

## Homework 1

### 6.872/HST.950 Problem Set 1

- *All answers need to be printed and not handwritten.*
- *There will be an SQL exercise on a small subset of the Children's Hospital database, from about five years ago. These data have been "scrubbed", so that specifically identifying information has been altered or removed. Nevertheless, please treat the data with sensitivity, because you will know from our discussion of medical data confidentiality that even innocuous-seeming data can help re-identify.*
- *Problems marked as bonus question are not required and you don't have to answer them, but I do suggest you take a look and think about them.*

1. You downloaded a program from the Internet that claims to predict whether a woman has breast cancer based on information such as age, weight, etc. To verify whether this program can benefit actual people, you collected a data set including 40 women under the age of 45 (none of which actually has breast cancer) and 10 women above the age of 45 (one of which actually has breast cancer). Only after you input all the data into the program did you realize that what the program actually does is to "predict" every woman above the age of 45 to have breast cancer but every woman under the age of 45 not to have breast cancer.

- a) What's the estimation of the sensitivity and specificity of the program based on the test on the data you collected?
- b) Suppose the prevalence of breast cancer is about 10% for women over 45, and 0.1% for women under 45, while about 60% women in the population are under 45. What will be the sensitivity and specificity of this program when used in the whole population for a long time?
- c) (Bonus question, no credit) Please comment on the difference between the results from a) and b), and make a suggestion on evaluating how accurate the estimation of sensitivity and specificity is.

2. SQL practice on CWS database: please use as few as possible queries.

- a) Please list the query to retrieve all disease names (no duplicated ones) in the database. You may either use the table PROBLEMS and/or any other tables you think appropriate.
- b) Please list the query or queries to retrieve the names of the doctors who have treated the disease DIABETES-INSIPIDUS. You may use the tables PPR, PERSNL\_PUBLIC and/or any other tables you think appropriate. (*If a doctor has treated a patient with DIABETES-INSIPIDUS, the doctor is considered as having treated this disease.*)

c) Give three different queries, each of which will estimate the total number of patients being tracked in the database. If they result in different numbers, discuss why. (*Ignore the possibility that the same patient is entered several times but with different identifiers.*)

3. We learned in class three different paper-based medical record systems: doctor's journal, patient's journal (e.g. Mayo), and problem oriented medical records (e.g. SOAP).

a) Please list the pros and cons of each system while comparing with the others.

b) Design a simple relational database table scheme of electronic medical record (EMR) system, from which the above three styles of medical records can be easily retrieved.

c) Please list the queries to create the table, and the queries to retrieve the three types of medical records.

d) Mr. Murphy was sent to the hospital with a broken leg because of a car accident, completely healthy otherwise. He was treated and discharged by Dr. Jenkins and Dr. Gehrke. Please list the query or queries to insert this new data into your database.

4. Describe an ideal Computer-based Patient Record (CPR) system based on the functional components and fundamental issues discussed in Shortliffe by listing at least two properties of each functional component and the three most important aspects of each fundamental issue. What kind of impediment would you expect to face to develop and use such a CPR?

5. (Shortliffe page 396, p4) Describe the key philosophies underlying the design of central, modular, and distributed HCISs. What are at least two advantages and two disadvantages of each architecture?

6. (Bonus question, no credit) Please list at least two possible explanations of the following fact: medicine spends 1-2% on IT, vs. 6-7% for business overall, vs. 10-12% for banking. Do you think information technology should play a more important role in medicine? Why?