

HAP720: PROJECT

RASHMI KONDAKINDI

PHASE 1

In this phase, please write SQL, Python, or other programming languages to integrate records from two source files (SEER and Claims) based on the MPI.

After you finish the integration, please answer the following three questions according to your results.

1. How many patients in the SEER do not have matching records in the Claims?
2. How many patients in the Claims do not have matching records in the SEER?
3. What does this mean? (i.e., why do patients not exist in SEER?)

---- PHASE 1

--- Merging SEER data with MPI data

```
select
x.member_id,x.diagnosis_year,x.diagnosis_age,x.sex,x.race,x.c_diagnosis,y.patient_id
into #temporary
from dbo.hap720_seer x FULL JOIN dbo.hap720_mpi y
on x.member_id=y.member_id

select * from #temporary
```

The screenshot shows the Microsoft SQL Server Management Studio interface. The query window displays the following SQL code:

```
---- PHASE 1
--- Merging SEER data with MPI data
select x.member_id,x.diagnosis_year,x.diagnosis_age,x.sex,x.race,x.c_diagnosis,y.patient_id into #temporary
from dbo.hap720_seer x FULL JOIN dbo.hap720_mpi y
on x.member_id=y.member_id

select * from #temporary
```

The Results window shows the output of the query, which is a table with 7 columns: member_id, diagnosis_year, diagnosis_age, sex, race, c_diagnosis, and patient_id. The table contains 18 rows of data, including some NULL values.

member_id	diagnosis_year	diagnosis_age	sex	race	c_diagnosis	patient_id
sp305446	2019	70	1	1	1742	NULL
sp345735	1992	83	1	1	1573	NULL
sp286783	2018	70	2	1	1605	NULL
sp579106	1986	80	2	1	1724	NULL
sp239104	2020	76	1	3	1639	NULL
sp333918	2008	91	1	1	2030	NULL
sp788730	2011	65	2	1	1748	NULL
sp579714	2006	74	1	2	1951	NULL
sp588812	2024	92	2	1	1515	NULL
sp129267	2022	84	1	3	2051	NULL
sp585534	1990	84	1	2	2301	NULL
sp419790	2003	80	1	1	1650	NULL
sp337795	2014	72	2	2	1599	NULL
sp215981	2014	69	2	1	1584	NULL
sp576005	2012	75	2	1	1640	NULL
sp411237	2015	86	2	2	1741	sp265449
sp616480	1992	91	1	1	1510	NULL
sp286821	1991	80	1	1	2072	NULL
sp45908	1989	92	1	1	2030	NULL

--- joining SEER data and claims_dgns_demo_dx_rnd_removed tables

SELECT

```
x.patient_id,x.claim_year,x.claim_date,x.death_date,x.sex,x.age,x.race,x.diagnosis9,x.
diagnosis10,claim_number,
y.member_id INTO #temporary1
FROM dbo.HAP720_claims_dgns_demo_dx_rnd_removed x FULL JOIN #temporary y
ON x.patient_id=y.patient_id
```

select * from #temporary1

SQL Query Analyzer - LAPTOP-239GV7\HAP720 (LAPTOP-239GV7\source (S1)) - Microsoft SQL Server Management Studio

Query: SQL Query 3 - 3GV7\source (S1) - SQL Query Analyzer - LAPTOP-239GV7\source (S1)

```
select * from #temporary
--- joining SEER data and claims_dgns_demo_dx_rnd_removed tables
SELECT x.patient_id,x.claim_year,x.claim_date,x.death_date,x.sex,x.age,x.race,x.diagnosis9,x.diagnosis10,claim_number,
y.member_id INTO #temporary1
FROM dbo.HAP720_claims_dgns_demo_dx_rnd_removed x FULL JOIN #temporary y
ON x.patient_id=y.patient_id
select * from #temporary1
```

patient_id	claim_year	claim_date	death_date	sex	age	race	diagnosis9	diagnosis10	claim_number	member_id
qp295379	2016	2016-08-21	NULL	2	70	1	20080	NULL	356	NULL
qp402236	2015	2015-04-23	NULL	2	70	1	4019	NULL	107	NULL
qp482215	2017	2017-03-19	NULL	1	52	3	71521	NULL	31	qp683006
qp482215	2017	2017-03-19	NULL	1	52	3	71521	NULL	31	qp683006
qp4313	2017	2017-02-15	NULL	2	74	1	3558	NULL	31	qp671185
qp717799	2017	2017-08-26	NULL	2	89	2	1889	NULL	22	qp36536
qp399212	2016	2016-08-15	NULL	1	76	1	V054	NULL	40	NULL
qp16081	2017	2017-02-09	NULL	2	71	1	7291	NULL	32	NULL
qp127395	2015	2015-10-25	NULL	2	88	5	1519	NULL	197	NULL
qp734589	2016	2016-09-21	NULL	1	88	1	71596	NULL	141	NULL
qp599500	2015	2015-01-17	NULL	2	81	2	2410	NULL	113	qp178676
qp114866	2015	2015-09-23	NULL	2	79	1	5821	NULL	2	NULL
qp588578	2015	2015-02-19	NULL	2	74	1	42769	NULL	88	NULL
qp494778	2015	2015-08-25	NULL	2	75	1	V5811	NULL	42	NULL
qp879395	2015	2015-07-28	NULL	1	75	2	8520	NULL	36	NULL
qp24385	2015	2015-07-25	NULL	2	89	1	73200	NULL	64	NULL
qp229227	2015	2015-01-11	NULL	1	70	1	20280	NULL	145	NULL
qp586022	2015	2015-09-27	NULL	1	79	2	4011	NULL	33	NULL
qp525543	2015	2015-04-15	NULL	2	84	1	3079	NULL	32	qp814230

Query executed successfully. LAPTOP-239GV7 (15.0 RTM) LAPTOP-239GV7\source... HAP720 00:00:38 3,511,943 rows

-----How many patients in the SEER do not have matched records in the Claims?

```
SELECT COUNT(DISTINCT member_id) FROM #temporary WHERE patient_id is null and
member_id is not null
```

SQL Query Analyzer - LAPTOP-239GV7\HAP720 (LAPTOP-239GV7\source (S1)) - Microsoft SQL Server Management Studio

Query: SQL Query 3 - 3GV7\source (S1) - SQL Query Analyzer - LAPTOP-239GV7\source (S1)

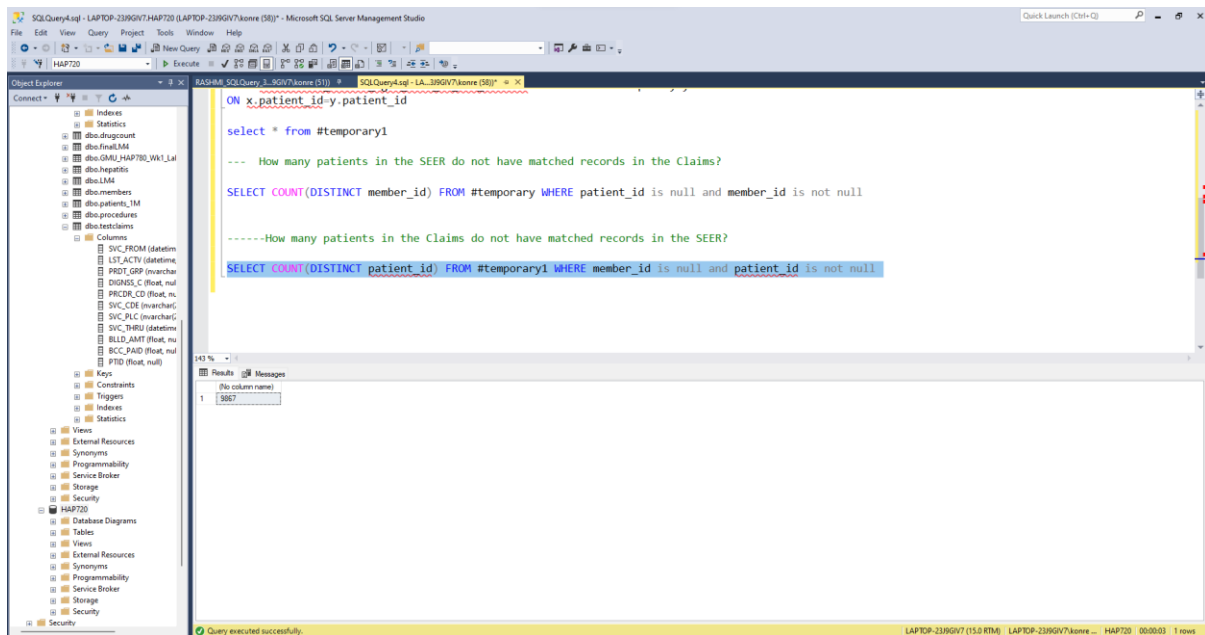
```
SELECT x.patient_id,x.claim_year,x.claim_date,x.death_date,x.sex,x.age,x.race,x.diagnosis9,x.diagnosis10,claim_number,
y.member_id INTO #temporary1
FROM dbo.HAP720_claims_dgns_demo_dx_rnd_removed x FULL JOIN #temporary y
ON x.patient_id=y.patient_id
select * from #temporary1
--- How many patients in the SEER do not have matched records in the Claims?
SELECT COUNT(DISTINCT member_id) FROM #temporary WHERE patient_id is null and member_id is not null
```

(No column name)
30081

Query executed successfully. LAPTOP-239GV7 (15.0 RTM) LAPTOP-239GV7\source... HAP720 00:00:00 1 rows

-----How many patients in the Claims do not have matched records in the SEER?

```
SELECT COUNT(DISTINCT patient_id) FROM #temporary1 WHERE member_id is null and  
patient_id is not null
```



-----What does this mean? (i.e., why do patients not exist in SEER?)

The patient may not be listed in SEER for a number of reasons; the cancer registry was established to compile information on all cancer patients in the SEER database. It's possible that the patient didn't have cancer or didn't match the criteria for inclusion if they aren't recorded in the SEER database. Other factors included errors or omissions during data collection or coding, which resulted in inaccurate or missing data.

PHASE 2

In this phase, please write SQL, Python, or other programming languages to map ICD9CM codes to ICD10CM in your merged file from Phase One. You can find the mappings between ICD9CM and ICD10 in the attached file.

----- PHASE 2

USE HAP720

```
SELECT * FROM dbo.HAP720_icd9_to_icd10
SELECT * FROM #temporary1
```

```
SELECT a.* INTO #temporary2 FROM #temporary1 a JOIN dbo.HAP720_icd9_to_icd10 b ON
a.diagnosis9=b.icd9
SELECT * FROM #temporary2
```

```
SELECT a.*,b.icd10 into #temporary3 FROM #temporary2 a JOIN dbo.HAP720_icd9_to_icd10 b
ON a.diagnosis9=b.icd9
SELECT * FROM #temporary3
```

SQLQuery_project_phase1.sql - LAPTOP-23HGV7-HAP720 (LAPTOP-23HGV7\saone (58)) - Microsoft SQL Server Management Studio

Object Explorer: Connect - HAP720

SQLQuery_project_p... (58) - SQLQuery_project_p... (58)

----- PHASE 2

USE HAP720

SELECT * FROM dbo.HAP720_icd9_to_icd10

SELECT * FROM #temporary1

SELECT a.* INTO #temporary2 FROM #temporary1 a JOIN dbo.HAP720_icd9_to_icd10 b ON a.diagnosis9=b.icd9

SELECT * FROM #temporary2

SELECT a.*,b.icd10 into #temporary3 FROM #temporary2 a JOIN dbo.HAP720_icd9_to_icd10 b ON a.diagnosis9=b.icd9

SELECT * FROM #temporary3

Results: 143 % - Messages

patient_id	claim_year	claim_date	death_date	sex	age	race	diagnosis9	diagnosis10	claim_number	member_id	icd10
cp705584	2015	2015-09-18	NULL	1	78	5	412	NULL	91	NULL	U52
cp571036	2015	2015-05-09	NULL	2	72	1	80701	NULL	32	ap705610	S2238XA
cp373815	2015	2015-07-04	NULL	2	66	1	71941	NULL	33	NULL	M05519
cp68213	2015	2015-02-16	NULL	2	68	1	29570	NULL	29	NULL	F259
cp507782	2015	2015-01-01	NULL	2	71	5	25002	NULL	223	NULL	E1185
cp950349	2015	2015-02-19	2015-02-14	1	74	1	V5052	NULL	110	ap705351	Z4602
cp587556	2015	2015-03-20	NULL	1	85	1	V5861	NULL	122	NULL	Z7901
cp584936	2015	2015-04-21	NULL	1	69	5	7881	NULL	43	NULL	R300
cp552270	2015	2015-03-05	NULL	2	69	1	5932	NULL	61	NULL	N281
cp244028	2015	2015-03-12	NULL	1	71	1	2720	NULL	23	NULL	E7609
cp232007	2015	2015-07-01	NULL	2	78	1	4539	NULL	106	ap716619	I8291
cp854515	2015	2015-03-08	2017-11-03	1	81	1	72979	NULL	74	NULL	M7949
cp584069	2015	2015-01-09	NULL	2	81	1	V7611	NULL	38	NULL	Z1231
cp919389	2015	2015-01-29	NULL	2	87	1	7028	NULL	15	ap571681	L580
cp536886	2015	2015-07-13	NULL	2	72	1	36612	NULL	30	NULL	H25099
cp607793	2015	2015-04-21	NULL	1	87	3	6269	NULL	42	ap833865	N926
cp117430	2015	2015-06-05	NULL	2	80	1	78907	NULL	114	NULL	R1084
cp779725	2015	2015-08-15	NULL	2	71	1	71946	NULL	78	NULL	H25569
cp321637	2015	2015-02-12	NULL	1	75	1	2891	NULL	223	NULL	D62

Query executed successfully.

LAPTOP-23HGV7 (15.0 KB) | LAPTOP-23HGV7\saone... | HAP720 | 00:00:46 | 1.4% 690 rows

PHASE 3

In this phase, we assume that we do not have MPI then we need to apply patient-matching algorithms. Please write SQL, Python, or other programming languages to integrate records from two source files (SEER and Claims) based on the patient's demographic (e.g., age, gender, etc.) and clinical information (e.g., ICD).

--- PHASE 3

USE HAP720

```
---- Merging tables based on patient records since
```

SELECT

```
a.patient_id,a.claim_year,a.claim_date,a.death_date,a.sex,a.age,a.race,a.diagnosis9,
a.diagnosis10,a.claim_number,b.member_id INTO dbo.HAP720_FINAL_Merged
FROM dbo.HAP720_claims_dgn_demo_distorted a FULL JOIN dbo.HAP720_seer_distorted b ON
a.claim_year = b.diagnosis_year AND
a.age = b.diagnosis_age AND a.sex = b.sex AND a.race = b.race AND a.diagnosis9 =
b.c diagnosis
```

```
SELECT * FROM dbo.HAP720 FINAL Merged
```

The screenshot shows the SQL Server Enterprise Manager interface. The top pane displays a query window with the following SQL code:

```

USE HAP720

----- Merging tables based on patient records since

SELECT a.patient_id,a.claim_year,a.claim_date,a.death_date,a.sex,a.age,a.race,a.diagnosis9,
a.diagnosis10,a.claim_number,b.member_id INTO dbo.HAP720_FINAL_Merged
FROM dbo.HAP720_claims_dgn demo distorted a FULL JOIN dbo.HAP720_seer distorted b ON a.claim_year = b.diagnosis_year AND
a.age = b.diagnosis_age AND a.sex = b.sex AND a.race = b.race AND a.diagnosis9 = b.c_diagnosis

SELECT * FROM dbo.HAP720_FINAL_Merged
  
```

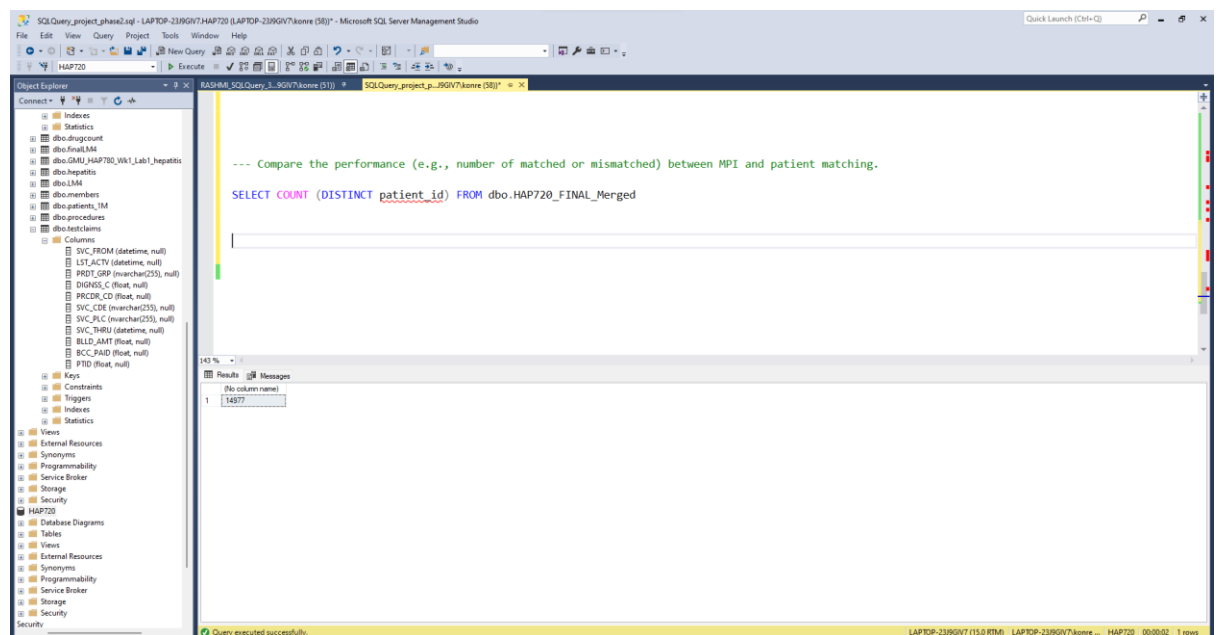
The bottom pane shows the results of the query, displaying a table with 14 columns and 18 rows of data. The columns are: patient_id, claim_year, claim_date, death_date, sex, age, race, diagnosis9, diagnosis10, claim_number, member_id, and 11 empty columns. The data is as follows:

patient_id	claim_year	claim_date	death_date	sex	age	race	diagnosis9	diagnosis10	claim_number	member_id							
1	2017	2017-08-10	NULL	2	72	1	7291	M609	1	NULL							
2	2017	2017-01-10	NULL	2	72	1	7291	M609	3	NULL							
3	2017	2017-08-28	NULL	2	72	1	7291	M609	13	NULL							
4	2017	2017-05-29	NULL	2	72	1	7291	M609	10	NULL							
5	2017	2017-03-21	NULL	2	72	1	7291	M609	103	NULL							
6	2017	2017-05-25	NULL	2	72	1	7291	M7969	28	NULL							
7	2017	2017-01-08	NULL	2	72	1	7291	M7969	83	NULL							
8	2017	2017-06-03	NULL	2	72	1	7291	M7969	19	NULL							
9	2017	2017-03-01	NULL	2	72	1	7291	M7969	61	NULL							
10	2017	2017-08-13	NULL	2	72	1	7291	M7969	12	NULL							
11	2017	2017-02-08	NULL	2	72	1	7291	M7969	31	NULL							
12	2017	2017-04-20	NULL	2	72	1	7291	M7969	93	NULL							
13	2017	2017-11-22	NULL	2	72	1	7291	M7969	57	NULL							
14	2017	2017-07-14	NULL	2	72	1	7291	M7969	30	NULL							
15	2017	2017-12-26	NULL	2	72	1	7291	M7969	47	NULL							
16	2017	2017-03-06	NULL	2	72	1	7291	M7969	79	NULL							
17	2017	2017-01-19	NULL	2	72	1	7291	M7969	164	NULL							
18	2017	2017-02-17	NULL	2	72	1	7291	M7969	84	NULL							

1. Compare the performance (e.g., number of matched or mismatched) between MPI and patient matching.

--- Compare the performance (e.g., number of matched or mismatched) between MPI and patient matching.

```
SELECT COUNT (DISTINCT patient_id) FROM dbo.HAP720_FINAL_Merged
```



2. Explain why the matching is not perfect?

Since neither table had a common unique ID, the matching was insufficient, thus we linked the patient characteristics tables. By first connecting the SEER database with MPI, and then the output table with the claims table, we were able to connect flawlessly in the first instance since we had an MPI table with columns for patient_id and member_id.

