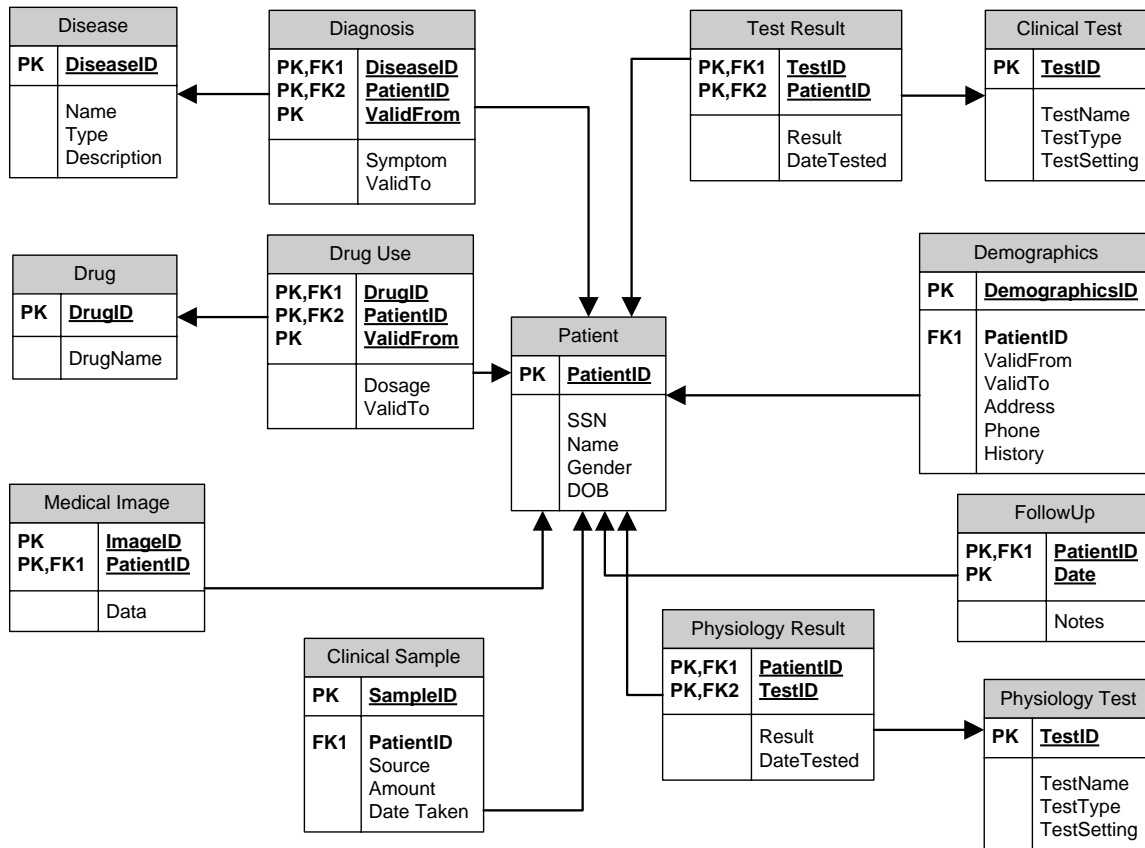


**Homework One: Schema Design for Biomedical Data  
Warehouse Report**

**CSE 601, Fall 2004**

**Amin Ghadersohi  
Stephanie H. Li  
Jian Liu**

The model given in the paper “Modeling clinical and genomic data for biomedical data warehousing and mining” has both its advantages and disadvantages. One of its advantages is that it reduces redundancies to a bare minimum. However, looked at another angle, one can pose the question, is this bare minimum data enough to keep all the possible information?



**Design 1 for Clinical Data Space**  
 This design solves the problem of a patient only being able to have a disease once, by adding **ValidFrom** as a key. Also the n to n problem with the Physiological data is also solved. E.g.: Electro Cardiogram, Polygraph, Fitness

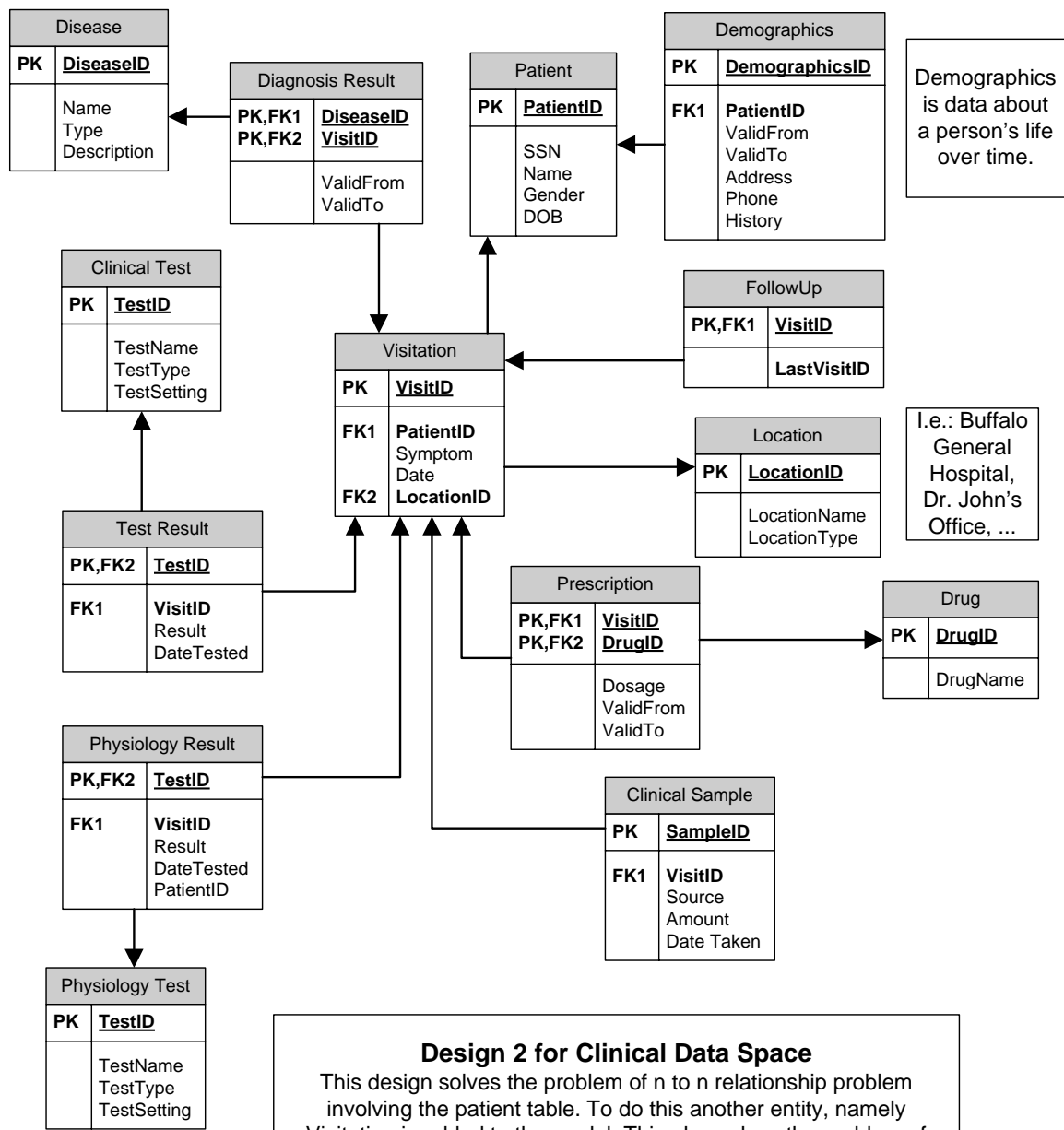
Having a closer look at the BioStar model schema for the clinical data space, one will notice that the Diagnosis table will not allow a patient to be diagnosed with the same disease twice. Without fully modifying this design, one can add a key to this table to solve

this problem. We proposed the addition of validFrom as a key. Here the assumption is that validFrom is accurate to the hour (maybe minute?).

Another problem that is visible is that it is not clear which diagnosis the drugs given to the patient belong to. In other words it is not easy to find what drugs were prescribed for what disease. Moreover, since a patient may be diagnosed with two or more different diseases at the same time, it would be useful to know in correlation to what disease the patient was tested (i.e. a blood test is a generic test given for many diseases. It would be useful to know what the doctor was looking for).

In our first re-design of the clinical data space schema, we added validFrom as a key to tables Diagnosis and DrugUse. This solves the correlation between drug and diagnosis, since even if a patient has two diseases at the same time and taking different drugs, it should be clear which is for what. Although, it would still be better to know exactly what the doctor was thinking, instead of having to guess.

As an alternative solution to the problems above, we have come up with a completely new design for the clinical data space schema (design 2). The new design solves the above problems, and gives us the power to deduce new facts from the data. For example, storing demographics data is now more efficiently accomplished.



**Design 2 for Clinical Data Space**

This design solves the problem of n to n relationship problem involving the patient table. To do this another entity, namely Visitation is added to the model. This also solves the problem of relating a prescription drug to a particular time when a patient had a disease. Also the patient can be tested and samples can be collected but there doesn't have to be a disease. Maybe it was just a regular checkup. The data can be thought of as control data.

The approach for this new design comes from the fact that for a patient to be diagnosed, tested, or sampled, they have to visit a place of healthcare. So we add a table called visitation. This table keeps track of every time a patient goes to visit a place of healthcare. So, now instead of relating a diagnosis, test, sample, medical image, follow-up,

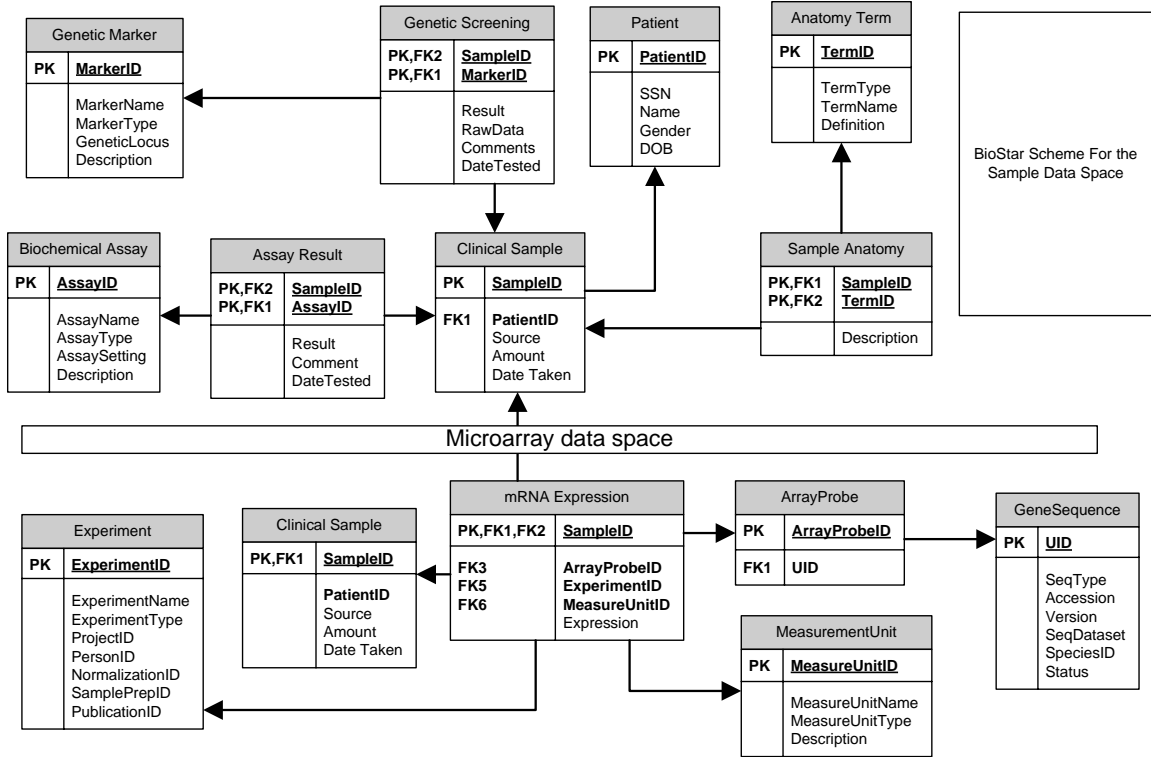
or a prescription with a patient, we shall be relating them with a visitation. This makes perfect sense, since usually the patient does not need to be known for such large scale data, but it is important to know what diseases a patient had when he/she was tested or donated a sample. This is accomplished through the visitation and follow-up tables and their relationship. Now we can tell how many diseases a patient had during a time period, how many times they visited, what they were prescribed, and so on.

Finally, the conceptual model asks for a one-to-many relationship between the patient and demographics. To store this data, a demographics table is added to the schema. Each row in the demographics table will refer to a patient. If a patient's information changes, a new demographics row is created. This table, along with the location table, will provide enough information about when and where a patient was sick, and when and where they were treated. Moreover, with this design all other data spaces stay the same except for the Clinical Sample table, which will have an extra field, namely visitID.

### **Design Explanations:**

Looking at our design 2, one may ask, isn't it redundant to store both a date for visitation, and one for sample and another for tests? The answer is simple, you may go to visit your doctor and the doctor gives you a blood test, for example, and they tell you to do it the next morning. So it won't have the same date.

We needed to store a one-to-many relationship between patient and demographics. Considering that demographics is a kind of data about a person's life, we added a demographics table, such that each record in this table is related to one patient only. Also, the location table provides more information about where the patient is receiving healthcare.



BioStar Scheme For the Sample Data Space

These two models have not been modified.