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 phyician 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 Graham[1]; Colin Barton[1]

[1]Geisinger 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 Fuller[1]; Deborah Seger[1]

[1]Group 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 Fuller[1]; David Carrell[1]; Beverly Green[1]

[1]Group 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