

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