A Phenome-Wide Association Study
Through Secondary Uses of Clinical and Research Data

Chunhua Weng¹, PhD, Richard Smiley², MD, PhD, Pamela Flood³, MD,
Bin Cheng³, PhD, Carol Friedman³, PhD, George Hripcsak⁴, MD, MS
¹ Department of Biomedical Informatics, ²Department of Anesthesiology, ³Department of Biostatistics
Columbia University, New York, NY, 10032 {cw2384@columbia.edu}

Introduction
In contrast to genome-wide association studies, we propose a new method to study genetic factors underlying diseases, which we call Phenome-Wide Association Studies. The idea is to use clinically relevant gene variants to simultaneously investigate phenotypes associated with the selected genes. If we use an M by N matrix to represent M phenotypes and N genes, a GWAS study preselects a phenotype (a row) and scan the N genes (columns) in the matrix for genes associated with the phenotype, while a phenome-wide association study preselects a gene (a column) and scan the M phenotypes (rows) associated with the gene. A similar idea (“clinical phenome scanning”) has been suggested¹, but no actual study has been reported. Electronic Health Records (EHR) are a useful data source for genetic association studies². In this poster, we outline the design of an initial phenome-wide association study on phenotypes associated with the human β2 adrenergic receptor (β2AR) through secondary uses of clinical and research data. In this study, the phenotypes were diseases and problems extracted from both structured and unstructured EHR data.

Methods
Coauthor Smiley had previously determined the genotype β2AR in a cohort of 811 pregnant women to study genetic factors connected to birth outcomes. We reused the genotype data for two functional Single Nucleotide Polymorphisms (SNP) at codon 16 (Arg16Gly) and 27 (Gln27Glu) of β2-Adenergic Receptor (β2AR) for his cohort, which collectively had 667 distinctive diagnoses coded in the 9th International Classification of Diseases (ICD9) in their entire medical records. We also processed those patients’ discharge summaries from years 2004 and 2005 using natural language processing software MedLEE³, and identified 196 distinctive UMLS-recognizable health problems, where 136 of which did not have corresponding ICD9 codes, demonstrating that we can capture richer phenotypes using narrative clinical notes. The extended Fisher’s Exact Test⁴ was used to identify phenotypes with associations. The resulting p-values were adjusted by the Benjamini and Hochberg (1995)⁵ method to ensure a false discovery rate of no more than 5%.

Results
We identified 42 associated diseases, which were divided into 4 groups: 4 diseases only associated with SNP Arg16Gly, 11 only with SNP Gln27Glu, and 24 only with the 2-SNP combo, and 3 with both the SNP Gln27Glu and the 2-SNP haplotypes. The 42 diseases included Diabetes mellitus, rheumatoid arthritis, cardiac dysrhythmias, and breast disorder.

Discussions and Conclusion
The β2AR adrenergic receptor is involved in cardiac contraction and conduction, and the regulation of metabolic control and bone biology. Several of the diseases for which we found associations would fit in these categories. Phenome-wide association studies bring new research challenges, including that phenotypes can be sparsely distributed across patients and coded at various granularity levels. In this study, the general good health of a cohort of young, pregnant women probably decreases the number of “clinical phenotypes” that arise from the EHR. However, it is still feasible to generate hypotheses about phenotypes which are associated with genes. Further research is needed to separate true from false positives, to compare association mining methods, to address phenotype representation variations, to investigate phenotype relationships, and to include richer phenotypes such as medications and symptoms.

Acknowledgment
This research was supported by grants LM06910, LM007659, LM00865, and HD048805. We thank Lyudmila Shagina for helping extracting phenotypes using MedLEE.

References