Faculty of Science
COMP-421 - Database Systems (Winter 2018)
Midterm Examination

March 14, 2018 Examiner: Joseph Dsilva
18:30 - 20:00 Associate Examiner:

Student Family Name:	**Student First Name:**

Student Number:

Instructions:

- You should have received this exam paper, a scantron sheet, and an exam booklet for the answers of the open questions.
- **DO NOT TURN THIS PAGE UNTIL INSTRUCTED**
- **INDICATE THE VERSION-0 ON THE SCANTRON!!!**
- **WRITE YOUR NAME AND STUDENT ID ON THE SCANTRON.**
- **WRITE YOUR NAME AND STUDENT ID ON THIS FIRST PAGE OF THIS EXAM PAPER**
- **You may split apart this exam paper, for example, to make it easier to read the background information about the example application. But you MUST WRITE YOUR NAME AND STUDENT ID on each of the separated sheets.**
- You have to return the scantron, ALL pages of this exam paper as well as the exam booklet.
- This is a **closed book** examination; only three letter-sized (8.5" by 11") **crib sheets** are permitted. This crib sheet can be single or double-sided; it can be handwritten or typed. Non-electronic translation dictionaries are permitted, but instructors and invigilators reserve the right to inspect them at any time during the examination.
- Additionally, only writing implements (pens, pencils, erasers, pencil sharpeners, etc.) and a simple calculator are allowed. The possession of any other tools or devices is prohibited.
- Answer **all** multiple choice questions on the scantron sheet.
- Answer open questions into the exam booklet.
- This exam paper has **11** pages including this cover page, and is printed on both sides of the paper.
- The Examination Security Monitor Program detects pairs of students with unusually similar answer patterns on mulitple-choice exams. Data generated by this program can be used as admissible evidence, either to initiate or corroborate an investigation or a charge of cheating under Section 16 of the Code of Student Conduct and Disciplinary Procedures.

Scoring

The exam is out of 90 points distributed as follows:

1. Section 1 (multiple choice with multiple answers): 8 questions; each 5 points for a total of 40 points
2. Section 2 (Open Questions): 5 questions for a total of 50 points

Version-0

This page is kept intentionally blank.

Background Information for this Exam

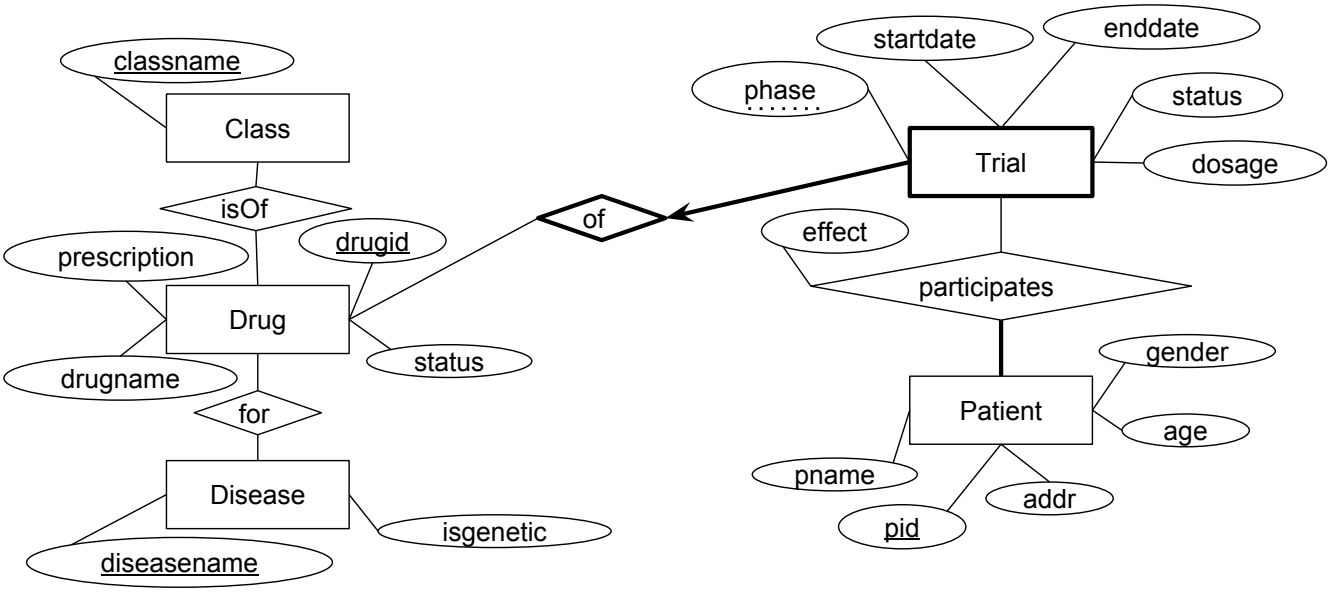
All questions on this exam are related to the following application.

BigPharma is a pharmaceutical company that is developing innovative drugs for treatment of various illnesses such as *Cancer*, *Alzheimer's* etc. They need a database to keep track of their drug testing progress and has hired a comp 421 student Debbie Tow to perform the task.

These are the data requirements that Debbie has gathered from BigPharama.

- The company develops drugs for both genetic disorders and non-genetic diseases. They keep track of each disease for which they are pursuing a drug and a flag (*Y/N*) to indicate if it is a genetic disorder.
- Drugs can be intended to be over-the-counter or prescription and is again kept track via a *Y/N* flag.
- A drug also goes through various status, viz. *research*, *trial*, *production*, and *abandoned*. The status can change back and forth, and we need to keep track of the current status of each drug. If a drug reaches the production or abandoned status, it will no more change status.
- A drug can also optionally belong to multiple classes of drugs, for example *Aspirin* is an *anti-pyretic* as well as a *salicylate*.
- A drug goes through multiple trials, (*Phase 0* through *Phase 4*) we need to keep track of all such trials, recording their start date, end date, dosage (recorded in grams), and the current status of the trial (*in-progress*, *completed*, *analysis*, *passed*, *failed*). If a trial fails, that drug should be abandoned.
- Patients are recruited for trials. We need to keep track of patients name, address, gender and age (this is a bad design, should be date of birth instead of age, but let us keep it simple).
- We also need to record the effect that the drug has on the patient (*neutral*, *positive*, *negative*).

Debbie has done a good job in producing an E/R diagram to capture this information, which is given below. Note that there might be some minor issues in Debbie's model. She may have also made some reasonable assumptions, where the requirements are not explicitly recorded.



She has also created a relational model from the ER, as given below.

`class(classname)`
`drug(drugid, drugname, prescription, status)`
`disease(diseasename, isgenetic)`
`trial(drugid, phase, startdate, enddate, status, dosage)`
$\longrightarrow$ `drugid` references `drug`
`patient(pid, pname, addr, age, gender)`
`drugclasses(drugid, classname)`
$\longrightarrow$ `drugid` references `drug`, `classname` references `class`
`druguses(drugid,diseasename)`
$\longrightarrow$ `drugid` references `drug`, `diseasename` references `disease`
`patienttrial(pid, drugid, phase, effect)`
$\longrightarrow$ `(drugid,phase)` references `trial`, `pid` references `patient`

This page is kept intentionally blank.

Multiple Answer Multiple Choice Question (40 points)

This section contains a set of multiple choice questions all referring to the example application given in the background information. Each question has **one or more correct answers**. **You have to select ALL correct answers**. If you miss a correct answer or you select a wrong answer, you will get 0 points for that question. For each question, we have indicated how many correct answers exists. Use this to guide your decision.
You can achieve a total of 40 points in this section (5 per each of the 8 questions).

1. (5 Points) Based on the ER translation of the requirements, which of the conclusions are correct ? (choose 3 correct answers).

   (A) A drug may be not associated with any disease.

   (B) Since `class` has only one attribute `classname`, it could have been made an attribute of `drug` entity set, discarding the `class` entity set.

   (C) `drugid` is an artificial key.

   (D) We cannot always tell what was the previous status of a drug using just the `drug` entity set.

   (E) A drug is useful only for one particular disease.

2. (5 Points) Based on the ER translation of the requirements, which of the conclusions are correct ? (choose 2 correct answers)

   (A) `trial` should not be a weak entity set as `phase` is unique to each `drug`.

   (B) A patient can participate on multiple trials (of different drugs) at the same time.

   (C) All patients in a given trial will have the same dosage.

   (D) It is not possible to compute the total number of *negative* effects in a given trial.

   (E) A patient must participate in more than one trial.

3. (5 Points) Which of the following is true with respect to the relational model that is given ? (choose 2 correct answers)

   (A) A patient may not participate in any trial.

   (B) The `patienttrial` relation is not required as we can include the `effect` attribute in the `patient` relation which will effectively capture the same information.

   (C) All trials for a given drug must have the same dosage.

   (D) `startdate` of a trial can be after its `enddate` since there are no constraints.

   (E) For `patienttrial` relation, `drugid` should be a foreign key to the `drug` relation, not `trial` relation.

4. (5 Points) Which of the following is true with respect to the relational model that is given ? (choose 2 correct answers)

   (A) Given the name of a disease, the number of drugs currently in production for that disease can be found by reading just one relation.

   (B) The total number of distinct drugs can be obtained either from the `drug` relation or the `drugclasses` relation.

(C) Given the name of a disease, to identify the classes of drugs associated with it (irrespective of their current status) requires at the least 3 relations to be joined.

(D) The number of tuples in the `drug` relation can be more than the number of tuples in the `class` relation.

(E) The number of tuples in the `class` relation can be more than the number of tuples in the `drug` relation.

5. (5 Points) Which of the following is true for the tables that will have to be created using SQL based on the relational translation that we saw? (choose 2 correct answers)

(A) `patienttrial` has to be the last table to be created because of its foreign key constraints.

(B) We will not be able to create `patienttrial` as-is since it has foreign key references to two different tables.

(C) We can impose a column level constraint on the `age` column of the `patient` to ensure it has no negative values.

(D) `patienttrial` table has two candidate keys, `(drugid,phase)` and `pid`.

(E) `patient` table has to be created before `patienttrial` table.

6. (5 Points) In relational algebra, we can find the names of drugs developed (irrespective of their status) to treat *cystic fibrosis* using the queries (choose 2 correct answers)

(A) $\Pi_{drugname}(\sigma_{diseasename=`cysticfibrosis'}(druguses \bowtie drugs))$

(B) $\Pi_{drugname}(drug) \bowtie \sigma_{diseasename=`cysticfibrosis'}(druguses)$

(C) $\Pi_{drugname}(\sigma_{diseasename=`cysticfibrosis'}(diseases \bowtie druguses \bowtie drugs))$

(D) $\Pi_{drugname}(drug) \cup \sigma_{diseasename=`cysticfibrosis'}(druguses)$

(E) $\sigma_{diseasename=`cysticfibrosis'}(druguses) \bowtie \Pi_{drugname}(drug)$

7. (5 Points) Assuming that the attributes in SQL tables directly correspond to the relational translation that we saw previously, we can rewrite an equivalent SQL query to address the previous question as (choose 2 correct answers)

(A) `SELECT diseasename FROM druguses, drugs WHERE diseasename = 'cycstic fibrosis'`

(B) `SELECT drugname FROM drugs WHERE drugid IN (SELECT drugid FROM druguses WHERE diseasename = 'cycstic fibrosis')`

(C) `SELECT drugname FROM druguses, drugs WHERE diseasename = 'cycstic fibrosis'`

(D) `SELECT drugname FROM drugs WHERE EXISTS (SELECT drugid FROM druguses WHERE diseasename = 'cycstic fibrosis' and drugid = drugs.drugid)`

(E) `SELECT DISTINCT drugname FROM druguses, drugs WHERE diseasename = 'cycstic fibrosis'`

8. (5 Points) Assume that `pid`, `drugid` are both 64 bit integers, `phase` is char(7), and `effect` is varchar(10) with an average size of 8. Further, 1 char occupies 1 byte. Ignore slot overheads. (this information can be used to compute the average length of a record). Assume that in our database, we have 8 drugs for which all 5 trials are conducted, with 150 patients participating in each trial. What is the approximate number of pages required to store the `patienttrial` table, assuming a page contains 4000 bytes and pages are filled 75% on an average? (Choose 1 correct answer)

(A) 62

(B) 2

(C) 47

(D) 6

(E) 66

Open Questions (45 points)

This section contains open questions based on our example application.

9. (5 Points) Using the relational model given in the background information, write a relational algebra query to find the name and address of patients who participated on both the *Phase 0* and *Phase 4* trials of a drug with `drugid` *1223*.
   You will receive partial credit (3 points) if instead you chose to write the SQL query.

10. (6 Points) Using the relational model given in the background information, write a SQL to find the names of all drugs that were tested on the patient with name *John Doe* that had a *negative* effect on him.

11. (8 Points) Using the relational model given in the background information, write a SQL query to find the name(s) of drug(s) in *production* `status` that can be used for treating *multiple sclerosis* and has at the least 100 patients with *positive* `effect` and less than 10 patients with *negative* `effect` across all the trials of the drug.
    If you are not able to solve this query, for 4 points, write a SQL query to find the name(s) of drug(s) in *production* `status` that can be used for treating *multiple sclerosis* and has at the least 100 patients with *positive* `effect` across all the trials of the drug.

12. (15 Points) One of the common queries in the application is to search for existing patients in the database that satisfy specific characteristics (such as age and gender) for recruiting to a new drug trial. Now consider the following information:
    `pid` is a 64 bit integer, `pname` is varchar(30) with average size 15 bytes, `addr` is varchar(50) with average size 30 bytes, `age` is a 16 bit integer, and `gender` is char(1) taking 1 byte. `age` has values from *20* through *69* (inclusive) and `gender` is either one of the values *(F, M, T)*. The `(age, gender)` data is uniformly distributed (i.e., the same number of patients for any age,gender combination). There is a multi attribute unclustered type II indirect index on `(age, gender)`. An rid takes 5 bytes. 1 page is 4000 bytes, filled 75%. (use this for both the data and index pages). We have 6000 patients.

    Note:- If you write down steps, you might get partial marks for them even if your numerical calculation is wrong.

    (a) (2 Points) What is the average length of record of the `patient` table ?

    (b) (2 Points) What is the average number of records of the `patient` table in a data page ?

    (c) (2 Points) What is the number of rids per data entry ?

    (d) (3 Points) What is number of leaf nodes in the index ?

    (e) (6 Points) For the SQL query
    `SELECT * FROM patients WHERE age BETWEEN X AND 50 AND gender = 'F'`
    find the smallest possible value for X, $X \leq 50$, such that the query is still guaranteed to benefit from using the index.

13. (16 Points) BigPharma has found the database to extremely useful and would like to enhance it to add more capabilities.

    - Each trial must have a single physician assigned to it. This person alone will be responsible to administer the drugs for that trial.
    - Further, the physician should be able to administer a different dosage to each patient, and even vary the dosage for the same patient on different dates.
    - You can assume that a drug is administered only once in a day to a given patient.
    - The dosage of the drug given to each patient on a day should be recorded, along with the date it is administered and its effect. The physician also writes a note (general observation details of the patient) when administering each dosage, which is also to be recorded.
    - Each dosage is applied from a single vial, which has a unique serial number. Vials are individual use, i.e., not shared between patients or used for multiple shots on the same patient. It is important to record the serial number of the vial used. We need not keep track of vials not (yet) administered.
    - Physician's name and license number should be stored. License numbers are unique to each physician. A physician may oversee multiple trials at the same time.

    (a) (11 Points) Debbie is unfortunately away for her ski trip, and you are the intern in charge. Draw a modified ER diagram to capture the additional information.
    Do you see any potential anomalies that cannot be avoided in the ER because of this change ? - write it down
    You **ONLY** need to draw the new entity sets, new relationships and their respective attributes and the existing entity sets (the later, without their attributes) to which they connect. **Furthermore, if you change an existing entity set or a relationship set, you should fully redraw it in the modified version with all its attributes.**
    Do not add any attributes other than what is given in the (old or new) data requirements. You may pick a reasonable name for your attributes and entity sets.
    Note:- Where required, make sure your "thick lines" in the diagram are evidently thicker than your thin lines. You may draw double lines if it is easier to do so.

    (b) (5 Points) Write the relational model for any new/modified relations resulting from your ER changes. Remember to also review existing relational model to see if you need to make any changes to it. If a relation changes, you need to write it down completely along with any foreign keys etc that it has (new and old). Do not forget to underline primary keys and write out the foreign key references. Relations with no changes to itself need not be written down. If you think any existing relation is to be removed or replaced by a new one, let us know.