International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2020.121712

Scopus

Research Article

IJCRR
Section: Healthcare
Sci. Journal Impact Factor: 6.1 (2018)
ICV: 90.90 (2018)

Data Management for Healthcare with a Focus on Privacy and Security for Cancer Patients

Hardeep Singh Sandhu[1], Daniel Mago Vistro[2]

School of Computing, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia.

ABSTRACT

Background: Despite the technological advancements in the medical field and patient care, a key area that is lacking for the healthcare sector is on patient privacy and security of the infrastructure enabling and managing patient data in digital format. Numerous security incidents such as ransomware and gross violations of patient security were observed in the healthcare sector, patient privacy must receive more attention from the medical sector. Furthermore, as the severity of the illness increases, it becomes paramount for the patient's privacy to be protected as there are socio-economic impacts on a patient's lifestyle. One such disease that is receiving greater attention and funding is cancer. With cancer-killing, some 8 million patients on an annual basis further research and diagnostics measures can leverage on various data management techniques to improve results accuracy and gain critical insights into the disease.

Methods: As such the cervical cancer dataset from "Hospital Universitario de Caracas" in Caracas, Venezuela is used to explore various data cleaning techniques for filling missing values such as global constants, proportion-based filling of missing values and using central tendency measures. Furthermore, as most data in this form of research tends to be skewed, data transformation techniques are also discussed to normalise the data. Another transformation which is applied extensively in this study is the discretisation methods that is used to bin continuous variables to qualitative groupings that are then used for machine learning techniques.

Results: As medical data can be extremely large, the Apache Hadoop framework is used to upload the dataset and Optimised Row-Column (ORC) is the most optimal way to store and read data is also demonstrated. Several hypotheses were developed and tested to gain some preliminary insights into cervical cancer.

Key Words: Data management, Privacy, Security, Healthcare, Hadoop, Optimized row-column (orc), Cervical cancer

INTRODUCTION

Patient confidentiality is the pillar upon which a patient establishes trust with their healthcare provider[1]. Violations of this trust not only has a repercussion on the healthcare provider but also jeopardises the lifestyle, employment, relationships and reputation of a patient. As for healthcare providers, a breach of trust not only has reputational consequences but also has a financial impact. For example, South Shore Hospital located in the State of Massachusetts, United States was legally obliged to pay $750,000 in damages for a data breach in the year 2010 that compromised the personal information of 800,000 patients[2]. Furthermore, erosion of patient trust arising from mismanagement of patient confidentiality may also damage the participation of patients into research efforts[3].

The digitisation of healthcare data has been occurring at rapid speed and scale, and this has brought about massive improvements in inpatient treatment. However, with digitisation, there is a need to focus on privacy and security in the healthcare sector. According to Leventhal[4], more than 27 million patient records were breached in 2016 for the United States. The stolen patient data was used for several nefarious purposes such as selling the data in the black market to foreign agencies and other criminals, engage in fraudulent activities, and perform illegal financial transactions[5]. The focus on security becomes even more paramount as the operations of the healthcare provider can be brought to a halt via various cyber-attacks. In 2016, ransomware attacks on hospitals were beginning to rise. A specific case of ransomware that is of interest was on Hollywood Presbyterian Medical Center in Los Angeles, California, United States. Hackers took con-

Corresponding Author:

Daniel Mago Vistro, School of Computing, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia;
Email: daniel.mago@apu.edu.my

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 23.06.2020 **Revised:** 29.07.2020 **Accepted:** 26.08.2020 **Published:** 08.09.2020

trol of the information technology systems of the hospital and demanded 9000 bitcoins, which amounted to $3.6 million to cease the attack[6].

As mentioned previously, privacy and security concerns have a big impact on the area of medical research. One research area of interest that has captured the attention of many researchers worldwide is cancer, which causes worldwide deaths of more than 8 million per annum[7]. Despite these morbid statistics, the cancer research has begun paying off with cancer survival rates improving worldwide[8]. Thus, to maintain this positive trend in cancer research, data management along with addressing privacy and security concerns in the medical community is a necessity.

As such, this report seeks to perform a cursory literature review of the privacy and security practices in the healthcare industry and on cancer in general. In addition to the literature review above, data management techniques such as exploratory data analysis, data pre-processing techniques for handling noisy data and data transformation will be done using SAS. The Hive technology is also explored as part of managing the large and unstructured datasets in the healthcare industry. The dataset to be used for this study is a cervical cancer dataset collected from "Hospital Universitario de Caracas" in Caracas, Venezuela.

Literature review on cancer

Cancer begins in a person when the cells become abnormal and begin dividing without control, thus affecting the function of one or more organ systems[9]. The type of cancer is named based on the organ or tissue affected by cancer and as such the types of cancers exceed more than 100 different types. Based on a study conducted by Allemani et al.[10], 75% of the 37 million patients sampled across 71 countries had one of the following 18 cancers; "oesophagus, stomach, colon, rectum, liver, pancreas, lung, breast (women), cervix, ovary, prostate, and melanoma of the skin in adults, together with brain tumours, leukaemia, and lymphomas".

In recent years, cancer research and treatment has begun to utilise vast amounts of data to sift through vast amounts of genetic data, looking for patterns to derive a cure, or to customise patient treatment[11].

The research study conducted by Allemani et al.[10] which involved 37 million patients is one of global scale, and for a study of this magnitude data management and security is crucial. As such, the study adhered to the data governance standards set by the Cancer Survival Group's System-Level Security Policy[10]. Also, to comply with the above policy, all 71 participating countries were required to transmit data only via a "specially configured file transmission utility with 256-bit Advanced Encryption Security". The data was also anonymised by removing patient identifying information such as name, telephone numbers, and addresses amongst others. This study, thus, demonstrates the seriousness of employing data management and security measures to protect the privacy of the patients in cancer research.

Literature review on privacy and security in healthcare

The enormous amount of data generated from the operations of healthcare providers such as patient medical data, transaction data, unstructured diagnosis notes by primary care providers and claims data is stored across various databases and enterprise data warehouses. As such, various information security measures such as access control, authentication and authorisation, cryptographic techniques and security policies are important to manage these databases. As for the organisation as a whole, Master Data Management and Data Governance are some of the ways a healthcare organisation may control its data.

Access control

The primary goal of access control in information security is to selectively restrict user access to data based on the authentication level of the user[12]. This means that an employee in the healthcare sector should only have sufficient access to data to perform their jobs. The primary issue with the healthcare industry is that most users from primary caregivers to non-medical staff members have unrestricted access to patient information[13] resulting in the inappropriate viewing of celebrity status patients[14] and sale of patient information by hospital staff[13].

Authentication

The easiest way to prevent unauthorised access to EHR medical systems and databases is via the use of user authentication. There are primarily two options available for authentication in healthcare which is single-factor authentication and multi-factor authentication, but the choice of authentication needs to be based on a risk analysis of the healthcare provider's system[15]. The type of authentication chosen can be implemented using various techniques such as inputting paraphrases and passwords, fingerprint, iris pattern or voice print matching and via the use of smartcard or a token, or a combination of any of these techniques[16].

Cryptographic Techniques

Ensuring the data that resides and moves around in the healthcare system's networks is protected is done using cryptographic techniques in the healthcare sector, primarily via the use of encryption. Two of the most commonly used cryptographic techniques is the Advanced Encryption Standard (AES); and Ron Rivest, Idi Shamir and Leon Adelman (RSA)[17]. The AES encryption is used by the United States government in all its healthcare data dealings and has proven to be safe and reliable in practice. The RSA is a more secure encryption method that makes deciphering encrypted data more difficult without the right security key, thus provides a higher level of security.

Despite all the privacy and security concerning patient medical data, encrypting medical data is still not a priority for many healthcare providers as it is seen as a hindrance to the workflow of medical professionals and is time-consuming and complex to implement[18]. As a result, 40% of the healthcare organisations in the United States have not yet implemented encryption[19].

Another area of interest is key management. Omotosho, Emuoyibofarhe, and Meinel[20] found that the weakest link in the overall encryption practice is the poor management of encryption keys, which could pose a security risk. As such, healthcare organisations must take a proactive approach to manage their encryption keys appropriately to ensure data breaches can be mitigated.

Master Data Management

Master Data Management (MDM) is an information management method employed to ensure high levels of data quality by addressing completeness, accuracy and timeliness of data[21]. Essentially, an organisation applying the principles of MDM seeks to clean, integrate and link data from many different information technology systems into a single, enterprise-wide point of reference[22]. One of the key ways that healthcare organisations are implementing MDM is by consolidating the information technology systems by migrating to EHR and ERP systems. If consolidation of IT systems is not desirable, healthcare providers can also use various third-party tools like Enterprise Master Patient Index (EMPI). The final approach to implement MDM is via an Enterprise Data Warehouse (EDW) which pulls information from various systems to standardise and tore the information in a central location.

Data Governance

Data governance is an important framework in the healthcare industry that manages the health information lifecycle of various stakeholders in a secure manner[23]. From a patient standpoint, data governance seeks to track various information across a patient's lifecycle such as treatment data, payments and reporting amongst others. As such the focus of data governance in healthcare is primarily on balanced and lean governance, ensuring high data quality, managing data access, improving data literacy amongst healthcare professionals and non-medical personnel, analytical prioritisation to increase the use of analytics in healthcare and MDM[24]. Thus, data governance seeks to guide data management and analytics via standardisation of relevant policies and practices.

RESULTS AND DISCUSSION

Before performing various data management tasks such as data pre-processing and data transformation, the target dataset should be explored to understand the nature and properties of the variables. The first and foremost task in the data exploration phase is to identify the measurement type of each attribute in the dataset. Identifying the level of measurement helps in the interpretation of the attributes so that appropriate summary statistics can be computed to identify the characteristics of the attributes. Each variable will be discussed where the measurement type and the summary statistics will be computed.

Age Attribute

The most appropriate measurement type of the Age variable is of type ratio since this attribute has a true zero (i.e. no age or newly born). Examination of the data reveals that all the values are discrete, but Age can also be treated as a continuous variable. The UNIVARIATE procedure was executed in SAS and corresponding results generated are displayed and discussed Figure 1.

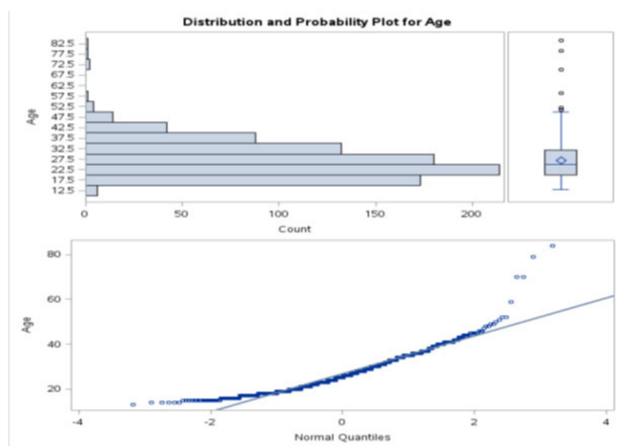

Figure 1: Distribution and Probability Plot for the Age variable.

Based on Figure 1, the distribution of the patients is positively skewed with a large number of patients falling within the range of 17 years old to 32.5 years old. The average age of the patient is 26.82 years old with a standard deviation of 8.5 years. The median of the age variable is 25 years old, and it has an interquartile range of 12 years which captures 50% of the instances in the Age variable.

The discretisation technique is used to convert the Age variable from a ratio measurement to an ordinal variable. The data is broken into various age categories which makes it easier for machine learning algorithms that use categorical variables for classification. However, converting data in this manner can result in loss of information so care must be taken to determine the optimal number of bins to categorise the data, assuming that the categorical ranges are to be of equal width. While there are no fast and hard rules on determining the number of bins, some commonly used general techniques to determine the number of bins is the Freedman-Diaconis rule and the Sturges rule. However, in this case, the histogram bin width of 5 calculated by SAS can be sufficient enough to derive the categories. As such, based on the histogram 8 categories can be made. These 8 bins are "< 18 years", "18 - 22 years", "23 – 27 years", "28 – 32 years", "33 - 37 years", "38 – 42 years", "43 – 47 years" and "> 47 years". The output of the binning in SAS is as per Figure 2 and it can be seen that the binning has preserved the original distribution of the dataset while allowing for various machine learning algorithms to work with it.

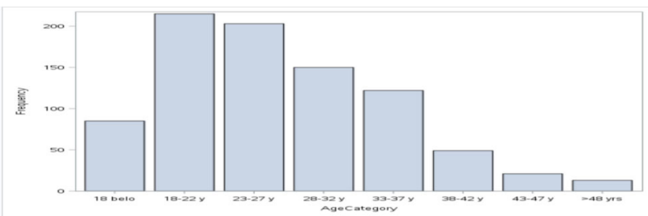

Figure 2: Discretisation of the Age variable into Age Category.

Number of Sexual Partners Attribute

The most appropriate measurement type of the Number of Sexual Partners variable is of type ratio since this attribute has a true zero (i.e. no sexual partners). Furthermore, this variable is also a discrete variable, which means it does not have a fractional component. The PROC UNIVARIATE and SGPLOT procedure were executed in SAS and corresponding results generated are displayed and discussed below.

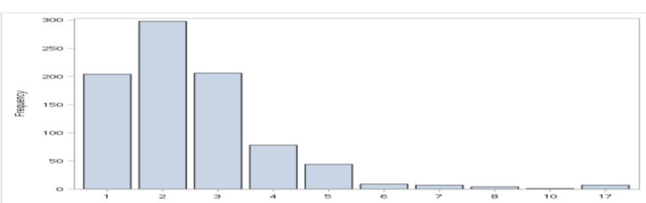

Figure 3: Bar chart depicting the distribution of the Number of Sexual Partners variable.

Based on Figure 2, the distribution of the patients is most likely to follow a poison distribution or negative binomial distribution. The mean of the distribution is 2.53 and the number of sexual partners that a person falls in between 2 to 4 people. The median of the variable is 2 sexual partners, and it has an interquartile range of 1 which captures 50% of the instances in the variable.

First Sexual Intercourse Attribute

The most appropriate measurement type of the First Sexual Intercourse variable is of type ratio. Examination of the data reveals that all the values are discrete, but First Sexual Intercourse represents the Age at first intercourse so it can also be treated as a continuous variable. The PROC UNIVARIATE was executed in SAS and corresponding results generated are displayed and discussed below.

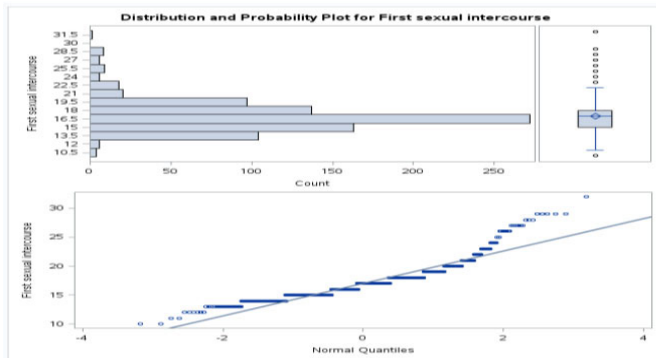

Figure 4: Histogram depicting the distribution of the First Sexual Intercourse variable.

Based on Figure 4, the distribution of the patients is positively skewed with an extreme departure from normality. The average age when a person starts intercourse is 17 years old. The median of the variable is 17 which is the same as the mean, and the variable has an interquartile range of 3 years which captures 50% of the instances in the variable.

The attribute can also be transformed from a ratio measurement type to a categorical variable. The approach used is the same as the Age attribute where the histogram bin width will guide the choice of the categories. As such the categories can be defined as "<13 years", "13 – 15 years", "16 – 18 years", "19 – 21 years", "22 – 24 years", ">25 years", and "No sexual intercourse". Note that the "No sexual intercourse" represents the missing values which are assumed to be genuine responses of persons that never had sexual intercourse. The output of the binning in SAS is as per Figure 4.3.2, and it can be seen that the binning has preserved the original distribution of the dataset while allowing for various machine learning algorithms to work with it.

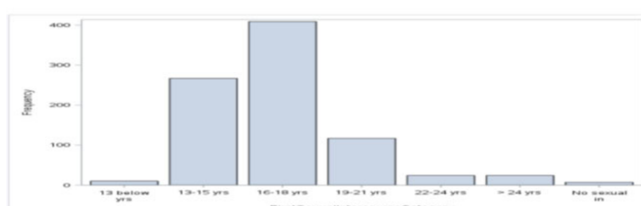

Figure 5: Discretization of the First Sexual Intercourse variable into First Sexual Intercourse Category.

Num of Pregnancies Attribute

The most appropriate measurement type of the Num of Pregnancies variable is of type ratio since this attribute has a true zero (no pregnancies). Furthermore, this variable is also a discrete variable, which means it does not have a fractional component. The PROC UNIVARIATE and SGPLOT procedure were executed in SAS and corresponding results generated are displayed and discussed below.

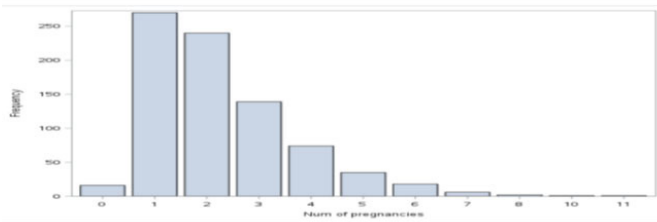

Figure 6: Bar chart depicting the distribution of the Num of Pregnancies variable.

Based on Figure 6, it is clear that the distribution of the patients is most likely to follow a poison distribution or negative binomial distribution. The average number of pregnancies is 2.27. The median of the variable is 2, and the variable has an interquartile range of 2 pregnancies which captures 50% of the instances in the variable.

Smokes(years) Attribute

The most appropriate measurement type of the Smokes(years) variable is of type ratio since this attribute has a true zero (never smoked). Examination of the data reveals that all the values are discrete, but Smokes(years) can also be treated as a continuous variable. The UNIVARIATE procedure was executed in SAS and corresponding results generated are displayed and discussed below.

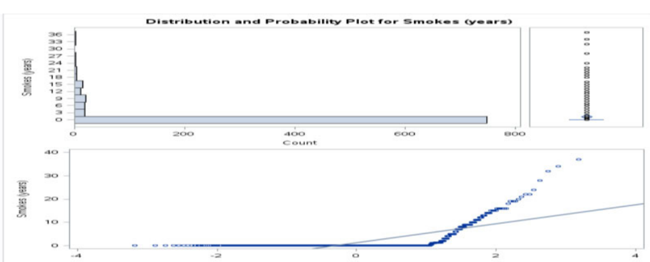

Figure 7: Distribution and Probability Plot for Smokes(years) variable.

Based on Figure 7, the distribution of the patients is similar to an exponential distribution and is highly positively skewed. The average number of years a patient has been smoking is 1.3 years. The median of the variable is 1 year, and the variable has an interquartile range of 0 years which means most of the patients are non-smokers.

IUD(years) Attribute

The most appropriate measurement type of the IUD(years) variable is of type ratio since this attribute has a true zero. The PROC UNIVARIATE procedure was executed in SAS and corresponding results generated are displayed and discussed below.

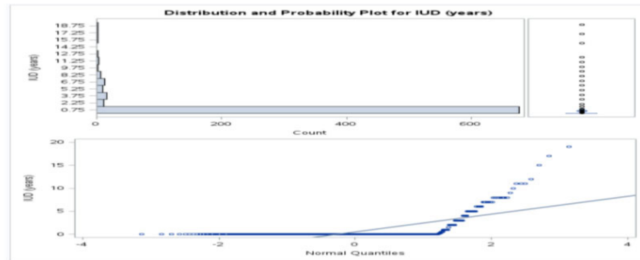

Figure 8: Histogram depicting the distribution of the IUD(years) variable.

Based on Figure 8, the distribution of the patients is similar to an exponential distribution and is highly positively skewed. The average years a patient was on IUD is 0.51 years. The median of the variable is 0 years on IUD, and the variable has an interquartile range 0 years which means most of the patients were not on IUD.

STDs(number) Attribute

The most appropriate measurement type of the STDs(number) variable is of type ratio since this attribute has a true zero (no STDs). Furthermore, this variable is also a discrete variable, which means it does not have a fractional component. The PROC UNIVARIATE and SGPLOT procedure were run in SAS and the corresponding bar chart and summary statistics generated are displayed and discussed below.

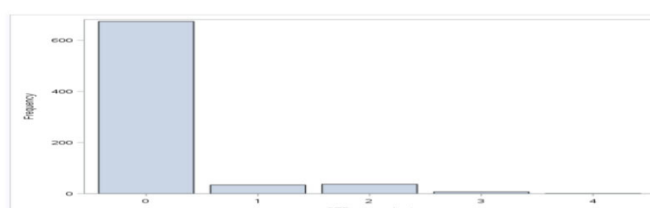

Figure 9: Bar chart depicting the distribution of the STDs(number) variable.

Based on Figure 9, it is clear that the distribution of the patients is most likely to follow a poison distribution or negative binomial distribution. The average number of STDs per patient is 0.17 years. The median of the variable is 0, and the variable has an interquartile range 0 years which means most of the patients do not have STDs.

STDs: Number of diagnosis Attribute

The most appropriate measurement type of the STDs: Number of diagnosis variable is of type ratio since this attribute has a true zero (no diagnosis). Furthermore, this variable is also a discrete variable, which means it does not have a fractional component. The PROC UNIVARIATE and SGPLOT procedure were executed in SAS and corresponding results generated are displayed and discussed below.

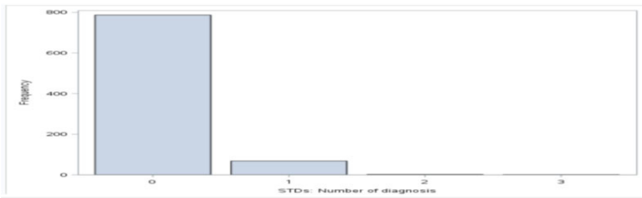

Figure 10: Bar chart depicting the distribution of the STDs: Number of diagnosis variable.

Based on Figure 10, the average number of diagnosis per patient is 0.17. The median of the variable is 0, and the variable has an interquartile range of 0 years most patients have never been diagnosed with an STD.

STDs: Time since first diagnosis Attribute

The most appropriate measurement type of the STDs: Time since first diagnosis variable is of type ratio since this attribute has a true zero (i.e. never had a diagnosis). Examination of the data reveals that all the values are discrete, but STDs: Time since the first diagnosis can also be treated as a continuous variable. The UNIVARIATE procedure was executed in SAS and corresponding results generated are displayed and discussed below.

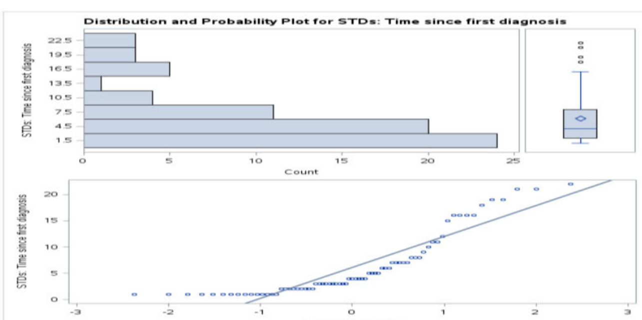

Figure 11: Distribution and Probability Plot for STDs: Time since first variable 35.

Based on Figure 11, it is clear that the distribution of the patients is positively skewed following the characteristics of the exponential distribution family. The average time since the last diagnosis is 6.14 years with a standard deviation of 5.9 years. The median of the variable is 4 years since the last diagnosis, and it has an interquartile range of 3 years which captures 50% of the instances in the STDs: Time since first diagnosis variable.

STDs: Time since last diagnosis Attribute

The most appropriate measurement type of the STDs: Time since last diagnosis variable is of type ratio since this attribute has a true zero (i.e. never had a diagnosis). Examination of the data reveals that all the values are discrete, but STDs: Time since the last diagnosis can also be treated as a continuous variable. The UNIVARIATE procedure was executed in SAS and corresponding results generated are displayed and discussed below.

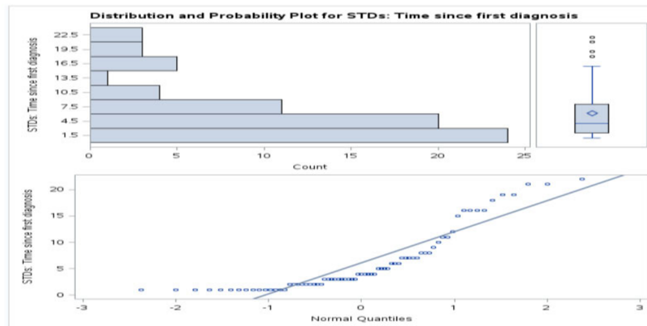

Figure 12: Distribution and Probability Plot for STDs: Time since last diagnosis variable 39.

Based on Figure 12, it is clear that the distribution of the patients is positively skewed following the characteristics of the exponential distribution family. The average time since the last diagnosis is 5.8 years with a standard deviation of 5.8 years. The median of the variable is 3 years since the last diagnosis, and it has an interquartile range of 2 years which captures 50% of the instances in the STDs: Time since last diagnosis variable.

Hormonal Contraceptives (years) Attribute

The most appropriate measurement type of the Hormonal Contraceptives (years) variable is of type ratio since this attribute has a true zero. The PROC UNIVARIATE procedure was executed in SAS and corresponding results generated are displayed and discussed below.

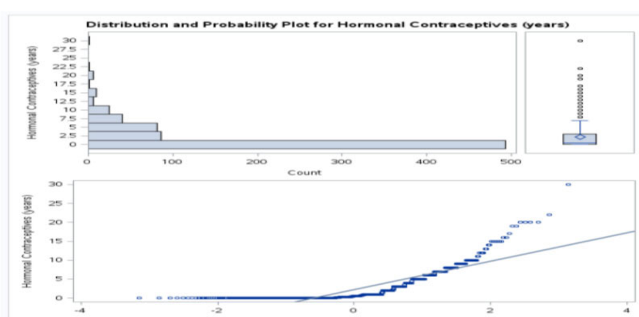

Figure 13: Histogram depicting the distribution of the Hormonal Contraceptives (years) variable.

Based on Figure 13, it is clear that the distribution of the patients is similar to an exponential distribution and is highly positively skewed. The average years a patient was on Hormonal Contraceptives is 2.25 years. The median of the variable is 0.5 years on Hormonal Contraceptives, and the variable has an interquartile range of 3 years.

As part of the data transformation process, the Hormonal Contraceptives (years) variable will be discretised into "Never used Hormonal Contraceptives", "Less than or equal 1 year" and "More than 1 year". This is because examination of the histogram and dataset shows that most of these patients have never used hormonal contraceptives and for those that did use, it is useful to inspect if short-term hormonal contraceptives usage has a similar impact with long-term hormonal contraceptives use on the risk of cervical cancer[25]. The output of the binning in SAS is as per Figure 14.

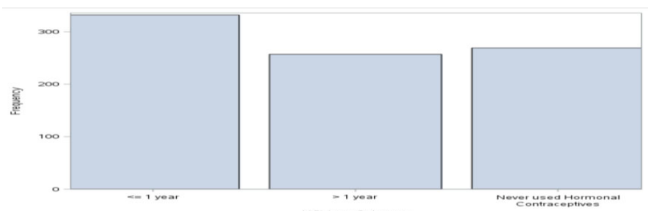

Figure 14: Discretization of the Hormonal Contraceptives (years) variable into HC Years Category.

Boolean Attributes

The Biopsy, Cytology, Schiller and Hinselmann attributes are the response variables in the "Hospital Universitario de Caracas" cervical cancer dataset. The remainder of the attributes is all explanatory variables. All of the variables are of type Boolean, which means the responses are No and Yes which have been encoded into 0 and 1. This means the Boolean attributes have a nominal measurement.

Hypothesis Formulation and Analysis

The dataset on cervical cancer obtained from 'Hospital Universitario de Caracas' in Caracas, Venezuela seeks to predict the indicators or diagnosis of cervical cancer given the risk factors of an individual. The first hypothesis formulated for this dataset is "Does using IUD for short-term (< 1 year) has the same effect at preventing a positive cervical cancer diagnosis as long-term IUD usage (> 1 year)?" The analysis of this hypothesis was conducted using an SQL query on Hive. The results of the query are as per Figure 15. Based on results it is clear that the use of IUD does provide some protection from the diagnosis of cancer itself, as well as in preventing the spread of the Human papillomavirus (HPV) and reduces the abnormal growth of cells in the cervix (detected using CIN). However, note that the occurrence of cancer is lower amongst short-term IUD users.

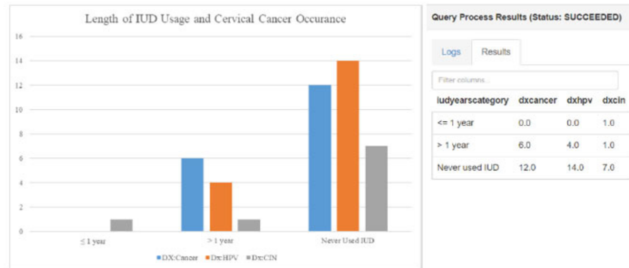

Figure 15: Does short-term IUD usage has the same effect as long-term IUD usage on cervical cancer diagnosis.

CONCLUSIONS

Privacy and security in the healthcare sector is an issue that needs to be taken seriously, but healthcare providers are not doing enough to ensure patient privacy where 27 million patient records were compromised in 2016. The privacy and security of patients become more paramount as the severity of the illness increases for cases such as cancer. As such, the cursory review on cancer and the multiple governance strategies an organisation can use were discussed with a focus on practice such as data governance, MDM, data encryption, authentication and access control to help healthcare providers manage the security of their systems and ensure the privacy of their patients.

Next, a cervical cancer dataset was used to explore various data management techniques such as data exploration, data cleaning and data transformation. For the data cleaning, note that no noisy data was encountered, but missing data was plenty for almost every attribute. The choice of the data cleaning approach taken was based on the characteristics of the data as well its inter-relationships with other variables. The methods used consisted of filling in missing values using a global constant or using the central tendency of the distribution. In the case of the latter, the median was used as the replacement values for the ratio measurement type as the attributes are heavily skewed. As for Boolean variables, the missing value was filled using the outcome with the highest percentage as these were the most likely outcome from the missing values. The data transformation was only performed on ratios exhibiting the ratio measurement type where log transformation was used to reduce the skewness of the distributions as most algorithms make assumptions of normality and several attributes were discretised for use in later analysis.

Finally, the data was uploaded into Hortonworks' implementation of the Apache Hadoop framework and stored in the ORC format to optimise read operations. Five different hypotheses were formulated to explore the likelihood of developing cervical cancer from a given risk factor. The biggest issue with this analysis was that it did not take into account the relationship between various risk factors.

ACKNOWLEDGEMENTS

The authors also wish to express gratitude to the management of Asia Pacific University of Technology and Innovation (APU) for their support.

Conflict of Interest: The authors involved in the current study does not declare any competing conflict of interest.

Funding and Sponsorship: No fund or sponsorship in any form was obtained from any organization for carrying out this research work.

REFERENCES

1. Nathanson, V. (2011) Patient confidentiality one of the most important pillars of medicine, says an expert. The Telegraph [Online] February 23rd. Available from: https://www.telegraph.co.uk/news/health/news/8344126/Patient-confidentiality-one-of-most-important-pillars-of-medicine-says-expert.html [Accessed: 21st March 2018]
2. Goldman, J. (2012) South Shore Hospital Fined $750,000 for Security Breach. eSecurityPlanet [Online] May 28th. Available from: https://www.esecurityplanet.com/network-security/south-shore-hospital-fined-750000-for-security-breach.html [Accessed: 5th April 2018]
3. Kalra, D., Gertz, R., Singleton, P. and Inskip, H.M. (2006) Confidentiality of personal health information used for research. The BMJ [Online] 333(7560). p. 196–198. Available from: https://dx.doi.org/10.1136%2Fbmj.333.7560.196 [Accessed: 25th March 2018]
4. Leventhal, R. (2017) Report: Healthcare Sector Hit Hard in 2016 by Data Breaches. Health Informatics [Online] January 24th. Available from: https://www.healthcare-informatics.com/news-item/cybersecurity/report-healthcare-sector-hit-hard-2016-data-breaches [Accessed: 20th March 2018]
5. Roohparvar, R. (2017) Why Is Information Security Important for the Healthcare Sector. Infoguard Cyber Security [Online] July 27th. Available from: http://www.infoguardsecurity.com/information-security-important-healthcare-sector/ [Accessed: 29th March 2018]
6. Becker's Hospital Review (2016) Hackers shut down Hollywood Presbyterian Medical Center IT systems, demand $3.6M ransom [Online] February 15th. Available from: https://www.beckershospitalreview.com/healthcare-information-technology/hackers-shut-down-hollywood-presbyterian-medical-center-it-systems-demand-3-6-million-ransom.html [Accessed: 28th March 2018]
7. Cancer Research UK (2014) Worldwide cancer mortality statistics [Online] Available from http://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer/mortality [Accessed: 5th April 2018]
8. Allen, K. (2018) Global Cancer Survival Rates Improve, but Wide Regional Gaps Remain. AARP [Online] February 2nd. Available from: https://www.aarp.org/health/conditions-treatments/info-2018/cancer-survival-rates-fd.html [Accessed: 26th March 2018]
9. National Cancer Institute (2015) What Is Cancer? [Online] February 9th. Available from: https://www.cancer.gov/about-cancer/understanding/what-is-cancer [Accessed: 5th April 2018]
10. Allemani, C., Matsuda, T., Carlo V. D., Harewood, R., Matz, M., Nikšić, M., et al. (2018) Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet [Online] 391(10125). p. 1023-1075. Available from: https://doi.org/10.1016/S0140-6736(17)33326-3 [Accessed: 5th April 2018]
11. W. Iftikhar, Z. Mahmood and M. Jayabalan, "A Review of Big Data in the Healthcare Sector: Evaluation and Analysis of Cervical Cancer Data", Journal of Advanced Research in Dynamical and Control Systems, vol. 10, no. 11, pp. 40-51, 2018.
12. Martin, J. (2018) What is access control? 5 enforcement challenges security professionals to need to know. CSO [Online] February 5th. Available from: https://www.csoonline.com/article/3251714/authentication/what-is-access-control-5-enforcement-challenges-security-professionals-need-to-know.html [Accessed: 23rd March 2018]
13. Jayabalan, M. and O'Daniel, T., 2016. Access control and privilege management in an electronic health record: a systematic literature review. Journal of medical systems, 40(12), p.261.
14. Menon, A. K., Jiang, X., Kim, J., Vaidya, J. and Ohno-Machado, L. (2014) Detecting Inappropriate Access to Electronic Health Records Using Collaborative Filtering. Journal of the American Medical Informatics Association [Online] 18(4). p. 498–505. Available from: https://dx.doi.org/10.1007/s10994-013-5376-1 [Accessed: 31st March 2018]
15. Indiana Security and Privacy Network (2016) What Type of Authentication is Right for you? [Online] November 8th. Available from: https://www.inspn.org/type-authentication-right/ [Accessed: 6th April 2018]
16. Jayabalan, M. and O'Daniel, T., 2017, November. Continuous and transparent access control framework for electronic health records: A preliminary study. In 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 165-170). IEEE.
17. Blume, A. (2012) Healthcare data encryption methods for healthcare providers. Health IT Security [Online] November 7th. Available from: https://healthitsecurity.com/news/healthcare-data-encryption-methods-for-healthcare-providers [Accessed: 26th March 2018]
18. Virtru (2015) Why Isn't Everyone Using Encryption in Health Care? [Online] Available from https://www.virtru.com/blog/encryption-in-health-care/ [Accessed: 25th March 2018]
19. Flexential (2016) Healthcare, Encryption and HIPAA Compliant Systems: How They're Connected, Why They're Crucial [Online] Available from http://www.peak10.com/healthcare-encryption-hipaa-compliant-systems-how-theyre-connected-why-theyre-crucial/ [Accessed: 31st March 2018]
20. Omotosho, A., Emuoyibofarhe, J. and Meinel, C. (2017) Ensuring patients privacy in a cryptographic-based-electronic health record using bio-cryptography. International Journal of Electronic Healthcare [Online] 9(4). p.227-254. Available from: https://dx.doi.org/10.1504/IJEH.2017.10003030 [Accessed: 27th March 2018]
21. Jone, D. (2011) Beginners Guide to Master Data Management (MDM). Data Quality Pro [Online] May 2nd. Available from: https://www.dataqualitypro.com/beginners-guide-to-mdm-master-data-management/ [Accessed: 18th March 2018]
22. Eliason, B., Burke, J. and Hess, P. (2014) Master Data Management in Healthcare: 3 Approaches. Health Catalyst [Online] Available from https://www.healthcatalyst.com/master-data-management-in-healthcare-3-approaches [Accessed: 29th March 2018]

23. University of Wisconsin (2017) What Is Data Governance in Healthcare? Available from: https://himt.wisconsin.edu/experience-uw-himt/data-governance-in-healthcare/ [Accessed: 1st April 2018]
24. Sanders, D. (2014) 7 Essential Practices for Data Governance in Healthcare. Health Catalyst. Available from: https://www.healthcatalyst.com/demystifying-healthcare-data-governance [Accessed: 29th March 2018]
25. American Cancer Society (2017) What Are the Risk Factors for Cervical Cancer? [Online] Available from https://www.cancer.org/cancer/cervical-cancer/causes-risks-prevention/risk-factors.html [Accessed: 23rd March 2018]