A large scale, cross-disease family health history data set
Hong Yu, George Hripcsak, MD
Department of Medical Informatics, Columbia University

Introduction: A family health history data set need to be evaluated before applying to the study of genetic diseases, genetic counseling, and epidemiological studies. We have obtained a large scale, cross-disease family health history data set (FhhDS) from electronic discharge summaries at Columbia Presbyterian Medical Center by using a pattern matching parser we have developed. Currently, FhhDS contains 22,292 patients' family health history. Here we have evaluated FhhDS by examining the scale, the coverage, and the completeness of the data set.

Background: The Columbia Presbyterian Medical Center (CPMC) has the Columbia Clinical Repository that contains many electronic discharge summaries. All electronic discharge summaries are inpatient and obtained by physician dictation. We have developed and evaluated a pattern matching parser--Family Health History Extraction Processor (FhhEP) that extracts family health history from narrative electronic discharge summaries and builds a database-ready output. The output is the health history findings containing ten attributes: unique identification number of the discharge summary, family member, number of the family members, disease, trait, presence or absence of the disease, certainty, date of developing disease or death, and location of the disease. From the unique identification number of the discharge summary, we can retrieve the patient's MRN and discharge date. Five physicians have evaluated the natural language processing parser. The majority physician opinion is the reference standard. Sensitivity is the number of positive answers in common between the parser findings and the reference standard divided by the number of positive answers in the reference standard. Precision is the number of positive answers in common between the parser findings and the reference standard divided by the total number of parser findings. The sensitivities in the presence of all ten attributes, five attributes (family member, number of family member, disease, trait, and presence of disease), four attributes (family member, disease, trait, presence of disease), and three attributes (disease, trait, presence of disease) are 75.6%, 85.3%, 85.7%, and 93.0%, respectively. The precision in the presence of all the attributes, five attributes (family member, number of family member, disease, trait, and presence of disease), four attributes (family member, disease, trait, presence of disease), and three attributes (disease, trait, presence of disease) are 72.0%, 72.3%, 72.3%, and 79.5%, respectively.

Methods: We obtained the FhhDS by running the FhhEP through all the electronic discharge summaries at CPMC, one year at a time, from 1992 to 1998. We measured the scale and the coverage of the data set by measuring the distribution of diseases (in ICD-9 code) associated with the number of patients in FhhDS. In order to identify whether the patients in FhhDS represent a normal inpatient population, we measured the distribution of the numbers of discharges, electronic discharge summaries, and electronic discharge summaries containing family history by year from 1992 to 1998. We defined a complete family health history as the one contains a complete set of blood-related family members for at least two generations and a complete set of diseases associated with each family member. We evaluate the completeness of FhhDS by comparing two or more discharge summaries that belong to the same patient. The rationale is that if a patient's family health history is complete in every discharge summary, all discharge summaries belonging to the patient should contain the same disease information. If two or more than two discharge summaries cover different diseases, then the discharge summaries are not complete. We thus manually compared the family health disease findings of twenty randomly chose patients who had more than one electronic discharge summary in FhhDS.

Results and discussion: FhhDS contains 22,292 inpatients from 1992 to 1998, 151,551 discharge diagnoses, 81,028 family health history findings, and 4,316 diseases associated with the patients. The patients in FhhDS represent a normal population of inpatients at CPMC. Twenty randomly chose patients who had more than one electronic discharge summary in FhhDS showed that 15% is completely identical (all the findings among different electronic discharge summaries are identical); 65% is partial in common (a part of the findings among different electronic discharge summaries is identical); 10% is not in common (all the findings are not in common); 5% is not consistent (the findings in different electronic discharge summaries conflict); and 5% is not valuable (all the findings are not related to any diseases). Thus, FhhDS is a large scale, cross-disease, partial complete and representative family health data set and may be applied to genetic disease study.

Acknowledgements: This work was supported by National Library Medicine grants LM05627 "Linking Knowledge-Based Systems to Clinical Databases" and LM07079 "Research Training Grant".

References: