Health Plan ICD-10 implementation team in the development of our project plans. The standard source to be used for mapping at the three Kaiser Research centers is the General Equivalence Mappings based Crosswalk Query Tool, which Kaiser is implementing to support transition mapping needs. **Results:** The CHR ICD-10 remediation team has completed the Framing and Planning phases including identifying programs/tables affected, sharing remediation solutions across the Kaiser regional research teams, and meeting with the Health Plan to coordinate efforts. We have developed a project remediation checklist and will report on the development of validation testing plans and the refinement of vanguard project code mappings. **Conclusions:** In the transition to ICD-10, it is essential to have a well-formed plan and to identify and include all stakeholders in the process. It is also beneficial to share tools and ideas between research centers to improve the overall methods.

Keywords: ICD-10; System-level changes

doi:10.3121/cmr.2014.1250.a4-2

A4-3:

Utilization-based Proxy Enrollment Versus Standard HMORN VDW Enrollment: A Pilot Validation Study

Irina Haller[1]; Brian Johnson[1]; Karen Riedlinger[2]; Pinky Barua[3]; Terese DeFor[4]; Paul Hitz[1]; Roy Pardee[5]; David Tabano[6]

[1]Essentia Institute of Rural Health; [2]Kaiser Permanente Northwest; [3]Geisinger Health System; [4]HealthPartners; [5]Group Health Research Institute; [6]Kaiser Permanente Colorado

Background/Aims: Most HMORN members offer both insurance coverage and health care. Health plan enrollment provides a well-defined population denominator for HMORN-based research. The HMORN common data model combines electronic data, routinely collected in health care delivery or claims processing, into the Virtual Data Warehouse (VDW). Sites with health plans capture insurance enrollment in the VDW enrollment file. Recently HMORN included sites that deliver health care without offering insurance. Since health plan membership is unknown, population denominators must be determined using alternate methods. This study validated a utilization-based proxy enrollment (PE) using standard VDW enrollment (SE) in 5 HMORN sites with Epic EHR and SE files. **Methods:** The utilization-based algorithm defines PE start at first of two non-ancillary health care visits (at least one being a primary care visit) separated by at least 90 days. PE ends at death or at the last qualifying visit if there is no utilization in the following 3-year period. PE files were built at each site by applying utilization-based algorithm to base tables created from Clarity. The PE and SE extracts included study ID, age, gender and the start/end of enrollment periods between 2000 and 2012, based on availability of Clarity data at the sites. The agreement between PE and SE was evaluated using differences in start/end of the first enrollment periods (days). **Results:** The differences between PE and SE starts varied by site with greater variation in children (<20 years) and older adults (64+ years). The differences varied by gender, but the differences between the genders were smaller in the young and the old. The differences were larger for males between ages of ~16 and 64, indicating less utilization by males than by females. The differences between PE and SE ends were generally negative across ages, indicating that PE extended beyond SE. Gender differences between PE and SE starts were similar to differences between PE and SE ends. These empirical results were confirmed by multivariate regression modeling. **Conclusions:** Agreement between PE and SE could be improved using additional parameters, as well as possible adjustment of the time lag for PE end in the utilization-based algorithm.

Keywords: Population denominator; Utilization-based enrollment

doi:10.3121/cmr.2014.1250.a4-3

A4-4:

Do Projects Have Complete Data Capture for Their Study Populations?

Donald Bachman[1]; Mark Hornbrook [1]; Debra Ritzwoller[2]; Roy Pardee[3]; Alan Bauck[1]

[1]Kaiser Permanente Northwest; [2]Kaiser Permanente Colorado; [3]Group Health Research Institute

Background/Aims: Investigators often assume that when patients are enrolled in a health plan, they have complete capture of utilization data from their health plan sources. This assumption may not always be true as patients have many incentives to choose care at multiple settings (convenience, price, residence, insurance type, drug coverage), which can give rise to missing data. Complete capture of medical data for population-based research is crucial to our ability to identify populations who have not had particular exposures or outcomes. In case-control designs, where the control group does not have a specified exposure, we measure this condition by absence of data. To address this problem, the VDW enrollment work group created a new enrollment variable, called "Outside_utilization", designed to identify members suspected of incomplete capture of encounters or pharmacy fills. This work reports on a quality assurance analysis of the extent and nature of the data gaps at different sites. **Methods:** For V3, HMORN sites added new VDW variables in their enrollment file including the "Outside_utilization" variable. This variable identifies populations suspected of having incomplete health care utilization capture. Since the reasons for incomplete data capture vary among the sites, the methods for identifying members with incomplete data capture were determined by the local site data managers. The authors distributed a program that computed utilization rates for specific cohorts for "complete" and "incomplete" data capture populations. We also computed rates for those on high deductible plans. We compared differences in these rates by year and site. In addition, we conducted a survey to determine how the incomplete populations were identified for each participating site. **Results:** Some sites clearly identify populations that have incomplete capture of data using the "Outside_utilization" variable. At other sites, the difference in rates is less apparent. The sites that can distinguish patients with incomplete versus complete data have certain common definitions used when designing the "Outside_utilization" variable. **Conclusions:** The "Outside_utilization" variable identifies populations with incomplete data capture at some sites. We recommend that projects use this variable to exclude populations suspected of incomplete data capture when computing population-based utilization rates.

Keywords: Data Population

doi:10.3121/cmr.2014.1250.a4-4

PS1-5:

Converting to ICD-10 Clarity Structures

Jamila Gul[1]; Srivardhan Chimmula[1]

[1]Kaiser Permanente Northern California

Background/Aims: The transition to the ICD-10 code represents a huge undertaking for every Health Care Organization including Analytic teams. In Kaiser Northern California we have effectively engaged KP HealthConnect Application Coordinators, Business Partners, and Regional IT Leads to create and build ICD-10 impact awareness through education and leveraging our partner's vast knowledge to identify critical areas of impact, issues, risks, assumptions and high level business objectives to support a successful transition to the ICD-10. **Methods:** To accommodate ICD-10, Epic has made significant changes to Clarity in their 2010 release including adding new fields, tables and views. These changes are designed to aid in reporting appropriate codes as well as allowing for historical reporting and trending. The new format of the codes will also be part of the changes to Clarity. Changes have been made to three main Epic Master Files: (1) Diagnosis masterfile - maintains a master set of diagnosis records. These diagnosis records can be industry-standard ICD codes used for billing or clinical terms used for clinical encounter documentation, but mapped to industry-standard ICD codes. The diagnosis code changes primarily impact clinical encounter documentation (problem list and a Dx association with encounters and orders), HIM coding, claims for all types of encounters. (2) ICD Procedure masterfile – maintains a master set of ICD hospital procedure records. ICD procedure codes in addition to supporting hospital inpatient billing. (3) Surgical Procedure masterfile - maintains a master set of records for each surgical procedure that can be performed with an operating room facility. It also contains procedure preference records, primarily used for scheduling surgical procedures. We would like to take a real study and switch the codes from ICD9 to ICD10. **Results:** By sharing a real life example and sample codes we are hoping to pass our learning and knowledge to others hoping to help eliminate some of the errors that we made. Also, discuss outstanding challenges and issues that Epic has not resolved. **Conclusions:** The purpose

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 Pardee[1]

[1]Group 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 Allison[1]; Shannon Leitch[1]; John Hall[1]

[1]Kaiser 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 Butani[1]

[1]HealthPartners

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 Sterrett[1]; Christina Clarke[1]; Carsie Nyirenda[1]; Komal Narwaney[1]; Glenn Goodrich[1]; Kris Wain[1]

[1]Kaiser 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