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

procedure once data entry is complete. 3b. A master stored procedure can call individual procedures in order (e.g., importing a positive lab result may trigger randomization). **Results:** The ETL process described above has proven itself to be robust and adaptable in a variety of study contexts. **Conclusions:** An automated system for adding clinical data to a patient registry or tracking database can save staff time, allow access to near real-time data, and facilitate integration of data from different sources.

Keywords: Patient Registry; Electronic Medical Record; Tracking Database
doi:10.3121/cmr.2013.1176.ps2-40

PS2-42:

Assessment of Drug Induced Liver Injury Using an Automated Causality Assessment Tool

T. Craig Cheetham¹; Janet Shin¹; Rich Murray²; Fang Niu¹; Steph Reisinger²; Robert Azadian³; Gregory Powell³

¹Kaiser Permanente Southern California; ²United BioSource Corporation; ³GlaxoSmithKline

Background/Aims: Drugs are a common cause of acute liver failure. However, confirming the diagnosis of drug-induced liver injury (DILI) is challenging due to its low incidence, lack of diagnostic markers and idiosyncratic nature. Our aim was to develop an automated scoring algorithm to identify potential DILI cases based on the Roussel Uclaf Causality Assessment Method (RUCAM) and test it in healthcare data. **Methods:** RUCAM includes seven criteria of causality assessment. Operational scoring definitions were developed for each of the criteria in collaboration with DILI experts and programmed into an automated algorithm. The automated algorithm was then tested on a retrospective cohort of patients at Kaiser Permanente Southern California. Study participants were ≥ 18 years old with twelve months of continuous membership plus drug benefit prior to exposure to one of 14 drugs commonly associated with DILI. Eligible patients filled at least one of the study drugs between January 1, 2003 and August 31, 2011. Patients were scored and cases were categorized into one of five categories ranging from 'Highly Probable' to 'Excluded'. Rates of DILI events per 10,000 exposures were calculated and frequency counts of possible scores for each of the seven criteria were analyzed. **Results:** We identified 14,925 potential DILI events following 3,321,835 study drug exposures. Four antibiotics accounted for 89.4% of events. The average (SD) patient age was 60 (16) years and 54.5% of potential DILI events occurred in females. Cholestatic injury (65.0%) was the most common type followed by hepatocellular 29.4%, and mixed 5.6%. DILI events were categorized as probable or highly probable in 15.5% of cases, 59.6% were identified as possible, and 24.9% were unlikely or excluded. The overall rate of probable or highly probable DILI events was 6.9 per 10,000 exposures. The scores for most criteria spanned the entire range of possible scores and there were no obvious floor or ceiling effects. However, Criteria 5 (other non-drug causes of liver injury) showed poor discrimination of possible scores. **Conclusions:** An electronic causality assessment algorithm was developed and successfully tested on healthcare data. Further validation is needed comparing these results with the 'gold standard' medical record review by DILI experts.

Keywords: Drug Induced Liver Injury; Causality Assessment
doi:10.3121/cmr.2013.1176.ps2-42

PS2-43:

Sensible Use of ICD-9 Diagnosis of Observational Data in Breast and Endometrial Cancer Research

Wen Feng¹; Xiaowei (Sherry) Yan¹; Ryan Colonie¹; Jeffrey Prichard¹; Radhika Gogoi¹; Azadeh Stark¹

¹Geisinger Health System

Background/Aims: Utilization of administrative data (such as EHR data) in population-based research is resource advantageous, despite its potential limitations. Few studies have fully assessed the validity and efficiency of EHR-retrieved data. We developed ICD-9 based algorithms and operational processes to evaluate applicability of EHR-derived cancer data. **Methods:** We retrieved data between 01/01/2002-12/30/2011 from 4 different EHR sources and developed 3 ICD9-based diagnostic algorithms (1+, 2+ and 5+). Women were classified into breast or endometrial cancers or benign breast

conditions (BBC). One trained abstractor manually reviewed medical records and recorded data into a structured database. Every 10 observations were selected and reviewed. Basic descriptive statistical analyses were conducted; observations with questionable values were flagged for re-evaluation. The final dataset was considered the "gold standard" and used to validate the algorithms and to assess the duration between the diagnostic and administrative dates. **Results:** A total of 1,056 women contributed to this study. Of these, 189 were diagnosed with breast and 40 with endometrial cancers. An additional 268 women had BBC. For breast cancer, using the first algorithm we calculated a sensitivity of 95.2% and specificity of 96.4%. Application of the second algorithm yielded a sensitivity of 94.2% and specificity of 97.6%. For the third one, our calculations indicated 89.0% sensitivity and 97.9% specificity. Our analyses based on the same algorithms yielded similar sensitivity and specificity for endometrial cancer. For BBC, we calculated a sensitivity of 82.5% and specificity of 56.7% based on the first, 73.1% sensitivity and specificity of 71.9% for the second, and 38.8% sensitivity and specificity of 90.5% for the third algorithm. The average duration between diagnostic and administrative dates for incident breast, endometrial cancers and BBC was 0.65, 0.01, and 0.31 years, respectively. **Conclusions:** Our initial findings confirm the validity and potential utility of EHR for population-based cancer research. The algorithm of "2+ ICD-9 Coding System" yielded the most efficient process. The observed lower sensitivity and specificity for BBC potentially can be attributed to the wider pathologic spectrum of BBC. The relatively short duration between the EHR and diagnostic dates suggests unbiased interchangeability of dates.

Keywords: Observational Data (EHR); Validation of ICD-9 Diagnostic Code; Breast and Endometrial Cancer

doi:10.3121/cmr.2013.1176.ps2-43

PS2-51:

Comparison of Two Different Methods of Classifying Periodontal Disease Severity

Jay Fuehrer¹

¹Marshfield Clinic / Security Health Plan of Wisconsin

Background/Aims: Accurately determining the severity of periodontal disease has historically been a challenge in the dental practice. The object of this pilot effort was to compare two different methods of classifying periodontal disease severity based on provider-documented data among a random sample of dental patients at Marshfield Clinic. The Marshfield Clinic's electronic dental record provides an opportunity to collect specific data points for each tooth and its different surfaces as part of the periodontal charting conducted by the dental hygienist. The dental hygienists also classify the overall periodontal condition based on the American Dental Association (ADA) classification of gum diseases. **Methods:** A random sample of 50 dental patients with at least one comprehensive periodontal screening was used to conduct this retrospective study. Using the Marshfield Clinic's electronic dental record, the study compared how the dental hygienist coded the periodontal condition based on ADA classification with that of a computer-based algorithm that was developed to classify periodontal condition based on criteria as described by the American Academy of Periodontology (AAP) and the Centers for Disease Control and Prevention (CDC). We then mapped the periodontal disease severity classifications from both methods to investigate how many of the sample of 50 patients were similar. **Results:** Out of the 99,255 dental patients in our data warehouse, 67,487 (68%) have had a visit in our medical system at least once in the last 3 years. Periodontal measures have been done on 44% (29,573/67,487) of these patients that have medical and dental data available. When comparing the hygienists' classification of periodontal disease severity (using ADA guidelines) to an electronic algorithm using clinical attachment and probing depth to classify periodontal disease, there was a 72% (36/50) agreement. **Conclusions:** Although one limitation of this pilot study is that the purpose of ADA classification of periodontal condition (meant for clinical classification) differs from that of the joint CDC/AAP classification (meant for population-based surveillance), the results encourage utilizing measurable data points from the periodontal chart to help providers make accurate diagnosis of the periodontal condition.

Keywords: Dental; Periodontal Disease

doi:10.3121/cmr.2013.1176.ps2-51