International Journal of Applied Business and Economic Research

ISSN : 0972-7302

available at http: www.serialsjournals.com

Volume 15 • Number 22 • 2017

Analyzing the Cancer to Improve the Health Using Big Data Analytics: A Review

Suraj Mehta[1] Rabindra Kumar Singh[2] and M. Sivabalakrishnan[3]

[1]*M.Tech Student, School of Computing Science Engineering, VIT University, Chennai. Email: mehtasuraj.2016@vitstudent.ac.in*
[2]*Assistant Professor, School of Computing Science Engineering, VIT University, Chennai. Email: rabindrakumar.singh@vit.ac.in*
[3]*Associate Professor, School of Computing Science Engineering, VIT University, Chennai. Email: sivabalakrishnan.m@vit.ac.in*

ABSTRACT

Cancer is disease of abnormal cell growth which invades or spread over whole or part of body which is major reason of causing death in medical science. The main reasons are limitation of medical sources, late detection of cancer and unable to use or not able to afford existing source efficiently. Currently to prevent the cancer, the best hope is early detection of cancer which can be archived by the knowing the cancer that co-occur with particular other disease in advance could be helpful to physician able to better tail or to treat cancer. Cancer is group of diseases often associated with other disease, thus the distribution of cancer that promoting their co-occurrence and co-relation could help in cancer predication and diagnostics. The aim of this paper is to study existing work has been done in this approach previously to know how the co-relation between cancer and other disease can be found by different methods efficiently. We review the different ways of statistical methods and machine learning algorithm of big data analytics which helps in to reduce risk factor of cancer and predict the cancer in early stage. We study on patients datasets how Chi square, Kaplan Meier, Proportion Confidence Interval methods and other methods are used to find out co-relation between cancer and other disease. We also research how big data analytics are useful in analyzing the chances of cancer occurring with other disease.

Keywords: Cancer, Big Data Analytics, Statistical Analysis, Cancer-disease relationship, co-relation

1. INTRODUCTION

Cancer is disease of growth of abnormal cells called tumor which is spread over whole or part of body that leads toward disability or sometimes early death. Cancer has many stages and many of times detecting cancer is complex, medical science shows detection in early stage can cure cancer completely. So to cure and for treatment early detection is very important. Cancer is more complex disease. According to Alexander Chang

a single tumor has more than 100 billion cells and every cell has its own mutation individually. Cancer is always changing, evolving and adapting, so finding right cancers mutation tumor is very important and also most difficult task. It is challenge to identify those cancers mutation which cannot be done by usual or regular techniques. This mutation data is very huge; to analyze these data big data comes into picture. There is no doubt to analysis and cure cancer has many ways but handle this kind of data is not easy.

The big data analytics and machine learning are gives some edge over here to analysis it. Usman Iqbal et.al proposed in "Cancer-disease associations: a visualization and animation through medical big data" 'Elsevier Journal 2016' that if we analysis and find association between cancer and other disease it helps to detect cancer and help in new treatment method which leads toward improve health care [1]. For example 20% incidence risk of breast cancer is found with type2 diabetes which is preventable disease [1]. This is one of the ways to analysis cancer with big data analytics. Our study is to understand and continue this concept. In this paper we review this concept more deeply and understand up till now how much work has been progress in this direction. Andreas Holzinger et.al also proposed same kind of concept for Rheumatic disease in "Disease-disease relationships for Rheumatic disease IEEE 2012" with help of Medline dataset [5]. This procedure maybe we can apply to find relationship between cancer and other disease. On this concept many researcher has been worked from all over the world like Usman Iqbal et.al study on Taiwan people[1], Andrea Gini et.al study on northern Italy people which they mention about their work in "Cancer among patients with type 2 diabetes mellitus: A population-based cohort study in northern Italy Cancer Epidemiology 2016" [4], Amrita Singh et.al did similar kind of study which they mention in "Association and multiple interaction analysis among five XRCCI polymorphic variants in modulating lung cancer risk in North Indian population" [12]. Hui Cao et.al proposed some association methodology in "Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics 2005" [2] and "A Statistical methodology for analyzing co-occurrence data from a large sample Science Direct Journal 2006" [3] which we elaborated below.

2. LITERATURE REVIEW

2.1. Usman Iqbal et.al proposes, "Cancer-disease associations: a visualization and animation through medical big data" in Elsevier Journal 2016 [1]

According to WHO organization cancer preclusion is important part of cancer control because up to 40 percent of all cancer death can be prevented. Larsson et.al observed that 20 percent occurrence risk of breast cancer is found with type2 diabetes which is preventable disease [1]. Furthermore studies show that hyper insulinemia with insulin resistance might raise the risk of breast cancer. This project they built in "Science and Technology, Taipei Medical University, Taiwan". They used datasets from Taiwan's national health Insurance of period 1st Jan 2000 to 31st Dec 2002. They start their study by assuming that 2 disease are related if they occur at least once in same individual in excess of above mention period. They observed and compute association between cancers and other disease. They made 100 subsets each for both genders. They made all possible pair wise combination of disease to disease association found in those patients who are observed over that mention 3 years of period. They made 3.9 million and 4.4 million exclusive associations with their co-relation between those pairs. If the association between disease with less than 5 co-occurrences or negative correlation they are excluded from datasets. They consider 0.001 as level of significance for hypothesis testing. After creating disease to disease association database they use 2D dynamic motion bubble chart technique to visualize the relation between 2 diseases which is helpful to

understand any medical related person visually. Limitation of their study is first of all them assuming that 2 diseases are associated with each other if they were found in same patient over that they observing for 3 years. And secondly they more concentrate on only 9 common cancer types which are stomach, colorectal, rectum rectosigmaid, liver, lung, breast, cervical and prostate cancer they not study more about other cancer categories.

2.2. Hui Cao et.al proposes, "Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics" in 2005[2]

In this paper they applied co-occurrence statistics to determine disease finding associations. For that they use two methods, chi square statistics and proportion confidence interval (PCI) technique to determine the dependence of pairs of disease and then used heuristic cutoff values for association selection. First they made list of similar disease for instance, a patient suffering with coronary artery disease and having history of "heart bypass surgery", "heart attack" and "persistent unstable angina" express as "Coronary Artery Disease (Myocardial Infarction, Coronary Artery Bypass Grafting Surgery, now with persistent unstable angina)" [2]. After that they eliminate very common finding such as pain, fever and cold which are clinically non-informative and also eliminate such underlying of diseases, for example chest pain in myocardial infarction patient. Limitation of this study is they are using "NLP tool" for extract disease and determine terms but disadvantage is NLP system can only extract the data which is in lexicon covers a large medical term. Their study report that association extracted 94 % by χ^2 method and 76.8 % by PCI are true which is statistically good but not clinically [2].

2.3. Hui Cao et.al proposes, "A Statistical methodology for analyzing co-occurrence data from a large sample" in Science Direct Journal 2006[3]

In medical field common problem is large amount of clinical data, so finding important associations with large dataset is most difficult task due to multiple concurrent hypotheses and selecting poor associations which are statistically not make such problem but it is not significant clinically. In traditional chi square test most null hypotheses rejected due large dataset. So for selecting clinically significant association they developed automatically calibrated cutoff. The problem is when huge dataset, test detect smaller deviation from null hypothesis. For instance there may be small association between kidney malfunction and cough except this association is not medically motivating [3]. So they made two sample table first one contain those entries which satisfy Cochran's rule and other one contain who not satisfy Cochran's rule and then uses modified chi square method. Drawback of this study is in this technique can not applicable for 3 or higher dimension table. Even if volume test worked as accepted for co-occurrences data they have applied carefully to finding cancer and disease association because of change in assumed sampling schema for 2 X 2 table.

2.4. Andrea Gini et.al proposes, "Cancer among patients with type 2 diabetes mellitus: A population-based cohort study in northern Italy" in Cancer Epidemiology 2016[4]

In this paper they study association between type-2 diabetes mellitus and cancer on patient's age between 40 and 84 from year 2002 to 2009 of northeastern Italy with 95 % of confidence Intervals [4]. They start follow up patient after 90 days of cancer diagnosis in type-2 diabetes and follow up more than 3 years. Then "SIR(standardized incidence ratio)" calculate as observed for expected number of cancer cases. Expected

number of cancer cases computed as per age, sex and calendar year for specific incidence rates. They find survival probabilities by Kaplan-Meier method .They found those female patients who have type 2 diabetes in them breast cancer risk factor increase. And also patient's age is between 65 and 74 they have high risk factor of lung and prostate cancer. Limitation of this study is there is nothing mention about some potential factor like body of mass index, obesity status, smoking and drug prescription. Study involved only those patients whose age is more than 40 year.

2.5. Andreas Holzinger et.al proposes, "Disease-disease relationships for Rheumatic disease" in IEEE 2012[5]

In this paper they extract data from MEDLINE and apply co-occurrence analysis on text mining techniques and statistical methods to evaluate importance of association between disease entities such as their names and drug names. They collect data from MEDLINE and build knowledge disease association using text-mining tool. Then they used Point wise Mutual Information (PMI) to measure strength of association. PMI provide an indication of how frequently query and perception co-occur than expected one by changes and ranked these values and frequency together for each disease. Yildirim et.al invent one method in which they found symptoms in various articles and then create dataset with number of occurrence of symptom of each and every disease then applied hierarchical clustering analysis in order to find similarity between the diseases [5]. For text mining they use FACTA which is developed by "NACTEM". It is "text search engine" for "MEDLINE" which mostly covers six categories of biomedical model: diseases, human genes/ proteins, symptoms, enzymes, drugs and chemical compounds [5].

2.6. Alison J. Price et.al proposes, "Circulating Folate and Vitamin B12 and Risk of Prostate Cancer: A Collaborative Analysis of Individual Participant Data from Six Cohorts Including 6875 Cases and 8104 Controls" in European urology Journal 2016[6]

Objective of this paper is to find association between circulating folate, vitamin B12 and threat of Prostate Cancer. For this study they collect blood samples from 1981 to 2008 and avg. follow up of 8.9 yr with 7.3 standard deviation of Prostate cancer in circulating folate and vitamin B12 [6]. This they calculate by multivariable adjusted conditional logistic regression method. In their study they uses 10^{th} and 90^{th} percentile of folate and vitamin B12 values. They consider age at blood collection, body mass index, height, marital status, educational status and cigarette smoking factors to associate with prostate cancer risk. They used χ^2 method values with and without study time's linear trend interaction term for comparison purpose of cancer risk. For testing heterogeneity of trends, in which each matched set were assigned a value with their matched case of subgroup using a method similar to a meta-analysis. For natural log transformation values they used approximate a normal distribution and geometric mean concentrations which is calculated by using analysis of variance. All statistical analysis done by using Stata at 0.05 level of signification.

2.7. Ashish Banerjee et.a.l proposes, "Discovering and Validating Breast cancer Treatment Correlation using an Associative Memory Model and Statistical methods" in 2015[7]

They use associative memory model for aggregating tumor registry and EMR datasets. First they store information of atoms with organizational vector space of 2^{120} distinct item location which is 4 32-bit characters with 8bits locations. That gives associative memory model of unified and compact representation of any data type, dynamic or stationary, structured or unstructured of arbitrary size and granularity. That

made mapping of anything to anything in any of billion associative dimension of associative memory model. Then they use scatter plot to show visualize graph for correlation between queried factors and uses statistical hypothesis testing to evaluate significance of correlations. First they test correlation between patient characteristic factors and administered treatment factor to evaluate observed frequencies of multiple pair wise combinations of factor categories are homogeneously distributed using chi-square method at 0.001 of level of significance. Then they take 2 samples, one of the samples consists of patient cancer stage and hormonal drugs and other sample consist of rest of population who received same identical set of drugs whose category is unknown. Then they use Kolmogorov-Smirnov test for normality. If both sample found normally distributed then they apply "t-test" and "non-parametric Mann-Whitney U test" (Wilcoxon rank sum test). The result obtained by this study is combination of clinically established and anomalous pattern between patient characteristics and administered intervention which help in future treatment plans. Drawback of this study is not sufficient data system and inexperience of clinical practice variation.

3. RELATED METHODS

3.1. Pearson's Correlation Method

In this method we measure linear dependence between two variables. Its values lie between -1 to +1. We can find Pearson's correlation coefficient by taking covariance of two variables divided by product of their standard deviation as shown below.

$$\rho_{x,y} = \frac{\mathrm{cov}(x, y)}{\sigma_x * \sigma_y}$$

Take cancer as one variable(x) and other disease as other variable(y) to find Pearson's correlation [1]. Like by all possible combination finding person's correlation and made database of cancer and disease association to visualize the relationship.

3.2. Chi square method (χ^2)

In χ^2 method, finding association strength between diseases and finding pair by comparing their likely and observed frequencies. If the difference is large then that pair are neglect as they reject null hypothesis. Then apply volume test over 2 X 2 sample table which has nearly 18,432 entries to interpret distance from independence [2]. For volume test p value is calculated as:

$$\varepsilon(\chi^2) = \frac{\left(\pi \times \frac{\quad}{\quad}\right)^{-} \left[\frac{\sqrt{\mathrm{IJ}}}{(\mathrm{IJ})}\right] \times \frac{\mathrm{IJ}-1}{\sqrt{\mathrm{IJ}}}}{\begin{pmatrix} \mathrm{IJ}-1 \\ \mathrm{IJ}-1 \end{pmatrix}}$$

Where n, D, I and J are the total case in the sample, degree of freedom, the number of rows and the number of columns respectively. For the stricter cutoff value they consider 96th percentile [2].

They set D(s) as set of diseases and F(s) is finding present. If d E D(s) a and f E F(s) then we said that they co-occur [2].

3.3. Modified Chi Square method

Chi square is corrected by following formula [3]:

$$\chi^2 \text{corr} = \frac{\left(|ad - bc| - \frac{1}{2}n\right)^2 n}{m_1 m_2\, n_1 n_2}$$

The hypothesis of not dependence between disease and finding is express as:

$$H_0 = \prod_{ij} = \prod_{i+} \text{X} \prod_{+j}$$

Then as Diaconis and Efron suggested adjusted in value test and made adjusted chi square test formula as [2]:

$$\varepsilon(\chi^2) = \frac{(|nS|)^{\frac{D}{2}} C_{IJ}(n-1)/\tau(IJ)}{\binom{n+IJ-1}{IJ-1}}$$

Then they apply conditional and fixed margin volume test as.

$$\in(\chi^2 | r) = \frac{(\pi S)^{\frac{D}{2}} r^I (J)\, r^J \left(\frac{I+1}{2}\right)}{\left(\frac{D}{2}\right)! \, r\,(J(I+1)/2)} \times \frac{(\pi_{i=1}^1 r_i)^{(J-1)/2}}{(\pi_{i=1}^1 r_i)^{(J-1)} \left(\frac{IJ}{2n}+1\right)^{I(J-1)}}$$

Where r_i are the components of vector

$$r = (1 - \square)/I + \square r, \square$$

$$= 1/(1 + (IJ/2n))$$

and $$I = (1,1\ldots1);$$

$$r = (r1, r2 == rI) \text{ are I X I vectors}$$

and $$S = (\chi^2/n).$$

The algorithm they developed for p-value [3]:

Step 1: Order the adjusted levels of χ^2 statistic from smallest to largest, *i.e.* $p(1), p(2) \ldots p(n)$ where $p(i)$ denotes the i^{th} ordered adjustment.

Step 2: Construct 1 $p(i)$ and N$p(i)$ where N$p(i)$ is the number of adjusted p-values greater than $p(i)$.

Step 3: Plot I $p(i)$ versus N$p(i)$ for all $i = 1,2,\ldots n$

Step 4: Fit no-intercept line to pairs(I $p(i)$, N$p(i)$), $i = 1,2,..n$

Their result shows 72.3 % was consider being direct associated, 21.9 % to be indirect association and 5.8 % to be false association.

3.4. Confidence Interval of a proportion

The proportion of diseases can be calculated using co-occurrence data. Due to some rare disease consider 95% confidence interval (CI). After experimenting cutoff value between 0.1 and 0.4 on small dataset they decide 0.15 values. While CI calculates as follow [2]:

$$CI = \frac{r}{n} + X_{1-\frac{\alpha}{2}} \sqrt{\frac{r \times (n-r)}{n^3}}$$

Where r is frequency of f_j in the d_i population using 5 % level of significance [2]. To finding disease co-occurrence, they set D(s) as set of diseases and F(s) is finding present. If d E D(s) a and f E F(s) then we said that they co-occur.

3.5. Point wise Mutual Information (PMI)

PMI gives discrepancy between probabilities of two variables coincidence given by joint distribution and individual distribution. For two words w_i and w_j have probabilities P (w_i) and P (w_j). Their mutual information PMI (w_i, w_j) is defined as[5]:

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)\,P(w_j)}$$

4. CONCLUSION

The relationship between cancer and other disease associations was accomplished by big data analytics. This analytical study has been done by using various statistical methods like Pearson's correlation, Chi square and Proportional Confidence interval method. Such information can be used to improve cancer treatment for medical researcher.

References

Usman Iqbal, Chun-Kung Hsua, Phung Anh (Alex) Nguyena, "Cancer-disease associations: A visualization and animation through medical big data", Computer Method and Programs in Biomedicine, I27, 44-5I, 2016.

Hui Cao, Marianthi Markatou, Genevieve B. Melton, "Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics", AMIA Symposium Proceedings, Page no. 106, 2005.

Hui Cao, George Heipcsak, Marianthi Markatou, "A statistical methodology for analyzing co-occurrence data from a large sample", *Journal of Biomedical Information*, Science Direct, Vol. 40, page no. 343-352,2007.

Andrea Gini, Ettore Bidoli, Loris Zaneir, "Cancer among patients with type 2 diabetes mellitus: A population-based cohort study in northeastern Italy", *IJCEDP*, vol.41, page no.80-87, 2016.

Andreas Holzinger, Klaus-Martin Simonic, Pinar Yildirim, "Disease-disease relationships for rheumatic diseases", *IEEE, ICCSA*, Vol. 36,Page no. 573-580,2012.

Alison J. Price, Ruth C. Travis, Paul N. Appleby, "Circulating Folate and Vitamin B12 and Risk of Prostate Cancer: A Collaborative Analysis of Individual Participant Data from Six Cohorts Including 6875 Cases and 8104 Controls", *European Urology*, 2016.

Ashis et. al, "Discovering and Validating Breast Cancer Treatment Correlations using an Associative Memory Model and Statistical Methods", International Conference on Healthcare Informatics, 2015.

Junqi Liu, Di Zhao, Ruitai Fan, "Shared and unique mutational gene co-occurrences in cancers", *Biochemical and Biophysical Research Communications*, Vol.465, Page no. 777-783, 2015.

Bao C. Q. Truonget.al., " High Correlation of Double Debye Model Parameters in Skin Cancer Detection", *IEEE*, 2014.

Atınç Yılmaz, Kür¸sat Ayan, Enes Adak," Risk Analysis in Cancer Disease By Using Fuzzy Logic", *IEEE*, 2011.

Hong-Jun Yoon, Georgia Tourassi, "Investigating the Association between Sociodemographic Factors and Lung Cancer Risk Using Cyber Informatics", *IEEE*, 2016.

Amrita Singh, NavneetSingh, Digambar Behera, Siddharth Sharma, "Association and multiple interaction analysis among five XRCC1 polymorphic variants in modulating lung cancer risk in North Indian population", *DNA Repair*, 2016.

Ik-Soo, Huh, Sohee-Oh, Taesung Park, " A Chi-square test for detecting multiple joint genetic variants in Genomewide association studies*", IEEE, ICBBW*, 2011.

Zhe Liu, Yuanyuan Shen, Jurg Ott, "P-value Distribution in Case-Control Association Studies", *IEEE, ICBBW*, 2010.