analytic data source optimized for research. Updates to the RDB can be reflected immediately in the VDW. **Conclusion:** The RDB, when it is complete, will be a unique and valuable resource for clinical,

Keywords: Building IT infrastructure, Project Ami/Method, Project benefits
doi:10.3121/cmr.2011.1020.ps1-39

PS1-44:

Comparing Measures of Disease Frequency Between Population Denominator Methods Using Data in the Electronic Health Record

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

¹Geisinger Health System

Background/Aims: The HMORN includes research centers that are part of integrated delivery systems where the clinical practice and insurance entities are independent (e.g., Geisinger, Marshfield) and where the primary care practice offers the larger population sample. Population denominators in these systems can be defined from insurance membership only, the primary care practice only (i.e., using electronic health records or EHR), or the overlap of the two (i.e., primary care patients may not be members of the insurance entity and visa versa). Methods for defining denominators and person-time for members of an insurance plan are well established. However, such methods have not been developed or validated for primary care practices using the EHR. The aim of this study was to use EHR data from a large primary care practice to derive measures of disease prevalence and incidence and to evaluate the validity of these measures using insurance claims data. **Methods:** We presented methods for defining population denominators from the EHR at HMORN 2010. One method performed better than others when validating against insurance enrollment. The current study will compare prevalence and incidence of hypertension and diabetes when using the various EHR denominator methods and validate them against incidence and prevalence as calculated from insurance claims. Results will be stratified by age and gender. Disease status will be defined by identifying subjects that had 2 or more outpatient office visits with a diagnosis or with the diagnosis appearing on the problem list. Ninety-five percent confidence intervals will be used to identify statistical differences. **Results:** A total of 164,030 patients have been identified for inclusion in the analysis. Analysis is ongoing and final results will be presented at HMORN 2011. **Conclusions:** The choice of denominator method may lead to differences in incidence and prevalence rates. More research is needed to determine which numerator and denominator methods most strongly agree with measures of disease frequency calculated from insurance enrollment and claims.

Keywords: Population denominators, Measures of disease frequency, Electronic health record

doi:10.3121/cmr.2011.1020.ps1-44

PS1-34:

Make Your Match and Capture the Data: Use of SAS Text Parsing Functions

Robert Rosofsky, BA, MA; Health Information Systems Consulting, LLC

Background: SAS has a set of RX ("regular expression") pattern matching and string manipulation functions. They provide a great deal more flexibility and power than string functions such as INDEX(), SCAN(), SUBSTR(), etc. in managing character strings. These functions enable one to locate, extract, and change patterns of character strings and are especially useful in situations of varying data patterns, formats, and placement within your source data. **Methods:** This presentation is an introduction to the power of SAS pattern matching functions and will make use of real-world examples to illustrate their utility. For extracting "poorly-formed" free text electronic medical data, such as notes, drug descriptions, and laboratory values into analytical data files, the presentation will enable programmers to get started in approaching their own data sources by using these functions. The presentation will present both the SAS RX functions as well as touch on the SAS implementation of Perl regular expression functions (PRX).

Keywords: SAS, Text extraction, Pattern matching

doi:10.3121/cmr.2011.1020.ps1-34

PS1-05:

Feasibility of Extracting Oncology Treatment Data from an Electronic Health Record

Nikki Carroll, MS¹; Capp Luckett, MS, CIS¹; Gwyn Saylor, BS¹; Debra Ritzwoller, PhD¹

¹Kaiser Permanente Colorado

Background/Aims: New source data systems almost always cause angst among programmers. Source data systems usually are built for user ease of use and are not built for ease of getting data out. A new Electronic Health Record (EHR) module designed specifically for Oncology treatment was no different. Having access to data that was previously unavailable caused excitement among researchers, so we conducted a project to explore extracting Oncology protocols, treatment plans and medications from this new module in order to: 1) determine the process needed to identify protocols, treatment plans and medications from the EHR, 2) validate the process with medical record review, and 3) build VDW tables that could logically hold this data and accommodate data from other HMOs. **Methods:** The study team: a) identified EHR tables and fields that contained medication data, protocols and treatment plans specific to Oncology, b) completed multiple rounds of validation through chart review, and 3) identified the structure and key variables needed to construct VDW tables of Encounters, Treatment Plans, and Medications. These three steps were then used to identify patients currently receiving cancer treatment in the Oncology department and data is being pulled to populate these VDW tables. **Results:** Multiple challenges were encountered and solutions identified. First, the EHR tables, fields and linkages required significant exploration to discern useful data elements and correct joins. For example, oral and infused medications are kept in separate tables and each table contains multiple and different date fields and status codes to determine if the drug was actually given to the patient. Other factors complicating identifying Oncology treatment included determining work flows and matching those work flows to data that was extractable from the EHR and poor documentation not specific to our HMO. An iterative process was used to validate each data pull. **Conclusion:** Identifying Oncology treatment data in the EHR was a process fraught with multiple challenges. We believe, however, that we have developed code that identifies protocols, treatment plans and medications used to treat cancer patients. This is an important first step in compiling data needed for future research on the treatment of various cancers.

Keywords: Electronic health record, Chemotherapy treatment

doi:10.3121/cmr.2011.1020.ps1-05

Plenary III-02:

Accuracy of Natural Language Processing to Identify Pneumonia from Electronic Radiology Reports

Sascha Dublin MD, PhD¹; Eric Baldwin, MS¹; Rod Walker, MS¹; Peter Haug, MD²; David Carrell, PhD¹; Wendy Chapman, PhD³

¹Group Health; ²Intermountain Health Care; ³University of California, San Diego

Background/Aims: Pneumonia is common and can be devastating in older adults. Health plan data hold promise for studying pneumonia, but ICD-9 codes have poor accuracy for this condition. Natural language processing (NLP) offers potential to accurately and efficiently identify pneumonia from electronic medical records (EMRs). Our aims were to train one NLP tool to identify pneumonia from electronic radiology reports and to assess its validity compared to manual review. **Methods:** ONYX is an NLP system that identifies clinical conditions (findings, symptoms, diagnoses) in free-text reports using knowledge about language in the reports and the specific medical domain. Building on a knowledge base from a prior NLP system, we trained ONYX using 1,100 chest radiograph reports from among 70,000 that were previously manually reviewed for pneumonia. We trained ONYX to classify reports into one of three mutually-exclusive categories: 1) consistent with pneumonia; 2) not consistent with pneumonia; and 3) requiring manual review (for example, containing conflicting statements about pneumonia). To assess validity, we ran ONYX on a "test set" of 5,000 randomly selected reports, oversampling reports showing pneumonia (based on manual review)

and those from people with comorbidities (cancer, chronic lung disease, or congestive heart failure). We estimated the sensitivity, specificity, and positive predictive value (PPV) of ONYX compared to manual review using multivariable logistic regression, accounting for clustering of reports within people. We examined how ONYX's performance varied by age and comorbidity. **Results:** ONYX classified 26% of reports (1,276/5,000; 38% [841/2,200] of true pneumonias and 16% [435/2,800] of non-pneumonias) as "requiring manual review" based on pre-defined criteria. Reports from older people were more likely to require manual review than those from younger people. Among reports that could be classified, ONYX had a sensitivity of 91% (1,242/1,359), specificity of 92% (2,170/2,365), and PPV of 81% (1,242/1,437, modeled based on pneumonia prevalence in the source database). Sensitivity and specificity were similar regardless of comorbidity. **Conclusions:** NLP offers potential for identifying pneumonia outcomes from EMR data. Next steps include 1) further training to decrease the proportion of reports requiring manual review and 2) evaluating the accuracy of ONYX in other health systems.

Keywords: Natural language processing, Pneumonia, Validity
doi:10.3121/cmr.2011.1020.plenary-iii-02

C-C4-02:

Using a Natural Language Processor to Remove All Elements of Personal Health Information (PHI) to De-identify Clinical Annotations for the Specimen Retrieval System (SRS)

Andrew Glass, MD¹; Sheila Taube, PhD²; Andrew McMurry, BA³; Chris Eddy, BS⁴; Pierre-Andre La Chance, BS⁴; Lisa McShane, PhD²; Mei-Yin Polley, PhD²; Isaac Kohane, MD, PhD³

¹Kaiser Permanente Northwest; ²National Cancer Institute; ³Harvard Medical School; ⁴Kaiser Center For Health Research

PHI in HMORN medical records must be systematically removed prior to use and sharing data with co-investigators at other institutions. The problem is particularly acute when one needs to link clinical and follow-up data with pathologic specimens from cancer patients in studies of prognostic and predictive tumor markers. Natural language processing (NLP) can be customized to identify and remove PHI in selected clinical records by substituting nonsense characters. Such a process of deidentification can proceed largely without human intervention and, if successful, can allow efficient linkage of clinical notes with similarly deidentified specimens for laboratory investigators at other institutions. The Shared Pathology Informatics Network (SPIN) was funded by the Cancer Diagnosis Program of the National Cancer Institute to develop a computerized program that would search pathology department text files and reports from several institutions and retrieve, in a database, all records that met search criteria. SPIN was designed to access information from each institution's electronic records without affecting the records and collating data from several sites into a single report. Our current project, the Specimen Retrieval System, requires the identification of large numbers of cancer specimens of specific types from Kaiser Permanente Northwest (KPNW) and, subsequently, other Cancer Research Network sites, retrieval of those specimens and linkage with clinical annotations that describe staging, treatment and outcome. We have used the existing databases of the KPNW Tumor Registry and Department of Pathology to identify cases. The EPIC electronic medical record and several other computerized text files provided the clinical notes for these patients. Using SPIN technology we then processed text files to 'scrub' them of PHI. All records were then manually inspected to assess the completeness of the process, which elements of PHI persisted and which had been successfully 'scrubbed' of PHI. We processed several hundred files from more than 100 patients and will report detailed statistical analysis. This technology has demonstrated significant capability to facilitate searches for pathologic specimens and clinical annotations using conventional reports from pathology departments. It is a valuable tool to remove PHI, deidentify medical records and ease sharing of clinical information with investigators at other sites.

Keywords: Personal Health Information, Deidentification of Clinical Annotations, Cancer Specimen Retrieval

doi:10.3121/cmr.2011.1020.c-c4-02

PS1-09:

Readmission and 30-day Mortality Among CAP Patients in Two Systems of Care

Laurel Copeland, PhD, MPH¹; Debbie Godwin, BS²; Dean Kjar, MS²; Daniel MacCarthy, BS³; Dennis Myers, MD²

¹Scott & White/Veterans Affairs ; ²Scott & White Healthcare; ³South Texas Veterans Health Care System

Objectives: Readmission and mortality among elderly patients admitted to inpatient care are critical outcome measures for healthcare systems. While the Veterans Health Administration (VA) has long made national administrative databases available for health services research, the HMORN Virtual Data Warehouse (VDW) at Scott & White (SWHP) has only been developed in the past year. The current study sought to utilize both data sources while examining outcomes of hospitalization for community-acquired pneumonia (CAP) in two systems of care in the same geographic region. **Methods:** Patients >65 years of age hospitalized with CAP (ICD9 codes 480-483, 485-487) in Central Texas in VA fiscal year 2009 (Oct 2008-Sep 2009) were identified. Dates of care and inpatient death were sought to determine outcomes subsequent to CAP admission. Covariates included sociodemographic measures (such as age, gender) and comorbidity. **Results:** There were 469 SWHP CAP patients and 152 VA CAP patients admitted in the year studied. The SWHP fiscal year differed from the VA fiscal year, potentially complicating the work of the SWHP data team. VA CAP patients averaged 77.2 years (SD 7.6), were 96% male, and experienced an inpatient death rate of 18%. SWHP CAP patients averaged 81.6 years (SD 8.2), and 56% were women. Date of death as an inpatient was not explicitly available in VDW data. Patients in the two systems varied significantly as VA patients were more likely to be male and younger. **Conclusions:** Patients in the federal VA system can be studied in conjunction with private-care patients in the same catchment area using newly available VDW data. The data presented required two teams of data analyst/programmers, one working within the VA system and the other working within the SWHP VDW system. Coordination was facilitated by the VDW Site Director Godwin working together with the jointly appointed VA/SWHP investigator Copeland. Sharing SAS code between VDW and VA programmers helped to lend consistency to the data analytic approaches across systems. Adjusted analyses employing high-dimensional propensity scores could be employed to assess whether inter-system differences in CAP-related outcomes can be validly estimated.

Keywords: Community acquired pneumonia, Healthcare systems, Epidemiology

doi:10.3121/cmr.2011.1020.ps1-09

C-B4-02:

Enhancing the Quality and Efficiency of Obstructive Sleep Apnea Screening Using Health Information Technology: Results of a Geisinger Clinic Pilot Study

Robert Moser, RRT, MHA¹; Elizabeth Boyer, BA¹; Diana Lupinski, RN¹; Jonathan Darer, MD, MPH¹; Tammy Anderer, PhD, CRNP¹; Alexander Villareal, MD¹; Peter Berger, MD¹

¹Geisinger Health Systems

Background: Health information technology (HIT) can improve the screening and delivery of care to patients with chronic illnesses. To date there are no published data describing the use of HIT for screening patients with obstructive sleep apnea (OSA), a prevalent under-diagnosed chronic condition representing a substantial and growing socio-economic burden. Geisinger Clinic is a physician-led, multi-specialty group practice in rural Central Pennsylvania with integrated electronic medical records. **Aim:** We sought to determine whether HIT could be used to identify OSA amongst patients at risk for the disease through the use of a simple questionnaire and a physician approved referral protocol. **Methods:** Using an internet-based portal through which patients can review their Electronic Health Record (EHR), schedule appointments, and receive communication, patients with a body mass index