CancerMAS Dashboard: Data Visualization of Cancer Cases in Malaysia

Saidatul Rahah Hamidi¹, Anitawati Mohd Lokman², Shuhaida Mohamed Shuhidan³, Muhammad Imran Mohd Nor Hilmi¹,²

¹²Advanced Analytics Engineering Centre (AAEC), Faculty of Computer and Mathematical Sciences, 40450 Universiti Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia.
³Accounting Research Institute (ARI), Faculty of Computer and Mathematical Sciences, 40450 Universiti Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia.

raha@tmsk.uitm.edu.my

Abstract. One of the most famous diseases to ever roam across the globe is cancer. There are many types of cancer but the most common sites are breast, trachea, bronchus, lung, and colon cancer. Cancer can also be found pretty much in every country on earth and our country, Malaysia is not to be missed. The number of mortality in Malaysia for cancer is one of the highest and steps can be taken to help reduce the mortality rate. CancerMAS is a web-based dashboard that visualizes cancer incidence in Malaysia using cancer data that can assist cancer researchers to conduct their studies. The dashboard displays visualizations on ten of the highest cancer sites in Malaysia for male and female along with age groups. The dashboard also presents visualizations of all cancer sites by gender and ethnic as well as the incidence by states in Malaysia. All of the visualizations are in the form of graphs. The dashboard allows the user to zoom into the graphs and isolates elements in the graphs including hovering over the graph to reveal information.

1. Introduction

According to American Cancer Society [1], cancer is a group of maladies that are elucidated by the anarchical growth and spread of anomalous cells. Death is one of the possible results if the barbarity of the cells does not stay in an acquiescent condition. The ruling of cancer comes from numerous internal and external facets and there are many possible treatment options such as surgery, chemotherapy, hormone therapy and targeted therapy. Cancer is a group of diseases “characterized by uncontrolled growth and spread of abnormal cells” [1]. It is the 2nd cause of death in the world in 2015 with 8.8 million deaths, as mentioned by the Director of General Health Datuk Dr Noor Ibrahim [2] and in Malaysia, it is the leading cause of death in government and private hospitals. Cancer comes in many varieties such as breast cancer, prostate cancer, lung cancer, colon cancer and many more. Cancer can also affect both male and female. Former Deputy Health Minister Datuk Seri Dr Hilmi Yahya stated that 100 000 Malaysians suffer from cancer every year, with women being the most contribution and by the age of 75 years old and beyond, approximately one out of four Malaysians will suffer from cancer due to the increase in population and prolonged life span.
This research it can assist researchers to make decisions efficiently and improve systematic outcomes within hospitals and their communities. This research aimed to visualize cancer data in Malaysia.

2. Research Work

2.1. Types and Incidence of Cancer

According to a study conducted from 2016 to 2017 by Mehrnoosh Akhtari-Zavare et al [3], out of 1620 participants, 698 are male and 922 are female, which is 43.1% male and 56.9% female. Table 1 shows a task summary on the types and incidence of cancer in Malaysia, depicted from a study [3].

| Type of Cancer | Number | Percentage (%) |
|---------------|--------|----------------|
| Breast        | 523    | 32.3           |
| Colorectal    | 450    | 27.8           |
| Others        | 179    | 11.1           |
| Stomach       | 161    | 9.9            |
| Cervical      | 143    | 8.8            |
| Lymphoma      | 102    | 6.3            |
| Ovarian       | 62     | 3.8            |
| **Total**     | 1620   | **100%**       |

2.2. Data Visualization

Data visualization necessitates data presentation in graphical or pictorial form which makes understanding information a lot easier. It helps to explain facts and dictate courses of action. Data visualization is concerned with the design, development and application of computer generated graphical data representation [4]. It provides effective representation of data that comes from different sources. This allows decision makers to observe analytics in graphical form. Hence, making it easier for them to make sense of the data to discover patterns, comprehended information and form an opinion. In data visualization, a dashboard will be developed on top of the web so that information can be displayed clearly and efficiently with different kinds of graphs and maps [5]. Patterns can be identified when presented in visualization. There are many types of patterns in data visualization and with patterns, they help in seeing comparisons, making connections and drawing conclusions from the datasets [6]. Comparisons help in revealing differences and similarities. Connections that are linked and related to data visualizations can reveal themes and associations within the datasets. Another method that combined from interactive information visualization and open data is known as narrative visualization [7].

2.3. Review on Existing Dashboard

This section discusses existing dashboard that relates to visualization of cancer data, using global data. Screenshots of the dashboard are in Table 2.
Table 2. Screenshots of Existing Visualization Dashboard

| Cancer Today | PeGIS | USCS (United States Cancer Statistics) |
|--------------|-------|----------------------------------------|
| ![Cancer Today Screenshot](image1) | ![PeGIS Screenshot](image2) | ![USCS Screenshot](image3) |

*Cancer today* – is an interactive dashboard that displays visualization on cancer cases in a global scale [8]. Data visualized can be sorted according to Incidence, Mortality, Prevalence and Gender. It also has a feature that allows file export of the summarized data either in CSV, PNG, JSON, or PDF format. It is also shareable through Twitter and LinkedIn.

*PeGIS* – is a dashboard that outputs visualizations on dengue cases in Malaysia [9]. It utilizes a lot map-based visualization that shows dengue’s hot spots and also bar charts and pie charts that shows the incidence and mortality due to dengue in different states within six years. The dashboard is also interactive as it shows additional information when hovered on the charts. It also provides news and facts about dengue in Malaysia.

*USCS (United States Cancer Statistics)* - is a dashboard that shows visualization of cancer cases in the United States of America (USA) that is also interactive [10]. It utilizes a lot of map-based visualization, tables and also bar plots that shows comparisons between cancer cases. Data visualized can be sorted according to Area, Incidence, Mortality, Gender, Cancer Type and also Year. The dashboard also allows the summarized data to be exported.

3. Methodology

This CancerMAS data visualization involve several phases which are preliminary study, analysis, development and testing phase. Activities, techniques and deliverables for each phase are shown in Table 3. Thus, details activities for acquiring data, analysis, until development discussed in sub section.

Table 3. Summary of Research Methodology

| Phases          | Activities                                      | Techniques                                       | Deliverables                      |
|-----------------|-------------------------------------------------|-------------------------------------------------|-----------------------------------|
| Preliminary Study | Understanding the idea of project. Explore to deeper comprehension of cancer and data visualization | Conduct preliminary studies on cancer and data visualization Do a literature review on cancer data visualization | Project title Problem statement Objectives Scope Significance Literature Review |
| Analysis        | Acquire cancer datasets from hospital and cancer | Send emails dataset request                     | Raw dataset of cancer data        |
3.1. Acquiring and Cleaning Data

The datasets were obtained through www.data.gov.my. It is Malaysia’s open data portal that stores a variety of datasets across different categories. The dataset acquired can be found in the health category. The cancer data obtained from year 2007 to year 2011 were published by the National Cancer Institute. It consists of a total of 18 dataset files, in which eight of the files are raw datasets with cancer incidence by site, age group and gender, another ten of files are summary datasets based on state, gender, cancer incidence and ethnic. Cancer site or cancer type that include in these datasets are more than 50 type of cancer from trachea, bronchus, lung cancer until other urinary organs cancer. It covers the majority ethnic (Malay, Chinese and Indian) in Malaysia and all states.

3.2. Analysis Data

Data obtained need to analyze for further understand what kind of data to be dealt with before proceeding with any development. If the data is wrongly used, the dashboard will be nugatory as it does not achieve the intended purpose. After understanding and analyzing the dataset files, it is found that all of the raw datasets consists of 22 columns and 52 rows of data. But, not all columns will be utilized and only the columns that meet this research requirement are chosen. Hence, 18 out of 22 of the columns in the dataset are put to use while the number of rows in the dataset remained intact. There are also three data formats or also known as dtypes which are float64, int64 and object. Before starting the data cleaning process, a few packages had been installed in the Python environment such as Pandas and xlrld. To perform data cleaning, strings of code are used to remove any unnecessary rows and columns to the datasets.
3.3. Paper Prototyping
The layout of the system will be in the form of a dashboard whereby the data visualization organized according to respected tabs. Refer figure 1 for paper prototype of CancerMAS (Cancer Malaysia Dashboard). It consists of three tabs or pages, which are the Overview page, Ethnic page and Site page. They are designed with the mind of what the pages are intended to look like. The overview page displays visualization of the cancer incidences occurs in Malaysia in a form of a geographical map that utilizes a heat map. The second page is the Ethnic page that presents visualization of cancer sites according to genders and ethnic group in Malaysia, which are Male, Female and Malay, Chinese and Indian respectively. Lastly, the third page is the Site page which shows cancer sites with the highest number of incidences for both male and female in Malaysia.

![Figure 1. Prototype of CancerMAS](image)

a) Overview page  b) Ethnic page  c) Cancer Type page

3.4. User Testing and Feedback
After completing the paper prototype, a user testing with potential user and visualization expert are commenced.

4. Result
This section will discuss the results from the analysis including the development of data visualization. Next, explanation of manage all visualizations on a dashboard with filtering option.

4.1. Top Ten Cancer Site in Malaysia according to Gender
Two DataFrames are created for both male and female that comprises of three major ethnic in Malaysia. With the help of Pandas, data was sort in descending order so that data starts from highest to lowest. After that, the DataFrame was slice to only retrieve the first 10 rows of data. The code is stored in a variable which later will be called in one of Dash Plotly’s components which is dcc.Graph().

![Figure 2. Top 10 Cancer Sites in Malaysia According to Gender](image)
Figure 2 shows that for male in Malaysia, highest number of cancer incidence for male is recorded as trachea, bronchus and lung cancer with more than 6000 cases. Next, colon, nasopharynx and rectum cancer are nearly 4000 cases recorded. Meanwhile, bladder and other skin cancer are having less than 2000 incidence for male gender in Malaysia. As compared to female, a vast difference with more than 15000 incidences recorded for breast cancer in Malaysia. Cervix and ovary cancer recorded nearly 5000 cases in Malaysia for female.

4.2. Top Ten Cancer Site in Malaysia by Age-Group and Gender
Figure 3 below display for top ten cancer sites according to age-group for male and female. The line graph shows that majority cancer incidence occurred for male is above 60 years old. Meanwhile, cancer incidence started early 40 years old for women in Malaysia.

4.3. Cancer Site for Ethnic and Gender in Malaysia
A scatter plot was created for all cancer sites according to genders (male and female) and ethnic, which are Malay, Chinese and Indian (Figure 4). This graph allows filtering based on gender and ethnic, which only selected data is visualize. Next, subgroup analyses can be conducted with user easily explore the data with a selection of any cancer type, gender or ethnic.
4.4. Number of Incidence in Malaysia by States Bubble GeoMap
Another type of data visualization, a map created by identifying Malaysia’s coordinates, which are the latitude and longitude. After that, a map can be generated with the help of Python’s package Folium. Then, bubble plots can be created for each state in Malaysia. Executing the code will show the Malaysia map along with the bubble plots that represents the number of incidence in each state as shown in Figure 5.

![Figure 5. Number of Cancer Incidence Bubble Map](image)

4.5. CancerMAS Dashboard
CancerMAS dashboard consists of three tabs or pages that show different kinds of data visualization that elicits information that are relevant to the tabs. The three tabs are Highest, Cancer Site and States in Malaysia.

![Figure 6. Sample of CancerMAS Dashboard](image)
The Highest tab consists of four graphs. The first two graphs are ten of the highest cancer types for male and female, the blue graph representing the male while the pink graph represents the female. The graph is in descending order and is in bar chart form to easily make comparison among each cancer site. Putting both graphs side by side allows a comparison between male and female as each gender has a different set of cancer site.

The next two graphs are ten of the highest cancer sites by age-group for male and female. The type of chart used for these graphs are line charts to identify the rise and fall of the cancer incidences as the age progresses. The age-group is 5-year intervals. Putting both graphs side by side allows a comparison between male and female as each gender have different patterns in accordance to the age-group and cancer sites. This interactive data visualization allows data filtering features.

4.6. User Testing and Feedback Result
Finding a potential user is not as easy as it seems due to a very specific stakeholder related to health information. Nevertheless, a potential user was able to identify for user testing. The user testing procedure went well as the user has a lot of positive feedback regarding the paper prototype and the CancerMAS dashboard but not forgetting to provide some suggestions such the implementation of cross filtering if possible.

5. Discussion
CancerMAS dashboard enables cancer researchers in identifying which cancer sites have the highest number of incidence for male and female as well as Malay, Chinese and Indian with the help of visualizations in the form of graphs. Visualizations help in eliciting the information easier than from the textual and tabulated form. So, with the information that they retrieved from the visualization, it can assist cancer researchers in decision making that are more effective to help reduce the number of cancer incidence. It can also enhance systematic outcomes within hospitals and their communities. As shown in figure 6, significance finding indicates that Malay female having the highest number of incidence specifically breast cancer with more than 8000 cases. Meanwhile, second highest is Chinese Female nearly 8000 cases. Another unique finding from the visualization, Indian population is having the lowest number of cancer incidence. This is due to the fact that possible the case was not recorded in IKN or the culture of Indian in life effects the occurrence of minimal number of cancer.

Some of the limitations are datasets coming from a single source only, unable to get a holistic view and also datasets that are not most recent. To conclude, developing CancerMAS dashboard may hopefully aid cancer researchers in continuing their studies in the respective area.

Acknowledgement
The authors gratefully acknowledge the financial grant 600-IRMI/PERDANA 5/3 BESTARI (094/2018) given by IRMI, Universiti Teknologi MARA, Malaysia and Faculty of Computer and Mathematical Sciences for all supports and resources.

References
[1] American Cancer Society. (2015a). *Cancer Facts and Figures 2015*. Atlanta, GA: American Cancer Society.
[2] Datuk Dr Noor H. A., (2018). Datuk Dr Noor H. A., (2018). Malaysian Study on Cancer Survival. National Cancer Registry
[3] Mehrnoosh A., Sherina M., Ummavathy P., Lekhraj R., Siti I. F., Rozi M., (2018). Determinants of quality of life among Malaysian cancer patients: a cross-sectional study. *BioMed Central*
[4] Sadiku, M., Shadare, A. E., Musa, S. M., & Akujuobi, C. M. (2016). Data Visualization. *International Journal of Engineering Research And Advanced Technology (IJERAT)*, 2(12), 11-16.
[5] Xin Y., Claude B., Rob V., Phillipe E, Alessandro B. (2018). Data Visualisation in Conceptual Design: Developing a Prototype to Support Decision. Conference: 12th International Conference on Modelling, Optimization and Simulation, At Toulouse, France.

[6] Julie R., Piotr K., (2016). Using Design Patterns to Find Greater Meaning to Your Data. O'reilly

[7] Ackermann P., Stockinger K. (2019) Narrative Visualization of Open Data. In: Braschler M., Stadelmann T., Stockinger K. (eds) Applied Data Science. Springer, Cham

[8] Cancer today. (n.d.). Retrieved July 29, 2019, from https://gco.iarc.fr/today/home

[9] Dengue Story Map Journal. (n.d.). Retrieved July 27, 2019, from http://pegis.maps.arcgis.com/apps/MapJournal/index.html?appid=cc7d8ecc38ad4335adef714bf3372136#

[10] U.S. Cancer Statistics Working Group. U.S. Cancer Statistics Data Visualizations Tool, based on November 2018 submission data (1999-2016): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; www.cdc.gov/cancer/dataviz, June 2019.