

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, MS<sup>1</sup>; Joseph Leader, BS<sup>1</sup>; Walter Stewart, PhD<sup>1</sup>

<sup>1</sup>Geisinger 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, MS<sup>1</sup>; Capp Luckett, MS, CIS<sup>1</sup>; Gwyn Saylor, BS<sup>1</sup>; Debra Ritzwoller, PhD<sup>1</sup>

<sup>1</sup>Kaiser 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, PhD<sup>1</sup>; Eric Baldwin, MS<sup>1</sup>; Rod Walker, MS<sup>1</sup>; Peter Haug, MD<sup>2</sup>; David Carrell, PhD<sup>1</sup>; Wendy Chapman, PhD<sup>3</sup>

<sup>1</sup>Group Health; <sup>2</sup>Intermountain Health Care; <sup>3</sup>University 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)