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