Breast Cancer analysis using visualization: A Swedish case study

Visual Data Analysis IT740A

Group 4

Hiwot G. Lassa

Welemhret W. Baraki

M. Usama Shehzad

Aimal Khan

2021-02-24

Contents

[1. Introduction 3](#_Toc61248378)

[1.1 Research questions/hypotheses 3](#_Toc61248379)

[2. Background 3](#_Toc61248380)

[3. Data 4](#_Toc61248381)

[4. Approach 4](#_Toc61248382)

[4.1 Data preparation 4](#_Toc61248383)

[4.2 Data analysis 4](#_Toc61248384)

[4.3 Visualization and interaction 4](#_Toc61248385)

[4.4 Evaluation 5](#_Toc61248386)

[5. Results 5](#_Toc61248387)

[5.1 Data preparation 6](#_Toc61248388)

[5.2 Data analysis 6](#_Toc61248389)

[5.3 Visualization and interaction 6](#_Toc61248390)

[5.4 Evaluation 6](#_Toc61248391)

[6. Discussion 6](#_Toc61248392)

[6.1. Limitation and Challenges 6](#_Toc61248393)

[7. Conclusions 7](#_Toc61248394)

[References 8](#_Toc61248395)

[Declaration of individual student efforts and time plan 9](#_Toc61248396)

[Time plan 9](#_Toc61248397)

# Introduction

Cancer is the name given to a collection of related diseases which are caused when some of the body’s cells begin to divide without stopping and spread into surrounding tissues and develop into tumors. Breast cancer is the most common cancer type and is worth studying and visualizing to be able to extract patterns and give insights for different audiences.

Breast cancer is the leading cause of cancer in women and according to the International Agency for Research on Cancer (IARC, 2021), it has overtaken lung cancer as the most commonly diagnosed cancer with about 2.3 million women being diagnosed with breast cancer in 2020. Breast cancer was also the most common cause of cancer death in women with about 685 000 deaths in 2020. Breast cancer was also the largest cancer diagnosed among women in Sweden in 2019 where 8,288 were diagnosed and 1,353 women died with breast cancer as the underlying cause of death in 2019 (Socialstyrelsen, 2020).

Although the diagnosis for breast cancer cases have increased, the mortality rates in developed countries have decreased with more advanced methods of treatment. However, early diagnosis, relevant awareness and information relating to the disease still plays a huge role. Visualization techniques can play eminent role to give insights and understandings for patients, care staff, researchers and decision-makers.

Analysis and visualization of statistical breast cancer data of Sweden will provide insight to health professionals, the government and individuals on coverage, Severity, and other breast cancer factors and indicators.

## Research questions/hypotheses

According to different research and statistical results breast cancer is commonly occurred in women is rare in men. The main problem, lack of coherent visualization in the health system to treat breast cancer patients which introduces problems and smart decision. Managements, professional and government officials of the health system face challenges in understanding the coverage and severity of the breast cancer due to lack of visual representations. As result human and capital resources can’t be efficiently distributed.

**Research Questions**

1. Which age group is most vulnerable to breast cancer?
2. How can we predict the breast cancer patients for next year based on the current statistics?
3. How to Analysis of different regions across Sweden based on age, incidence, mortality rate, diagnosis, biological subtype, TNM[[1]](#footnote-1) factor, distant metastases, and treatment pathways for understanding of the Breast Cancer Disease?
4. Which countries have high incidence and mortality rates?
5. Has early diagnosis improved over the years? Is this one the same as ‘Has early diagnosis increase life expectancy of the patients?’
6. *What are the key focus groups in terms* ***coverage, and age range*** *of government officials and healthcare professionals?*

# Background

Fundamentally visual data analysis is critically important and plays an eminent role in the health sector.

The data in the health sector that is exponentially growing from time to time needs special tool to analyse and manage for smart decision making for doctors and clinical personnel’s. Sanyour, et al [1] designed and developed a real time tool for cancer disease data analysis and visualization using K-nearest Neighbours, Support Vector machine and Naïve Bayes machine learning algorithms for classification to determine whether the breast lump is benign or malignant.

As a result many researches and technological developments have been adopted and used from recent years. In the era of Big data in which huge amount of data is accumulated in the health repositories so that utilizing and visualizing the historical data of patients and develop inferences and statistical data is difficult for health professionals and decision makers. To support health professionals and decision maker’s smart data analytics and visualization [2] makes health professional draw smart and accurate decisions and conclusion for different problems. Health related data can be analysed using different methods, approaches and tools like Hidden Markov Models[3], deep learning [1] and other machine learning algorithms. The bio-psycho-social model [1] designed was to consider not only the biological aspects of patients but also their psychological and social aspects that matters(affects) the biological aspect of a patient. Understanding the psychological and social aspects of the patients designed to help and visualize the patients as well as the health professional for effective health treatment through effective interactive visualization. Our visual design can be standalone application, or web-based[4] according the requirements.

We planned to solve the problem by collecting and cleaning the data set from different sources, designing and developing modern dashboards, bar charts, pie charts and others. The data repositories are International agency for research on cancer and Swedish statistical repositories. The Spotfire data visualization and analysis software will be used. The prototypical representation of the system will be designed with the spotfire. For the analysis part, regression analysis will be conducted and show current and future predicted trends of breast cancer incidents in Sweden. The visualization process starts with pre-processing cancer data, transform the data and create visual mappings, finally create views and presentations for the health professionals and decision makers to be able to extract knowledge.

This visualization project on breast cancer is required for health professionals to pay attention to the most vulnerable ages, regions and give insight for their future tasks and actions to minimize and mitigate breast cancer.

# Data

The statistical data sets of breast cancer are extracted from the national quality register for breast cancer (NKBC)[[2]](#footnote-2) and Socialstyrelsen[[3]](#footnote-3) which are collected from all of the Sweden’s healthcare providers. Data sets from the International agency for Cancer research (IARC) [[4]](#footnote-4) for different countries gathered for comparison. The swedish breast cancer data set are in the National Quality Register for Breast Cancer (NKBC), come from all of Sweden's healthcare providers. The purpose is to be a source of knowledge about Swedish care for patients, care staff, researchers and decision-makers. It has many indicators and variables like gender,, age ranges, region, health care, tumor size, lymph node status,, biological subtype, invesivity, data after and before diagnosis, survival rates of each year according to each regions and other vairables. The data sets are produced from 2008 -2019.

A few important attributes from the Swedish Dataset are listed below

|  |  |
| --- | --- |
| Attribute | Value |
| Region | Blekinge, Dalarna, Gävleborg, Gotland, Halland, Jämtland, Jönköping, Kalmar, Kronoberg, Norrbotten, Örebro, Östergötland, Skåne, Södermanland, Stockholm, Uppsala, Värmland, Västerbotten, Västernorrland, Västmanland, Västra Götaland |
| Diagnosis Years | 2008 - 2019 |
| Women diagnosed | count |
| Age at diagnosis | count < = 65 years, count > 65 years |
| Screening detected breast cancer | count = yes / count = total |
| Invasive cancer at diagnosis | count = yes, count = No |
| Spread to lymph nodes at diagnosis | count(cN-) = No, count(cN+) = yes |
| Distant metastases at diagnosis | count(M0) = No, count(M1) = yes |
| Tumour size at diagnosis | count <=20mm (T0/T1) , count >20mm (T2-T4) |
| Biological subtype at diagnosis | Triple Negative = count  HER 2 positive = count  Luminal = count |
| Chemotherapy | Pre-Operative = count  Pre and post operative = count  Post operative = count |
| Types of surgery | Breast conserving = count  Breast reconstruction = count  Mastectomy = count  Axillary = count |
| 5 year survival | count = yes / count = total |

# Approach

## 4.1 Data preparation

After collecting data from different sources we are going to clean and transform the data into suitable data features and formats for the Spotfire software package.

The following activities performed during the data preparation process:

**Step-1- Data Collection** –The breast cancer data sets are collected from National Quality Register for breast cancer, socialsysterlsen and International Agency for Cancer Research (IARC).

**Step-2- Pre-process and clean** the collected data sets from different sources, identify and filter relevant features for the breast cancer analysis and visualization.

**Step-3-Transformation** and merging data the pre-processed and cleaned data.

After conducting the data preparations process, the data analysis phase will come next and described below.

## 4.2 Data analysis

The Spotfire tool will be used for data analysis and visualization of the breast cancer data set. The spotfire software package tool has plenty features which allows users to integrate data into a single framework and to achieve a cohesive view with an integrated visualization. It speeds up visualization across an organization for faster, confident and much accurate decision making. Additionally, the Spotfire software also have built-in implementation features for different machine learning algorithms like regression analysis, classification, clustering and many more features. We will use regression modelling to predict future expected breast cancer patients in each region or the country as a whole.

## 4.3 Visualization and interaction

The collected data should have to visualize correctly to avoid biases and minimize risks by easily supporting the different stakeholders. Visualizing different hidden patterns, trends and information in the form of dashboards, line graphs, barcharts, maps, scatterplots, pie charts, tree maps and others.

As shown in figure 1 below the number of patients in some hospitals is really high. Especially in stockhom, vastra gotaland and skane which needs great attention in resource mobilization and taking care of breast cancer patients.

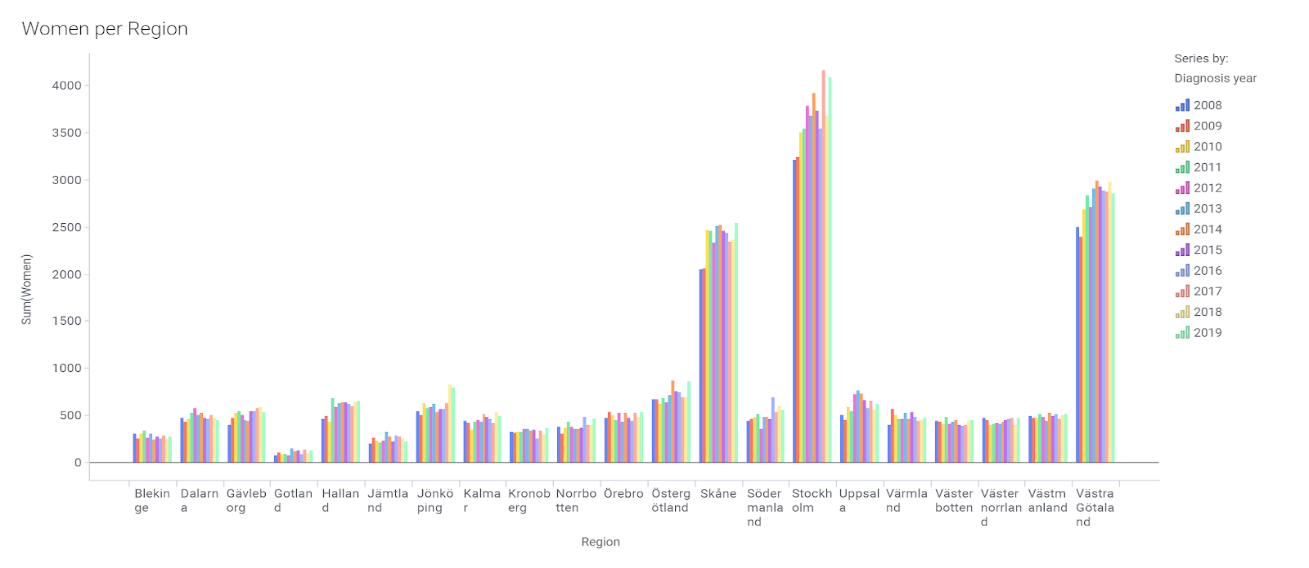


Figure 1 Women Breast Cancer Patients distribution accross hospitals in Sweden

In the figure 2 below the percentage of coverage of diagnosis is decreasing. By observing the droping in percentage on the coverage of breast cancer diagnosis that will trigger the health professionals to take measures against.

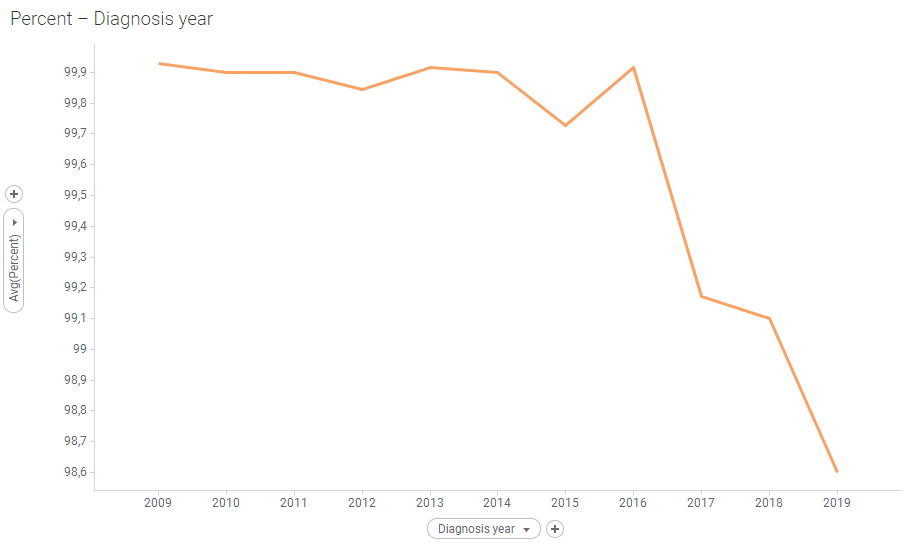


Figure 2 Coverage of diagnosis each for breast cancer patients.

## 4.4 Evaluation

The evaluation strategy will be based on usability testing and heuristic evaluation techniques. Usability and heuristic evaluation[5] will be performed based on the results. Heuristic evaluation is conducted according to the system specifications and the problems associated are identified to check if the system functionalities are included.

Finally Usability testing will be conducted to evaluate the system on the visualization of the breast cancer by distributing some questions to express the users how and what they feel based on System Usability Scale distributed from strongly disagree to strongly agree. The Users responses will be analyzed and present the result.

# References

[1] W. So, E. P. Bogucka, S. Scepanovic, S. Joglekar, K. Zhou, and D. Quercia, “Humane visual AI: Telling the stories behind a medical condition,” *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 678–688, 2021, doi: 10.1109/TVCG.2020.3030391.

[2] C. Leung, Y. Zhang, C. S. H. Hoi, J. Souza, and B. H. Wodi, “Big data analysis and services: Visualization on smart data to support healthcare analytics,” *Proc. - 2019 IEEE Int. Congr. Cybermatics 12th IEEE Int. Conf. Internet Things, 15th IEEE Int. Conf. Green Comput. Commun. 12th IEEE Int. Conf. Cyber, Phys. So*, pp. 1261–1268, 2019, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00212.

[3] B. C. Kwon *et al.*, “DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways,” *IEEE Trans. Vis. Comput. Graph.*, pp. 1–15, 2020, doi: 10.1109/TVCG.2020.2985689.

[4] E. Polychronidou, I. Kalamaras, K. Votis, and D. Tzovaras, “Health Vision: An interactive web based platform for healthcare data analysis and visualisation,” *2019 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2019*, 2019, doi: 10.1109/CIBCB.2019.8791462.

[5] B. S. Santos, S. Silva, and P. Dias, “Heuristic evaluation in visualization: An empirical study : ition paper,” *Proc. - 7th Bienn. Work. Eval. Beyond Methodol. Approaches Vis. BELIV 2018*, pp. 78–85, 2019, doi: 10.1109/BELIV.2018.8634108.

# Declaration of individual student efforts and time plan

## Time plan

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # | Activity/Milestone | Responsible |  |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |
|  |  |  | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | | 11 | | 12 | | 13 | |  | |  | |  |
|  |  |  |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |
| A1 | Proposal Writing | All members |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |
| A2 | Data Collection | All members |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |
| A3 | Data Preparation | All members |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |
| A4 | Data analysis | All members |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |
| M1 | Data Visualization | All members |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |
| A5 | Evaluation | All members |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |
| A6 | Results and conclusions | All members |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |

1. TNM –stands for Tumor Nodes Metastases which indicates whether or not the tumor has spread into neighboring lymph nodes. **Source:** *Cancer:What do the codes in the doctor’s letter mean?* [*https://www.ncbi.nlm.nih.gov/books/NBK279426/*](https://www.ncbi.nlm.nih.gov/books/NBK279426/) *accessed on 23-02-2021)* [↑](#footnote-ref-1)
2. National quality register for breast cancer (NKBC): <https://statistik.incanet.se/brostcancer/> (accessod on Feb 2021) [↑](#footnote-ref-2)
3. Socialstyrelsen: <https://www.socialstyrelsen.se/statistik-och-data/statistik/statistikamnen/cancer/> (accessed on Feb 2021) [↑](#footnote-ref-3)
4. IARC (International Agency for Cancer Research) for world countries: <https://gco.iarc.fr/today/home> (accessed on Feb 2021) [↑](#footnote-ref-4)