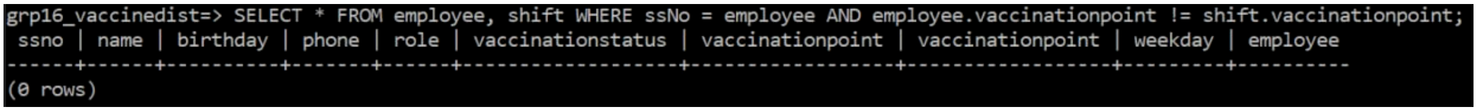
Group 16: Areeb Shahid 100280089 Nhut Cao 906939, Sahar Shaban 903372, Kirill Eliutin 890870

CS-A1153 Databases

Part 2

In this part we changed our schema slightly to accommodate to the differences between part 1 and the actual data:

1. In part 1 we assumed staff members aren't attached to a specific workplace (nothing was said about it in part 1, so we went with our assumption). We saw it's not the case from data, so we added an association between employee and hospital (Employee now has "vaccinationPoint" attribute). We removed "vaccinationPoint" from Shift to avoid redundancy and inconsistencies
   Before deleting "vaccinationPoint" from Shift We used the following query to verify that each employee belongs to only one workplace

   ```
   grp16_vaccinedist=> SELECT * FROM employee, shift WHERE ssNo = employee AND employee.vaccinationpoint != shift.vaccinationpoint;
    ssno | name | birthday | phone | role | vaccinationstatus | vaccinationpoint | vaccinationpoint | weekday | employee
   ------+------+----------+-------+------+-------------------+------------------+------------------+---------+----------
   (0 rows)
   ```

2. We wanted to track data about the batch's location in only one place - the TransportationLog. We wanted to treat the last arrival point from the transportation log as the current location of the batch - this would not allow for inconsistencies like the one in query 3. From data (and query 3) we saw that we're supposed to have the "location" as an attribute of the batch, so we added it to our schema (we wanted to just remove it at first, but then decided that inconsistencies present in data can't be easily resolved, and so we should just represent them in our database (also for purposes of query 3)). This also allowed us to put NOT NULL on departurePoint and departureDate (previously we wanted to put a record with no departure data to record the location of a batch that never moved between hospitals)

Otherwise the tables design was fairly straightforward from UML we got, with same primary keys and quite obvious foreign keys, for which we also defined sensible ON UPDATE/ON DELETE policies (for example, we don't allow to delete a vaccine if there's any batches referencing it; we can update ID for anything that has an ID and CASCADE it to anything referencing it; we can delete the patient's data, and all the appointments/diagnoses for that patient would be gone too (to allow people to delete their data on request)).
We executed the table creation code on the database with the psql command-line tool manually, and then put it into table.sql for evaluation .
For cleaning we mostly used pandas: the only change we made manually on excel is removal of a row in Diagnosis due incorrect date "44237" (row 92); We saved the excel file with the row removed as data/vaccine-distribution-data-cleaned.xlsx, which we load and then clean in pandas.
In python script ("code/part2.py") we clean data using pandas: we load it from excel, strip the spaces off of the end of columns names, convert data to proper types (int/float/bool/str), put date in the required format ("YYYY-MM-DD"), split data according to how it should be in the tables, and then push it to database.

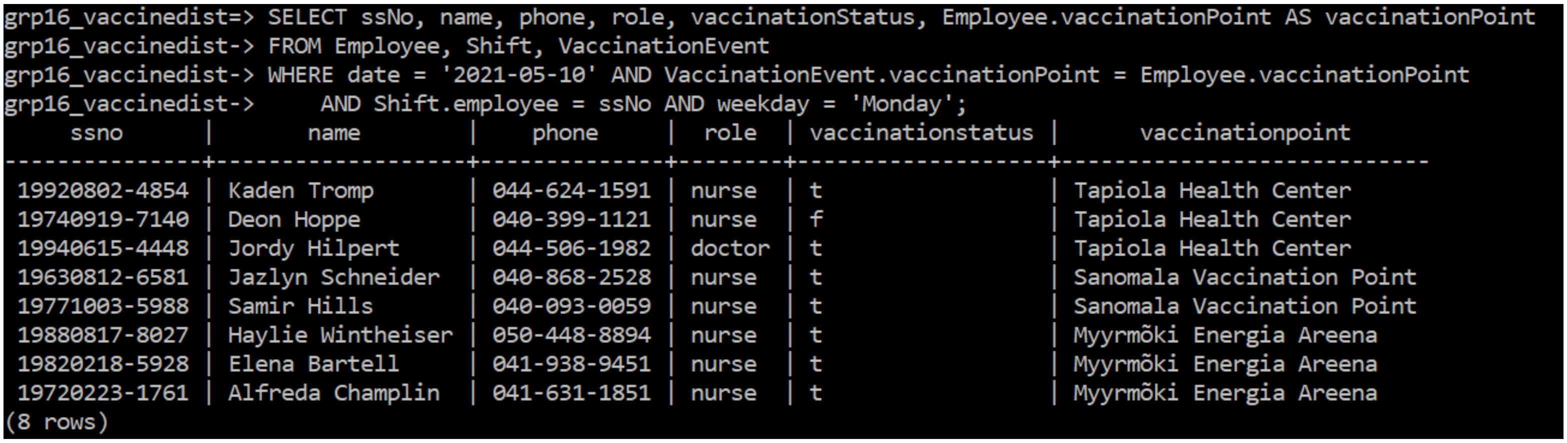
We added three check constraints :

1. In table Employee we check that the role is either nurse or doctor, we assumed these are the only 2 roles since there were no other roles present in the data

2. In table shift we check that the weekday is Monday-Friday
3. In table patient we check that the genfeder is F, M or O which also complies with data

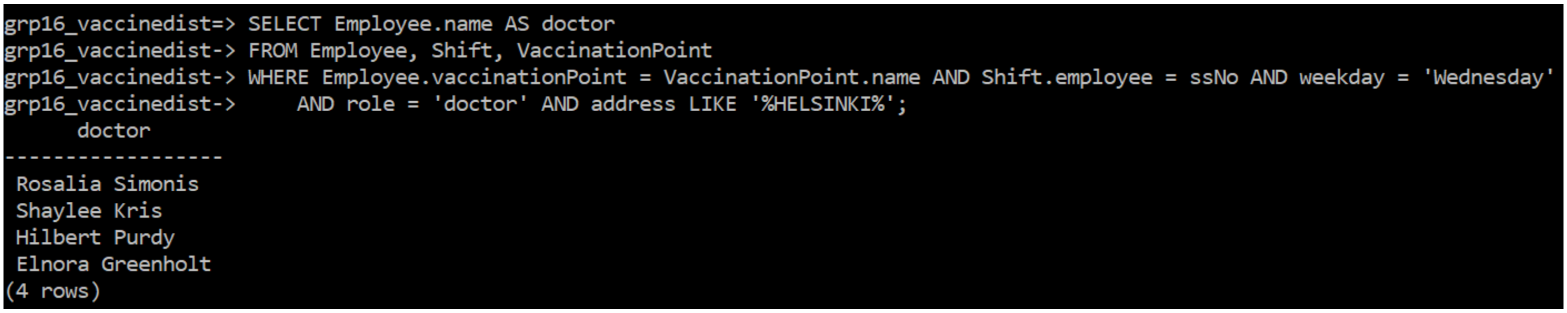
Queries:

1) 10.05.2021 was a monday, so we find out all the employees who have shifts on Monday in hospitals that had a vaccination event on 10.05.2021



```
grp16_vaccinedist=> SELECT ssNo, name, phone, role, vaccinationStatus, Employee.vaccinationPoint AS vaccinationPoint
grp16_vaccinedist-> FROM Employee, Shift, VaccinationEvent
grp16_vaccinedist-> WHERE date = '2021-05-10' AND VaccinationEvent.vaccinationPoint = Employee.vaccinationPoint
grp16_vaccinedist->     AND Shift.employee = ssNo AND weekday = 'Monday';
     ssno      |       name        |    phone     |  role  | vaccinationstatus |      vaccinationpoint
---------------+-------------------+--------------+--------+-------------------+---------------------------
 19920802-4854 | Kaden Tromp       | 044-624-1591 | nurse  | t                 | Tapiola Health Center
 19740919-7140 | Deon Hoppe        | 040-399-1121 | nurse  | f                 | Tapiola Health Center
 19940615-4448 | Jordy Hilpert     | 044-506-1982 | doctor | t                 | Tapiola Health Center
 19630812-6581 | Jazlyn Schneider  | 040-868-2528 | nurse  | t                 | Sanomala Vaccination Point
 19771003-5988 | Samir Hills       | 040-093-0059 | nurse  | t                 | Sanomala Vaccination Point
 19880817-8027 | Haylie Wintheiser | 050-448-8894 | nurse  | t                 | Myyrmõki Energia Areena
 19820218-5928 | Elena Bartell     | 041-938-9451 | nurse  | t                 | Myyrmõki Energia Areena
 19720223-1761 | Alfreda Champlin  | 041-631-1851 | nurse  | t                 | Myyrmõki Energia Areena
(8 rows)
```

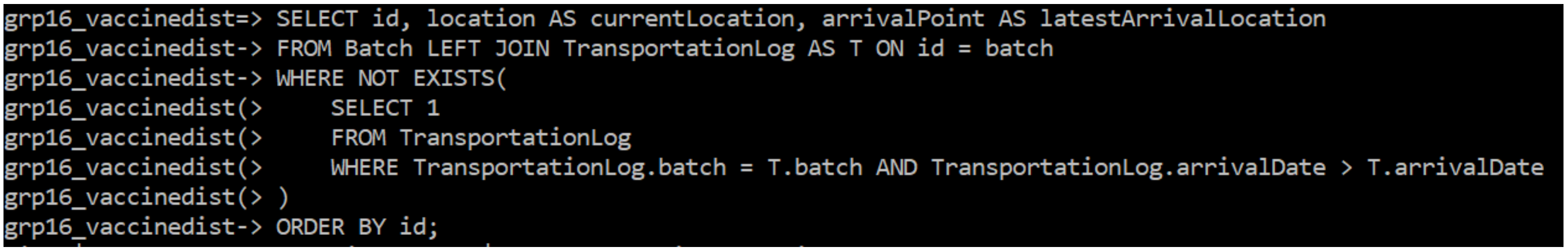
2) We select all employees with role “doctor” who work in hospitals with “HELSINKI” in name and have a shift on Wednesday



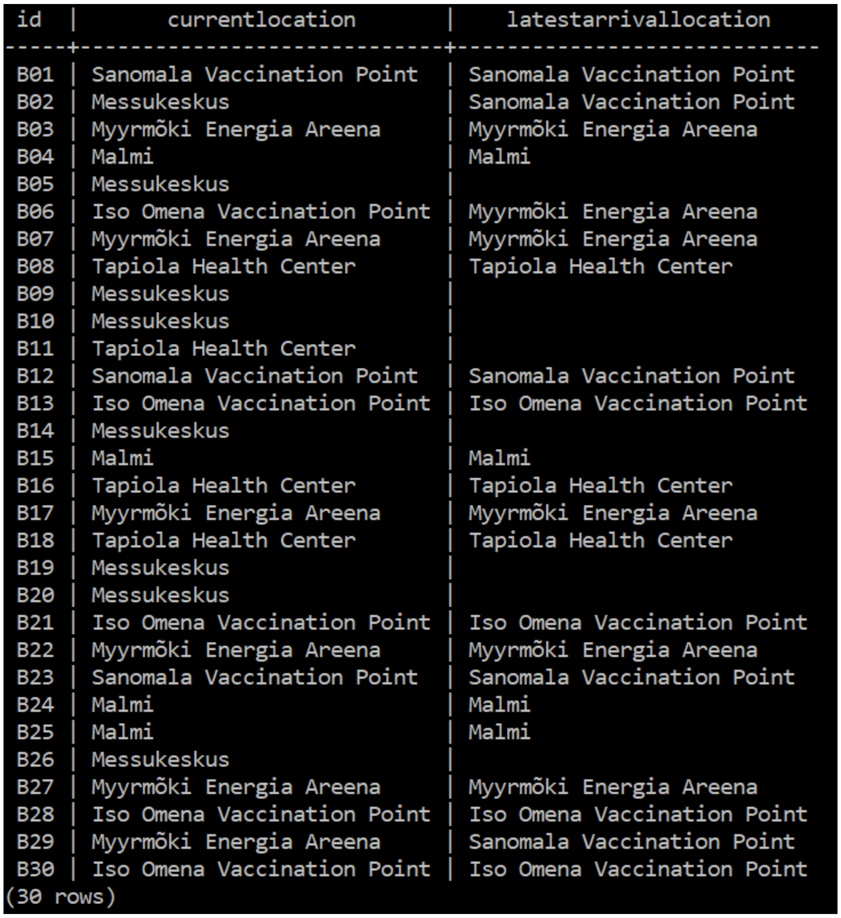
```
grp16_vaccinedist=> SELECT Employee.name AS doctor
grp16_vaccinedist-> FROM Employee, Shift, VaccinationPoint
grp16_vaccinedist-> WHERE Employee.vaccinationPoint = VaccinationPoint.name AND Shift.employee = ssNo AND weekday = 'Wednesday'
grp16_vaccinedist->     AND role = 'doctor' AND address LIKE '%HELSINKI%';
      doctor
------------------
 Rosalia Simonis
 Shaylee Kris
 Hilbert Purdy
 Elnora Greenholt
(4 rows)
```

3)

This query is split into 2 parts, the first one for finding the location for each batch and it’s last location in the transportation log, “currentlocation” is the location in the batch table and “lastestarrivallocation” is the last arrival location in the transportation log table

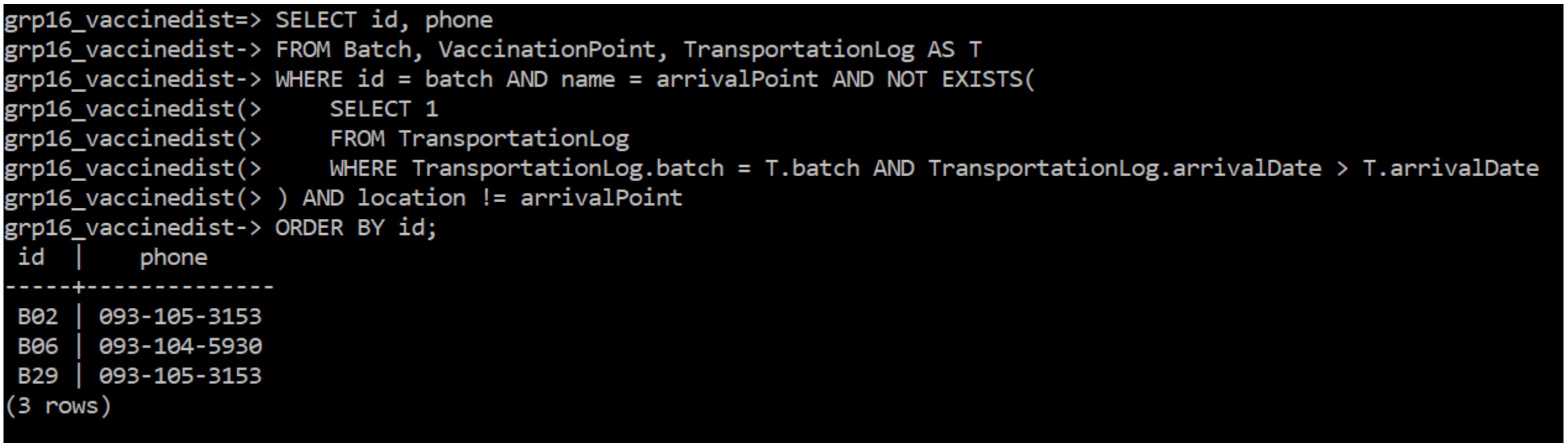


```
grp16_vaccinedist=> SELECT id, location AS currentLocation, arrivalPoint AS latestArrivalLocation
grp16_vaccinedist-> FROM Batch LEFT JOIN TransportationLog AS T ON id = batch
grp16_vaccinedist-> WHERE NOT EXISTS(
grp16_vaccinedist(>     SELECT 1
grp16_vaccinedist(>     FROM TransportationLog
grp16_vaccinedist(>     WHERE TransportationLog.batch = T.batch AND TransportationLog.arrivalDate > T.arrivalDate
grp16_vaccinedist(> )
grp16_vaccinedist-> ORDER BY id;
```



```
 id  |       currentlocation       |    latestarrivallocation
-----+-----------------------------+----------------------------
 B01 | Sanomala Vaccination Point  | Sanomala Vaccination Point
 B02 | Messukeskus                 | Sanomala Vaccination Point
 B03 | Myyrmõki Energia Areena     | Myyrmõki Energia Areena
 B04 | Malmi                       | Malmi
 B05 | Messukeskus                 |
 B06 | Iso Omena Vaccination Point | Myyrmõki Energia Areena
 B07 | Myyrmõki Energia Areena     | Myyrmõki Energia Areena
 B08 | Tapiola Health Center       | Tapiola Health Center
 B09 | Messukeskus                 |
 B10 | Messukeskus                 |
 B11 | Tapiola Health Center       |
 B12 | Sanomala Vaccination Point  | Sanomala Vaccination Point
 B13 | Iso Omena Vaccination Point | Iso Omena Vaccination Point
 B14 | Messukeskus                 |
 B15 | Malmi                       | Malmi
 B16 | Tapiola Health Center       | Tapiola Health Center
 B17 | Myyrmõki Energia Areena     | Myyrmõki Energia Areena
 B18 | Tapiola Health Center       | Tapiola Health Center
 B19 | Messukeskus                 |
 B20 | Messukeskus                 |
 B21 | Iso Omena Vaccination Point | Iso Omena Vaccination Point
 B22 | Myyrmõki Energia Areena     | Myyrmõki Energia Areena
 B23 | Sanomala Vaccination Point  | Sanomala Vaccination Point
 B24 | Malmi                       | Malmi
 B25 | Malmi                       | Malmi
 B26 | Messukeskus                 |
 B27 | Myyrmõki Energia Areena     | Myyrmõki Energia Areena
 B28 | Iso Omena Vaccination Point | Iso Omena Vaccination Point
 B29 | Myyrmõki Energia Areena     | Sanomala Vaccination Point
 B30 | Iso Omena Vaccination Point | Iso Omena Vaccination Point
(30 rows)
```

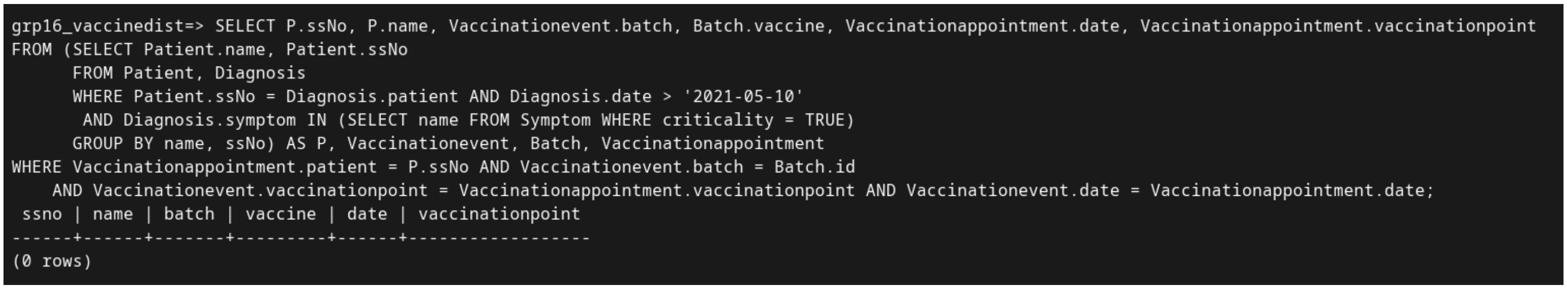
The second part finds the batches with inconsistent location data (where the batch's currentlocation is different than the latest arrival location in transportation log) and lists the phone number of the clinic where the batch should actually be.



```
grp16_vaccinedist=> SELECT id, phone
grp16_vaccinedist-> FROM Batch, VaccinationPoint, TransportationLog AS T
grp16_vaccinedist-> WHERE id = batch AND name = arrivalPoint AND NOT EXISTS(
grp16_vaccinedist(>     SELECT 1
grp16_vaccinedist(>     FROM TransportationLog
grp16_vaccinedist(>     WHERE TransportationLog.batch = T.batch AND TransportationLog.arrivalDate > T.arrivalDate
grp16_vaccinedist(> ) AND location != arrivalPoint
grp16_vaccinedist-> ORDER BY id;
 id  |    phone
-----+-------------
 B02 | 093-105-3153
 B06 | 093-104-5930
 B29 | 093-105-3153
(3 rows)
```

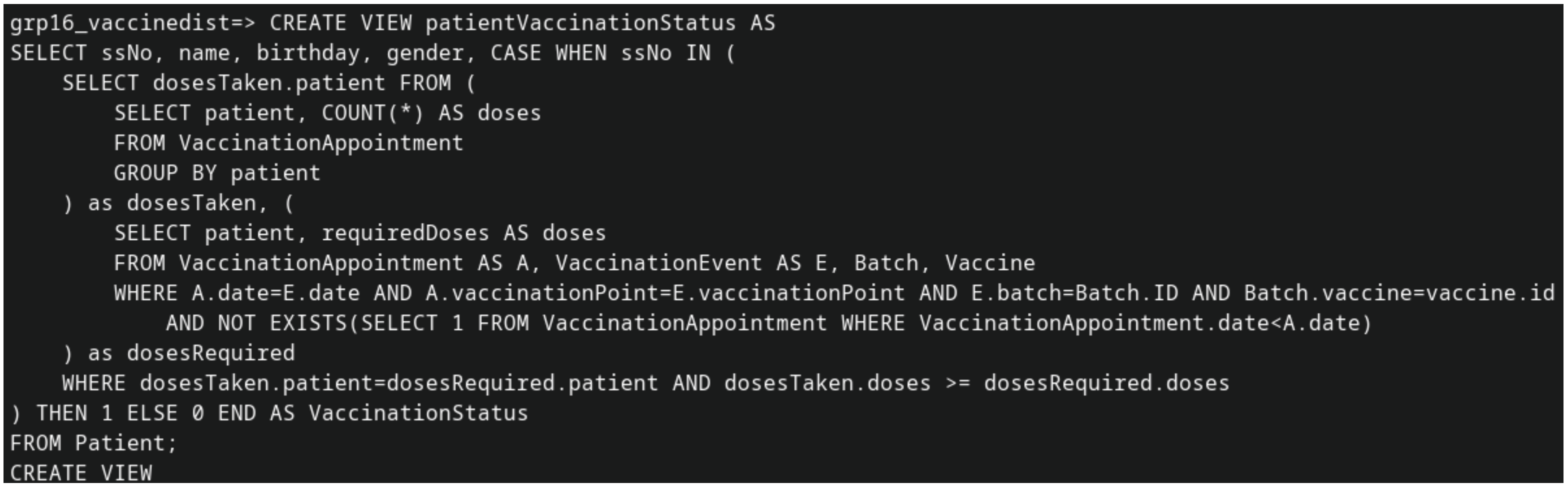
4) This query finds out all the diagnoses of critical symptoms after 10.05.2021, and then finds out which vaccine, and which batch caused each diagnosis through a chain of relations (Diagnosis - Patient -

VaccinationAppointment - VaccinationEvent - Batch)

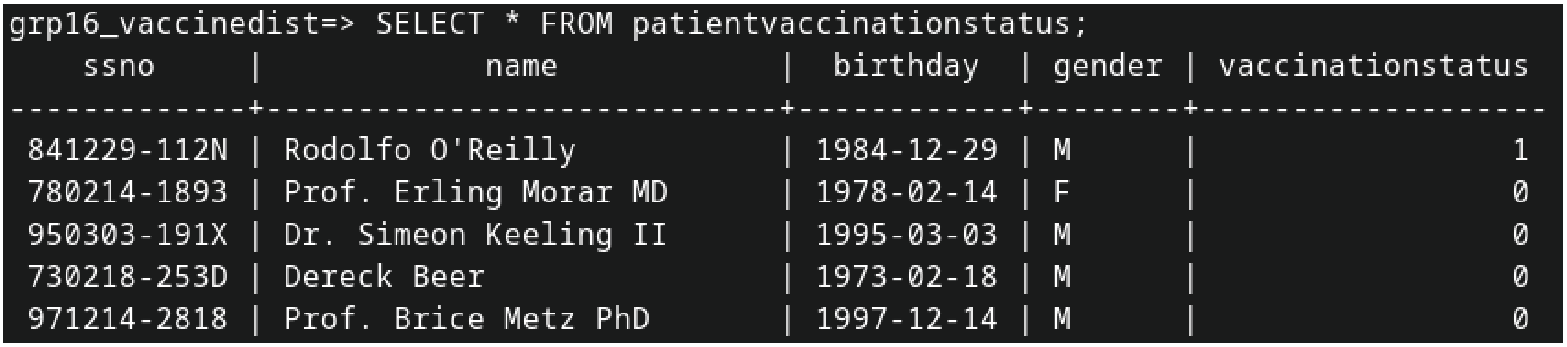


```
grp16_vaccinedist=> SELECT P.ssNo, P.name, Vaccinationevent.batch, Batch.vaccine, Vaccinationappointment.date, Vaccinationappointment.vaccinationpoint
FROM (SELECT Patient.name, Patient.ssNo
      FROM Patient, Diagnosis
      WHERE Patient.ssNo = Diagnosis.patient AND Diagnosis.date > '2021-05-10'
       AND Diagnosis.symptom IN (SELECT name FROM Symptom WHERE criticality = TRUE)
      GROUP BY name, ssNo) AS P, Vaccinationevent, Batch, Vaccinationappointment
WHERE Vaccinationappointment.patient = P.ssNo AND Vaccinationevent.batch = Batch.id
    AND Vaccinationevent.vaccinationpoint = Vaccinationappointment.vaccinationpoint AND Vaccinationevent.date = Vaccinationappointment.date;
 ssno | name | batch | vaccine | date | vaccinationpoint
------+------+-------+---------+------+------------------
(0 rows)
```

5) This query gets the number of required doses for each patient’s first vaccine, finds out how many doses each patient currently had, find all the patients for who second number is bigger than the first, and then puts 1 as vaccinationStatus for those who are in that list, 0 otherwise.

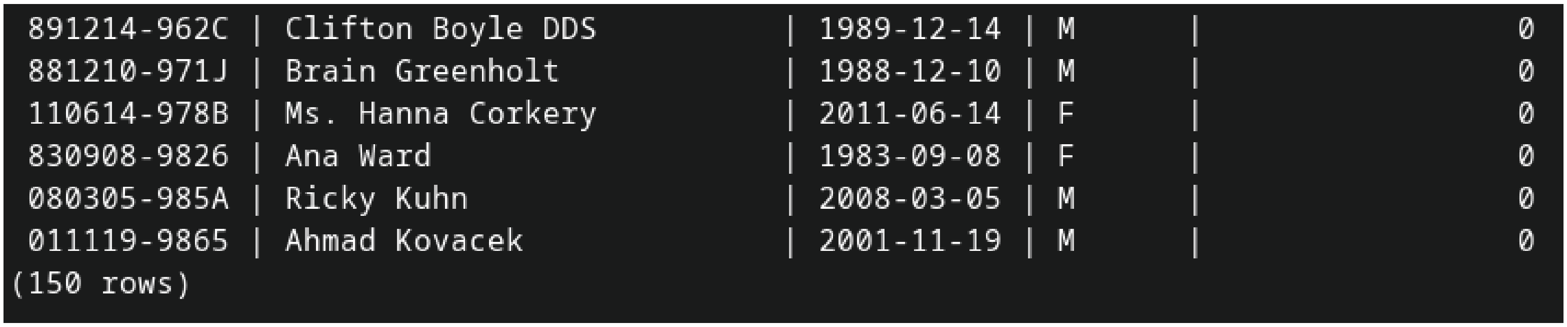


```
grp16_vaccinedist=> CREATE VIEW patientVaccinationStatus AS
SELECT ssNo, name, birthday, gender, CASE WHEN ssNo IN (
    SELECT dosesTaken.patient FROM (
        SELECT patient, COUNT(*) AS doses
        FROM VaccinationAppointment
        GROUP BY patient
    ) as dosesTaken, (
        SELECT patient, requiredDoses AS doses
        FROM VaccinationAppointment AS A, VaccinationEvent AS E, Batch, Vaccine
        WHERE A.date=E.date AND A.vaccinationPoint=E.vaccinationPoint AND E.batch=Batch.ID AND Batch.vaccine=vaccine.id
            AND NOT EXISTS(SELECT 1 FROM VaccinationAppointment WHERE VaccinationAppointment.date<A.date)
    ) as dosesRequired
    WHERE dosesTaken.patient=dosesRequired.patient AND dosesTaken.doses >= dosesRequired.doses
) THEN 1 ELSE 0 END AS VaccinationStatus
FROM Patient;
CREATE VIEW
```

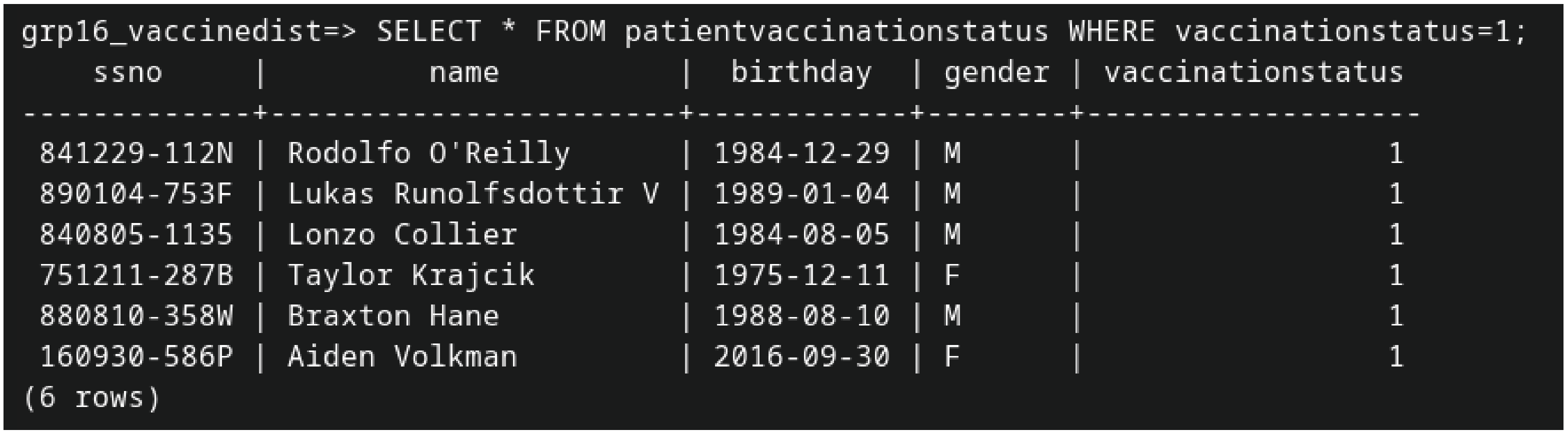


```
grp16_vaccinedist=> SELECT * FROM patientvaccinationstatus;
    ssno     |            name             |  birthday  | gender | vaccinationstatus
-------------+-----------------------------+------------+--------+-------------------
 841229-112N | Rodolfo O'Reilly            | 1984-12-29 | M      |                 1
 780214-1893 | Prof. Erling Morar MD       | 1978-02-14 | F      |                 0
 950303-191X | Dr. Simeon Keeling II       | 1995-03-03 | M      |                 0
 730218-253D | Dereck Beer                 | 1973-02-18 | M      |                 0
 971214-2818 | Prof. Brice Metz PhD        | 1997-12-14 | M      |                 0
```

….

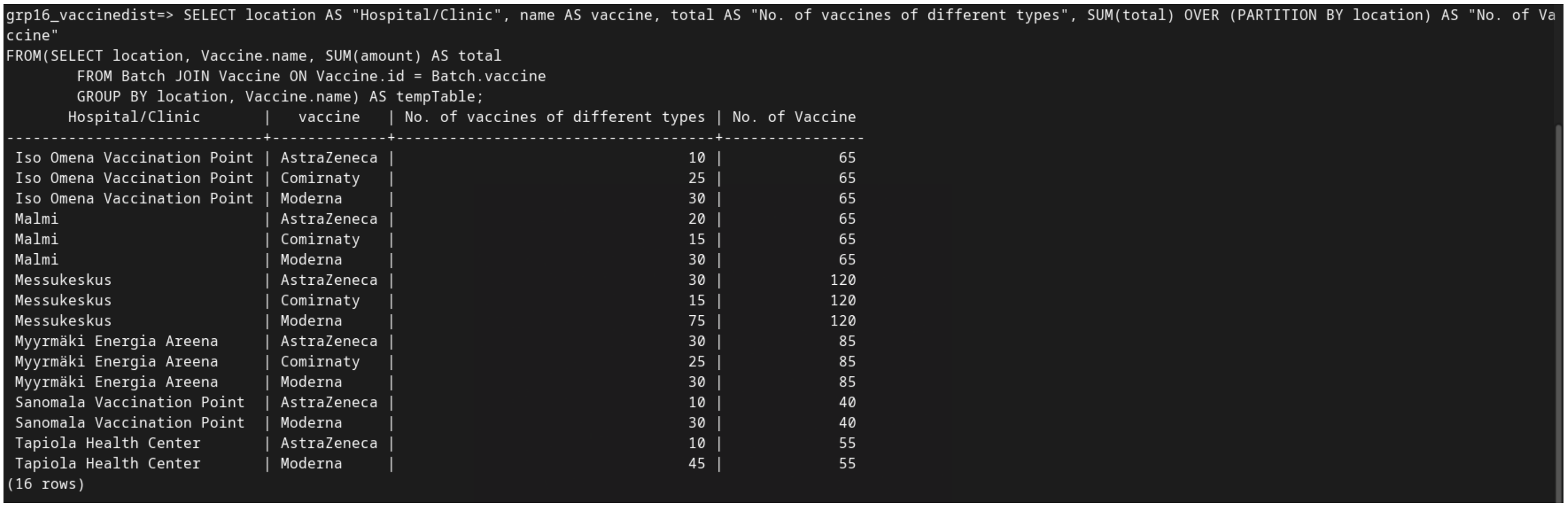


```
 891214-962C | Clifton Boyle DDS           | 1989-12-14 | M      |                 0
 881210-971J | Brain Greenholt             | 1988-12-10 | M      |                 0
 110614-978B | Ms. Hanna Corkery           | 2011-06-14 | F      |                 0
 830908-9826 | Ana Ward                    | 1983-09-08 | F      |                 0
 080305-985A | Ricky Kuhn                  | 2008-03-05 | M      |                 0
 011119-9865 | Ahmad Kovacek               | 2001-11-19 | M      |                 0
(150 rows)
```



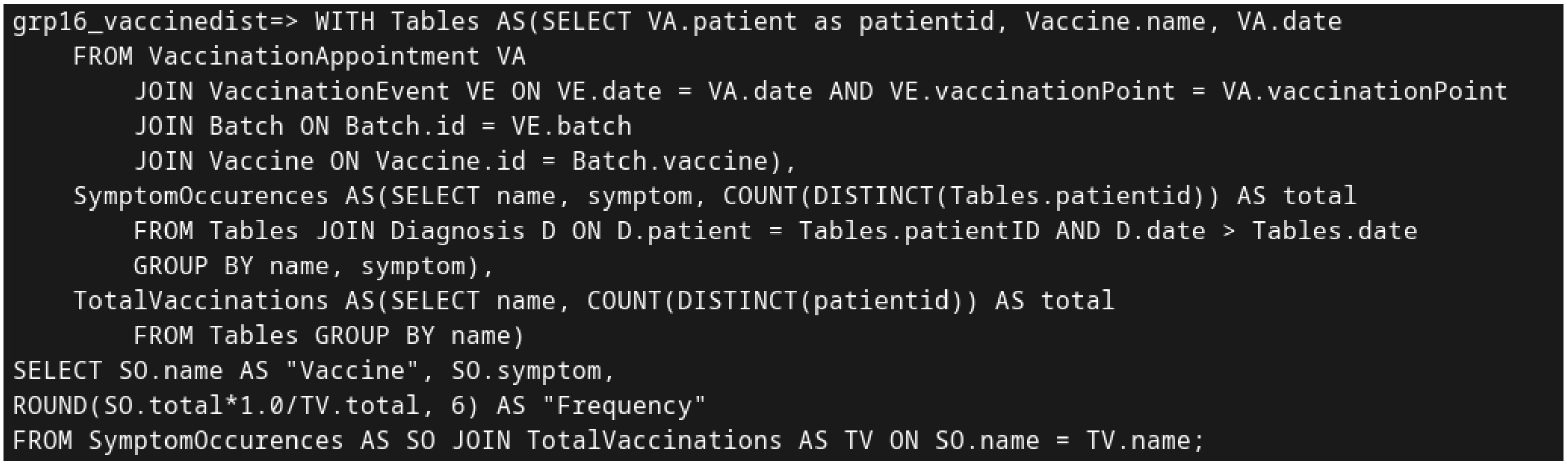
```
grp16_vaccinedist=> SELECT * FROM patientvaccinationstatus WHERE vaccinationstatus=1;
    ssno     |         name          |  birthday  | gender | vaccinationstatus
-------------+-----------------------+------------+--------+-------------------
 841229-112N | Rodolfo O'Reilly      | 1984-12-29 | M      |                 1
 890104-753F | Lukas Runolfsdottir V | 1989-01-04 | M      |                 1
 840805-1135 | Lonzo Collier         | 1984-08-05 | M      |                 1
 751211-287B | Taylor Krajcik        | 1975-12-11 | F      |                 1
 880810-358W | Braxton Hane          | 1988-08-10 | M      |                 1
 160930-586P | Aiden Volkman         | 2016-09-30 | F      |                 1
(6 rows)
```

6) We find the total amount per hospital and vaccine with GROUP BY, and then add the column with the total amunt per hospital using PARTITION BY location.

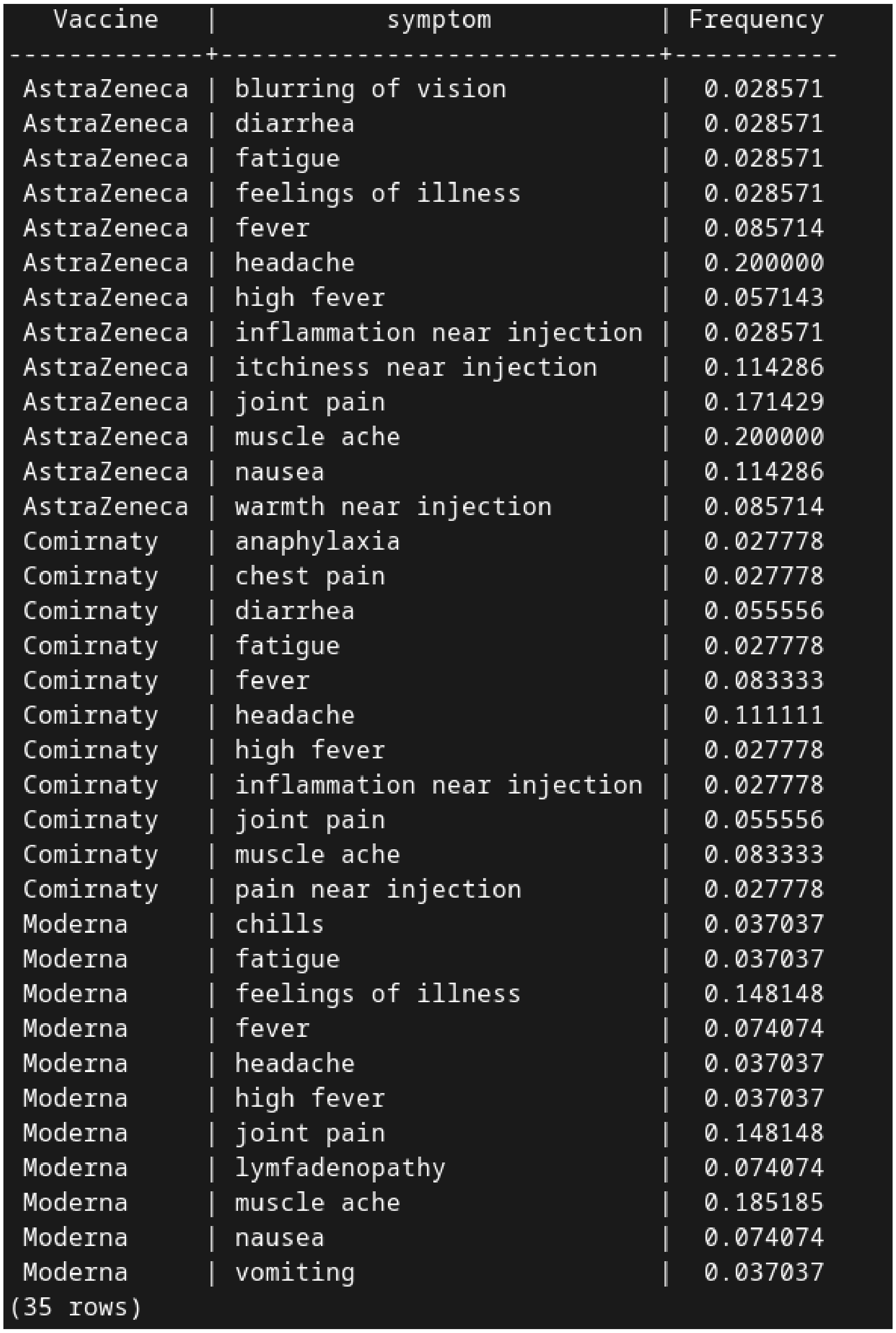


```
grp16_vaccinedist=> SELECT location AS "Hospital/Clinic", name AS vaccine, total AS "No. of vaccines of different types", SUM(total) OVER (PARTITION BY location) AS "No. of Va
ccine"
FROM(SELECT location, Vaccine.name, SUM(amount) AS total
       FROM Batch JOIN Vaccine ON Vaccine.id = Batch.vaccine
       GROUP BY location, Vaccine.name) AS tempTable;
      Hospital/Clinic       |   vaccine   | No. of vaccines of different types | No. of Vaccine
----------------------------+-------------+------------------------------------+----------------
 Iso Omena Vaccination Point | AstraZeneca |                                 10 |             65
 Iso Omena Vaccination Point | Comirnaty   |                                 25 |             65
 Iso Omena Vaccination Point | Moderna     |                                 30 |             65
 Malmi                       | AstraZeneca |                                 20 |             65
 Malmi                       | Comirnaty   |                                 15 |             65
 Malmi                       | Moderna     |                                 30 |             65
 Messukeskus                 | AstraZeneca |                                 30 |            120
 Messukeskus                 | Comirnaty   |                                 15 |            120
 Messukeskus                 | Moderna     |                                 75 |            120
 Myyrmäki Energia Areena     | AstraZeneca |                                 30 |             85
 Myyrmäki Energia Areena     | Comirnaty   |                                 25 |             85
 Myyrmäki Energia Areena     | Moderna     |                                 30 |             85
 Sanomala Vaccination Point  | AstraZeneca |                                 10 |             40
 Sanomala Vaccination Point  | Moderna     |                                 30 |             40
 Tapiola Health Center       | AstraZeneca |                                 10 |             55
 Tapiola Health Center       | Moderna     |                                 45 |             55
(16 rows)
```

7) We find out number of occurrences for every symptom per vaccine, find out total number of doses per vaccine, and compute the frequency.



```
grp16_vaccinedist=> WITH Tables AS(SELECT VA.patient as patientid, Vaccine.name, VA.date
    FROM VaccinationAppointment VA
        JOIN VaccinationEvent VE ON VE.date = VA.date AND VE.vaccinationPoint = VA.vaccinationPoint
        JOIN Batch ON Batch.id = VE.batch
        JOIN Vaccine ON Vaccine.id = Batch.vaccine),
    SymptomOccurences AS(SELECT name, symptom, COUNT(DISTINCT(Tables.patientid)) AS total
        FROM Tables JOIN Diagnosis D ON D.patient = Tables.patientID AND D.date > Tables.date
        GROUP BY name, symptom),
    TotalVaccinations AS(SELECT name, COUNT(DISTINCT(patientid)) AS total
        FROM Tables GROUP BY name)
SELECT SO.name AS "Vaccine", SO.symptom,
ROUND(SO.total*1.0/TV.total, 6) AS "Frequency"
FROM SymptomOccurences AS SO JOIN TotalVaccinations AS TV ON SO.name = TV.name;
```



```
   Vaccine   |           symptom           | Frequency
-------------+-----------------------------+-----------
 AstraZeneca | blurring of vision          |  0.028571
 AstraZeneca | diarrhea                    |  0.028571
 AstraZeneca | fatigue                     |  0.028571
 AstraZeneca | feelings of illness         |  0.028571
 AstraZeneca | fever                       |  0.085714
 AstraZeneca | headache                    |  0.200000
 AstraZeneca | high fever                  |  0.057143
 AstraZeneca | inflammation near injection |  0.028571
 AstraZeneca | itchiness near injection    |  0.114286
 AstraZeneca | joint pain                  |  0.171429
 AstraZeneca | muscle ache                 |  0.200000
 AstraZeneca | nausea                      |  0.114286
 AstraZeneca | warmth near injection       |  0.085714
 Comirnaty   | anaphylaxia                 |  0.027778
 Comirnaty   | chest pain                  |  0.027778
 Comirnaty   | diarrhea                    |  0.055556
 Comirnaty   | fatigue                     |  0.027778
 Comirnaty   | fever                       |  0.083333
 Comirnaty   | headache                    |  0.111111
 Comirnaty   | high fever                  |  0.027778
 Comirnaty   | inflammation near injection |  0.027778
 Comirnaty   | joint pain                  |  0.055556
 Comirnaty   | muscle ache                 |  0.083333
 Comirnaty   | pain near injection         |  0.027778
 Moderna     | chills                      |  0.037037
 Moderna     | fatigue                     |  0.037037
 Moderna     | feelings of illness         |  0.148148
 Moderna     | fever                       |  0.074074
 Moderna     | headache                    |  0.037037
 Moderna     | high fever                  |  0.037037
 Moderna     | joint pain                  |  0.148148
 Moderna     | lymfadenopathy              |  0.074074
 Moderna     | muscle ache                 |  0.185185
 Moderna     | nausea                      |  0.074074
 Moderna     | vomiting                    |  0.037037
(35 rows)
```