The dataset came in a flat schema, with all the records in one table.

METABRIC; Molecular Taxonomy of Breast Cancer International Consortium

<https://www.cbioportal.org/study/clinicalData?id=brca_metabric>

The METAABRIC project was funded by Cancer Research UK, the British Columbia Cancer Foundation and the Canadian Breast Cancer Foundation BC/Yukon.

The dataset provides information about breast cancer patients, including clinical, treatment, and survival data.

**Type of Breast Surgery**

<https://www.cancerresearchuk.org/about-cancer/breast-cancer/treatment/surgery/types-surgery>

Breast Conserving: removing an area of cancer aka lumpectomy

Mastectomy: removing the whole breast

**Cancer Type**

[**https://www.brighamandwomens.org/surgery/surgical-oncology/breast-sarcoma#:~:text=Breast%20sarcomas%20are%20a%20very,and%20lobules%20of%20the%20breast**](https://www.brighamandwomens.org/surgery/surgical-oncology/breast-sarcoma#:~:text=Breast%20sarcomas%20are%20a%20very,and%20lobules%20of%20the%20breast)

Breast Cancer:

Breast Sarcoma: Unlike most breast cancers that begin in the milk ducts, breast sarcomas begin in the connective tissue that supports the ducts and lobules of the breast. Because breast sarcoma cells are more like the connective tissue in the breast than the ductal breast tissue, they act differently than more common kinds of breast cancer. They are often high-grade, that is, their cells are rapidly dividing and very abnormal-looking. They tend to be larger at diagnosis than other types of breast tumors.

Breast sarcoma and age of diagnosis?

**Cancer Type detailed**

<https://www.macmillan.org.uk/cancer-information-and-support/breast-cancer/types-of-breast-cancer>

<https://www.cancercenter.com/cancer-types/breast-cancer/types#:~:text=But%20generally%20speaking%2C%20the%20most,to%20be%20more%20slow%2Dgrowing>.

Breast Angiosarcoma

Breast Invasive ductal carcinoma

Breast Invasive lobular carcinoma

Breast Invasive mixed mucinous carcinoma

Breast Mixed ductal and lobular carcinoma

Invasive breast carcinoma

Metaplastic breast cancer

**Cellularity**

High

Low

Moderate

**Chemotherapy**

Yes

No

**Pam50**

PAM50 is a 50-gene signature that classifies breast cancer into five molecular intrinsic subtypes: Luminal A, Luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, Basal-like and Normal-like. Each of the five molecular subtypes vary by their biological properties and prognoses. Luminal A generally has the best prognosis; HER2-enriched and Basal-like are considered more aggressive diseases. Less common subtypes, such as Claudin-low, Interferon-rich and Molecular Apocrine, have also been identified using other gene expression profiling assays

Normal

Basal

Claudin-low

LumB

LumA

Her2

NC

**Cohort**

1-9

**ER Status measured by IHC**

[**https://www.cancercenter.com/cancer-types/breast-cancer/types#:~:text=But%20generally%20speaking%2C%20the%20most,to%20be%20more%20slow%2Dgrowing**](https://www.cancercenter.com/cancer-types/breast-cancer/types#:~:text=But%20generally%20speaking%2C%20the%20most,to%20be%20more%20slow%2Dgrowing)**.**

Positive

Negative

**Neoplasm Histology Grade**

1-3

**HER2 status measured by SNP6**

Gain

Loss

Neutral

Undef

**HER2 Status**

[**https://www.cancercenter.com/cancer-types/breast-cancer/types/breast-cancer-molecular-types/her2**](https://www.cancercenter.com/cancer-types/breast-cancer/types/breast-cancer-molecular-types/her2)

Negative

Positive

**Tumor Other Histologic Subtype**

Ductal/NST

Mixed

Lobula

Metaplastic

Medullary

Mucinous

Tubular/cribriform

Other

**Hormone Therapy**

Yes

No

**Inferred Menopausal State**

Pre

Post

**Integrative Cluster**

The integrative clusters classification is based on patterns of CNAs across the genome and assigns breast cancers into 11 clusters (clusters 1 to 10 with cluster 4 being subdivided into a 4ER+ and a 4ER- group). The 11 integrative clusters are characterized by variable patterns of chromosome segment alterations that include regional gains and losses.

1-10 exc. 4

4ER+

4ER-

**Primary Tumor Laterality**

Right

Left

**Lymph nodes examined positive**

Integers

**Mutation Count**

Integers

**Nottingham prognostic index**

**Oncotree**

IDC

MDLC

BRCA

BREAST

ILC

IMMC

MBA

PBS

**Overall Survival (Months)**

**Overall Survival Status**

Living

Deceased

**PR Status**

**RadioTherapy**

**Relapse Free Status (Months)**

**Relapse Free Status**

**Sex**

**3-Gene classifier subtype**

**Tumor Size**

**Tumor Stage**

**Patient's Vital Status**

QUESTIONS

1. How many patients underwent each type of breast surgery?
2. What percentage of patients received chemotherapy or radio therapy?
3. What are the most common breast cancer types?
4. What is the survival rate for each cancer type or subtype?
5. What are the average relapse-free months by type of therapy (e.g., chemotherapy vs. no chemotherapy)?
6. How does age at diagnosis vary across different cancer subtypes?
7. **Recurrence Analysis**: What are the recurrence rates by subtype, stage, and type of therapy?

This analysis is aimed at beginners and intermediate learners on the clinical side of breast cancer study. It goes beyond demographic data to the nuances of histology and cancer genomes to get a wholistic and insightful picture of breast cancer. This analytical study offers more depth than is readily available and less jargon found in the literature in easy-to-understand visuals.

The visuals of the missing data indicate MAR, meaning that the missingness in the data is as result of the observed data. This can be inferred from the high correlation in missingness. Due to the multivariate nature of the dataset and its missingness, KNN imputation is the best method to handle the missing data. This is the preferred method as opposed to the traditional mean or mode imputation because of the high correlation in missingness. KNN imputation accounts for relationships in the other variables.

A bar code with text overlay

Description automatically generated with medium confidence

The focus of the dashboard would be on survival based on the patient, clinical, and treatment features.

This drives the design of the data model. The first page of the dashboard would consist of understanding the data. Then, we progress to the different features and survival. Finally, we progress to predictive modelling, survival analysis, and other more advanced techniques.

Fact Table: Patient Outcomes

Includes numeric and measurable data such as:

Overall Survival (Months)

Relapse Free Status (Months)

Tumor Size

Mutation Count

Lymph nodes examined positive

Dimension Tables:

Patients: Patient-specific details

Patient ID, Age at Diagnosis, Sex, Cohort, Inferred Menopausal State

Cancer Type: Cancer-specific attributes

Cancer Type, Cancer Type Detailed, Pam50 + Claudin-low subtype

Treatment: Treatment information

Type of Breast Surgery, Chemotherapy, Radio Therapy, Hormone Therapy

Clinical Details: Clinical characteristics of the tumor

Tumor Stage, Tumor Other Histologic Subtype, Neoplasm Histologic Grade, Cellularity

Survival Status: Information on survival and recurrence

Overall Survival Status, Relapse Free Status

1. What is a Foreign Key?

A foreign key is a column in one table that references the primary key in another table, establishing a relationship between the two tables.

In your case:

Patient ID is the primary key in all your dimension tables because it uniquely identifies each patient.

In the fact table, Patient ID acts as a foreign key, linking the numeric or measurable data in the fact table to descriptive data in the dimension tables.

2. How Patient ID Works as a Foreign Key in the Fact Table

The fact table contains quantitative data (e.g., Tumor Size, Overall Survival (Months)), and it includes Patient ID as a column.

Each Patient ID in the fact table matches a Patient ID in the dimension tables (e.g., Patient Dimension, Clinical Dimension).

This connection allows Power BI to "lookup" descriptive attributes from the dimension tables when analyzing numeric data.

In a star schema, the fact table typically does not have a primary key that uniquely identifies each row because:

Fact Tables Store Measurable Data:

The fact table is designed to hold quantitative, measurable data (e.g., counts, sums, or other metrics) along with foreign keys that reference dimension tables.

It often contains aggregated or repeated data for analysis (e.g., multiple rows for the same patient if they have different events or measurements).

Multiple Records for the Same Entity:

In your case, the fact table might include multiple rows for the same Patient ID because measurable data (e.g., Tumor Size, Survival (Months)) could vary by context or time.

For example, a single patient might have data across multiple rows for different stages of their treatment or diagnostic results.

Role of Foreign Keys:

The fact table uses foreign keys (e.g., Patient ID) to link to primary keys in dimension tables. These links allow analysis and filtering based on the attributes stored in the dimension tables.

In the fact table, Patient ID doesn’t need to be unique—it just needs to connect to the dimensions.