of this abstract is to share information, concerns, issues when switching to ICD10.

Keywords: ICD9 to ICD10; Clarity; Masterfile structural changes
doi:10.3121/cmr.2014.1250.ps1-5

PS1-6:

Big Data, Data Science and You--Demystifying Some Big Ideas

Roy Pardee¹

¹Group Health Research Institute

Background/Aims: Lately the popular press is rife with tantalizing references to coming advances brought by “Big Data” methods and software. Modern living--mobile and social computing particularly--emits enormous plumes of data. This data, we are told, can be analyzed in real-time to spot trends and yield important insights, to the benefit of business and mankind generally. The general idea of incidentally-produced data that can be exploited to produce valuable insights is one an HMORN audience is eminently comfortable with--it describes quite a lot of our research. But where do we fit in with these new trends? Have we all been “Data Scientists” doing “Big Data” for years now, and the rest of the world is just now catching up to us? Or, are these things really different and new? Are there things we should be appropriating from this “new” field to make our own work stronger? **Methods:** The proposed talk will describe and define several commonly-cited ideas and methods--to wit: (a) big data; (b) map/reduce; (c) no SQL; and (d) data science. **Results:** The talk will locate these in a larger technological context, list synonyms and closely related technologies and describe situations where expanding into less-familiar tools may well bear fruit for research data projects. **Conclusions:** While much of the tools and methods of big data are squarely addressed to problems we don't frequently encounter in HMORN research, it is good to have a conceptual understanding of them so that when we do hit the limits of conventional methods and resources we have “someplace to go” before giving a project up as infeasible. **Keywords:** Big data; Data science

doi:10.3121/cmr.2014.1250.ps1-6

PS1-7:

An Automated Metadata-driven Extract, Transformation and Load (ETL) Process: Using Code and Data to Generate Code

Michael Allison¹; Shannon Leitch¹; John Hall¹

¹Kaiser Permanente Northwest

Background/Aims: A stable, automated ETL environment is essential to maintain a consistent virtual data warehouse (VDW) refresh schedule. This stability is threatened at each refresh cycle because VDW source table schemas can vary over time, causing extract programs to fail when they encounter a schema that is different from what is expected. Such a failure requires manual intervention to change the extract program to conform to the new schema, disrupting the refresh schedule. **Methods:** We are initiating a process to dynamically modify extract code when source table schemas change, through a real-time comparison of current source table schemas against previously extracted table schemas. The process requires a metadata database (MDDb) containing schema metadata from all the sources, as well as metadata about our ETL packages. The first step in a refresh cycle is to acquire the current source schema metadata in real time. Then, using a combination of T-SQL and .Net code, we compare the current schema metadata against the metadata from the prior extract to identify source table changes. If there are discrepancies, the metadata about our extract packages provides the means to identify those packages of extract code requiring modification. We have created a maintenance program that uses the MDDb information to identify those packages requiring update, generate the needed data definition language (DDL) code segment changes, and dynamically modify the data flows within those extract packages. This is essentially application code that uses metadata to generate and modify other application code. Built into this process are logging, notification, historical archiving, and other documentation for process assurance. **Results:** This methodology allows us to generate corrections and changes to ETL packages quickly and easily without manual intervention. Since all tables are stored in the MDDb, the extract program can be constructed according to the schema specifications

found there. It then can be dynamically updated should the schema change, saving valuable programming time in developing new extract packages for data mining/exploration. **Conclusions:** This approach saves both programming time and computer resources, and allows us to execute more frequent ETL refresh cycles.

Keywords: Extract, transfer, load (ETL); Virtual data warehouse
doi:10.3121/cmr.2014.1250.ps1-7

PS1-8:

EMR v. Insurance Type Data in the VDW across HMORN Sites

Amy Butani¹

¹HealthPartners

Background/Aims: To describe the various organization models across the HMORN and the resulting data. **Methods:** Each HMORN site has patients and/or insurance members. The data available for these two populations are different; the differences will be demonstrated. **Results:** The VDW has been designed to accommodate data for both member and patient populations. **Conclusions:** Understanding the data available for each population will help in designing studies, especially multi-site studies.

Keywords: Virtual data warehouse; Data
doi:10.3121/cmr.2014.1250.ps1-8

PS1-9:

Data Issues and Meta-content to Improve Analysis of Electronic Health Data

Andrew Sterrett¹; Christina Clarke¹; Carsie Nyirenda¹; Komal Narwaney¹; Glenn Goodrich¹; Kris Wain¹

¹Kaiser Permanente Colorado

Background/Aims: KPCO continuously checks our electronic health data for unexpected values. These values may be errors or legitimate reflections of clinical procedures and data collection rules. Rigorous data analysis requires an understanding of these unexpected values as well as methods to address them. We do not yet have an organized way to document unexpected data values. **Methods:** We followed a multi-phase plan to develop a database that describes data issues from HMORN data partners. In Phase 1 (Requirements), we met with stakeholders at KPCO to determine needs. For Phase 2 (Design), we drafted a data collection survey and built an initial prototype of the database. In Phase 3 (Development), we improved upon multiple versions of the database and documented procedures to enter new data items, create reports, and review the accuracy of existing data items. When the database framework was sufficiently developed, members of the Analytic Team at KPCO entered issues based on their experience with HMORN studies. In Phase 4 (Testing), we evaluated our documentation and ran sample reports against the new data content. We also assessed the degree of overlap with the VDW Issue Tracker. During Phase 5 (Implementation), we presented the database to key stakeholders at KPCO and began discussions to build similar databases at other HMORN sites and/or enhance the Issue Tracker as a multi-site centralized database. **Results:** The Analytic Team at KPCO recorded a wide range of data issues occurring in projects from 2010-2013. Data items came from general VDW quality assessments as well as study-specific data checks. Database content included data errors as well as real phenomena attributable to clinical practice or data collection. **Conclusions:** Analysts can better anticipate problems and nuances in electronic health data with a database that describes the past data anomalies and methods for using the data.

Keywords: Data; Meta-content
doi:10.3121/cmr.2014.1250.ps1-9