A proactive, best practice approach to systems architecture can create benefits for research databases, particularly in performance speed. Improvements in performance can be achieved through software, by adopting a relational database management system such as Oracle, that allows table indexing and partitioning, and through investment in high-performance hardware. The two strategies can be implemented alone or in conjunction for maximum benefits.

Keywords: Virtual Data Warehouse, Health data improvement, Health data performance analysis

doi:10.3121/cmr.2010.943.c-a5-03

C-A5-04:

A Simple, Accurate SAS Algorithm for Electronic Abstraction of Race from Digitized Progress Notes

Douglas Roblin, PhD¹; Peter Joski, MSPH¹; Junling Ren, MD¹; Robert Farmer, PhD²; David Baldwin, PhD³; David Carrell, PhD³; Gene Hart, MS³; Roy Pardee, JD³; Donald Bachman, MS⁴

¹Kaiser Permanente Georgia; ²University of Massachusetts; ³Group Health; ⁴Kaiser Permanente Northwest

Background and Aims: Individual-level race/ethnicity is important for research into causes and consequences of health disparities. For various non-research reasons, it has rarely been collected on enrollees in integrated delivery systems. Individual-level race/ethnicity can be found in medical record documentation. Manual abstraction on large numbers of medical records is costly. We developed a simple SAS algorithm for electronic abstraction of white and African American race from digitized progress notes and evaluated its accuracy by comparing electronically abstracted race with other data sources. **Methods:** A simple SAS algorithm, based on text search strings (e.g. white male, African American woman), scanned digitized progress notes for provider face-to-face visits from 2005 through July 2009 in Kaiser Permanente Georgia's (KPG) and Group Health Cooperative's (GHC) electronic medical record systems. White and African American race was abstracted. If the patient had more than 1 visit with abstracted race, the patient was classified using the earliest visit. Abstracted race was linked at the individual-level to survey datasets with self-reported race (2005 survey of working age adults, 2007 survey of adults with hypertension, 2000-2005 Medicare surveys) and mother's race on 2000-2006 birth certificates. White and African American race was abstracted from GHC progress notes from 2005 through July 2009 using the same algorithm and compared to self-reported race on health risk appraisals. Accuracy of the SAS algorithm was assessed by overall proportion matching race from the other datasets, Cohen's kappa, and McNemar's test. **Results:** White or African American race was electronically abstracted for 56,261 KPG and 6,427 GHC enrollees. Abstracted race matched race from the other datasets in 97-99% of enrollees. Cohen's kappas were highly significant ($p < 0.05$), ranging from 0.939 ± 0.013 ($N=657$ matches with hypertension survey records) to 0.994 ± 0.006 ($N=518$ matches with Medicare surveys). McNemar's tests were marginally significant for several datasets; and, misclassification was not systematically biased toward white or African American race. **Conclusions:** The SAS algorithm was highly accurate in electronically abstracting white and African American race from digitized progress notes of provider visits at KPG and GHC. We are expanding the evaluation to include additional sites and additional race/ethnic categories (e.g. Asian, Hispanic).

Keywords: SAS Algorithm for race abstraction, Race classification, Research and race

doi:10.3121/cmr.2010.943.c-a5-04

C-C1-01:

Building the Hypertension Registry: Adventures in Using the VDW Model

Heather Tavel, BS¹; Teri DeFor, MS²; Bix Swain, MS³; Nicole Schneider, BA²; Janelle Lee, MHA, DrPH³; Susan Shetterly, MS¹; Patrick O'Connor, MD, MPH²; Joseph Selby, MD, MPH³; David Magid, MD, MPH¹

¹Kaiser Permanente Colorado; ²HealthPartners; ³Kaiser Permanente Northern California

Background/Aims: The Hypertension Registry was built to support research on the recognition, detection, and treatment of hypertension at three HMO Research Network (HMORN) sites. The goal was to build a data mart modeled after the Virtual Data Warehouse (VDW) containing patients identified as hypertensive. Two of the sites already had VDWs in place at the start of the project, while the third site used the project as a springboard to build a VDW. **Methods:** A team consisting of data managers and analysts from each site met weekly in order to 1) determine exactly which existing VDW elements and possibly new fields would be required to support

hypertension studies, 2) help the VDW development efforts by clarifying data element creation rules and interpretations, and 3) document any new columns or tables in a Hypertension Registry data dictionary. All three sites had equal responsibility for registry development, documentation, and writing distributed code for quality checks rather than a lead site being designated for these efforts. **Results:** An algorithm for defining hypertension was used to create a “superset” of hypertensive patients. Using the protocol as a guide, the data team examined existing VDW tables to determine if all aims could be met with current data elements and developed solutions for additional variables or tables necessary to support planned papers. Challenges to harmonization of data elements were encountered in relation to data quality, data availability, and differing business models. Data team members gained a better understanding of the strengths and benefits of having a VDW in place, but also learned of areas where differing interpretations of the VDW data dictionary necessitated rework in the VDW or the need for better documentation. This presentation or poster will share how the team attempted to resolve these challenges in order to aid similar studies in the future. **Conclusions:** The Hypertension Registry is currently supporting at least 12 different hypertension papers and has been fully documented so that Hypertension Registries could potentially be built at additional sites. Continued discussion of VDW supported research efforts within the HMORN community is needed to learn how to best identify and create efficiencies. **Keywords:** Hypertension Registry, Virtual Data Warehouse, Algorithm for defining hypertension

doi:10.3121/cmr.2010.943.c-c1-01

C-C1-02:

Creating a GIS Infrastructure to Evaluate Air Quality's Effect on Health Outcomes

Rick Krajenta, BS¹; Lois Lamerato, PhD¹; Jason Booza, PhD²; Alice Grgicak-Mannion³; Helen Krause²; John Reiners, PhD²; Delbert Raymond, RN, PhD⁴; Linda Weglicki, PhD, RN⁵

¹Henry Ford Health System; ²Wayne State University; ³University of Windsor; ⁴Eastern Michigan University; ⁵National Institutes of Health

Background and Aims: Air pollution exposure is a complex and heterogeneous condition. The association of air pollution with disease is consistently established in many epidemiological studies. This creates a need for an infrastructure to measure both air pollution levels and health outcomes geospatially. **Methods:** As part of the North American Public Health Institute, we developed a multidisciplinary, international research team encompassing Windsor, Ontario, Canada and Detroit, Michigan, USA. While this geographic region consists of a single air-shed it exhibits different socioeconomic conditions and health delivery systems. We deployed 100 active and/or passive pollution samplers across this common air-shed based on U.S. Environmental Protection Agency guidelines for ambient air monitoring to determine levels of gaseous (i.e., NO₂, SO₂), volatile and particulate pollutants at a resolution greater than what is currently available. We are collecting asthma-related health outcomes for the populations residing within this region using emergency room, clinic, and hospital encounters from American and Canadian health care providers. Utilizing global positioning satellite for pollution sampler allocation and geocoding technique for patient address location, we created a complex mappable system to relate air pollution components and levels to health outcomes. **Results:** Air pollution levels can be linked to asthma events by several mechanisms. The first step is to use direct measurement value of the closest sensor by creating Voronoi polygons and mapping the asthma events to that geographic area. Associations between pollution level and asthma exacerbation can be examined between these data set. Models of the pollution levels using land use regression can increase the resolution beyond the point values. Longitude/latitude coordinates can be mapped to a pollution level in the surface generated by the model. **Conclusion:** Establishing a geospatial information system (GIS) integrated database with both pollution measurements and health outcomes creates a comprehensive resource. This infrastructure is flexible enough to investigate the effects of air pollution on

many diseases, such as respiratory and cardiovascular disease, by physical and mathematic modeling. We can also compare health outcomes by health delivery system.

Keywords: Air pollution exposure, Geospatial Information System, Asthma-related Health outcomes

doi:10.3121/cmr.2010.943.c-c1-02

C-C1-03:

Using the Electronic Health Record to Create Population Denominators: Optimization Using Insurance Enrollment

G. Craig Wood, MS¹; Jason Roy, PhD¹; Joseph Leader, BS¹; Walter Stewart, PhD¹

¹Geisinger Health System

Background and Aims: The HMORN includes research centers that are part of integrated delivery systems where the practice and insurance entities are independent (e.g., Geisinger, Marshfield). Rules for defining population denominators at these centers require acknowledging that patients in the practice population may not be members of the insurance entity. Methods for defining a population denominator for primary care patients from the electronic health record (EHR) can be validated in the subset of patients who are in the primary care system and members of the insurance entity. The aim of this study was to validate and optimize a method for calculating population denominators from the EHR. **Methods:** We proposed a method for defining population denominators from the EHR data at HMORN 2009. This method was based on describing utilization patterns of primary care patients' overtime. For this study, the cohort was limited to the subset of Geisinger primary care patients who were enrolled in the Geisinger Health Plan. Survival analysis was used to minimize bias in person-time estimates and incidence estimates. The aims were to identify optimal times from insurance enrollment to first utilization and time from final utilization to insurance disenrollment. Since the results are likely dependent upon gender and age (i.e. a typical gap in utilization will tend to be longer for young males as compared to older males), the analyses were compared across patient demographics. **Results:** To define cohort entry, the time between initial encounter and insurance enrollment was estimated using survival analysis. EHR enrollment was considered active until the patient failed to have any encounters with a primary care clinic for an age/gender specific cutoff of time. If the patient became inactive, the end date was imputed forward in time based on optimization from survival analysis using the insurance enrollment. These estimates were used to create an age/gender specific algorithm for calculating population denominators from the EHR. **Conclusions:** EHR utilization can be used to define population denominators. Validation of the proposed method was conducted by comparing results to insurance enrollment spans. This application is limited to clinical areas where there is evidence of relatively complete capture.

Keywords: Electronic Health Record, Defining population denominators, Data utilization

doi:10.3121/cmr.2010.943.c-c1-03

C-C1-04:

How to Win Friends and Influence People with the SAS Output Delivery System

William Tolbert, MA; Group Health Cooperative

Background and Aims: Long-time SAS users remember the days when SAS output was embarrassingly ugly. Version 7 saw the introduction of the Output Delivery System (ODS). ODS has matured into a very capable subsystem that gives users powerful reporting options. This presentation will highlight useful features and outline a macro-based system for handling multiple ODS destinations simultaneously. Nowadays there is no excuse for ugly SAS output! When building reports, SAS users should think about the needs of those using the reports. Some people just want to review frequency tables, and are happy to do so on a monitor. Others want to be able to print data for review in a meeting. And, there are always those that want to work with the data in a spreadsheet. Consider the ideal formats for each of the users