

Abstract PS2-37

An N-Gram Approach to Predicting Health Services and Generating Realistic Simulated Data

Tyler R. Ross, MA, *Center for Health Studies, Group Health Cooperative*;
Roy E. Pardee, JD, MA, *Center for Health Studies, Group Health Cooperative*

Background/Aims: N-Grams are an established statistical tool for use in processing streams of tokens found in nature. "Tokens" can be words in speech or a document (as in natural language processing applications), nucleotide sequences (for genetic applications) or, we will argue, health service events (HSEs) such as diagnoses, procedures, and pharmacy fills occurring over time. We propose applying N-gram analysis to streams of HSEs observed in HMO administrative data, both as a method for analyzing actual health services, and also for generating realistic simulated data that are not subject to privacy or IRB concerns. Such simulated data could be used for applications where using real data would be too risky, or for developing code prior to actual IRB approval. **Methods:** N-grams are based on counting token co-occurrence in a reference dataset. The n=2 case is easiest to explain. By counting the number of times each token X is preceded by token Y in the reference dataset, a matrix of conditional frequencies is generated. This matrix is then normalized and smoothed, resulting in a matrix of conditional probabilities that any given X is preceded by Y. **Results:** The resulting matrix will be interesting to study in its own right. While many of the high-probability co-occurrences will be obvious and well-known, there certainly will be some (particularly those where the time intervening between events X and Y is long) that are not, and that will bear investigation. Thus, the matrix is a tool for hypothesis generation. Further, the matrix can be used to generate realistic streams of simulated HSEs. A starting token is chosen at random, using the matrix probabilities as weights (e.g., if a checkup visit occurs twice as frequently as a starting HSE than a flu diagnosis, then the checkup visit would be approximately twice as likely to be chosen for a starting token). Once the starting token is chosen, a next token is randomly drawn from a weighted sample based on the starting token, and so on, each token providing the sample weights for the choice of the next token. In this way, any amount of data can be manufactured. **Conclusions:** Applying N-grams from HSE streams will offer unique opportunities to investigate event co-occurrence, and provide simulated data useful for testing purposes.

Abstract PS2-38

Formalization of the Laboratory Result Content Area of the VDW

Gwyn Saylor, BA, *Institute for Health Research, Kaiser Permanente Colorado*; Jennifer L. Ellis, MSPH, *Institute for Health Research, Kaiser Permanente Colorado*; Marsha A. Raebel, Pharm D, *Institute for Health Research, Kaiser Permanente Colorado*

Background: The availability of comparable laboratory (lab) results across sites will expand the abilities of the HMO Research Network (HMORN) to perform collaborative research. The Coordinated Clinical Studies Network (CCSN) funded research to define the process of adding a lab result content area to the virtual data warehouse (VDW). We describe the development efforts and implementation steps required to create and maintain a laboratory result content area that is equivalent in definition across HMORN sites. **Methods:** Existing information about lab data and coding systems at HMORN sites was reviewed, and we surveyed sites about current lab data. Sites were specifically asked about the availability of Logical Observation Identifiers Names and Codes (LOINC), which are a set of universal names and ID codes for identifying lab and clinical test results. To enable sites to participate in lab research studies, written instructions for incorporating lab tests, including a data dictionary, were prepared. The accuracy of the instructions was assessed for selected tests: serum creatinine and potassium, international normalization ratio, glycosylated hemoglobin, and fasting blood glucose. When a lab test is to be added to the laboratory content area, the lead site determines a list of LOINC codes that are interchangeable at a component level for that lab test. Sites that have incorporated LOINC into their lab database use this list to pull data. Each site without LOINC and each site where data are needed for time periods prior to LOINC availability will determine a list of local codes. Data pulled by LOINC and data pulled by local codes are tied together by a common test type. **Results:** Most HMORN sites indicated they had the LOINC code set available. Kaiser Permanente

Colorado is working with several other HMORN sites on hypertension studies. These sites are beginning to implement this process, utilizing LOINC and local codes from 2002 forward. This will be the first large-scale test of the lab component of the VDW. **Conclusions:** Implementing the lab content area requires upfront effort from the lead site for each lab test. Each participating site must also invest time determining its local codes. The aggregate information from all sites participating in a study will enhance the usefulness of the VDW to epidemiological, observational, and interventional research in the HMORN and in other collaborative efforts.

Abstract PS2-41

Tumor Registry Content Area of the Virtual Data Warehouse

Karen E. Wells, BA, *Henry Ford Health System/Department of Biostatistics and Research Epidemiology*; Richard Krajenta, BS, *Henry Ford Health System/Department of Biostatistics and Research Epidemiology*; Lois Lamerato, PhD, *Henry Ford Health System/Department of Biostatistics and Research Epidemiology*; Donald J. Bachman, MS, *The Center for Health Research, Kaiser Permanente Northwest, Portland, OR*

Background: The virtual data warehouse (VDW) was created as a mechanism to produce comparable data across health systems for the purpose of facilitating multi-site research within the Cancer Research Network (CRN). Each site maintains its own standardized files according to mutually agreed upon dataset definitions. The common structure of the VDW files enables a site programmer to distribute a SAS program to all participating sites that, with minimal modifications, can run against their local VDW. The program produces de-identified summary results that can be transferred to the coordinating site programmer. The Tumor content area is just one of eight standardized files maintained at each of the local CRN sites with tumor registries or access to tumor registry data. **Methods:** The Tumor content area was one of the first to be developed by the Scientific & Data Resources Core. The common format for each element of the Tumor content area, variable name, label extended definitions, code values and value labels, was largely driven by the data standards as defined by the North American Association of Central Cancer Registries. The Tumor content area contains detailed information on patient demographics and incident primary tumors such as date of diagnosis, ICD-O site, morphology, stage at diagnosis and first course of treatment. Eleven CRN sites maintain a Tumor content area for the VDW. **Results:** The Tumor content area has been used for case identification in several funded multi-site CRN studies. Additionally, the Cancer Counter on the CRN website uses aggregate data from the Tumor content areas of all CRN sites to provide counts of primary tumors for every combination of selected variables. The Cancer Counter has proven to be a valuable tool in facilitating proposal development. **Conclusions:** The standardization of the Tumor content area across CRN sites enables the sharing of compatible data in multi-site studies by improving programming efficiency, accuracy and completeness of data.

Abstracts received on September 11, 2008

doi:10.3121/cmr.2008.849