Relational Calculus & Relational Algebra

1. **Give one reason why MRN is a poor choice for the primary & foreign keys?**
   One single patient may come to see clinicians for more than one time, so the same MRN could appear on multiple tuples in the relation VISIT.

2. **Write Relational Calculus expressions for the following:**

   (a) Who is 25 years old?

   $$\{pa.\,FIRSTNAME, pa.\,LASTNAME \mid PATIENT(pa) \wedge pa.\,AGE = '25'\}$$

   (b) Who had a medical visit in December, 2017?

   $$\{ pa.\,FIRSTNAME, pa.\,LASTNAME \mid PATIENT(pa) \wedge \exists(v)(VISIT(v) \wedge v.\,DATETIME = 'December, 2017' \wedge pa.\,MRN = v.\,MRN)\}$$

   (c) Who got a 'flu shot'?

   $$\{pa.\,FIRSTNAME, pa.\,LASTNAME \mid PATIENT(pa) \wedge \exists(v, pr)(VISIT(v) \wedge PROCEDURE(pr) \wedge pr.\,NAME = 'flu\ shot' \wedge v.\,VISIT\_ID = pr.\,VISIT\_ID \wedge pa.\,MRN = v.\,MRN)\}$$

   (d) Who did NOT get a 'flu shot'?

   $$\{pa.\,FIRSTNAME, pa.\,LASTNAME \mid PATIENT(pa) \wedge \neg\exists(v, pr)(VISIT(v) \wedge PROCEDURE(pr) \wedge pr.\,NAME = 'flu\ shot' \wedge v.\,VISIT\_ID = pr.\,VISIT\_ID \wedge pa.\,MRN = v.\,MRN)\}$$

   (e) What is the first and last name of all patients who have seen MD Paula Jones?

   $$\{pa.\,FIRSTNAME, pa.\,LASTNAME \mid PATIENT(pa) \wedge \exists(v, pr, c)(VISIT(v) \wedge PROCEDURE(pr) \wedge CLINICIAN(c) \wedge c.\,CERT = 'MD' \wedge c.\,FIRSTNAME = 'Paula' \wedge c.\,LASTNAME = 'Jones' \wedge pr.\,CLIN\_ID = c.\,CLIN\_ID \wedge v.\,VISIT\_ID = pr.\,VISIT\_ID \wedge pa.\,MRN = v.\,MRN)\}$$

3. **Write Relational Algebra expressions for the following questions. Return the relation primary key to identify the tuples, unless otherwise specified.**

   (a) Who is 25 years old?

   $$\pi_{MRN}(\sigma_{AGE = '25'}(PATIENT))$$

   (b) What is the first and last name of all patients who have seen a Physician's Assistant (cert is 'PA')

   $$C \leftarrow \rho_{FIRSTNAME\_C/FIRSTNAME,\ LASTNAME\_C/LASTNAME}(CLINICIAN)$$

   $$\pi_{FIRSTNAME,\ LASTNAME}(\sigma_{CERT = 'PA'}(PATIENT * VISIT * PROCEDURE * C))$$

(c) Which patients have the same names and age?

$\pi_{P.MRN,\ P1.MRN}(\sigma_{P.MRN\ !=\ P1.MRN}(\rho_{P(\cdots)}(PATIENT) \bowtie_{P.FIRSTNAME=P1.FIRSTNAME \wedge P.LASTNAME=P1.LASTNAME \wedge P.AGE=P1.AGE} \rho_{P1(\cdots)}(PATIENT)))$

(d) Which patients who got a flu shot also got a measles immunization during the same visit?

$R \leftarrow \pi_{VISIT\_ID,\ NAME1}(\rho_{NAME1/NAME}(\sigma_{NAME = \text{'flu shot'}}(PROCEDURE)))$

$S \leftarrow \pi_{VISIT\_ID,\ NAME2}(\rho_{NAME2/NAME}(\sigma_{NAME = \text{'measles immunization'}}(PROCEDURE)))$

$\pi_{MRN}(R*S*VISIT*PATIENT)$

(e) Which patients who have seen an MD have not seen a PA?

$C \leftarrow \rho_{FIRSTNAME\_C/FIRSTNAME,\ LASTNAME\_C/LASTNAME}(CLINICIAN)$

$R \leftarrow \pi_{FIRSTNAME,\ LASTNAME}(\sigma_{CERT = \text{'MD'}}(PATIENT * VISIT * PROCEDURE * C))$

$S \leftarrow \pi_{FIRSTNAME,\ LASTNAME}(\sigma_{CERT = \text{'PA'}}(PATIENT * VISIT * PROCEDURE * C))$

$\pi_{MRN}((\pi_{FIRSTNAME,\ LASTNAME}(PATIENT) - S) \cap R)$

Queries

1. **List the nct id and study type from the study whose brief title is "Autologous Cell Therapy After Stroke".**

```
SELECT nct_id, study_type
FROM studies
WHERE brief_title = 'Autologous Cell Therapy After Stroke';
```

Data Output | Explain | Messages | Notifications

	nct_id character (11)	study_type character varying (35)
1	NCT00908856	Interventional

2. List the different values for study type, in alphabetical order.

```
SELECT DISTINCT study_type
FROM studies
ORDER BY study_type ASC;
```

Data Output | Explain | Messages

	study_type character varying (35)
1	Expanded Access
2	Interventional
3	N/A
4	Observational
5	Observational [Patient Re...

3. How many terminated studies that started and completed in 2016 have reported events?

```
SELECT COUNT (DISTINCT studies.nct_id)
FROM studies, reported_events
WHERE studies.nct_id = reported_events.nct_id
        AND start_date >= '2016-01-01'
        AND completion_date < '2017-01-01'
        AND overall_status = 'Terminated';
```

Data Output

	count bigint
1	15

4. How many of the studies that started in February 2016, but on or after the 15th, are expected to complete (or have completed) within 6 months of their start date?

```
SELECT COUNT(DISTINCT nct_id)
FROM studies
WHERE start_date >= '2016-02-15'
        AND start_date < '2016-03-01'
        AND completion_date <= start_date + INTERVAL '6 MONTH';
```

Data Output

	count bigint
1	10

Reading/Short answer

1. **List three reasons the database was "normalized."**
   (1) Normalization makes the organization of data more efficient.
   (2) Reduce redundancy.
   (3) Ensure logical data dependencies by storing only related data within a given table.

2. **On page 4, the authors list a number of data elements that are contained in the XML download files from ClinicalTrials.gov. How are these different elements implemented in the design table in the AACT database?**

The values associated with each data element are tested for correctness and completeness by comparing them with the original source data from downloaded XML files.

3. **Which non-description field in the design table is least populated? That is, which field is most often left blank?**

observational_model

4. **In the conditions table there is an attribute called "name" and an attribute called "downcase name". What is the difference? Why might a database provide both of these fields? What trade-offs are involved?**

(1) For the attribute "name", the first letter of the contents is uppercase or the contents are uppercase acronyms. While in "downcase name", all letters are lowercase.
(2) It's more convenient for a user to query on database.
(3) It requires more space to store the information.

5. **Look at some of the name, downcase name pairings. Do you see any anomalies? Give 2 examples of what challenges might these anomalies pose to a user of the database.**

(1) Some values of the "name" do not have pairing in "downcase name". If users want to run some queries using "downcase name", they might miss some of the data they want. For example, from row 29 to 36, the values of their "downcase name" is null. But all of these rows have their corresponding value in attributes "name".
(2) The ambiguity of name between uppercase and lowercase. For example, the attribute "name" of row 2 is "Hiv". However, the name of row 19 is "HIV". Actually both of them refer to the same disease. So a user looking for the study of this disease might miss one of the two HIV.