Health Informatics

C-C1-01:

Strategic Plan for Developing Natural Language Processing Capacity at HMORN Sites

David Carrell, PhD, *Group Health, Center for Health Studies*; Virginia Immanuel, *Group Health, Center for Health Studies*

For the specific purpose of crafting a strategic plan for developing scientific and technical capacity for applying natural language processing (NLP) techniques to clinical text in HMO Research Network (HMORN) sites we convened a working conference in Seattle, December 17–18, 2008. Expert consultation in the strategic planning process was provided by six NLP experts from three institutions considered national leaders in the development and application of NLP technologies: the Mayo Clinic, the University of Pittsburgh Medical Center, and Vanderbilt University. Conference participants included representatives of five HMORN sites, Northwestern University, and the University of Washington. Specific aims of the conference were: 1) To develop a strategic plan for acquiring natural language processing capacity including, potentially, development of in-house capacity, subcontracting NLP tasks to one or more external collaborators/contractors, some combination of these approaches, or none at all; and 2) To educate investigators and technical staff about the processes of developing, testing, and applying NLP algorithms for extracting data from clinical text in the context of typical HMORN-style research programs. Traditional conference presentations addressed "What is natural language processing," "Illustrations of potential NLP projects," "Considerations when developing and applying NLP algorithms," and "Comprehensive, large-scale information extraction systems." A round-table-style discussion sessions focused on issues of human capital requirements for NLP projects, subcontracting NLP tasks to outside entities, multi-site collaboration on NLP projects, and a strategic plan for developing NLP capacity. This presentation summarizes the strategic planning products of the conference, currently in draft form and undergoing edits by a working group of the expert consultants and conference participants. Preliminarily, the strategic plan recommends incremental development of NLP capacity in HMORN-like sites through sequential execution of thoughtfully-selected small-scale NLP and NLP-like projects that represent key technical aspects of what can become a mature program of NLP capacity within 3–5 years. Initially, small-scale projects would be conducted in close collaboration with outside informatics experts to provide scientific and technical guidance to the immediate project and to assure fidelity with a longer-term locally-adapted strategic plan.

C-C1-02:

Data Extraction From Text, Step 1: Preparing Text for Machine Processing

David Carrell, PhD, *Group Health, Center for Health Studies*

Background: Natural language processing (NLP) uses software to assist in the extraction of information from clinical text, a process usually performed entirely by chart abstractors. Before NLP can be applied the text in question must be prepared for machine processing. In research settings this pre-processing work often involves several successive and related tasks, requiring substantial amounts of time and attention from people representing various types of clinical, scientific and technical expertise. Appreciating the tasks and participants involved in pre-processing clinical text can make the work more manageable, efficient, and effective. **Methods:** The information presented here comes from case study analyses of three small-scale projects involving preparation of clinical text (pathology reports, radiology reports, and progress notes) for processing by the Cancer Text Information Extraction System. Supplementing these experiences is information from anecdotal conversations with natural language processing experts. **Results:** Ten separate pre-processing tasks were identified: 1) obtaining source feeds, 2) assessing completeness, 3) de-duplication, 4) universe description, 5) cleaning and formatting, 6) de-identification, 7) database loading, 8) sampling, 9) preparation of the NLP system input feed, and 10) quality assurance. Nine types of expertise or task participants required for pre-processing were identified: 1) IRB representative, 2) source-system manager, 3) network/dbase administrator, 4) programmer, 5) statistician, 6) investigator, 7) informaticist, 8) clinical domain expert, and 9) manual chart abstractor.

Conclusions: Pre-processing clinical text is an important phase and potentially challenging aspect of extracting information from clinical text using NLP. Because researchers require accurate information about the larger universe of documents or patients represented by the sampled and processed text, pre-processing can present numerous challenges, the solutions to which draw on many areas of expertise in a multi-step and sometimes iterative process.

C-C1-03:

Identifying Respiratory-Related Clinical Conditions From ED Reports With Topaz

Wendy W Chapman, PhD, *University of Pittsburgh*; Henk Harkema, PhD, *University of Pittsburgh*

Background: Identifying patients for research studies often requires timely and expensive manual chart review. Automated searching on textual records results in many false positive cases, because a large proportion of clinical conditions described in reports occurred in the past history (past history of pneumonia), were experienced by a relative and not the patient (father suffered from CHF), or are described as being absent (patient denies chest pain). Natural Language Processing (NLP) promises to automatically encode clinical conditions and modifiers necessary for accurate identification of patients matching a given case definition. **Methods:** We developed an NLP system called Topaz that currently identifies 55 respiratory related clinical conditions from clinical reports, assigning the value acute, chronic, or absent to each of the conditions in a three-step process. Module 1: Find instances of the conditions in the text. Module 2: For each condition marked by Topaz, assign the following values: Existence (present, absent); Temporality (recent, historical, hypothetical); Experiencer (patient, other). Module 3: Integrate information from individual annotations to assign a single value to each of the 55 conditions. **Results:** We evaluated each module on a set of 60 emergency department reports by comparing against physician annotations. Sensitivity and positive predictive value for Modules 1 and 2 are as follows: Module 1: Mark conditions: 99%, 89%; Module 2: Negation (98%, 96%), Historical (76%, 50%), Hypothetical (100%, 81%), Experiencer (100%, 33%). Weighted kappa score between values assigned to the 55 conditions by Topaz (Module 3) compared to physician annotations was 0.85, indicating high agreement between Topaz and a physician. **Conclusion:** In spite of mistakes in marking conditions and assigning contextual features, Topaz performed similarly to a physician in assigning 55 conditions the values acute, chronic, or absent based on the ED report. Topaz could improve sensitivity of case retrieval by addressing linguistic variation common in clinical reports (e.g., synonymy, abbreviations, etc). Topaz could also improve precision of case retrieval by addressing temporality, experiencer, and negation of clinical conditions.

C-C1-04:

Building a Health Services Information Technology Research Environment

David W Gehrum, BS, *Geisinger Information Technology*; JB Jones, PhD, MBA, *Geisinger Center for Health Research*; Gregory J Romania, MS, *Geisinger Information Security Office*; David L Young, BS, *Geisinger Information Technology*; Virginia R Lerch, MPH, *Geisinger Center for Health Research*; Christa A Bruce, MPH, *Geisinger Center for Health Research*; Diane Donkochik, BA, *Geisinger Information Technology*; Walter F Stewart, PhD, MPH, *Geisinger Center for Health Research*

Background: The electronic health record (EHR) has opened a new era for health services research (HSR) where information technology (IT) is used to re-engineer care processes. While the EHR provides one means of advancing novel solutions, a promising strategy is to develop tools (e.g., online questionnaires, visual display tools, decision support) distinct from, but which interact with, the EHR. Development of such software tools outside the EHR offers an advantage in flexibility, sophistication, and ultimately in portability to other settings. However, institutional IT departments have an imperative to protect patient data and to standardize IT processes to ensure system-level security and support traditional business needs. Such imperatives usually present formidable process barriers to testing novel software solutions. We describe how, in collaboration with our IT department, we are creating an environment and a process that allows for routine and rapid

testing of novel software solutions. **Methods:** We convened a working group consisting of IT and research personnel with expertise in information security, database design/management, web design, EHR programming, and health services research. The working group was tasked with developing a research IT environment to accomplish two objectives: 1) maintain network/data security and regulatory compliance; 2) allow researchers working with external vendors to rapidly prototype and, in a clinical setting, test web-based tools. **Results:** Two parallel solutions, one focused on hardware, the second on oversight and management, were developed. First, we concluded that three separate, staged development environments were required to allow external vendor access for testing software and for transitioning software to be used in a clinic. In parallel, the extant oversight process for approving/managing access to internal/external personnel had to be altered to reflect the scope and scale of discrete research projects, as opposed to an enterprise-level approach to IT management. **Conclusions:** Innovation in health services software development requires a flexible, scalable IT environment adapted to the unique objectives of a HSR software development model. In our experience, implementing the hardware solution is less challenging than the cultural change required to implement such a model and the modifications to administrative and oversight processes to sustain an environment for rapid product development and testing.

C-D3-01:

Developing and Implementation of Secure Linear Regression on Distributed Databases

Fang Zhang, PhD, *Harvard Pilgrim Health Care*; Lingling Li, ScD, *Harvard Pilgrim Health Care*; Ken Kleinman, ScD; Jeff Brown, PhD, *Harvard Pilgrim Health Care*; Kimberly Lane, MPH, *Harvard Pilgrim Health Care*; Richard Platt, MD, MS, *Harvard Pilgrim Health Care*

Background/Aims: Improving access to multiple sources of electronic health information is crucial for population-based health care research. But it also raises serious issues relating to data confidentiality and security. We propose to develop methods and tools (e.g., SAS macros) to enable multivariate analyses to be conducted using multiple datasets that are stored separately without exposing any patient level information. Phase I of this project focuses on linear regression. **Methods:** A two-step process is carried out by SAS macros. The macros will be installed on both the central and site computers that hold the analytic datasets. After the central computer initiates the process, each site computer executes the SAS program locally and submits intermediate summarized statistical results to the central computer. The central computer then combines the intermediate results, and computes the parameter estimates. A second round of program execution delivers the variance/covariance estimates of the parameter estimates. **Results and Conclusions:** Simulated separated data sets were used to test the macros. The results were identical to the outputs from the SAS PROC REG procedure when the data sets were combined. The macro enables remote execution of multivariate linear regression analysis within the context of a multi-institutional distributed research network. The successful implementation of the secure statistical analysis within a distributed database model will overcome one of the largest hurdles in fully developing such a model, namely, the current need to combine data for multivariate analyses. Obviating this need will save the time and resources currently spent developing data-sharing agreements and will greatly minimize the patient confidentiality and proprietary concerns

C-D3-02:

Syndromic Surveillance Reporting Via CDC's Public Health Information Grid

Michael Klompas MD, MPH, *Harvard Pilgrim Health Care*; Ross Lazarus, MBBCh, MPH, *Harvard Pilgrim Health Care*; Gillian Haney, MPH, *Harvard Pilgrim Health Care*; Katherine Yih, PhD, *Harvard Pilgrim Health Care*; Richard Platt, MD, MSc, *Harvard Pilgrim Health Care*

Background: CDC is building a public health information grid to connect public health stakeholders nationwide. The grid will enable efficient and secure sharing of data and applications between public health authorities at the national, state, and local levels as well as researchers and other interested parties (JAMIA 2008;15:705). The grid may also support federated analysis of distributed data sources to detect clusters that might be invisible to smaller, isolated systems. Grid

protocols allow data owners to control data access. Broad sharing of data, however, requires a model to customize the level of patient identifiers seen by users depending upon their role. We describe here a work-in-progress implementation of an ambulatory syndromic surveillance module on the public health grid, including a model for sharing reports with different levels of granularity depending upon user privileges. **Methods:** We are installing a grid node at Atrius Health, a large multisite, multispecialty ambulatory medical practice serving over 600,000 patients predominantly in eastern Massachusetts. Standard CDC-defined ICD-9-and temperature-based syndrome algorithms developed for the National Bioterrorism Demonstration Project (MMWR 2004;53 Suppl:43) will be applied to ambulatory encounter data gathered nightly by the Electronic medical record Support for Public Health system (ESP, esphealth.org). ESP consists of a dedicated server populated with nightly extracts of comprehensive encounter data from the electronic medical record system of Atrius Health (JAMIA 2009;16:18). ESP currently analyzes this data to submit fully-identifiable case reports on patients with notifiable conditions to the state health department (MMWR 2008; 57:373). The syndromic surveillance addition to ESP will modify ESP's reporting module to vary the level of report granularity according to users' access privileges. ESP reports will include fully de-identified aggregate counts stratified by time period and zip code for basic grid users; semi-identified unit records with age, gender, and geo-code for intermediate users; and fully identifiable unit records with patient names and addresses for highly privileged users such as the state health department. **Results:** Installation is currently underway. **Conclusions:** Controlled distribution of syndromic surveillance data through the CDC's nascent grid will help develop the emerging paradigm for sharing public health information and applications.

C-D3-03:

Developing a Distributed Research Network to Conduct Population-based Studies and Safety Surveillance

John H Holmes, PhD, FACMI, *University of Pennsylvania School of Medicine for the HMO Research Network CERT DeCIDE Center Distributed Research Network project team*

Introduction: The overall objective of this project is to design a scalable distributed research network (DRN) that supports secure data analyses for research on the comparative effectiveness and safety of therapeutics. The plan for achieving this objective includes designing an information system architecture, implementation of a prototype of the specified system architecture, and completion of a proof of concept pilot project conducted in partnership with the Agency for Healthcare Research and Quality (AHRQ). Methods already implemented by the HMORN provide the foundation for the development of this DRN. These include established procedures for: 1) use of decentralized data files that remain in possession of the covered entities; 2) governance methods emphasizing local control and decision-making about participation in research projects; 3) transparency in choosing projects, protocol development, and communication of results; and 4) best practices for preserving privacy, ensuring data integrity, and accurate data interpretation. **DRN System Architecture:** The DRN architecture is being modeled as a federation of separate, heterogeneous databases that will be accessible through a common exchange model, implemented as a common query system (CQS). Under the common exchange model, heterogeneous data and systems do not need to be physically integrated, but the model affords interoperability, such that users can access these data resources using a single interface using a common syntax and semantic representation of data elements and coding schemes. This is a distinct advantage for the DRN, since there will be different data resource providers, some of which place strict constraints on the use of data. The common exchange model obviates the need for users to learn the syntactic and semantic characteristics of the resources available across the entire DRN. **Prototype Development:** The DRN is being developed using a rapid prototype life-cycle. This user centered approach employs incrementally-developed system prototypes such that all participants become stakeholders in the development process and have access to increasingly complex, concrete representations of the final system.