Background/Aims: We assessed the effectiveness of eCare for Moods™ (patent pending), a patient-centered, highly interactive, web-delivered patient self-management and care management program on patients treated for recurrent or chronic depression in specialty psychiatry through a randomized clinical trial with two-year follow-up. **Methods:** Patients with recurrent or chronic depression were randomly assigned to eCare (N = 51) or usual specialty mental health care (N = 52). The 12-month eCare program provided patients with individualized self-monitoring, tailored patient education and training in depression self-management including relapse prevention. eCare was integrated with participants' ongoing depression care, linked to their electronic medical records. It provided clinicians with panel management and clinical decision support. Participants were interviewed at baseline and 6, 12, 18, and 24 months after enrollment. Telephone interviewers blind to treatment assignment used a timeline follow-back method to estimate depression severity on a 6-point scale for each of the 105 study weeks (including the baseline). Differences between groups in weekly severity over two years were examined by generalized estimating equations. **Results:** Participants in eCare experienced more reduction in depressive symptoms (estimate = -0.74 on the 6-point scale over two years; 95% confidence interval [CI] = -1.38 to -0.09, $P = .025$) and were less often depressed (-0.24 over two years; CI = -0.46 to -0.03, $P = .026$). At 24 months, 43% of eCare and 30% of usual-care participants were depression free; the number needed to treat to attain one additional depression-free participant was 8. eCare participants had other favorable outcomes: improved general mental health ($P = .002$), greater satisfaction with specialty care ($P = .003$) and with learning new coping skills ($P < .001$), and more confidence in managing depression ($P = .006$). **Conclusions:** Patient-centered, web-delivered care management improves outcomes in patients treated for recurrent and chronic depression. **Keywords:** Internet-delivered care; Depression

doi:10.3121/cmr.2014.1250.b1-4

C3-1:

Footprints in the Sand: Tracking Physician Work Efforts in Primary Care Using Access Logs in an Electronic Health Record

Ming Tai-Seale¹; Caroline Wilson¹; Cliff Olson²; Meg Durbin¹; Criss Morikawa¹; Hal Luft¹

¹Palo Alto Medical Foundation Research Institute; ²Palo Alto Medical Foundation for Healthcare, Research and Education

Background/Aims: Using EpicCare Electronic Health Record (EHR) data in a large multispecialty ambulatory delivery system, we explore a unique opportunity in which existing EHR data may offer clues on how clinicians use time, a scarce yet critical resource in health services delivery. Traditional means of studying physician time use during clinical encounters (e.g., direct observation) are costly and ignore pre-service and post-service work of physicians' services. The EpicCare EHR offers an alternative, unobtrusive portal to study time use through analysis of access logs. **Methods:** We used EHR access log data for one month in 2013 from 49 physicians in two primary care departments who cared for 22,174 patients in a large multispecialty ambulatory delivery system. Over 3 million EHR transactions are examined to explore individual physicians' style of time use on different tasks, as reflected by the access log. In-depth key informant interviews are used to complement the access log data on how physicians use the EHR and the activities that are more or less likely to be captured by the access log. **Results:** About 43.7% of physicians' total time for the month involved in-person face-to-face visits, 33.8% involved pre and post visit time, 11.4% telephone calls, 5.6% secure messaging to patients, 2.6% prescription refills, and 1.6% on orders for labs, medications or referrals. The earliest EHR access in the office occurred at 12:00 am and the latest logging out time in the office was at 11:59 am the following day. For each patient visit, an average of 16.7 minutes was logged in the exam room and 7.9 minutes logged outside of the exam room. **Conclusions:** The access log is a valuable tool for studying physician work efforts. Our findings highlight the significant amount of time clinicians spend outside of office visits. Unless there is a fixed ratio of in-office to total time, visit-centric FFS payment may undercompensate the significant efforts outside of visits. As "desktop medicine" (e.g., via phone, messaging) increases in the age of the Internet, smart phones, and EHRs, reforming provider payment mechanisms to

account for work outside of office visits is warranted.

Keywords: Physician work efforts; EHR access log

doi:10.3121/cmr.2014.1250.c3-1

C3-2:

All-In-One: How Group Health Organizes Clinical Text from Clarity for Research

Scott Halgrim¹; David Carrell¹; Diem-Thy Tran¹

¹Group Health Research Institute

Background/Aims: Group Health Cooperative runs Epic as its Electronic Medical Record. The clinical text stored in Clarity, Epic's relational reporting database, is valuable to research at Group Health Research Institute (GHRI). However, due to a number of factors, GHRI's access to this data was limited. These factors included: 1) a limited window during the day allotted to GHRI due to higher priority reports on Group Health's care delivery side, 2) Clarity splitting text notes into lines of about 5,000 characters, and 3) restrictions on managing the database itself, like adding a full-text index. We sought to make the clinical text more valuable for research by making it available at all times, combining all the content of a note into one record, and allowing for more database management options. **Methods:** We have developed a nightly Python process that moves clinical text from four Clarity tables into one full-text-indexed table on our own server. In addition, we store metadata about each note—including note type, encounter date, department, and provider—in a parallel table. **Results:** This conversion process, begun in 2010, converts about 60,000 notes per night and has converted every extant note in Group Health's Clarity database for a total of 123 million notes as of October 2013. The notes' availability has sped development of sophisticated NLP algorithms in the years since its inception. Another benefit is nightly automated status e-mails sent to the developers. When there was a recent import of several years of historical notes from legacy systems into Epic, GHRI knew immediately that a greater history of notes was available for research. **Conclusions:** The text store at GHRI has strengthened research and grant submissions. Due to Clarity's consistent data model and large footprint throughout the nation's medical community, the solution should be easily transferable to other sites wishing to realize the same advantages. The solution is amenable to enhancement as needs arise for more metadata or for clinical text from other parts of Clarity. Remaining challenges are tracking changes to notes in Clarity and improving performance.

Keywords: Natural language processing; Clarity

doi:10.3121/cmr.2014.1250.c3-2

C3-4:

An Algorithm to Combine Machine Learning and Structured Data to Automate De-identification of Clinical Text

Diem-Thy Tran¹; Scott Halgrim¹; David Carrell¹

¹Group Health Research Institute

Background/Aims: Clinical text is an important resource for research. To maintain patient privacy when researching this text, we use de-identification. The hiding in plain sight (HIPS) method is promising; it replaces personally identifiable information (PII) with realistic surrogates so any remaining real PII would be hard to distinguish from the fake information. However, there remain some challenges with HIPS, such as overlooked PII. We explored these challenges and hypothesized that we could find more PII by combining structured data with a machine learning algorithm. **Methods:** The machine learning de-identification software we used, developed by MITRE, is the MITRE Identification Scrubber Toolkit (MIST). Trained chart abstractors annotated Family Practice notes with the following PII types: address, age, date, provider name, email, IP address, consumer number, organization name, other id, phone, patient name, room id, social security number, and URL address. Structured data included in this experiment are patient's address, age, date of birth, email, phone, consumer number, social security number, the visit provider name, visit date, and visit location. We queried this data from Clarity, a relational reporting database for Group Health's electronic health record (EHR) system. Our first test experiment used MIST to train a model on 100 documents then tested on 10 notes. We reviewed the remaining PII and determined if they are available in the structured data.

Results: MIST's precision was 0.93 and recall was 0.77. MIST left 13 leaks and incorrectly identified two instances of blood pressure numbers as dates. We can reduce 3 out of 13 leaks using structured data obtained for that visit note. The remaining leaks are: 3 locations, 1 age which belonged to the patient's child, 1 visit date pre-EHR system, 3 pieces of historical visit data, and 2 mentions of 'sheriff' because the chart abstractor determined it to be an "other" PII type. **Conclusions:** Our results show that combining structured data with MIST can potentially improve de-identification of Group Health clinical text. Future work is to build an automated system that include historical visit data, use gender, race, and language to create more realistic surrogates and to evaluate the loss of important clinical information in de-identified documents.

Keywords: Natural language processing; De-identification
doi:10.3121/cmr.2014.1250.c3-4

PS1-45:

Get Smart: Finding Structured Data Using SmartTools in Clarity

Sharon Fuller¹

¹Group Health Research Institute

Background/Aims: Chart notes and other text blocks in Epic's electronic medical record system generally appear to be free text, requiring natural language processing (NLP) for analysis. In fact, though, many notes contain boilerplate text that can be traced back to structured data created using various Epic "SmartTools." We investigated Clarity, Epic's reporting database, to learn how to locate these structured data. **Methods:** Through trial and error we developed a data map to guide searches for structured data in note text. Starting with a known block of text, we found the underlying boilerplate, traced the sources of the individual data elements, and finally located the values recorded for those elements in the encounter. **Results:** Smart phrases, lists, links and forms are a few of the commonly used Smart elements in Epic. Each of these elements may be composed of other elements, increasing the difficulty of charting a straight path. So a SmartPhrase may include the option to choose from a SmartList, and those choices may, in turn, reference SmartLinks. However, if one can locate, for example, a definitive statement of pathology results within the structured Smart data of a results letter to the patient, it may be possible to avoid much of the expense and uncertainty involved in using NLP to parse the same information. Similarly, if some note text is known to be a direct copy of structured data from another part of the chart, one can safely exclude that text from further examination. Even if one does not analyze the structured data directly, discovering which text is "Smart" and the expected values for the included data elements can greatly facilitate the development of NLP algorithms. It is important to work with clinicians at this first level of discovery, to learn the mnemonics or "dot-phrases" they use to invoke SmartTools and the text that results. **Conclusions:** Epic's SmartTools hold great promise for teasing structured data out of apparently unstructured text. However, unravelling the complex interrelations among SmartTools requires a good deal of motivation and ingenuity. Further investigation is needed to determine the most appropriate and effective uses of these data.

Keywords: Clarity; Structured data
doi:10.3121/cmr.2014.1250.ps1-45

PS1-13:

Probabilistic Linkage (Also Known as "Fuzzy Matching"): the Theoretical Foundations of Modern Record Linkage

Christopher Mack¹

¹Group Health Research Institute

Background/Aims: Data joins have historically been done using deterministic linkage techniques which require exact data matches to succeed. For data extracted from disparate sources, requiring an exact match may be too limiting. Probability-based linkage techniques have gone from an arcane art involving expensive third-party software to having multiple free and commercial options available for use. We will walk through the theoretical underpinnings of the probability-based linkage theory used by many record linkage applications today. **Methods:** Describe the history of record linkage, the cause of linkage errors, the need for probability-based

techniques for match assessment and the mathematical framework of probabilistic linkage. Discuss matching best practices and limitations in matching ability and match assessment. **Results:** The audience will better understand how to choose linking variables, ways to limit the search space for comparisons, and the importance of pre-processing linkage variables, including comparison by permutation of variables, phonetic codes and edit-based measures. **Conclusions:** The audience will be able to better evaluate, and to use more successfully, the available freeware and commercial linkage software given a better understanding of their inner workings, limitations and potential pitfalls. An advanced audience will be in a better position to create their own probabilistic matching algorithms.

Keywords: Linkage; Matching

doi:10.3121/cmr.2014.1250.ps1-13

PS1-14:

Pregnancy Data in Clarity - Stay Tuned for Our Next Episode

Sharon Fuller¹

¹Group Health Research Institute

Background/Aims: For a condition that is so easily defined, pregnancy can be remarkably difficult to locate with any precision in existing structured data sources, especially while a woman is still pregnant. Researchers may require pregnancy information: (a) to avoid recruiting pregnant women to clinical studies; (b) to recruit women in a certain trimester or meeting certain pregnancy-related criteria; (c) to exclude measures taken during pregnancy; (d) to evaluate prenatal care across populations; (e) to provide conditional variables for other analyses, including neonatal outcomes. In 2004, Group Health (GH) adopted Epic's electronic medical record system. In this poster, we investigate Clarity, Epic's reporting database, to determine its usefulness as a source of structured pregnancy data. **Methods:** We examined all OB episodes of care at GH clinics that included any part of 2012. We also examined records of all completed visits in 2012 with an encounter type of OB Visit. **Results:** Almost all OB visits recorded in Epic at GH are part of an OB episode and pregnancy dating information is almost universally recorded. OB episodes assemble a large amount of useful data, with more variables anticipated following the recent implementation of Epic 2012 at GH. It remains to be seen how many of these will be populated, and if any will be backfilled from other sources. Although prenatal episode data are quite complete, parity, delivery and outcomes information at GH require more complex querying and result in less definitive data. Nor are episode data available for patients who obtain prenatal care outside of the GH network of clinics. Episodes may also be useful for examining the overall care picture for other health conditions. **Conclusions:** OB episodes in Clarity can be a rich source of structured data for current and past pregnancies, although implementations of Epic and Clarity at different sites may vary in the specific data elements captured.

Keywords: Pregnancy; Clarity

doi:10.3121/cmr.2014.1250.ps1-14

PS1-15:

Pre-filling Breast MRI Abstraction Forms Using Natural Language Processing

Hongyuan Gao¹; Karen Wernli¹

¹Group Health Research Institute

Background/Aims: Information in breast MRI reports is valuable for breast cancer research, 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 information and pre-fill abstraction form from free-text breast MRI reports. **Methods:** We identified 465 reports for women receiving breast MRI at Group Health between 2010-2012. We developed an NLP algorithm in SAS v9.2. The algorithm extracts information of reading radiologist, laterality, parenchymal enhancement, whether computer-aided technique is used, comparison exams, clinical indications and assessment from breast MRI reports. The NLP results are compared with manual abstraction from an experienced abstractor. **Results:** The algorithm correctly extracts reading radiologist, laterality and whether computer-aided technique for all 465 breast MRI reports, except 1