Data Visualization on Global Trends on Cancer Incidence An Application of IBM Watson Analytics

Kelvin KF Tsoi[1,2], Felix CH Chan[1], Hoyee W Hirai[1], Gary KS Leung[1], Yong-Hong Kuo[1], Samson TAI[4], Helen ML Meng[1,3]

[1]Stanley Ho Big Data Decision Analytics Research Centre
[2]Jockey Club School of Public Health and Primary Care
[3]Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
[4]IBM China/Hong Kong Limited

E-mail: kelvintsoi@cuhk.edu.hk

Abstract

Visual analytics is widely used to explore data patterns and trends. This work leverages cancer data collected by World Health Organization (WHO) across over a hundred of cancer registries worldwide. In this study, we present a visual analytics platform, IBM Watson Analytics, to explore the patterns of global cancer incidence. We included 26 cancers from different geographic regions. An interactive interface was applied to plot a choropleth map to show global cancer distribution, and line charts to demonstrate historical cancer trends over 29 years. Subgroup analyses were conducted for different age groups. With real-time interactive features, we can easily explore the data with a selection of any cancer type, gender, age group, or geographical region. This platform is running on the cloud, so it can handle data in huge volumes, and is assessable by any computer connected to the Internet.

Keywords: visual analytics, global comparison, cancer incidence, Choropleth map, interactive Interface

1. Introduction

Visual analytics have been shown to be effective for data exploration [1], but the requirement of computational power is high for global comparisons of disease trends. Therefore, cloud computing is one of major enablers for explorations of data. This is a general shift of computer processing, storage, and software delivery away from the traditional desktop computers and local servers towards the cloud [2].

Cancer is one of the leading causes of morbidity and mortality worldwide, with approximately 14 million new cases and 8.2 million cancer related deaths in 2012. It is expected that annual cancer cases will rise to 22 million within the next 2 decades [3]. The commonness of cancer varies by gender, age, ethnicity, geographical location, economic status, and so on. Generally, the cancer causes of death were common on breast, lung, liver, stomach, colon and rectum [4]. Although the age-standardized incidence rates on some cancer showed stable trends, but the prevalence of cancer has grown along with the ageing population. To better understand the progression of cancer, cancer registries were set up in different countries. The first population-based cancer registry was in Germany Hamburg in 1926 [4]. Cancer registries have been widely used in epidemiological research, so the World Health Organization (WHO) formed an International Agency for Research on Cancer (IARC) to collect cancer registry data across different countries. Descriptive studies use the registry database to examine differences in the incidence of cancer for different patient characteristics [5]. The data volume of the global cancer incidence is huge, so visual analytics can help to enhance the data interpretation on disease distribution and trends.

The main contribution in this study is to use a visual analytics platform (IBM Watson Analytics) to distinguish cancer trends and patterns from WHO cancer registries. We wish to answer several major

URI: http://hdl.handle.net/10125/41555
ISBN: 978-0-9981331-0-2

questions presented to us from cancer researchers, including: (i) What are the top-ranking cancers (ii) across different regions, (iii) across gender, (iv) over the years, (v) across high- and mid-income regions, and (vi) across different age groups? In the following, we provide a visualization of the WHO data that enables us to intuitively answer these questions. In the future, such intuition can guide us in further explorations. For example, we can use the patterns to show the effective of colorectal cancer screening programme in US, and compare the results with some regions that do not have screening programme. We selected the Choropleth map and the traditional line charts for demonstration. The Choropleth map used to present the population density in different geographical regions [6]. In this study, the regions with high volume of cancer incidence will appear with darker colors on the map. The line chart is the traditional way of presentation which shows data trends along timeline. In this study, line charts are used to demonstrate cancer trends across regions and different population groups, such as for different gender. A matrix of line charts are also developed for cross-sectional comparison between different age groups across the regions. The structure of this paper is organized as follows: related work in the academic field, data structure of cancer registry from WHO, data categorization for subgroup comparison, visual analytics environment, application scenario, and conclusion.

2. Related Work

Data visualization is important to enhance the understanding of the overall dataset. It is widely applied to different academic areas. Fan et al. used different color regions to show the spatial distribution between weather temperature and light intensity, and developed a color coding scheme and showed on a map [8]. MuCusker et al. applied motion charts to demonstrate the association between tax payments and smoking prevalence. Visualization can initiatively show that when the taxes go up, the prevalence of smoking goes down [9]. Torres et al. designed a matrix of scatter plots to help researchers explore data patterns and formulate research hypotheses within the National Health and Nutrition Examination Survey [10]. Yung et al. developed an interactive platform to visualize multiple sets of proteomic data in huge data volumes [11]. Researchers can quickly judge the quality of selected proteomic features. One of the recent publications by Blevins presented an interactive data visualization application for human immunodeficiency virus (HIV) cohorts [12]. The platform was used to present the longitudinal plots, bubble plots and choropleth maps. Data visualization can demonstrate disease trends and distribution. The above platforms for data visualization were mainly conducted in the local computers. In the era of big data, the data volume will be a technical challenge to most of the existing platforms. Therefore, visual analytics on the cloud will be an upcoming trend of application, and this study is a demonstration of visual analytics on the cloud platforms with global cancer incidence data using IBM Watson Analytics.

3. Data Structure of Cancer Registry

The International Agency for Research on Cancer (IARC) is a specialized cancer agency of the WHO in order to promote international collaboration in cancer research. It has an important role in describing the burden of the global cancer through co-operation with cancer registries worldwide, monitoring geographical variations and assessing trends over time. IARC has developed a cancer database, called Cancer Incidence in Five Continents Time Trends (CI5 plus), providing access to detailed information on the incidence of cancer recorded by regional or national registries.

Variable	Detail
REGISTRY	4 digit number represent different cancer registries around the world
ETHNIC_GROUP	Code represent different race groups
YEAR	Year of the incidence record
SEX	Gender groups
CANCER	Cancer type groups
TOTAL	Total number of incidence from all age group
{age_group}	Age are divided into 5-year intervals. The age groups are divided into 0-4, 5-9, …, 75-79, 80-84, 85+
N_unk	Total number of incidence of patients with unknown age

Table 1. Details Cancer Incidence Data in Five Continents Time Trends (CI5 plus)

The database contains annual data of the population size, source of cancer registry, ethnic, gender and incidence data for 102 cancer registries from 118 regions. All histological data were available until 2007. The cancer site dictionary contains 181 diagnostic units, which comprise cancer sites at the

third digit level of the International Classification of Disease, 10th revision categories. All data were reported by gender and by age at 5-year interval. A 5-digit code has been assigned for each cancer registry. Registries from the same country can be identified from the first 2-digits of the registry code. The data structure is shown in Table 1.

We selected the regions which reported 20 years of cancer incidence between 1988 and 2007. A total of 8 regions were selected, including the United States, the United Kingdom, Costa Rica, Sweden, Croatia, Japan, Hong Kong and China (Shanghai).

4. Data Categorization

In the database, we classified the data into different subgroups for comparison, including cancer type, gender, age group, and economic status. For cancer type, a total of 26 cancers were available in the registries with combining colon and rectum as colorectal cancer. Some of the cancers are gender specific, such as breast cancer for female and prostate cancer for male. Therefore, separate analyses were performed. Some cancers may be age dependent, so subgroup analyses on different age groups can help to better understand the overall picture. A lot of evidence showed that the cancer trends are different across economic status of a country [12]. Therefore, the economic status of each country was stratified based on the World Bank Atlas method with adjustment for exchange rate, local and international inflation. Low- to middle-income economies were defined as those with a gross national income (GNI) per capita of USD $6,000 or less in 1988 and USD $11,455 or less in 2007. All other countries were defined as high-income economies with the same cut-off [13]. Although Shanghai should be regarded as a well-developed city nowadays, this study used the data before 2007, and therefore, Shanghai was not classified as a high-income region.

5. Visual Analytics Environment

IBM Watson Analytics was used as the interactive tool for visual analytics, as it is constructed on the cloud platform that we can handle the dataset in large volume. It can be used with any internet-connected browsers. The following are the steps to construct the interactive platform.

5.1. Major Cancer Types across Regions

To answer the first research question about the major cancer types across regions, we use a word cloud to display the statistics of the eight regions with complete data. Cancer types with higher incidences show up proportionally which higher font sizes. We can see that the major cancer types across regions seem to remain fairly consistent – including lung cancer, colorectal cancer, prostate cancer and breast cancer.

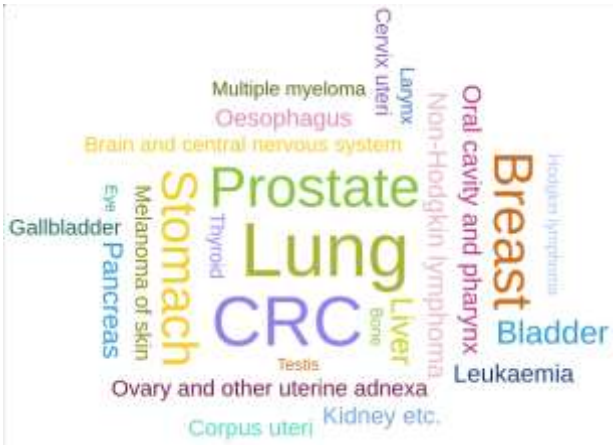

Figure 1. Word Clouds showing major cancer types across regions.

5.2. Dashboard Presentation

After the preliminary data exploration, we chose the choropleth map to show the density of cancer distribution, line charts to show the trends of cancer history over the past 20 years, and a pie chart to show the gender distribution. A global choropleth map is embedded in the platform; and there is an auto-detection function for all country names (the left upper corner in Figure 2) so that the statistics for cancer incidences per 100,000 population in each region can also be presented. All cancer incidences for the eight regions were combined together for the line chart. All these are integrated in a dashboard as shown in Figure 3. At the bottom of the dashboard we provide a selection menu showing 26 different cancer types. Clicking on any specific cancer type will enable immediate updates of the statistics displayed on the dashboard. This enables interactive

exploration of the data. Also, the data were further categorized into four age groups, i.e. age 0-19, 20-39, 40-64, and 65+. In each group, cancer data from the eight regions were plotted together against the years. The trends in the US, UK and Hong Kong were plotted as a demonstration (Figure 4). This data visualization presents intuitive answers to the 6 research questions stated earlier. Essentially, the top-ranking cancers sites are lung, colorectum, prostate and breast. The cancer incidence was dominated by lung cancer; it has been the leading cancer since the earliest data available. Colorectal cancer was the second most common cancer; its incidence has been increasing rapidly and was approaching the rate of lung cancer. For those gender-specific cancers, prostate and breast cancers for male and female respectively were leading cancers as well. Increasing trends could be observed in most of the cancers except stomach cancer. As cancer is an ageing-related illness, the cancer incidence skewed to age group older than 40.

6. Application Scenario

When colorectal cancer (CRC) was selected in the selection menu, all graphs automatically update and reflect the incidences for CRC. In the choropleth map (Figure 2), the CRC incidences were shown to be higher in the high-income regions, including the US, the UK, Sweden, Hong Kong and Japan. Costa Rica showed the lowest CRC incidences in the past 20 years. However, as shown in the line chart, the low- to middle-income countries had an increasing trend of CRC, whereas the high-income countries showed stable but progressive trends. Besides, as shown in the pie chart, a higher proportion of male was suffered from CRC. Subgroup analyses were performed according to the age groups (Figure 3). Although the patients aged below 40 showed different trends on CRC, the actual incidences were far from the group aged over 65, as the scales in y-axis are different across the figures. The colorectal cancer incidences gradually increased in most of the countries, except the US. It may be attributed to the fact that CRC screening programme is effective to reduce the risk of this cancer in the US.

When we switched CRC to other cancers in the selection menu, we can immediately explore the global cancer trends. We also prepared a matrix to show the five common types of cancer (breast, colorectal, liver, lung, and prostate cancer) across different age groups (Figure 4). Cancer incidences were generally increased in breast, colorectal and

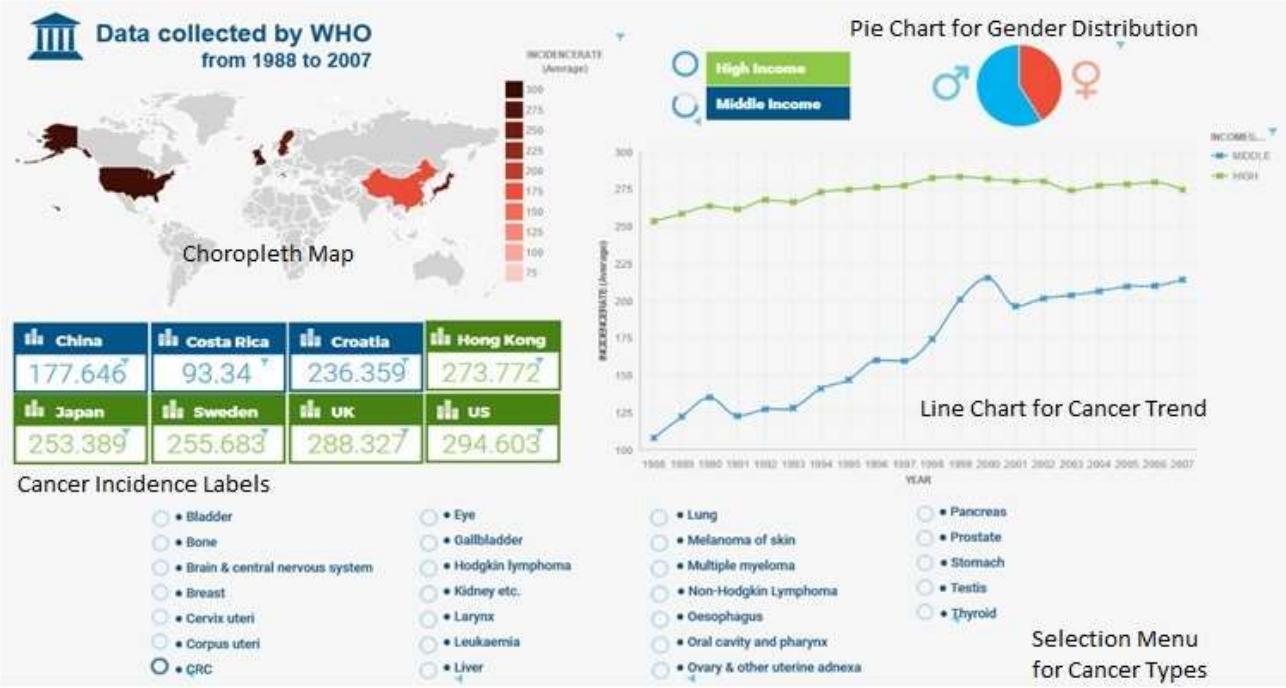

Figure 2. Overview of the Interactive Functions on the Plots

prostate, but not in liver and lung in the ageing

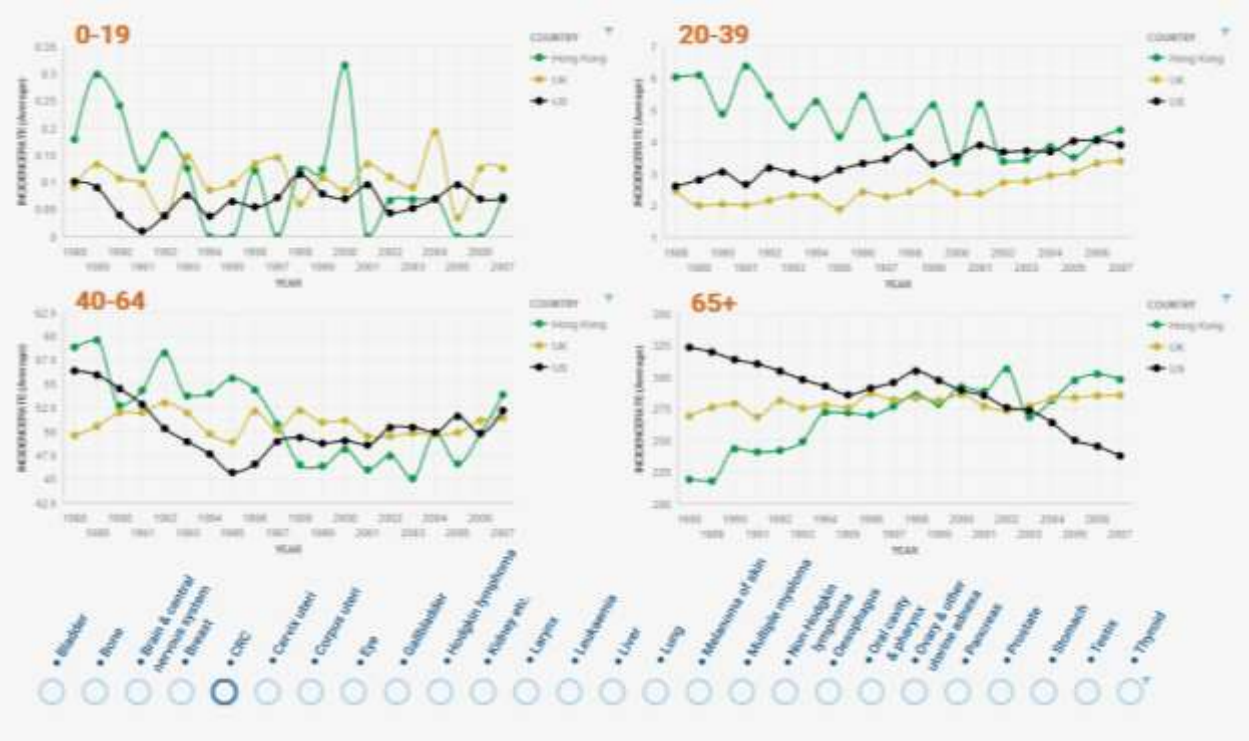

Figure 3. Cancer Incidence Trends across Different Age Groups

population. In the US and Sweden, the incidences of prostate cancer among the subjects aged between 40 and 65 increased dramatically compared with other regions. Further clinical investigation should be conducted to confirm the trends.

8. Conclusion

In this study, we have described data visualization with the IBM Watson Analytics platform to explore the open-sourced data on global cancer trends. The system makes use of an interactive interface to demonstrate the data distribution and trends. Future work can be extended to the data projection on the cancer incidence.

9. Acknowledgement

The authors gratefully acknowledge the collaboration and support of Mrs. Mary Law and her team from IBM Hong Kong for this research work, both providing the IBM Watson Analytics Platform and also offering valuable insights and suggestions about it.

10. References

[1] D.A.Keim, "Information Visualization and Visual Data Mining", IEEE Transactions on visualization and computer graphics, 2002, 8(1):1-8.

[2] E. Saranya, A. Sunitha. "Identifying data integrity in Cloud Storage" International Journal of Computer Science, , Mar 2012, 9: 1694-14.

[3] B. Stewart and C.P. Wild (eds.), International Agency for Research on Cancer, World Health Organization, (2014) World Cancer Report 2014 [Online], Available from: http://www.iarc.fr/en/publications/books/wcr/wcr-order.php [Accessed: 10th June 2016].

[4] G. Wagner, History of cancer registration. In: O.M. Jensen, D.M. Parkin, R. MacLennan, et al, eds. Cancer registration: principles and methods. IARC Scientific Publications no 95. Lyon: International Agency for Research on Cancer, 1991: 3–6.

[5] D.M.Parkin, The evolution of the population-based cancer registry, Nat Rev Cancer, 2006; 6: 603–12.

[6] A.M. MacEachren, C.A. Brewer and L.W. Pickle, "Visualizing Georeferenced Data: Representing Reliability of Health Statistics", Environment and Planning, 1998, 30:1547-61.

[7] F. Fan and E.S. Biagioni, "An approach to data visualization and interpretation for sensor networks", Proceeding of the 37th HICSS, 2004.

[8] J.P. McCusker, D.L. McGuinness, J. Lee, C. Thomas, P. Courtney, Z. Tatalovich, N. Contractor, G. Morgan and A. Shaikh, "Towards next generation health data exploration: a data cube-based investigation into population statistics for

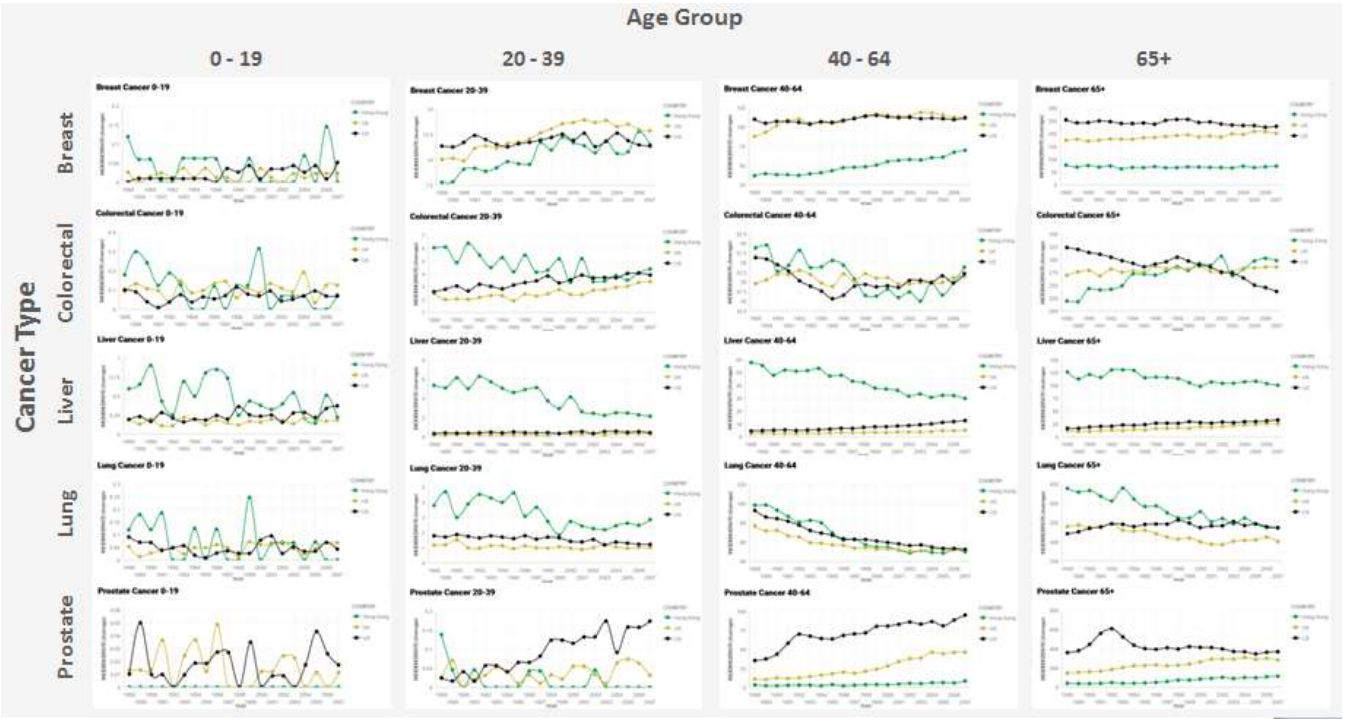

Figure 4. Matrix of Five Cancer Incidence in Different Age Groups

tobacco", InSystem Sciences (HICSS), 2013 46th Hawaii International Conference on, 2013, 2725-2732 IEEE.

[9] S.O. Torres, H. Eicher-Miller, C. Boushey, D. Ebert and R. Maciejewski, "Applied visual analytics for exploring the national health and nutrition examination survey", Proceeding of the 45th HICSS, 2012.

[10] L.S. Yung, C. Yang and M. Dakna, "SyncPro: A synchronized visualization tool for differential analysis of proteomics data sets", InBioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference, 2010, 95-100 IEEE.

[11] M. Blevins, F.H. Wehbe, P.F. Rebeiro PF, C.C McGowan and B.E. Shepherd, "Interactive Data Visualization for HIV Cohorts: Leveraging Data Exchange Standards to Share and Reuse Research Tools", PloS one, 2016, 11:e0151201.

[12] C.R. Baquet, J.W. Horm, T. Gibbs and P. Greenwald, "Socioeconomic factors and cancer incidence", J Natl Cancer Inst, 1991, 83:551-57.

[13] O.B. Ahmad, C. Boschi-Pinto, A.D. Lopez, C.J. Murray, R. Lozano and M. Inoue. Age standardization of rates: A new WHO standard. World Health Organization 2001. GPE Discussion Paper Series: No.31. Available from: http://www.who.int/healthinfo/paper31.pdf, accessed [6 Jun 2016]