CS 3200: Database Design
Homework 2 : Genes and Diseases – Single Table Queries
Prof. Rachlin

PROBLEM DESCRIPTION:

In this assignment, we will explore GAD, the Genetic Association Database (Becker *et al*., 2004) and in the process we'll re-discover some possible biological connections between seemingly disparate diseases. While we sometimes think of genetic diseases as being associated with a single faulty gene, the truth is much more complex. Most diseases are "multi-genic" meaning that there are many genes which have been linked to the disease or disease phenotype. To say that there is a link or *association* means that some particular research study found a statistically-significant connection between some variation of a gene, and the occurrence of a disease phenotype – without necessarily identifying the underlying biological basis for the connection. The association may be limited to a particular population (Japanese, American Indian, etc.) and the association may only suggest some increased *probability* of acquiring the disease. There may be other, as yet unknown connections with environmental factors such as diet which have yet to be teased out.

GAD, the Genetic Association Database, is a catalog of research studies reporting an association (or lack of association) between genes and a disease. The data has the following columns:

gad_id – The table's primary key

association – Whether there is a positive association ('Y')

phenotype – The name of the disease or "phenotype"

disease_class – The disease class (e.g**.,** NEUROLOGICAL)

chromosome – The chromosome on which the gene resides

chromosome band – A descriptor for the chromosomal region

dna_start – Where on the chromosome the gene begins

dna_end – Where on the chromosome the gene ends

gene – The official gene symbol

gene_name – The gene's full name

reference – The research paper where the association was reported

pubmed_id – Publication identifier (PubMed)

year – Year of publication

population – The population associated with study

Import the GAD data (gad.csv) into new schema called gad with a single table, also called gad. Then open the attached script and answer each of the first 12 questions with a SQL query. (Your answer to the 13'th question can be typed into the script as well as comments.)

SUBMIT: A .SQL script with the answers to each query. Please rename your script file: CS3200_HW2_*lastname*.sql