# Understanding the Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| P-0000616 | 58 | Female | Deceased | 1:DECEASED | 19.43 |  | Mod\_Diff | P-0000616-T01-IM3 |
| P-0000616 | 58 | Female | Deceased | 1:DECEASED | 19.43 |  | Mod\_Diff | P-0000616-T02-IM3 |
|  |  |  |  |  |  |  |  |  |

clinicalPatient has 1099 rows while clinicalSample has 1134 rows

The reason for this is that clinicalPatient only contains demographic information pertaining to one patient while clinicalSample varies based on the SAMPLE\_ID column.

This can be demonstrated by the following query

**SELECT COUNT(\*)**

**FROM (SELECT DISTINCT PATIENT\_ID**

**FROM patients) T;**

in MySQL (on the result of the merge of the two tables). The whole table has 1134 rows but finding only distinct Patient IDs, you get only 1099 rows; hence patients can have multiple tumors.

Using a subsequent query,

**SELECT t.Total, COUNT(\*) AS NumberPerCategory**

**FROM (SELECT PATIENT\_ID, COUNT(PATIENT\_ID) AS Total**

**FROM patients**

**GROUP BY PATIENT\_ID) t**

**GROUP BY t.Total**

**ORDER BY t.Total;**

we can find out how many people have multiple tumors:

* 1 tumor = 1068 patients
* 2 tumors = 28 patients
* 3 tumors = 2 patients
* 4 tumors = 1 patient

Finding the MSI status was also indicated as a topic of priority, hence a query

**SELECT MSI\_STATUS, COUNT(MSI\_STATUS)**

**FROM patients**

**GROUP BY MSI\_STATUS;**

indicates the split of people among the different statuses.

* 105 MSI
* 326 MSS
* 1 Inconclusive
* 327 Unknown

Merged tables of clinicalPatient and clinicalSample can be merged with the mutations file to see all of the mutations related to a patient, in this case this can be done in SQL using the following query:

**SELECT COUNT(\*)**

**FROM mutationsdf, patients**

**WHERE Tumor\_Sample\_Barcode = Sample\_ID;**

with Tumor\_Sample\_Barcode coming from mutationsdf and Sample\_ID coming from patients, as this creates a lossless join (mutations has 14628 rows and combining it with patients yields the same number of rows).

For the CNA table, it displays different values, as they do not add up with the values presented in the mutations tables. For example, all of the mutations listed in the mutations table are not listed in the CAN table for patient 'P-0000119-T01-IM3'; in fact, they seem to list different values.

Here are the values in the CNA table as well as what they mean.

Values: -2 = homozygous deletion; -1 = hemizygous deletion; 0 = neutral / no change; 1 = gain; 2 = high level amplification.