Columbia University

George Hripcsak, MD, MS, Chunhua Weng, PhD, Richard Smiley, MD

We are creating a phenotypic framework to support genome-wide association studies (GWAS) based on our 20 years of experience mining clinical data. This framework will employ a pipeline architecture to derive clinical phenotypes from electronic health records. It entails four components. First, data elements are mapped to standardized terminologies such as SNOMEDCT and ICD9, and continuous values are normalized to standard units. Second, much of the useful narrative phenotypic data, such as symptoms and signs, are converted via natural language processing to terminology-coded data. Third, these facts are then assembled into a coherent longitudinal clinical record that accounts for the temporal relationships among data elements. Finally, each patient's record is characterized to ensure that a sufficient amount of data is available that the derived phenotype is likely to be representative of the patient. The result is a standardized (coded, organized, and verified) phenotype for each patient's record that may be associated with the genotype as part of a GWAS.

The Phenotypic Pipeline Architecture

A Case Study on β2AR's Clinical Effects

The prototype of this phenotypic system is being applied to accelerate a genotype-phenotype association study led by Dr. Smiley based upon β2AR genotype for 2200 patients. Of 400 enrolled subjects, 150 underwent vaginal deliveries ≥ 34 weeks gestational age with sufficient intrapartum documentation to allow for meaningful analyses of labor progress. Progress was assessed using linear equations for both early (0 to 4 cm) and active labor (4 to 10 cm). β2AR haplotype was considered the independent variable. Active labor plots were compared between haplotype groups with linear regression using Medcal Origin. 9.3% of subjects delivered at < 37 weeks gestational age. Of 9 potential haplotypes involving codons 16 and 27, 6 occurred with sufficient frequency to allow for analysis. Double-homozygous subjects for β2AR haplotype Gly16Glu27 (n=16) had significantly faster active labor (0.95 ± 0.05 cm/hr) than subjects possessing other genotypes. Double-homozygous subjects for β2AR haplotype Arg16Gln27 (n=30) had significantly slower labor (0.53 ± 0.03 cm/hr) than subjects possessing other genotypes. Subjects heterozygous at either or both codons demonstrated intermediate phenotypes that were not significantly different between groups. These findings shown below are consistent with recent findings concerning preterm labor and delivery and, if confirmed in prospective studies, would mean that β2AR genetic variation has a significant clinical effect on labor progress.

References & Demos

2. A Temporal Tagger for Clinical Text: http://www.dbmi.columbia.edu/~hripcsa/cgi-bin/tcstagger_demo.cgi
4. MED: The Medical Entity Dictionary http://med.dmi.columbia.edu/