Internship Training

At

Karkinos Healthcare

KARKINOS

“*Use of ETL process for identifying factors related to breast cancer: A Case Study*

By

Nidhi Chaudhary

PG/20/044

Under the guidance of

Prof. Divya Aggrawal

PGDM (Hospital & HealthManagement)

2020-2022

IIHMR
DELHI

International Institute of Health Management Research New Delhi

KARKINOS

Date: 3-May-2022

Internship Completion Certificate

To Whom It May Concern

This is to certify that **Nidhi Chaudhary,** has worked as **"Tech Intern"** with Karkinos Healthcare Private Limited and has successfully completed the internship under the guidance of Arup Ghosh.

Internship Duration: 1st February-2022 to 2nd May-2022.

We wish all the best.

Karkinos Healthcare Pvt. Ltd.

Pooja Sharma

Vice President-HR

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Nidhi Chaudhary** student of PGDM (Hospital & Health Management) from International Institute of Health Management Research, New Delhi has undergone internship training at **Karkinos Healthcare** from **1st February 2022 to 31st April 2022 .**

The Candidate has successfully carried out the study designated to him during internship training and his/her approach to the study has been sincere, scientific and analytical. The Internship is in fulfillment of the course requirements. I wish him all success in all his/her future endeavors.

Dr. Sumesh Kumar
Associate Dean, Academic and Student Affairs
IIHMR, New Delhi

Mentor

IIHMR, New Delhi

Certificate of Approval

The following dissertation titled **“Use of ETL process for identifying factors related to breast cancer : A Case Study”** at **“Karkinos Healthcare** is hereby approved as a certified study in management carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of **PGDM (Hospital & Health Management)** for which it has been submitted. It is understood thatby this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the dissertation only for the purpose it is submitted.

Dissertation Examination Committee for evaluation of dissertation.

Name Signature

Dr. S.B Gogia

Certificate from Dissertation Advisory Committee

This is to certify that **Ms. Nidhi Chaudhary,** a graduate student of the **PGDM (Hospital & Health Management)** has worked under our guidance and supervision. He/ She is submitting this dissertation titled “ **Use of ETL process for identifying factors related to breast cancer : A Case Study** at “Karkinos Healthcare” in partialfulfillment of the requirements for the award of the **PGDM (Hospital & Health Management).**

This dissertation has the requisite standard and to the best of our knowledge no part of it has beenreproduced from any other dissertation, monograph, report or book.

anjali

Institute Mentor Name,

Name : Prof. Divya Aggrawal
Designation: Associate Dean
Organization: IIHMR,Delhi

Organization Mentor :

Dr. Anjali Kulkarni
Designation: Vice president
Organization: Karkinos Healthcare

INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH, NEW DELHI

CERTIFICATE BY SCHOLAR

This is to certify that the dissertation titled … **Use of ETL process for identifying factors related to breast cancer : A Case Study…**…….. and submitted by (Name) …**Nidhi Chaudhary** ….. Enrollment No. …PG/20/044……under the supervision of ……**Prof. Divya Aggrawal** ………. for award of PGDM (Hospital & Health Management) of the Institute carried out during the period from .1st February 2022. to ..30th April 2022…embodies my original work and has not formed the basis for the award of any degree,diploma associate ship, fellowship, titles in this or any other Institute or other similarinstitution of higher learning.

Signature

FEEDBACK FORM

Name of the Student: Nidhi Chaudhary

Name of the Organization in Which Dissertation Has Been Completed:

Karkinos Healthcare . Mumbai

Area of Dissertation: ***Use of ETL process for identifying factors related to breast cancer : A Case Study***

Attendance: 99%

Objectives achieved: Completed 3 months internship in the K cloud team . Nidhi was part of a team working on the clinical data and building analytics platform in Karkinos and has completed all essential training like SNOMED , NLP tools , Google analytics for the project works assigned. Presented in the weekly technology club .

Deliverables: 1 Data team project workflow for the pipeline

2 Data analysis summary document

3 Clinical validation test report

4 Mapping between SnoMED and clinical data

5 Requirement document for the new Clinical pathways

Strengths: Good understanding of the healthcare domain and ready to learn new skills and tools. Quick learner and timely submission of project deliverables.

Suggestions for Improvement: Can have more research mindset and read related new developments in the field . Should develop skills to write and publish articles related to work .

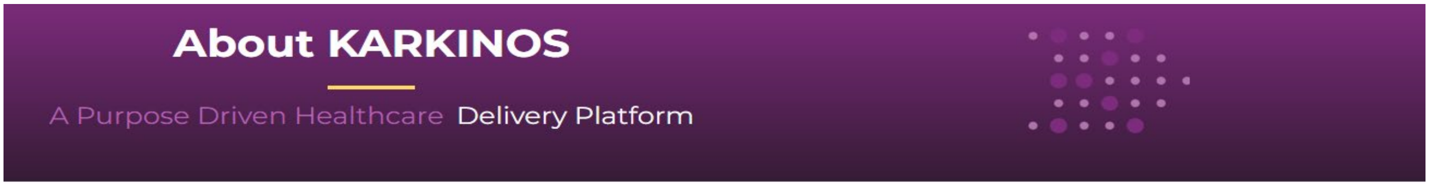
Suggestions for Institute (course curriculum, industry interaction, placement, alumni):

Can have more close industry and academia collaboration

Signature of the Officer-in-Charge/ Organisation Mentor (Dissertation)

Date:06-06-2022

Place:Karkinos Healthcare, Bangalore



About KARKINOS

A Purpose Driven Healthcare Delivery Platform

Every year, India is predicted to have 2.25 million cancer cases, which doubles every 10 years. Because of an inaccessibility to standardized cancer care, three-quarters of these malignancies are discovered late in their progression, with significant fatality rates. While Indians in the West are now at the forefront of the latest medical research, India as a nation is lagging behind in investigating and treating the disease.

Karkinos Healthcare intends to solve these challenges with an end-to-end technology platform that intercepts the cancer care continuum; a medical centre for complex cancer treatment; and a research centre that uses technologies like sensors, genomic information, synthetic biology , and AI to evaluate data and develop affordable cancer interventions. Karkinos Health's guiding idea is the democratisation of cancer care through collaboration with current health practitioners, researchers, and technologies.

Karkinos Healthcare is an oncology-focused platform that includes a knowledge network with surgical skills ,medical procedures, a digital pathology centre, and dispersed care centres. The majority of the accumulated economic interest will be used to support comparable societal purposes as well as research, education, and patient care.

Mr. R Venkataramanan, Advisor to Reliance Industries Limited's Chairman and former Managing Trustee of the Tata Trusts, launched KARKINOS HEALTHCARE (KH). The company wants to create an end-to-end technology-driven oncology-focused managed healthcare platform that ensures that nearly no one is denied care due to a lack of access or affordability.

The design and implementation will be done using tailored cancer care solutions as a one-stop shop in terms of experience, answering key market demands for this specialised health care. It will leverage technology and Al Based Continuous Feedback to tailor care to our specific needs, with the results being scaled out across India and beyond.

OUR VISION IS DRIVEN BY THE 4D'S:

Detect & Diagnose – Establishment of participatory systems and near-home care, Genomic research as a basis for prevention, Invention and game-based outreach for early diagnosis and wellbeing

Data & Research -Contribute to Atmanirbhar Bharat via drug discovery research and treatment innovation, large-scale screening, and longitudinal data to construct strong AI/ML analytics, predictive models, and clinical decision support systems based on real-world evidence.

Deliver-Managed health-care delivery – Annually, more than 2 million patients are serviced, and more than 10 million patient hours are saved.

KH will spend the next 18 months laying the groundwork for the following components: • a cutting-edge technology platform selected for oncology. • fifty Level 4, fifteen Level 3, five Level 2, and one Level 1 centres, as well as a knowledge network with medical procedures, surgical skills, and a digital pathology centre. The majority of the founders' economic interests will be kept in a Charitable trust, and the proceeds will be used to support comparable societal goals and to improve education, research ,and patient care.

In the state of Kerala, a proof-of-concept implementation of the KH roadmap is proposed. For early detection and day care services, a Memorandum of Understanding was signed with Mar Baselios Medical Mission Hospital, and a Memorandum of Understanding with Chottanikkara Medical Relief Society is in the final stages. A pipeline of 12-15 brownfield assets has been identified. Under the DBFO (Design, Build, Finance, and Operate) investment model, an Expression of Interest was submitted to the Manipur Government for the establishment of a state-of-the-art 150-bed multi-specialty cancer institute with complete infrastructure and solid technical support.

Use of ETL process for identifying factors related to breast cancer : A Case Study

By Nidhi Chaudhary

JUNE 21, 2022

KARKINOS HEALTHCARE
Bangalore

ACKNOWLEDGEMENT

First and foremost, we would like to thank **Almighty,** who gave us strength and courage to accomplish our dreams and kept us upright even in the deepest of our sorrows and pain.

I would like to thank our beloved parents and all our family members for their un-conditional love and immense support throughout the journey of our life.

I would like to express my sincere gratitude towards my guide **Dr.Anjali Kulkarni ,** VP-Clinical Informatics, who helped me immensely throughout the tenure of my dissertation. She inspired me greatly to work in this project with her valuable guidance, support, interest, encouragement, involvement and advice.

I would like to take immense pleasure to express my sincere and deep sense of gratitude to my mentor **Prof. Divya Aggrawal** (Associate Dean Academics) for her valuable suggestions, motivation and guidance throughout the course of project work. Without her guidance and encouragement, it would not have been possible to complete this study.

I would like to thank **Arup Ghosh,** Vice President - Projects Engineering and whole Karkinos Healthcare team for allowing me to experience such great opportunities.

<u>Use of ETL process for identifying factors related to breast cancer : Case Study</u>

ABTRACT

- *Introduction*: Data analytics can be very beneficial in cancer care An oncologist spends a significant amount of his/her time reading the medical history of the patients, to solve this problem data analytics comes in handy. It can help understand the association between different risk factors of cancer and help in the treatment of cancer . The data analysis process comprises numbers of steps . The major step is ETL (Extract, Transform, And Load) . Since ETL consumes up to 80% of the time of the data analysis process, so it's important to carry out it efficiently and quickly .

- *Aim* :To study use of ETL process for identifying factors related to breast cancer : A Case Study

- *Methodology:* This is a cross-sectional study conducted to address the influence of different factors such as demographic or clinicopathological on the status of patient . The breast cancer dataset used for analysis is from Kaggle. Chi square test was performed on excel to study the association .

- *Result :*1. Out of 334 patients, maximum number of patients had HER2 negative and were alive (59.2%) followed by patients who were dead and had HER2 negative (17.9%) . No patient had HER2 positive and unknown patient status ($p<0.05$),

2. Maximum number of patients had tumor grade 2 and were alive (43.1%) followed by Grade 3 (17.9%)($p>0.05$) .
3. maximum number of patients belonged to the age group (50-64) and were alive (38.9%) followed by age group (25-49) (19.4%)($p<0.05$) .

Discussion:

1. We found that HER2 status has clear influence on the patient status in breast cancer patients . Similar results were obtained in Liang et.al., study , they found that HER2 status had a clear influence on overall survival in patients with breast cancer.
2. In the present report , we found that tumor grade had no clear influence on the patient status which is different from the results obtained in Ablavi A et.al., study , they found that there was a significant association between histologic grade and breast cancer subtypes .Such difference is due to regional and cultural differences
3. Age is a demographic factor which has a clear influence on the patient status in breast cancer patients in our study but it differs from Ablavi A et.al, study where they found that there was no significant association between age and breast cancer .

.

LIST OF FIGURES

LIST OF TABLE

List of Abbreviations
ETL – Extract, Transform and Load
CDSS – Clinical Decision Support System
KH -Karkinos Healthcare
SQL – Standard Querying Language
GCS- Google Cloud Services
SCD – Slowly Changing Dimension

Background – what's cancer, why cancer data?

Cancer is defined as a category of disorders characterized by abnormal cell proliferation with the ability to infiltrate and spread to other sections of the body. Benign tumors, on the other hand, do not spread. A lump, unusual bleeding, a persistent cough, unexplained loss of weight, and a shift in bowel motions are all possible indications and symptoms. Cancer is a huge problem for a country like India. The number of persons suffering from the disease is believed to be over 2.25 million, with over 11,57,294 new cancer cases being reported each year. Cancer is responsible for 7,84,821 deaths. Breast cancer is the most commonly diagnosed cancer in women **(24.2%**, i.e. about one in 4 of all new cancer cases diagnosed in women worldwide are breast cancer). These statistics are evident enough that a concrete and effective solution is needed to deal with this problem. The cancer care sector has historically generated huge volumes of data,while the majority of data is still saved on paper, the current tendency is for massive volumes of data to be quickly digitized. These large amounts of data offer the promise of facilitating a wide variety of medical functions, driven by statutory requirements and the ability to enhance the quality of healthcare services while lowering costs.. Using analytics on this data can help discover numerous insights regarding patterns of disease risk factors, influence of certain risk factors and infective treatment. Organizations like karkinos healthcare are desirous to build an end-to-end technology-driven oncology-focused managed healthcare platform.[1]

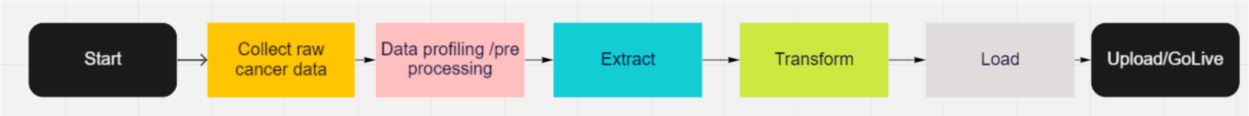
When dealing with big data ETL is a crucial part of the process, considering it consumes 80% of the time.

Introduction:

Breast cancer arises in the lining cells (epithelium) of the ducts (85%) or lobules (15%) in the glandular tissue of the breast. The malignant development is initially contained inside the duct or lobule ("in situ"), where it often exhibits no symptoms and has a low risk of spreading (metastasis).Breat Cancer care is a great illustration as to how the three vs of data variety, velocity (the rate at which data is generated), and volume, are all inherent characteristics of the data it generates. Speaking of velocity **every** day, **every** hour, **every minute someone** is getting **diagnosed** with **cancer** adding up to this ‘big data ‘, every patient’s cancer journey is different (variety) and lastly, around 2.25 million people in India are suffering from cancer which is sufficient to understand the volume of data present. The biggest question is what to do with the ginormous data? The answer is data analysis. Data analytics can be very beneficial in cancer care An oncologist spends a significant amount of his/her time reading the medical history of the patients, to solve this problem data analytics comes in handy. It can summarize the patient’s medical history highlighting the major events and can also help in creating a clinical decision support system. CDSS helps doctors to avoid medical errors and saves a lot of time in the long run. Systems can also be trained using unsupervised and supervised machine learning to suggest treatment and generate alerts in cases of errors . When dealing with patient data two things are very crucial and non-negotiable - firstly : this data is highly confidential and patient’s privacy should not be sacrificed and secondly , no matter how many processes data goes through the data quality and authenticity should be maintained. The data analysis process comprises numbers of steps . The major step is ETL (Extract, Transform, And Load) . ETL takes up to 80% of the time of the data analysis process .so , its really important to do it efficiently and quickly .

ETL Process Steps :

Figure 1: ETL process



- The first step for any ETL process is to collect raw data . For cancer care studies this data is provided by the hospitals or physicians so as to get insights. Insights can be reason of loss of follow up ,survival analysis etc. .The data should be authentic and credible and that should be maintained throughout the process . The data should not have any personal information of the patients.
- Second step of this process is Data profiling which means to understand the structure from the source's information . Data profiling consists of the following tasks :

Column Null Proportion: This operation shows the null ratio % for each table column; it's essential when you need to alter these sorts of numbers into something more relevant.

- *Column Profile Pattern*: Based on regular expressions, it displays the pattern or group of patterns that contain the column's contents

Candidate Key: A profile that assesses if a column or combination of columns is suitable to act as a key for the specified table, including whether another candidate's key exists and whether there are any violations. Distribution of Column Lengths: It shows the row proportions for each column with a specified pattern, as well as the min and max column lengths. Column Value Distribution: The number of various entries for each table column, and also the proportion of rows that comprise them, are displayed in this column.

Column Statistics Profile: Only works with numeric data (integer, float, decimal, and so on) and dates in the Column Statistics Profile (dates only allow minimum and maximum calculations). Numerical computations are used to determine the maximum, minimum, standard deviation and average

- *Functional Dependency*: A Functional Dependency profile depicts how entries in one column (the dependent column) are influenced by values in another column or set of columns (the determinant column). This assessment can also help you spot data problems, such as inaccurate numbers. Take a look at the link between the State and the Postal Code columns. In this profile, the very same Postcode should have always had the same state, however the profile detects violations of the dependence.

Fact table and dimensions data storage

Dimensions provide context to fact tables and, as a result, to all metrics in the data warehouse. Dimension table characteristics play a significant role in data warehouses. Because they are the source of almost all useful limitations and report labels, they are crucial to making the data warehouse usable and understandable. In numerous respects, the data is intriguing.

A warehouse's size is determined only by its dimension characteristics .The terms "dimensions" and "hierarchy" are interchangeable in some business intelligence systems (especially in multidimensional databases). As a result, a geographical dimension encompasses continents, nations, regions, provinces, and municipalities at various levels..

B. Facts

The ETL process is nowhere as plainly visible as in the fact table; this phase of the process involves a significant number of actions, including data quality verification, transformation to meet data warehouse needs, and finally loading into the destination.

- The third step is Extraction . Data received can be from different sources hence, different formats . Data can be unstructured such as raw patient notes or structured data .All this data should be stored in a warehouse . Example : Google Cloud Storage (GCS) etc
- Cleaning and transformation are the fourth and last processes in the data warehouse transformation process. The following actions were taken to complete the needed conversions.

Null conversions : In most circumstances, transformations are connected to null conversion; a null value simply indicates that a value was not stored in the database; as a best practise, a default value should be inserted instead of a null value whenever a row lacks a value for a column. This type of transition has occurred in a number of integration efforts. A default value was obtained for each column that had no value, and the null value was substituted with this. For example, the null value of the age attribute was updated with a zero, which implies "no registered age" at the time.

Data Conversion: Transforming data types to adapt and change them to better match the data warehouse is a regular activity. We may use SQL standard for this step (the natural language for querying databases)

- The fifth and sixth step is loading and going live with the data and insights . It has few subparts

Uploading and updating

Before uploading data into the data warehouse, an activity called as updating must be completed. During this process, two sorts of data will be verified: those that have been updated and those that have been added. As a result, two procedures will be carried out:

- For altered data, such as a SCD, after the rows that have been modified are discovered, they are marked as non valid, giving you a time stamp for a certain date. - The new ones are uploaded to the target table.

When dealing with cancer care data, this study highlights a crucial feature of an ETL procedure.

ETL process differs on the basis of data as well . Unstructured data involves supervised machine learning training during the transformation process where all the concepts are coded with a UMLS concept ID and extracted using data pipeline through Airflow tool. The training is done on tools like MedCat trainer

Literature review:

A descriptive study was done by wullianullar r et.al., to describe the promise and potential of big data analytics in healthcare.this study found big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. [2]

Another descriptive study was done by ashwin b et.al.,to discuss recent research which targets utilization of large volumes of medical data while combining multimodal data from disparate sources potential areas of research within this field which have the ability to provide meaningful impact on healthcare delivery are also examined.[3]

Descriptive research was done by roblero j et.al., to study important considerations of an ETL process in cfe's health and security area. The report concluded that in order to create ETL's projects with high performance is important to follow a methodology without avoiding any step: extract, clean, transform, integration and update data, must be done as an automatic process for each of them.[4]

A cross sectional study was done by ablavi a et.al., to study to evaluate the expression of ER, PR, HER2, and molecular subtypes of breast cancer receptors in Togolese patients and to establish the correlation between clinical and histological data and molecular types. The report found that There was a significant association between stage and breast cancer subtypes, histologic grade, and subtype but no correlation was found with age, menopausal status, and tumor size.[5]

Another cross-sectional study was done by Liang c et.al., to analyze the effects of human epidermal growth factor receptor-2 (HER2) status on the prognosis of male breast cancer (MBC). The study concluded that HER2 status had a clear influence on overall survival in patients with MBC.[6]

A cross-sectional study was done by fei ji et.al., to analyze the survival of patients with breast cancer diagnosed after a prior cancer and identify risk factors of breast cancer death in this population and concluded that most of the breast cancer risk factors were similarly distributed among the four major breast cancer subtypes; commonality is predominant.[7]

A retrospective case control study was done by yan l et.al., to investigate the clinical and pathological features and risk factors for primary breast cancer treated at our center in order to provide a reference for the prevention and treatment of breast cancer in the Zhuhai-Macao region and concluded that the differences in age at disease diagnosis, age at menarche, and history of surgery for benign breast lesions were statistically significant.[8]

A descriptive study was done by vijayalakshmi m et.al., to study the Dynamic multi-variant relational scheme-based intelligent ETL framework for healthcare management and concluded a method that improves the performance in ETL at different test cases.

Methodology :

Type of Analysis : This is a cross-sectional study conducted to address the use of ETL process for identifying factors related to breast cancer

Type of data: Secondary Data

Source of data : Kaggle, a Google LLC subsidiary, is an online platform of data scientists and deep learning practitioners that provided the data for this investigation. Kaggle is a web-based data science platform that allows users to search and post data sets, explore and construct models in a web-based data science environment, collaborate with other data scientists and deep learning experts, and compete in data science challenges. 334 breast cancer patients are represented in the data

.**Data Analysis** : For data analysis different tools were used such as

Google sheets- Data profiling

WinMerge-Data validation

Microsoft Excel – Chi square test and visualizations

Statistical Analysis : Chi square test

Operational definition : The extract, transform, and load (ETL) tool is used to extract data from a variety of heterogeneous data sources, process it, and then load it into a data warehouse. Any data warehouse, data mining, or business intelligence system is built on top of it. In simpler terms , Data integration is referred to as ETL. This process entails extracting the appropriate data from the data source, processing the data, and loading the data into the data warehouse according to the data warehouse's predetermined model.

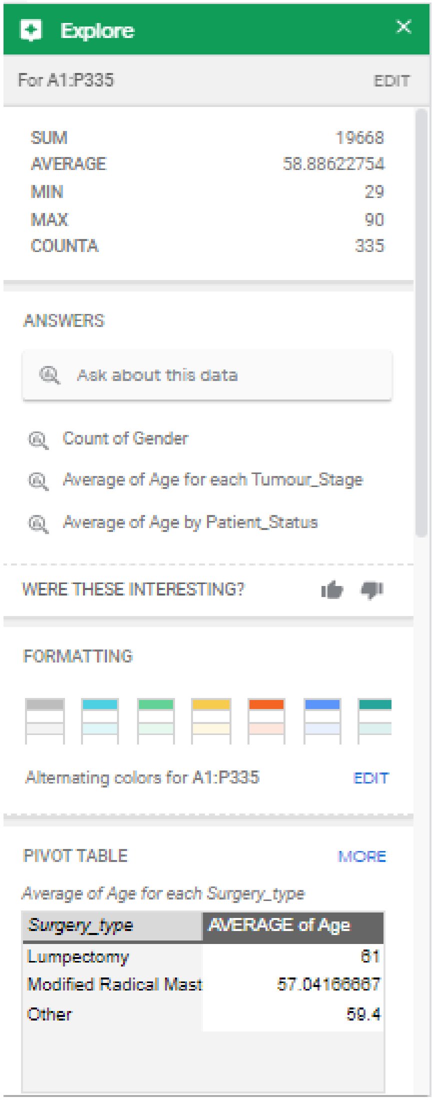
Process : Step 1 : Data was downloaded from Kaggle and it was made sure that it had no personal information about the patients. This raw data file was saved on our local system .This ensures that our raw data files are not hampered during the process and we need be we can always start afresh .

Step 2: Data profiling : For this step we used two different tools . First one was google sheets : we uploaded the data into the google sheets and used the explore function so as to get analysis of each and every column (refer figure 2a and 2b). For more detailed insights about columns we looked at column stats.(refer figure 3a and 3b)

We searched for a specific kind of analysis or chart for any column in explore function using search function .For our analysis we are concerned about 4 columns i.e Age , tumor grade, patient status and HER2 status .

Figure 2a : Explore function used for column-Age showing sum ,average ,count, pivot table etc

2b: Autogenerated analysis for column -Age .



For A1:P335 EDIT

SUM	19668
AVERAGE	58.88622754
MIN	29
MAX	90
COUNTA	335

ANSWERS

Count of Gender

Average of Age for each Tumour_Stage

Average of Age by Patient_Status

WERE THESE INTERESTING?

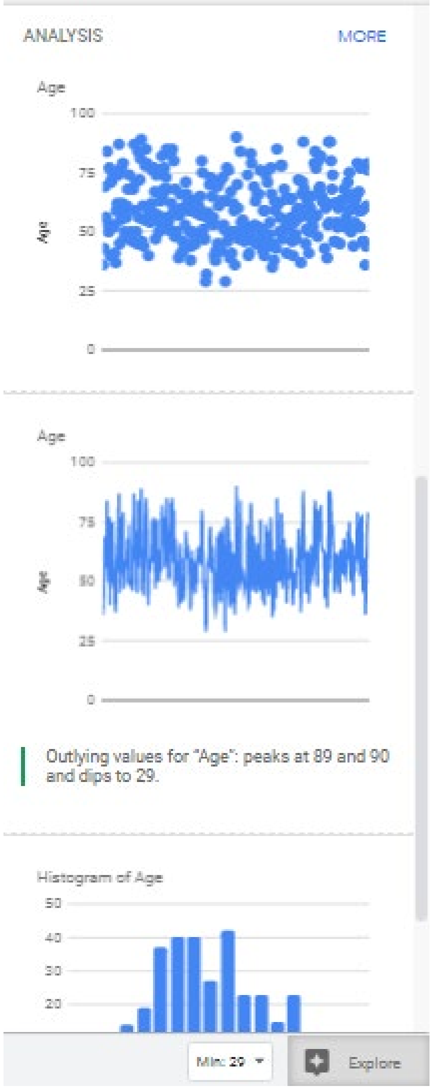
FORMATTING

Alternating colors for A1:P335 EDIT

PIVOT TABLE MORE

Average of Age for each Surgery_type

Surgery_type	AVERAGE of Age
Lumpectomy	61
Modified Radical Mast	57.04166667
Other	59.4

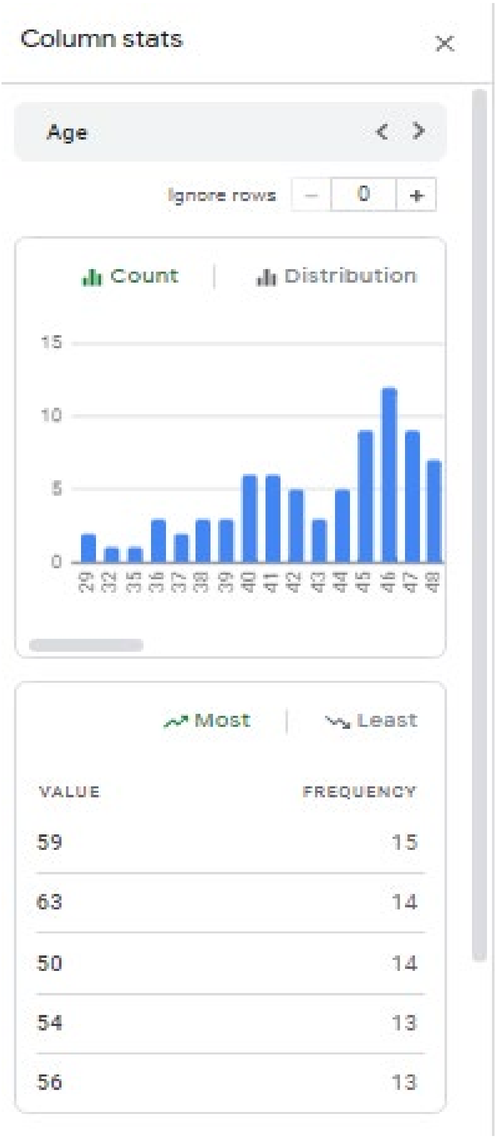


Outlying values for "Age": peaks at 89 and 90 and dips to 29.

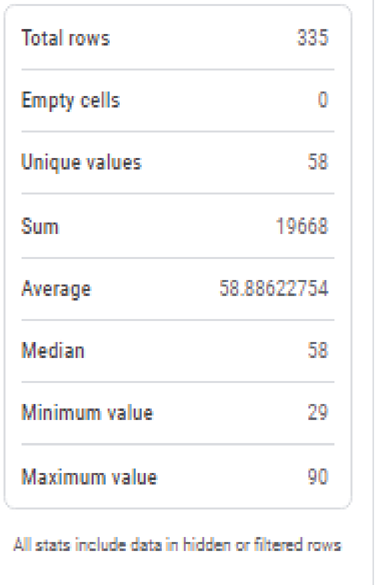
Column Stats is another function of google sheets that gave us the information about the type of values present in the column and the number of null values i . (refer figure 3a and 3b)

Figure : 3a: Column Stats for the column -Age. It shows the count and distribution of the values. It also gives information about the frequency of the values.

3b: Column stats inform about the empty cells, unique values of the columns.



VALUE	FREQUENCY
59	15
63	14
50	14
54	13
56	13



Total rows	335
Empty cells	0
Unique values	58
Sum	19668
Average	58.88622754
Median	58
Minimum value	29
Maximum value	90

All stats include data in hidden or filtered rows

Step 3: Transformation: In this step, we did cleaning and transformation of the data. The major step of transformation was null conversions in our four columns . All the nulls were marked as a default value when a row hasn't a value for a column, as a best practice.

1. For instance date of the last visit had 17 null values which were marked as NULL.

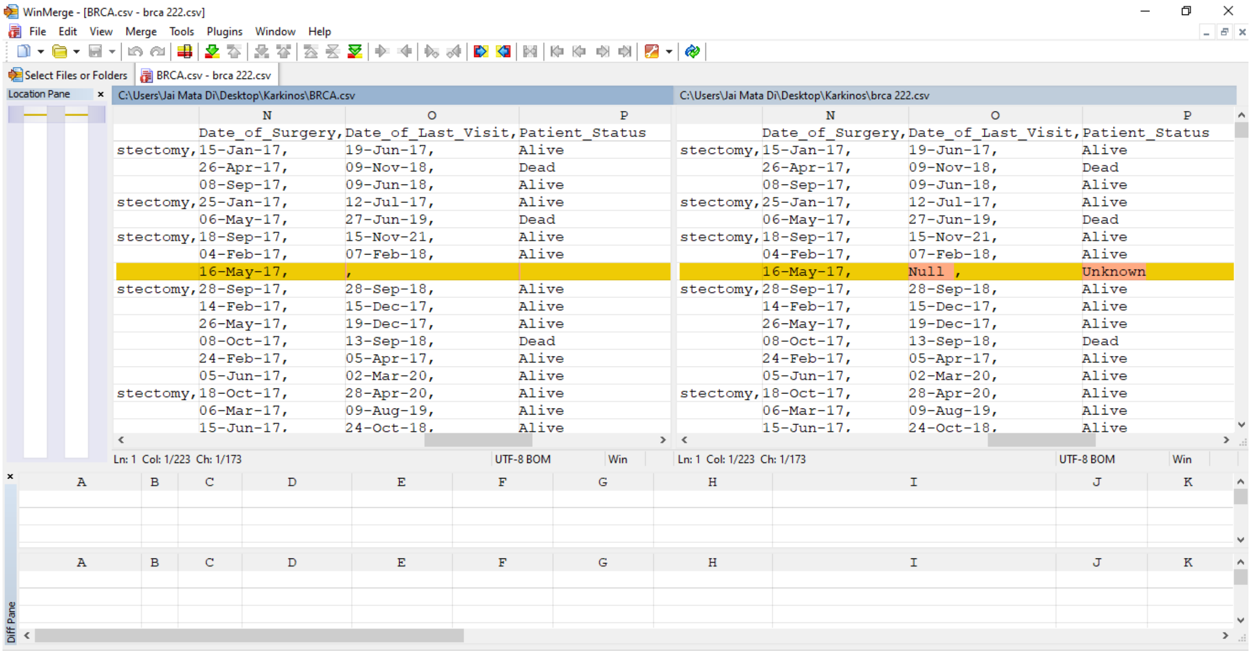
Data transformation :

- Few dates were not in the format of dd-mm-yyyy they were transformed so as to maintain uniformity .
- Patient status had one null value which was marked as 'Unknown'.

.

After every step of transformation, we used a diff tool (WinMerge) to make sure there was no unexpected change from the previous version of the data. (refer to figure 4) The changes are highlighted in pink. Since, these changes were intentional we moved on to the next step of the process.

Figure 4: Comparison of datasets using the WinMerge tool



C:\Users\Jai Mata Di\Desktop\Karkinos\BRCA.csv

```
                  N                 O                    P
         Date_of_Surgery,Date_of_Last_Visit,Patient_Status
stectomy,15-Jan-17,      19-Jun-17,         Alive
         26-Apr-17,      09-Nov-18,         Dead
         08-Sep-17,      09-Jun-18,         Alive
stectomy,25-Jan-17,      12-Jul-17,         Alive
         06-May-17,      27-Jun-19,         Dead
stectomy,18-Sep-17,      15-Nov-21,         Alive
         04-Feb-17,      07-Feb-18,         Alive
         16-May-17,      ,
stectomy,28-Sep-17,      28-Sep-18,         Alive
         14-Feb-17,      15-Dec-17,         Alive
         26-May-17,      19-Dec-17,         Alive
         08-Oct-17,      13-Sep-18,         Dead
         24-Feb-17,      05-Apr-17,         Alive
         05-Jun-17,      02-Mar-20,         Alive
stectomy,18-Oct-17,      28-Apr-20,         Alive
         06-Mar-17,      09-Aug-19,         Alive
         15-Jun-17,      24-Oct-18,         Alive
```

Ln: 1 Col: 1/223 Ch: 1/173 UTF-8 BOM Win

C:\Users\Jai Mata Di\Desktop\Karkinos\brca 222.csv

```
                  N                 O                    P
         Date_of_Surgery,Date_of_Last_Visit,Patient_Status
stectomy,15-Jan-17,      19-Jun-17,         Alive
         26-Apr-17,      09-Nov-18,         Dead
         08-Sep-17,      09-Jun-18,         Alive
stectomy,25-Jan-17,      12-Jul-17,         Alive
         06-May-17,      27-Jun-19,         Dead
stectomy,18-Sep-17,      15-Nov-21,         Alive
         04-Feb-17,      07-Feb-18,         Alive
         16-May-17,      Null ,             Unknown
stectomy,28-Sep-17,      28-Sep-18,         Alive
         14-Feb-17,      15-Dec-17,         Alive
         26-May-17,      19-Dec-17,         Alive
         08-Oct-17,      13-Sep-18,         Dead
         24-Feb-17,      05-Apr-17,         Alive
         05-Jun-17,      02-Mar-20,         Alive
stectomy,18-Oct-17,      28-Apr-20,         Alive
         06-Mar-17,      09-Aug-19,         Alive
         15-Jun-17,      24-Oct-18,         Alive
```

Ln: 1 Col: 1/223 Ch: 1/173 UTF-8 BOM Win

Once all the transformations were validated, our dataset was ready to generate insights.

Step 4: For this specific dataset, we wanted to study the influence of different factors on breast cancer using Chi-square test on Microsoft excel and go live with our insights .

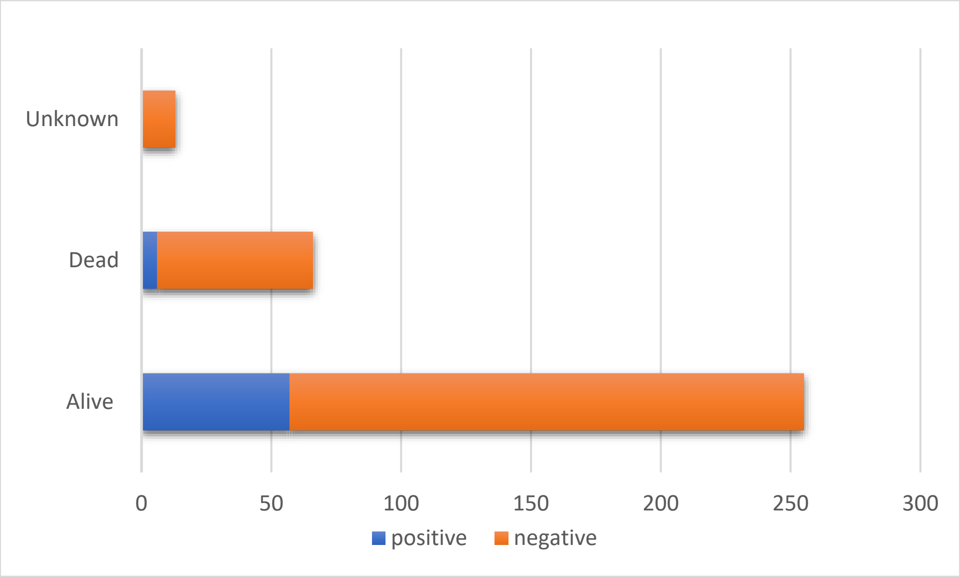
Statistical Analysis: The data analysis of association between the demographic and clinicopathological factors such as Age ,tumor status ,HER2 status and patient status were done by Chi-square test. All statistical calculations were done through Microsoft Excel 2019. A p value less than 0.05 was considered statistically significant.

Results:

After above transformations the total number of patients is 334 . Table 1 shows baseline demographic and clinicopathological characteristics of the included patients according to patient status

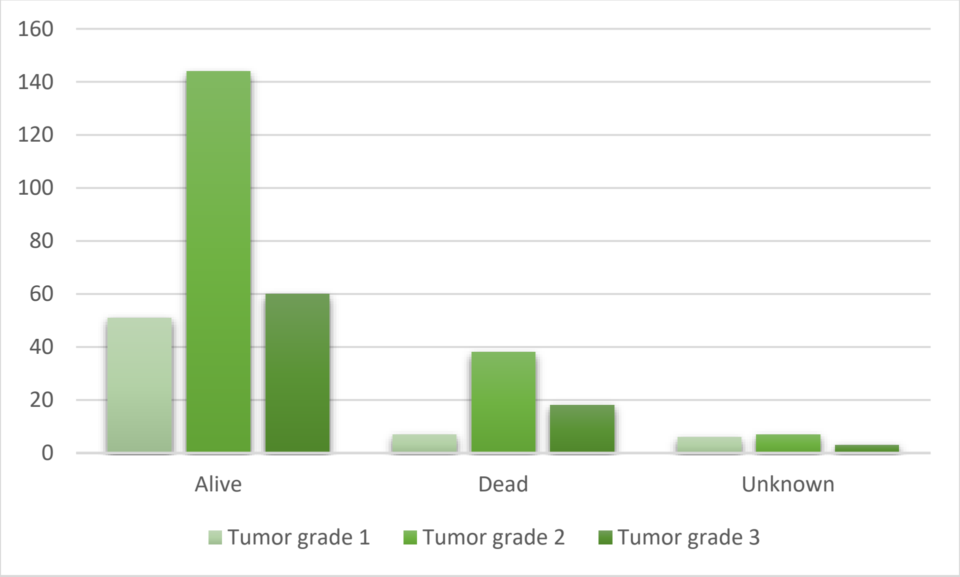
1) Out of 334 patients , maximum number of patients had HER2 negative and were alive (59.2%) followed by patients who were dead and had HER2 negative (17.9%) . No patient had HER2 positive and unknown patient status ($p<0.05$) . (Refer figure 5)

Figure 5: Patient Status and HER2 Status



2) Maximum number of patients had tumor grade 2 and were alive (43.1%) followed by Grade 3 (17.9%).Minimum number of patients had tumor grade 3 and unknown patient status (0.8%)($p>0.05$) .(refer figure 6)

Figure 6: Patient Status and Tumor grade



3) Out of these 334 patients , maximum number of patients belonged to the age group (50-64) and were alive (38.9%) followed by age group (25-49) (19.4%) . Least number of patients belonged to the age group (25-49) and had unknown patient status (0.8%) (p<0.05) .(refer figure 7)

Figure 7: Patient Status and Age groups

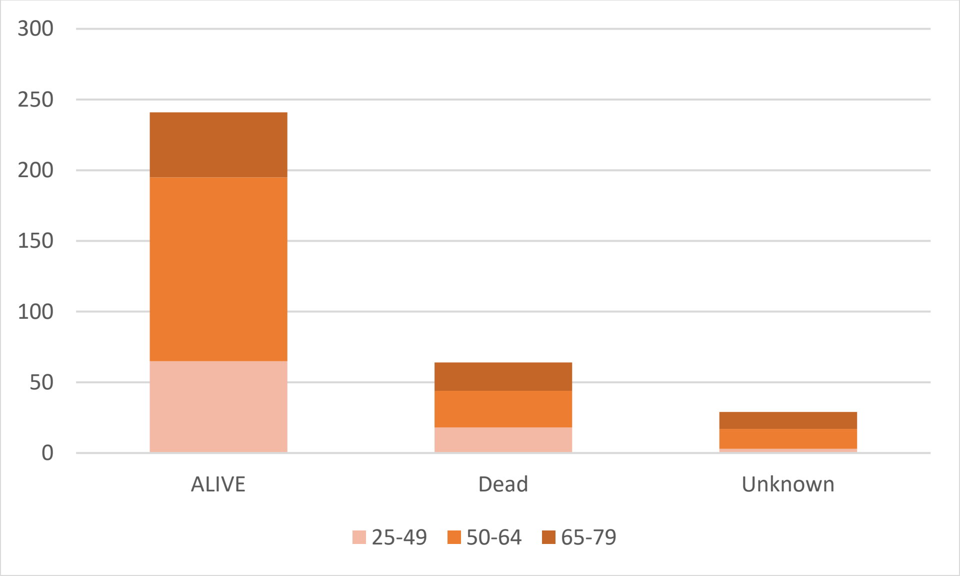


Table 1: Patient characteristics

Characteristic	***Patient Status***			***P value***
	Alive	*Dead*	*Unknown*	
HER2 Status				*<0.05*
Positive	57(17%)	61(1.7%)	0	
Negative	143(59.2%)	60(17.9%)	13(3.8%)	
Tumor Grade				*<0.05*
I	51(15.2%)	7(2%)	6(1.7%)	
II	144(43.1%)	38(11.3%)	7(2%)	
III	60(17.9%)	18(5.3%)	3(0.8%)	
Age				
25-49	65(19.4%)	18(5.3%)	3(0.8%)	*0.171*
50-64	130(38.9%)	26(7.7%)	14(4.1%)	
65-79	46(13.7%)	20(5.9%)	12(3.5%)	

Discussion :

1. In the current study, we analyzed the clinicopathological characteristics of breast cancer according to patient data using chi square test .A p value less than 0.05 rejects the null hypothesis i.e variables are independent of each other and accepts the alternate hypothesis . We found that HER2 status has clear influence on the patient status in breast cancer patients . Similar results were obtained in Liang et.al., study , they found that HER2 status had a clear influence on overall survival in patients with breast cancer.
2. In the present report , we found that tumor grade had no clear influence on the patient status which is different from the results obtained in Ablavi A et.al., study , they found that there was a significant association between histologic grade and breast cancer subtypes .Such difference is due to regional and cultural differences
3. Age is a demographic factor which has a clear influence on the patient status in breast cancer patients in our study but it differs from Ablavi A et.al, study , they found that there was no significant association between age and breast cancer .Such difference is due to lesser number of patients in our study or regional differences .

Recommendations :

The efficiency of ETL integration can make or break the rest of the data management workflow . Few of the ETL best practices are as follows :

- Your outputs will be faster and cleaner if you send less data into the ETL process. As a result, you should delete any unnecessary data as early as possible in the ETL process. If a database contains redundant items, for example, eliminate them before beginning the ETL process rather than processing of the data only to delete it afterwards..
- Automated data quality solutions can help with this by finding missing and inconsistent data in your data sets.
- Data cleansing prior to ETL integration, as well as ongoing, continuous data quality management, are required to obtain the highest-quality data.

  If you want your ETL integration operations to be rapid and efficient, you should automate them. However, given that we live in a period where complete automation is challenging to achieve, particularly for teams working with legacy infrastructure, tools, and procedures, it's important to remind ourselves of the value of automation.
- In addition to avoiding unnecessary data input, you may speed up the ETL integration process by using incremental data updates. When your data sets change, instead of changing all of the current data and starting over, you just add

the new data to your ETL pipeline. As component of an ETL integration system, the time it takes to conduct incremental data changes is worth it .

In practice, ETL integration automation entails relying only on tools to clean data, transport it through the ETL pipeline, and check the outcomes.

Conclusion:

This study identified that HER2 status and age had a clear influence of the status of the patient . whereas tumor grade did not show any association .In addition to this study also highlights the different tools that can be used during the ETL (Extract , Transform , Load) process and reduce the time consumption .

Bibliography

1. Venkatraman R. Karkinos – A purpose driven oncology platform by R VenkataramananR [Internet]. Karkinos Healthcare. 2021 [cited 2022Jun6]. Available from: https://www.karkinos.in/karkinos-a-purpose-driven-oncology-platform-by-r-venkataramanan/
2. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems [Internet]. 2014 Feb 7;2(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/
3. Belle A, Thiagarajan R, Soroushmehr SMR, Navidi F, Beard DA, Najarian K. Big Data Analytics in Healthcare. BioMed Research International [Internet].

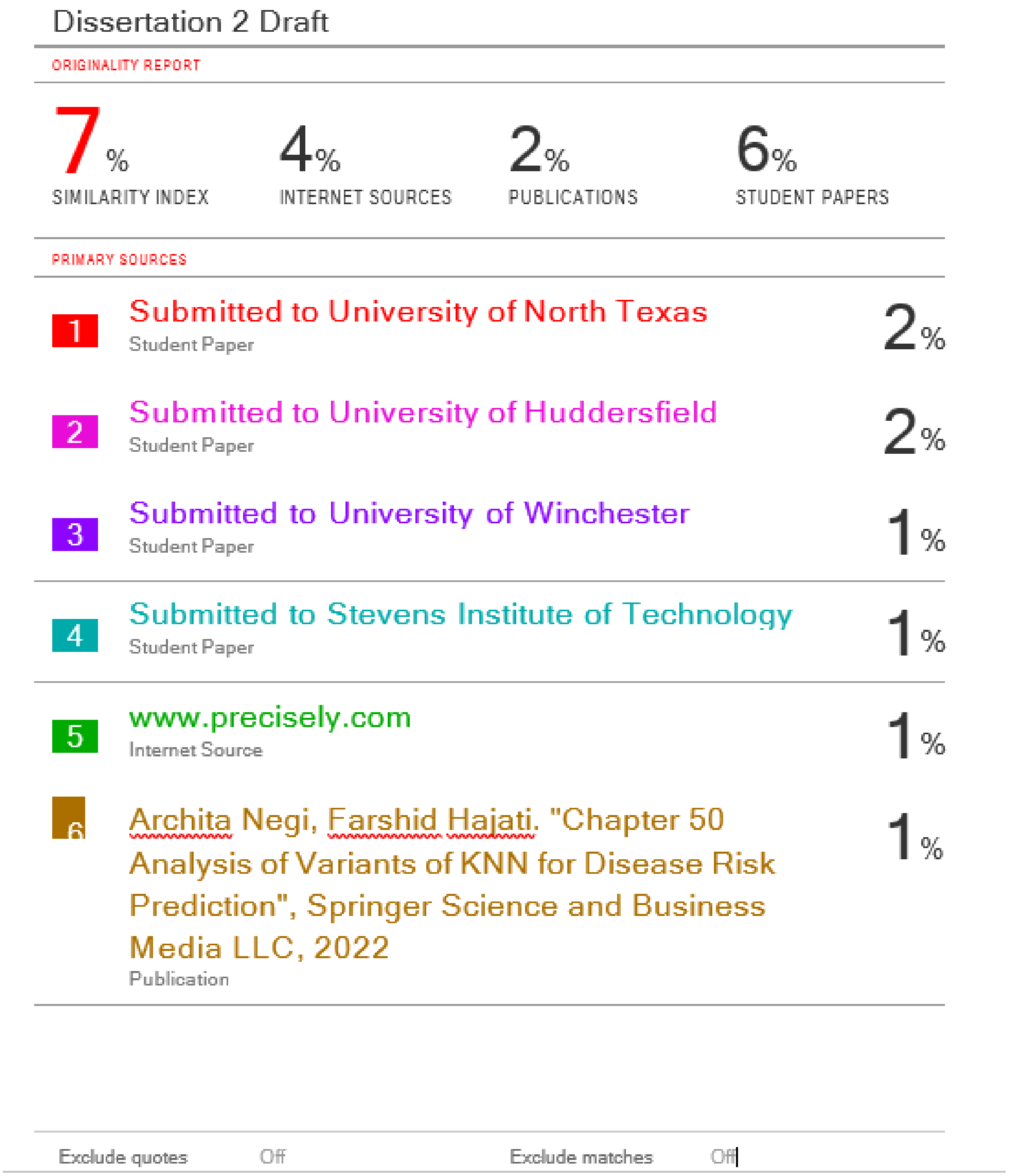
2015;2015:1–16. Available from: https://www.hindawi.com/journals/bmri/2015/370194/abs/

4. Roblero, J., Jácome, N. and Argotte, L., 2022. *Important Considerations for an ETL Process in CFE's Health and Security Area*. [online] Iaeng.org. Available at: <http://www.iaeng.org/publication/WCECS2012/WCECS2012_pp528-533.pdf> [Accessed 4 June 2022].
5. Adani-Ifè A, Amégbor K, Doh K, Darré T. Breast cancer in togolese women: immunohistochemistry subtypes. BMC Women's Health. 2020 Nov 23;20(1).

6. Chen, Liang, et al. "Effects of HER2 Status on the Prognosis of Male Breast Cancer: A Population-Based Study." *OncoTargets and Therapy*, vol. Volume 12, 5 Sept. 2019, pp. 7251–7260, 10.2147/ott.s209949.
7. Ji, Fei, et al. "Risk of Breast Cancer-Related Death in Women with a Prior Cancer." Aging (Albany NY), vol. 12, no. 7, 6 Apr. 2020, pp. 5894–5906, www.ncbi.nlm.nih.gov/pmc/articles/PMC7185107/, 10.18632/aging.102984. Accessed 1 Mar. 2021.

8. Lei, Ye-Yan, et al. "Clinical and Pathological Features and Risk Factors for Primary Breast Cancer Patients." *World Journal of Clinical Cases*, vol. 9, no. 19, 6 July 2021, pp. 5046–5053, 10.12998/wjcc.v9.i19.5046.
9. Manickam, Vijayalakshmi, and Minu Rajasekaran Indra. "Dynamic Multi-Variant Relational Scheme-Based Intelligent ETL Framework for Healthcare Management."

Soft Computing, 21 Mar. 2022, 10.1007/s00500-022-06938-8. Accessed 28 Apr. 2022.



Dissertation 2 Draft

ORIGINALITY REPORT

7%	4%	2%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of North Texas Student Paper	2%
2	Submitted to University of Huddersfield Student Paper	2%
3	Submitted to University of Winchester Student Paper	1%
4	Submitted to Stevens Institute of Technology Student Paper	1%
5	www.precisely.com Internet Source	1%
6	Archita Negi, Farshid Hajati. "Chapter 50 Analysis of Variants of KNN for Disease Risk Prediction", Springer Science and Business Media LLC, 2022 Publication	1%

Exclude quotes	Off	Exclude matches	Off