PS2-22:

Accuracy of Diagnostic Codes to Identify Rheumatoid Arthritis in Archived Electronic Health System Data: Support for Future Cancer Research Network Studies of Lymphoma Risk Pathways

Robert Greenlee¹; Jennifer Drahos²; Jeffrey VanWormer¹; Ola Landgren²; Jill Koshiol²

¹Marshfield Clinic / Security Health Plan of Wisconsin; ²National Cancer Institute

Background/Aims: In preliminary work toward Cancer Research Network-based studies of lymphoma risk pathways, we undertook an evaluation of the utility of ICD-9 CM diagnostic codes to accurately identify potential risk factor conditions, including rheumatoid arthritis (RA). **Methods:** Using the enrolled Virtual Data Warehouse (VDW) cohort at Marshfield Clinic Research Foundation, we ascertained a set of potential RA cases diagnosed between 2000 and 2010 by the presence of one or more diagnostic codes for RA (ICD-9 CM 714-714.89). Medical records were abstracted by trained staff on a random sample of 206 cases. Cases were adjudicated into categories of confirmed, probable, documented physician diagnosis only, equivocal, and non-case using a literature-based gold standard definition scheme. Outcome measures included positive predictive value (PPV) and relative sensitivity, with the confirmed, probable and physician diagnosis categories considered valid cases for the main analysis. **Results:** Upon review, 25 subjects did not have sufficient medical records available for evaluation, leaving 181 subjects for analysis, including 57 with only one RA code and 124 with 2 or more instances of an RA code. Overall PPV for patients with one or more RA codes was 56%. Subjects with only one diagnostic code had very low PPV (12%), while PPV was higher among those with 2 or more (76%). PPV improved to 91% when requiring a rheumatologist diagnosis, but only 40% of the true cases in the set were detectable in this way. The one algorithm that provided acceptably high PPV and relative sensitivity required 2 or more RA diagnostic codes and history of a rheumatoid factor test (PPV 83%, relative sensitivity 82%). **Conclusions:** A single ICD-9CM diagnostic code for RA was not highly predictive of a true diagnosis, but PPV was enhanced when requiring multiple codes and incorporating other VDW data elements. With one exception, algorithms with higher PPV had strong reductions in relative sensitivity. Limitations include non-systematic collection of RA medication data preventing the use of treatment in the algorithms, and the inability to evaluate absolute sensitivity. Next steps include a joint RA analysis with investigators at Kaiser Permanente Southern California, and possible inclusion of RA treatment data into the Marshfield algorithms.

Keywords: Rheumatoid Arthritis; Diagnostic Codes; Lymphoma
doi:10.3121/cmr.2013.1176.ps2-22

PS2-25:

Using Natural Language Processing to Extract Findings from Mammography Reports

Hongyuan Gao¹; Erin Aiello Bowles¹; David Carrell¹; Diana Biust¹

¹Group Health

Background/Aims: Mammographic findings such as a mass may be associated with breast cancer risk, but these data are only available in free-text reports and require resource-intensive manual abstraction. We developed and tested a Natural Language Processing (NLP) algorithm to extract mammographic findings (mass, calcification, asymmetric density, and architectural distortion) from free-text mammography reports. **Methods:** We identified 92,947 reports for women receiving screening and diagnostic mammography at Group Health between 2007-2008. We developed an NLP algorithm based on Perl Regular Expressions in SAS v9.2. The algorithm identifies words indicating mammography findings (mass, distortion, asymmetry and calcification) and their related words denoting laterality, negation, family history, personal history and uncertainty. Three flags are made indicating possible errors of the NLP algorithm. An experienced abstractor manually reviewed a random sample of 50 mammography reports to test and refine the NLP algorithm. **Results:** The algorithm correctly identified a mass on 46/50 reports, calcifications on 48/50 reports, asymmetric density on 50/50 reports, and architectural distortion on 48/50 reports. The

NLP algorithm misinterprets sentences such as, "there are calcifications with no other asymmetry." The NLP algorithm incorrectly associated the negation word "No" with the key word "calcifications." Building more refined rules on association between negation words and key words will improve the accuracy. **Conclusions:** This NLP algorithm holds promise for accurate and fast identification of findings from free-text mammography reports. It can be shared across institutions and is an example of what can be done with free-text radiology reports, in addition to mammography. Manual review may still be necessary for some reports with a high probability of error, depending on resources available.

Keywords: Natural Language Processing; Validation; Mammography Report
doi:10.3121/cmr.2013.1176.ps2-25

PS2-26:

Coordinating Heterogeneous Data and Mixed Collection Methods to Support Population-Based Cancer Screening Research

Aruna Kaminen¹; Scott Halgrim¹; Gabrielle Gundersen¹; Sharon Fuller¹; Gene Hart¹; David Carrell¹; Carolyn Rutter¹

¹Group Health

Background/Aims: The central goal of Population-Based Research Optimizing Screening through Personalized Regimens (PROSPR), a recently-funded NCI initiative, is to develop multi-site, transdisciplinary research to improve the screening process for breast, colon, and cervical cancer. To support this goal, we aim to collect, document, and manage data for the entire colorectal cancer (CRC) screening process at Group Health (GH), an integrated health system and PROSPR Research Center. We describe the data sources, types, and collection methods being used to assemble the breadth of relevant information on patients, providers, tests, pathology, treatment, and outcomes this effort requires. **Methods:** To characterize the CRC screening process for GH members enrolled from 1993-2015, we employed administrative databases, previous CRC studies, data partnerships, and GH's EpicCare-based electronic medical record (EMR). These resources contain both structured data and unstructured text requiring the use of multiple collection methods, including programmatic extraction, natural language processing (NLP), and manual abstraction. **Results:** We are programmatically extracting demographic information on patients and providers from well-established administrative databases. Information on stool-based tests is extracted from lab databases and EpicCare. Colonoscopy and corresponding pathology notes are available as unstructured text in EpicCare for GH-performed procedures, and we are employing NLP to extract information on family history, test indication, and results from these notes. Scanned notes from contracted colonoscopy providers require manual abstraction; however, through partnership with our largest contracted provider, we receive electronic transfers of this information as structured data, minimizing manual review. For colonoscopies occurring prior to GH's 2005 implementation of EpicCare, we rely on data from five previously-conducted CRC studies. Treatment information is extracted from pharmacy and utilization databases, and CRC outcomes are available as structured data through partnerships with our local cancer registries. **Conclusions:** Under the auspices of an ambitious initiative such as PROSPR, documenting the entire screening process can be achieved by creating a comprehensive data collection system that coordinates all available data sources and maximizes their value with appropriate collection methods. Efficiencies can be gained by using data from prior studies and developing external data partnerships for access to higher-quality data.

Keywords: Data Collection System; Screening Process
doi:10.3121/cmr.2013.1176.ps2-26

PS2-37:

Development and Use of a Predictive Analytics Tool in a Large Healthcare Organization

Isaac Hoch¹; Tomas Karpati¹

¹Maccabi Healthcare Services

Background/Aims: Health organizations are beginning to apply predictive analytics as a central and critical tool for more effective healthcare management. However, the art is still far from maturity, and it is necessary to

develop and perfect the requisite analytic tools. A need exists for methods to measure illness burden and identify patients for targeted interventions. Most commercial programs are unable to use all of the data we have available for analysis. Their input is limited to age, gender, diagnoses and medications, while our database also contains a wide range of demographic, socio-economic, clinical and financial data at the patient level. We hypothesized that utilizing the richer data would generate robust analytic and predictive capabilities. We then developed a predictive analytical system that accesses our entire database. The design requirements included flexible and generic database mapping and transparency of any algorithm's internal processes. In addition, the system has embedded quality assurance processes and maintains an historical record of all analytical models and results. **Methods:** Data sources included approximately 15 years of history of physician and other medical professional visits, hospitalizations, emergency room visits, diagnoses, medications, laboratory results, imaging studies, pathology results, and extensive socio-economic, demographic data and associated costs of all medical expenditures. Analytics techniques used included linear regression, classification trees, and additional data mining methods. Models developed included: predicted annual cost and prediction of re-hospitalization within 30 days. Models were validated using R^2 , C-statistics and Positive Predictive Value (PPV). **Results:** The first model ($R^2 \sim 0.36$) was used to create reports for risk adjustment and physician profiling. The second model (PPV 54%) was incorporated into an existing program for preventing re-hospitalization. **Conclusions:** The Maccabi analytical tool has a robust predictive ability and has been successfully used for physician profiling and predicting re-hospitalization. We suggest evaluating this tool on different databases to yield insight into its transferability and robustness. The minimal required data set for use in other organizations needs to be determined.

Keywords: Predictive Analytics; Risk Adjustment; Modeling
doi:10.3121/cmr.2013.1176.ps2-37

PS2-38:

Sensitivity to Different Criteria for Identifying Diseased Patients in the Electronic Health Record

Jove Graham¹; Colin Barton¹

¹Geisinger Health System

Background/Aims: One of the primary challenges to observational research based on electronic health record (EHR) data is judging whether or not a patient had a given disease as of a specific date. The problem list, encounter record, and prescription medication orders are all sources of coded diagnoses, but their use varies across providers and diseases. The objective of this study was to test the sensitivity of using different diagnosis criteria for classifying patients with specific diseases in the EHR. **Methods:** A deidentified dataset was assembled for a randomly selected population of 10,000 patients (age 18+) who had at least one encounter per year from 2007-11, including all problem list, encounter, and prescription diagnoses. For each of the 17 diseases in the Charlson Comorbidity Index, we counted the total number of patients that had 1 or more problem list diagnoses, encounter diagnoses, prescription diagnoses, and combinations of the three. Because many past projects at our institution have used a criterion of 1 problem list entry or 2 encounter diagnoses ("P1/E2") as an inclusion criterion, this was highlighted as a benchmark for comparison. **Results:** While every patient with a disease should have it documented everywhere in the EHR, this study confirmed that this ideal scenario is not the case. The criterion of "P1/E2" was usually ranked between 5th and 9th (out of 20 criteria tested), suggesting that it casts a fairly wide net for most diseases. We were surprised to find that "time elapsed between two events" did not generally impact the results; i.e., requiring patients to have two encounter diagnoses >90 days apart did not yield substantially fewer patients than requiring them >1 day apart. Overall sensitivity varied dramatically by disease: for example, for Neoplasms, the 4 least restrictive criteria yielded similar numbers of patients, but then there was a significant drop-off. For Heart Failure, there was a much more gradual decline in the number of patients as criteria became more restrictive, suggesting greater consistency across the EHR. **Conclusions:** In conclusion, we hope empirical data like these can aid researchers in better understanding how diagnosis criteria affect their cohorts in retrospective studies.

Keywords: Diagnosis; Electronic Health Record; Charlson Comorbidity Index
doi:10.3121/cmr.2013.1176.ps2-38

PS2-39:

Family History Data from Clarity and Other Sources

Sharon Fuller¹; Deborah Seger¹

¹Group Health

Background/Aims: Family history information is an important tool to identify potential risks for many diseases like cancer or diabetes. Group Health Cooperative (GHC) collects family history at a variety of patient encounters. In 2004, GHC adopted Epic's electronic medical record system. We investigated Clarity, Epic's reporting database, to determine its usefulness as a new source of structured family history data. **Methods:** We compared several data sources for family history of four conditions: colon cancer, colon polyps, breast cancer, and ovarian cancer. We used proportional Venn diagrams for visualization of results. We examined all patients with a known birth date and sex, and any encounter recorded in Clarity between Jan 1, 2004 and Jun 30, 2012. For breast and ovarian cancer, we reviewed females only. **Results:** Among patients with any evidence of family history of a condition, Clarity's FAMILY_HX table independently contributes as many as 75% of patients with available family history; our VDW DX dataset adds an additional 2-20%, while Clarity's PROBLEM_LIST table adds no more than 2-3%. For female GHC patients, the Breast Screening Recruitment and Reminders survey (BSRR), completed at every screening mammogram, remains a key source, being the sole contributor of 25% (21,961 of 86,368) of patients with a family history of breast cancer and 77% (47,487 of 61,484) for ovarian cancer. When comparing information available from different data sources, it is important to be aware of varying definitions. For example, ICD-9 family history diagnosis codes are general, specifying only the primary site with none of the restrictions on relationship degree or age that may be intended in Epic. A key advantage of Clarity family history data and the BSRR survey over ICD-9 codes is that they provide detailed information about exactly which relatives are affected, including affirming a negative history. On the other hand, VDW data are available from both internal and external providers – the remaining sources are available only for GHC group practice patients. **Conclusions:** Examining records from Clarity's FAMILY_HX table may significantly augment, in both numbers and detail, conventional clinical and survey sources of family history data.

Keywords: Family History; Venn Diagram; Clarity
doi:10.3121/cmr.2013.1176.ps2-39

PS2-40:

Using Automation to Add Electronic Clinical Data to a Research Patient Registry

Sharon Fuller¹; David Carrell¹; Beverly Green¹

¹Group Health

Background/Aims: Clinical research trials often rely on automated data systems, such as the electronic medical record to create and maintain a patient registry, to determine a participant's progress through the study, or for analytic purposes. Data might come from a variety of sources and include such elements as procedures, lab results, hospitalization, vital status, enrollment and participant responses from survey software. An automated process for adding outside data to a study tracking database can save significant time over abstraction or manual data import processes, and can allow determination of an individual's study status in close to real time. **Methods:** There are a number of ways of setting up an automated import process, depending on data needs and availability. Over the course of several clinical trials, we have developed a process with the following steps, all occurring automatically on a scheduled basis: 1. A SAS program performs Extract/Transform/Load (ETL) tasks: 1a. Extract and transform the desired data from each source. 1b. Load new data into a SAS dataset for historical purposes and into a staging table in the tracking database preparatory to actual use. 1c. Generate automated emails when specific milestones occur (e.g., study end date reached or first use of a given lab test code). 2. Another SAS program checks the ETL logs and sends email about any errors occurring in the ETL process. 3. A database job calls stored procedures to insert data from the staging tables into the main study events table and send email detailing how many records were processed. 3a. If manual intervention is needed before the final data is loaded, the user can call the stored