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, BA[1]; David Carrell, PhD[1]

[1]Group 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, MS[1]; G. Craig Wood, MS[2]; Terry Hartman, PhD[1]; Christopher Still, DO[2]; Gordon Jensen, MD, PhD[1]

[1]The Pennsylvania State University; [2]Geisinger 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, PhD[1]; Jamila Gul, BS[1]

[1]Kaiser 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