THE UNIVERSITY of EDINBURGH

Usher Institute
Population Health
Sciences & Informatics

Data Science in Medicine: Tutorial 4

Relational model & SQL Querying

Semester 1, 2020-2021

- Please attempt all questions on this worksheet in advance of the tutorial, and bring with you all work, including printouts of code and other results. Tutorials cannot function properly unless you do the work in advance.
- You are welcome to bring along any questions you may have from the lectures, textbook, etc.
- Assessment is formative, meaning that tutorials do not contribute to your final grade.
- Attendance is compulsory. If you have good reasons to miss a session, you should contact your year coordinator in advance to arrange to attend a different session.

Introduction

In this tutorial you will try your hand at working with relational databases for medicine and healthcare. In the first part of the tutorial you will familiarise yourself with a set of table declarations from a drug prescription database, while in the second part of the tutorial you will query the database to answer questions of interest. In the third part of the tutorial you will reflect on differences between relational and flat databases.

Part 1: Mapping the ER model to the relational model

In this part of the tutorial you will familiarise yourself with a relational database for keeping track of patients, their GPs and drugs prescribed by GPs to patients. Have a close look at the relational database specification provided below, and answer the following questions.

Relational database specification

```
CREATE TABLE Patient (
    chi             CHAR(10),
    name            VARCHAR(60),
    email           VARCHAR(60),
    postcode        CHAR(6),
    year_of_birth   INTEGER,
    gp_id           CHAR(10) NOT NULL,
    PRIMARY KEY (chi),
    FOREIGN KEY (gp_id) REFERENCES General_Practitioner
 )


CREATE TABLE General_Practitioner (
    id                     CHAR(10),
    name                   VARCHAR(60),
    email                  VARCHAR(60),
    current_practice       VARCHAR(60),
    years_of_experience    INTEGER,
    PRIMARY KEY (id)
 )


CREATE TABLE Drug (
    id              CHAR(10),
    brand_name      VARCHAR(60),
    generic_name    VARCHAR(60),
    company         VARCHAR(60),
    PRIMARY KEY (id)
 )


CREATE TABLE Prescription (
    pr_id           CHAR(10),
    p_id            CHAR(10),
    gp_id           CHAR(10),
    d_id            CHAR(10),
    quantity        INTEGER,
    date            DATE,
    PRIMARY KEY (pr_id),
    FOREIGN KEY (p_id) REFERENCES Patient,
    FOREIGN KEY (gp_id) REFERENCES General_Practitioner,
    FOREIGN KEY (d_id) REFERENCES Drug
 )
```

Example (fictitious) data

Patient

chi	name	email	postcode	year_of_ birth	gp_id
0103624538	Alastair Brown	a.brown@example.com	EH89FK	1962	gke8849340
1208783406	Amy Murray	a.murray@example.com	AB83KL	1978	vnn8458554
3005402592	Fiona Campbell	f.campbell@example.com	LO43PR	1940	asw2213032
0812965634	Julia Clark	j.clark@example.com	SD34TR	1996	asw2213032
1411845100	Rhona Wilson	r.wilson@example.com	SD98VF	1984	kwr9852345
3101974980	Andrew Ross	a.ross@example.com	SH51MN	1997	fcv0949043
2208663398	Hamish Walker	h.walker@example.com	EH24DX	1966	sdf2939475
1909793256	Iain Scott	i.scottt@example.com	EH56FF	1979	fcv0949043

General_Practitioner

id	name	email	current_practice	years_of_ experience
sdf2939475	Charlotte Aitken	c.aitken@example.com	Meadows Clinic	4
gke8849340	David Taylor	d.taylor@example.com	Rose Clinic	23
vnn8458554	Lucy Taylor	l.taylor@example.com	Rose Clinic	35
fcv0949043	Jack McGregor	j.mcgregor@example.com	Talbot Practice	12
asw2213032	Kyle Russell	k.russell@example.com	Earth Practice	26
kwr9852345	Hannah Mclean	h.mclean@example.com	Foster Clinic	8

Drug

id	brand_name	generic_name	company
gf23496889	Humolin R	Minocycline	PharmaWorld
po50094505	Novalin R	Minocycline	GrecoGen
mq95032359	Precoz	Acarbose	PharmaWorld
op99823820	Glucabay	Acarbose	HealthRight
kr87019382	Mycabutin	Rifabutin	GrecoGen
zg93055406	Zagan	Sparfloxacin	HorizonMed

Prescription

pr_id	p_id	gp_id	d_id	quantity	date
dfgkj38392	3005402592	asw2213032	gf23496889	1	20-01-2006
pepro83321	3005402592	asw2213032	gf23496889	1	29-11-2007
merer11760	3101974980	fcv0949043	po50094505	3	10-06-2014
mettr44039	2208663398	sdf2939475	gf23496889	5	08-01-2015
plote50975	3005402592	asw2213032	op99823820	1	08-01-2015
clarw81294	2208663398	sdf2939475	zg93055406	2	18-05-2015
bfhoo06912	0812965634	asw2213032	mq95032359	4	20-01-1999

Questions

(1) How many fields (i.e. columns) are there in the General_Practitioner table, according to the table declaration? What are their names, and in what order do they appear?

(2) What is the primary key of Patient, and what does "primary key" mean in practice here? Are there any other appropriate candidate keys, and why do you think that this particular field was chosen as primary key?

(3) Suppose that you are adding a tuple to the Patient table, as per the DDL code below. What error messages or warnings would you expect to get from the relational database management system, based on the table declarations and the example data provided above?

```
INSERT
  INTO Patient (chi, name, email, postcode, year_of_birth, gp_id)
  VALUES ('1208712406', 'Alex Getty', 'a.getty@example.com', 'AB83KL', 12-08-1978,
'bb8458599')
```

Part 2: SQL querying

In this part of the tutorial you will get to practise with SQL queries for the database presented in Part 1. For each of the following questions, formulate the corresponding query in SQL. You are expected to write them by hand. If you want to try running them (this is entirely optional) to see what results you get and test whether your queries work, please follow the instructions in the section "Running your SQL queries (for the curious ones)".

Questions: Queries in SQL

(1) Retrieve the details of all GPs. The schema of the output table should be the same as that of the General_Practitioner table.

(2) Retrieve the names and emails of all GPs that have more than 10 years of experience.

(3) Retrieve the names of all patients that are registered with a GP that has more than 10 years of experience.

(4) Retrieve all drugs produced by PharmaWorld and HealthRight.

(5) Retrieve the names of all patients that have been prescribed a drug with generic name ‘Minocycline’.

(6) *Optional* Retrieve the names of all patients that have been prescribed a drug with generic name ‘Minocycline’ by a GP that has more than 10 years of experience.

Optional: Running your SQL queries (for the curious ones)

You can run your queries with the use of the freely available online tool SQL Fiddle at http://www.sqlfiddle.com/. To use the tool, please copy and paste the content of the text file tut4.txt into the Schema Panel on the left, and then click the “Build Schema” button. You only need to do this once (to create the database and populate it with data). Once this is successfully done, you can write your SQL queries inside the Query Panel on the right, and then click the “Run SQL” button. The results should appear on the bottom part of the screen. Note that you should only include one SQL query in the Query Panel, and you should click the “Run SQL” button every time you want to run a query.

Part 3: Discussion

Suppose that, instead of the relational database presented in Part 1, you were provided with a flat file database (e.g. csv file) containing the same data. A simplified version is presented in Table 1. Note that in order to make Table 1 fit this page, we had to ignore quite a few fields from different tables in Part 2 (e.g. patient email, GP current practice, etc.), but hopefully you get the idea.

pr_id	p_id	p_name	gp_id	gp_name	d_id	brand_name	company	quantity	date
dfgkj38392	3005402592	Fiona Campbell	asw2213032	Kyle Russell	gf23496889	Humolin R	PharmaWorld	1	20-01-2006
pepro83321	3005402592	Fiona Campbell	asw2213032	Kyle Russell	gf23496889	Humolin R	PharmaWorld	1	29-11-2007
merer11760	3101974980	Andrew Ross	fcv0949043	Jack McGregor	po50094505	Novalin R	GrecoGen	3	10-06-2014
…	…	…	…	…	…	…	…	…	…

Table 1

(1) How can one read Table 1? Which fields from which tables in Part 1 do the different columns come from?

(2) Which of the two approaches (i.e. relational database vs. flat file database) do you find easier for eyeballing the data?

(3) According to the Patient table in Part 1, gp_id should not be null. Is there a way to enforce this constraint in a flat file database (e.g. when working with a csv file using Excel or a text editor)?

(4) Are there any advantages of using a relational database (as in Part 1) vs. a flat file database (as in Table 1) for this particular scenario?

(5) In general, when would you choose to use a relational database and when a flat file database?