well as emerging functions. These transformational trends can be summarized as the use of evolving collaborative technologies and social networks, development of tools to support mobile access, an emphasis on instructional technology services and information literacy, the promotion of embedded librarians as members of research teams, reconfiguration of library space to support learning and research, digitalization of shared library collections, the assumption of leadership in scholarly communication and publication issues, and the development of services supporting data management and curation. **Conclusions:** With the advent of digital technologies, libraries are proving to be successful partners and influential leaders in the research process. Emerging resources and services will continue to create new opportunities, challenges and expectations and will foster innovation and improvement in research performance and effectiveness.

Keywords: Information technology, Collaboration, Research libraries
doi:10.3121/cmr.2011.1020.ps1-36

C-C4-01:

Rapid Exploration of Large Clinical Text Corpora for Information Extraction Feasibility Studies

Sharon Fuller, BA¹; David Carrell, PhD¹

¹Group Health Research Institute

Background/Aims: Large amounts of information are “buried” in unstructured clinical text such as chart notes, pathology reports and radiology reports. Through electronic medical record systems, much of this text is available for computer-aided analysis. Determining the specific language used in clinical text to express content of interest is an important early step in text-mining efforts. **Methods:** We copied ~29.2 million clinical documents from our Epic Clarity database and other data sources to a secure SQL Server 2008 database, adding a full-text index to the textual content. Details will be discussed. To query and view clinical text we developed a Clinical Text Explorer application using Microsoft Access. None of the clinical text documents are de-identified; IRB approval is required for use. Features include: intuitive interface for testing and refining search schemes; quickly returns chart documents containing specified text; user can review either random or “best match” samples of documents to inform estimates of sensitivity and specificity of the search; highlighting marks facilitate visual scanning for terms of interest. **Results:** Clinical Text Explorer allows researchers to quickly and easily identify patients who could not have been identified reliably by searching only on structured data. The following example describes the iterative process of defining the best search terms to find records mentioning results of the Oncotype DX test for breast cancer. In less than an hour we determined that the search “oncotype dx” was too narrow, while “oncot*” was too broad (drawing in records with words like “oncotech”). The search “oncotyp*” was the most comprehensive without losing specificity. When limiting the search to test results, we found that adding the additional criterion that “oncotyp*” occur near “score” eliminated most irrelevant documents, while adding “recurrence” narrowed the results too far. This search returned substantially more records than were discovered by searching structured data from lab results alone. **Conclusions:** The ease with which complex searches over large amounts of clinical text can be executed by this application eliminates barriers to text exploration posed by conventional methods such as regular expressions in SAS, allowing the domain expert (epidemiologist, physician, chart abstractor, etc.) to directly evaluate the results and refine the search.

Keywords: Abstraction, Electronic medical record
doi:10.3121/cmr.2011.1020.c-c4-01

PS1-19:

Using Electronic Data Extraction to Identify Subjects with Metabolic Syndrome: A Validation Using Manual Chart Review

Pao Hsiao, MS¹; G. Craig Wood, MS²; Terry Hartman, PhD¹; Christopher Still, DO²; Gordon Jensen, MD, PhD¹

¹The Pennsylvania State University; ²Geisinger Health System

Background/Aims: Metabolic syndrome (MetS) increases risk for developing serious health conditions, but MetS is rarely recorded as a diagnosis in clinical practice. The purpose of this study was to compare an electronic data extraction process to a manual chart review in identifying patients with MetS. **Methods:** Electronic health records (EHR) of 48 randomly selected Geisinger Rural Aging Study (GRAS) participants (24 males, 24 females; age =65) were selected for review. A trained auditor collected information for each participant on biochemical measurements (including triglycerides, HDL-cholesterol, glucose); blood pressure and measured height and weight. MetS was defined as having three of the five criteria based on ATP III guidelines. However, since waist circumference was not available for all participants, BMI >30 kg/m² was used. Records were reviewed for a 48-month period, starting at date of GRAS study entry. Demographics, diagnosis codes, laboratory data, past medical history, medications, progress notes, and physician comments were included in the audit. Independently, an electronic data extraction of the EHR was used to identify subjects with MetS. Rates of agreement between the manual chart review and data extraction process were calculated. Discrepancies were examined for sources of disagreement and statistical agreement was assessed with Cohen's Kappa. **Results:** As expected, very few subjects with an ICD-9 diagnosis code for MetS were identified (n=2). The manual chart review identified 27 subjects (56%) that met criteria based on ATP III guidelines. The independent electronic process identified 25 of these 27 subjects as having MetS and found an additional 2 subjects (inadvertently missed during manual chart review). The two cases that were missed using the electronic process had abnormal glucose values that were not available from the electronic data extraction. Agreement between the two methods was almost perfect (kappa=0.92). **Conclusions:** This validation study demonstrates that an electronic data extraction process for identifying MetS has substantial agreement with the gold standard, a manual review of health records for clinical data. This process has the advantage of quickly querying large amounts of data that may be missed by manual chart review and can aid in collection of data for health outcomes research.

Keywords: Electronic data extraction, Metabolic syndrome
doi:10.3121/cmr.2011.1020.ps1-19

PS1-39:

The Kaiser Permanente Northern California Oracle Research Database

Joe Terdiman, MD, PhD¹; Jamila Gul, BS¹

¹Kaiser Northern California

NCAL developed a pilot Research Database (RDB) in KPNC that contains clinical and administrative data for the 3.3 million current and 10 million past Health Plan members in the KPNC region, with some data that spans a 40 year period. The primary reason to create such a data warehouse is that researchers are often frustrated by the difficulty of finding, extracting, cleaning, interpreting and transforming the data into useful analytical datasets. This is due to the absence of an efficient research data infrastructure that is optimized for the analytical needs of researchers. In the RDB, we have aggregated clinical data from legacy systems, the current electronic health record (EHR) system, various public use datasets, and research datasets into a research data warehouse. **Aims:** The specific aims of the project are: 1) to develop a uniform schema with consistent data definitions and coding, using standardized terminology, that will harmonize aggregated legacy and current KP electronic health records, 2) develop interfaces and extract, transform and load (ETL) processes for the KP electronic medical record system and other clinical systems that will feed data to the NRDB, 3) implement a security model that is compliant with HIPAA and State regulations and KP policies, and 4) develop federated schema with the other KP regions, so that it will be possible to do federated queries of each region's databases and aggregate the results, while enabling each region to control access to its data. **Methods:** We developed the RDB in Oracle. ETLs to various data sources were developed in SAS, Informatica and PL/SQL. **Results:** The database is available to over 150 KP researchers and programmers in the Division of Research and other KPNC researchers, who are currently engaged in over 250 research projects. The KPNC Virtual Data Warehouse (VDW) has been incorporated into the RDB as a materialized view or data mart. It is a primary example of an

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)