Advances in Medical Informatics
ISSN: 2249-9466 & E-ISSN: 2249-9474, Volume 3, Issue 1, 2013, pp.-037-040.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000189

USING DATA MINING FOR ORAL CANCER RISK STRATIFICATION IN TERMS OF AGE, GENDER AND SOCIOECONOMIC STATUS

SHARMA N.[1]* AND OM H.[2]

[1]Padmashree Dr D.Y. Patil Institute of Master of Computer Applications, Pune- 411 044, MS, India.
[2]Computer Science and Engineering Department, Indian School of Mines, Dhanbad- 826 004, Jharkhand, India.
*Corresponding Author: Email- nvsharma@rediffmail.com

Received: November 28, 2013; Accepted: December 18, 2013

Abstract- Oral cancer is a major concern in India as the literature shows large number of cases getting registered every year. The predominant medical model approach to research and prevention on the risks of the disease gives very little attention to the effect of demographic information. There is uncertainty and limited recognition of the relationship between age, gender, socioeconomic inequalities and oral cancer, therefore we aim to quantitatively assess the association between various demographics and oral cancer incidence risk. Data mining is applied to oral cancer database in order to understand association. The results show that females have less probability to develop oral cancer in comparison to male. The people in the age group of 50-55 years are more prone to build up malignant cells. The mean age of diagnosis is 58.7 years. It is observed that the patients of lower socioeconomic status have shown the significant increase in incidences.

Keywords- Socioeconomic Status, Data Mining, Age, Association, Oral Cancer, Weka, Gender

Introduction

Data collection using computer technology has existed since 1960s, and the gathered data subsequently was used by large corporations to improve their business. The development of more sophisticated computer databases in 1980s allowed the business to flourish manifold [1]. The concept of data mining is relatively new and was initially considered as more of a hyped up idea. Nevertheless, nowadays it is well thought-out to be a useful and reliable tool for businesses and organizations, as it assists to analyze and make right decision. Data mining is the process of congregation of information from various sources and perspectives, and subsequently evaluating it and presenting it in the most useful form for the task at hand [2]. Data mining applications are frequently designed around the unambiguous requirements of an industry sector or even tailored and built for a single organization. This is because the data patterns of one organization may vary than that of other. For example, an organization might need data mining application to track client spending habits in order to detect unusual transactions that might be fraudulent or a government body uses data mining application to detect association between individuals who may be involved in terrorist activities [1]. However, it had a limited use in medical science until recently [3]. But, now the medical community has also become aware that they can be greatly benefitted by applying data mining and extracting different knowledge from long-term health data collection. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to preconceived notions, miscalculation and excessive medical costs which affects the quality of service provided to patients. Integration of clinical decision support with computer-based patient records could decrease medical blunders, improve patient wellbeing, reduce unwanted practice variation, and enhance patient outcome [4]. Therefore, we attempt to forward the benefits of data mining to healthcare industry as well.

Data Mining is the science of extracting critical information from the large amount of existing raw data and deploying that information across the organization [1,5,6]. Patterns and trends discovered go beyond straightforward examination and can answer inquiries that cannot be answered through basic question and reporting systems. However, data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Hidden patterns, outcome prediction, generation of valuable information and focus on large datasets and databases are the key properties of data mining [7-9]. However, it doesn't dispense with the need to know the business, to comprehend the data, or to comprehend analytical techniques. Data mining discovers hidden information in data; but cannot tell the value of the information to your organization. Important patterns might already be known as a result of acquaintance to business domain and working with data over time. Notwithstanding, data mining can confirm or qualify such empirical observations in addition to finding new patterns that may not be immediately apparent through simple observation. The different pattern that can be uncovered using data mining includes association analysis, characterization, cluster analysis, discrimination, classification and regression, outlier analysis and evolution analysis.

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is used to identify strong rules discovered in databases using different measures of interestingness [10], which can be applied in different areas such as e-commerce, sports, census analysis, healthcare, etc. An association rule put in the picture about the association between two or more items. These rules are if/then statements that help uncover relationships between apparently unnecessary data in a relational database or other information repository [11-17]. An association rule has two parts: an antecedent (if) and a subsequent or consequent (then). An antecedent is an item found in the data and a consequent is an item that is found in combination with the antecedent. The usefulness of the association rule is determined by two popular interestingness criteria - support and confidence.

Support: The rule holds with support 'supp' in T (the transaction dataset) if supp% of transactions contain $X \cup Y$ [18].

$Supp(X \rightarrow Y) = P(X \cup Y)$.

Confidence: The rule holds with confidence 'conf' in T if conf % of transactions that contain X also contain Y [18,19].

$$Conf\ (X \rightarrow Y) = P(Y \mid X) = Supp(X \cup Y) / Supp(X)$$
$$= P(X \text{ and } Y) / P(X)$$

The objective of this paper is to apply data mining to oral cancer database for oral cancer risk stratification in terms of age, gender and socioeconomic status. Oral cancer was estimated to be the 8th most common cancer worldwide in 2000, with approximate 267,000 new cases and 128,000 deaths, and with the maximum burden in developing countries [20]. In spite of availability of literature on the effects of poverty and inequality on health [21], the effect of demographic information on oral cancer is given little attention in a predominant medical model approach to research and prevention on the risks of the disease [22]. Since there is uncertainty and limited recognition of the relationship between age, gender, socioeconomic inequalities and oral cancer, we aim to quantitatively assess the association between various demographics and oral cancer incidence risk. The rest of the paper is organized in the following manner: section 2 discusses materials and methods and section 3 presents the experimental results. In section 4, the result is discussed briefly and section 5 presents the conclusion. Finally, at the end, references are mentioned.

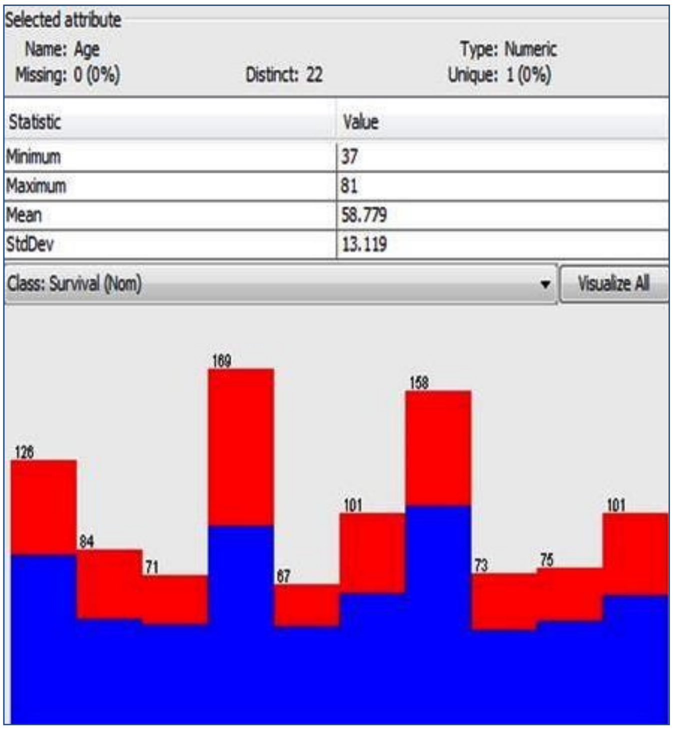
Materials and Methods

Oral cancer data is collected for the period of five years in non-randomized or non-probabilistic method. A retrospective chart review from ENT and Head and Neck Department, the records of the cancer registries of Tertiary Care Hospitals, OPD data sheet and from archives of departments of Histopathology, Surgery and Radiology was carried out for collecting the data related to oral cancer and for creating the database for this research work. The dataset is based on the records of all the patients who reported with a lesion and treated at the centre from Jan 2004 and June 2009. The clinical details, personal history and habits were collected manually from the records to complete the datasheet of the patients. The complete process of data preparation, data integration and data cleaning (i.e. removing missing values, noisy data and inconsistent data) was strictly followed to create the database of oral cancer patients [23]. The database contains the records of 1025 oral cancer patients, which has been described with the help of 33 data columns (variable).

Results

The association among various valuable data pertaining to demographic information and survivability of the cancer patients is derived using data mining tool WEKA3.7.9. This is Java based open source tool created by researchers at the University of Waikato in New Zealand [24]. It is a collection of many data mining and machine learning algorithms. It also includes data pre-processing, clustering, classification and association rule extraction. The oral cancer data is initially stored in MS Excel sheet, which is converted into comma separated values (.csv) file format and subsequently into attribute relation file format (.arff) as .arff is accepted by the WEKA tool.

[Fig-1] shows age as a numerical attribute. Minimum age that can be affected by oral cancer is 37 years, maximum is 81 years and mean is 58.779 years, with standard deviation of 13.119. [Fig-2] shows age as a discrete attribute. Continuous variable age is converted in to discrete variable with 10 age brackets, which are 37.0-41.4, 41.4-45.8, 45.8-50.2, 50.2-54.6, 54.6-59, 59-63.4, 63.4-67.8, 67.8-72.2, 72.2-76.6, 76.6-81. The age bracket which has more number of incidences of oral cancer is 50.2 - 54.6, followed by 63.4 - 67.8. [Fig-3] reveals that 636 patients out of total are male where as 389 patients are females. [Fig-4] presents that 690 patients are in low socioeconomic status, whereas 218 patients are in middle and 117 patients are in high socioeconomic status.

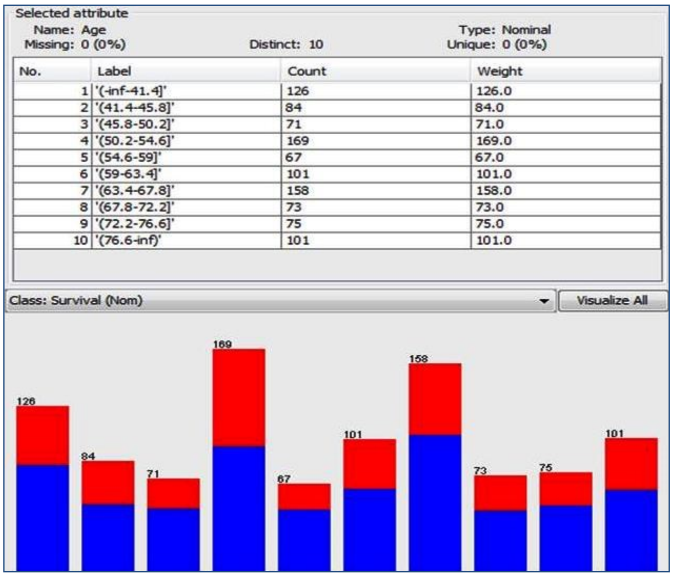


Selected attribute
Name: Age Type: Numeric
Missing: 0 (0%) Distinct: 22 Unique: 1 (0%)

Statistic	Value
Minimum	37
Maximum	81
Mean	58.779
StdDev	13.119

Class: Survival (Nom)

Fig. 1- Oral Cancer Stratification on Age (Numeric)

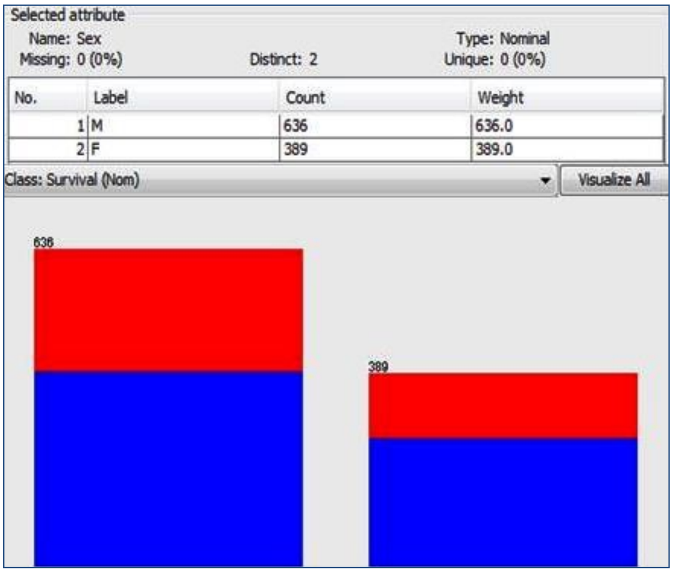


Selected attribute
Name: Age Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-41.4]'	126	126.0
2	'(41.4-45.8]'	84	84.0
3	'(45.8-50.2]'	71	71.0
4	'(50.2-54.6]'	169	169.0
5	'(54.6-59]'	67	67.0
6	'(59-63.4]'	101	101.0
7	'(63.4-67.8]'	158	158.0
8	'(67.8-72.2]'	73	73.0
9	'(72.2-76.6]'	75	75.0
10	'(76.6-inf)'	101	101.0

Class: Survival (Nom)

Fig. 2- Oral Cancer Stratification on Age (Discrete)

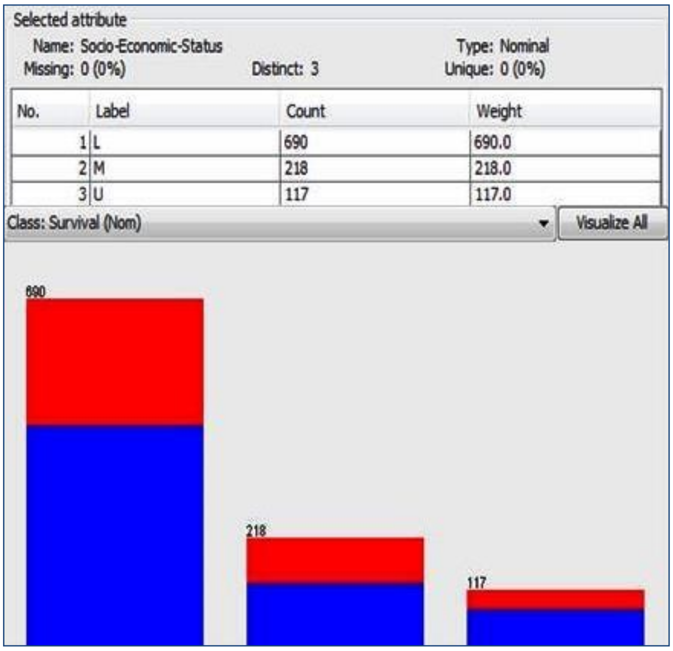


Selected attribute
Name: Sex　Type: Nominal
Missing: 0 (0%)　Distinct: 2　Unique: 0 (0%)

No.	Label	Count	Weight
1	M	636	636.0
2	F	389	389.0

Class: Survival (Nom)

Fig. 3- Oral Cancer Stratification on Gender



Selected attribute
Name: Socio-Economic-Status　Type: Nominal
Missing: 0 (0%)　Distinct: 3　Unique: 0 (0%)

No.	Label	Count	Weight
1	L	690	690.0
2	M	218	218.0
3	U	117	117.0

Class: Survival (Nom)

Fig. 4- Oral Cancer Stratification on Socio-Economic-Status

Discussion

Various research works on oral cancer is carried out and published to understand development, control and treatment of the disease. However, most of them have used the contemporary analysis methods. Yeole, et al. [25] have studied the data on survival of oral cancer patients registered by the Bombay population-based cancer registry, India, during 1992-1994. They have used cox model for analysis and results clearly shows that detecting oral cancer in premature stage, when these are acquiescent to single modality therapies, offers the greatest chance of long-term survival. Misra, et al. [26] performed a prospective clinic-histological study of premalignant and malignant lesions of the oral cavity using histological correlation and compared it with a 10-year retrospective data related to age distribution, incidence, site, personal habits and type of lesion. The study shows that the histology along with a detailed clinical workup is found to be a valuable, dependable, and precise diagnostic technique for lesions of the oral cavity. Warnakulasuriya [27] and Conway, et al. [28] have performed meta-analysis on subgroup like SES measure, age, sex, global region, development level, time-period and lifestyle factor adjustments. The result showed that low SES was significantly associated with increased oral cancer risk in high- and lower-income countries, across the world, and remained when adjusting for potential behavioral confounders which proves that oral cancer risk is significantly associated with low social-economic status and lifestyle. Hashibe, et al. [29] presented the evidence to focus on lifestyles factors, higher SES index, education and income and its association with decreased risk of oral premalignant lesions. Conway, et al. [30] assessed the socio-economic inequalities, pattern, magnitude, and time trends of the distribution of oral cancer in Scotland. Ramchandran, et al. [31] reviewed the past studies on oral cancer and compared the same with current trend. It was found that apart from chewing habits, there are few other factors like illiteracy, poverty, low caloric diet and non-availability of free medical facility are the cause for rise in oral cancer incidences.

Data mining is applied in the field of healthcare since many years in order to improve the quality of services and treatment offered. There are many researchers and authors who have used data mining techniques and algorithms on oral cancer data for helping the practitioners and benefit of the society at large. Exarchos, et al. [32] monitored the oral cancer evolvement and progression during the whole follow-up period (i.e. 24 months) so as to evaluate the post treatment condition of a patient and also to infer about the probability as well as approximate timing of a potential reoccurrence. Dynamic Bayesian Networks (DBNs) are used to capture the temporal dimension of the disease and procure new and informative biomarkers which correlate with the progression of the disease and identify early potential relapses. HariKumar, et al. [33], compares the classification accuracy of the TNM (tumour, lymph nodes, metastatis) staging system with that of Chi-Square Test and Neural Networks on oral cancer data. When TNM classification and Chi-Square methods were compared, it was observed that Chi-Square classification closely followed that of clinical investigation. Artificial neural networks (MLP and RBF) are significantly more accurate than the TNM staging system. Kaladhar, et al. [34] predicted oral cancer survivability using classification algorithms. The various algorithms used for classification are Random Forest, CART, LMT, and Naïve Bayesian. 10 fold cross validation is used by the algorithms to classify the cancer survival using training data set. Out of the other techniques, the Random Forest classification technique correctly classified the cancer survival data set. The absolute relative error is less when compared to other methods. Nahar, et al. [35] attempt to extract the significant prevention factors for particular types of cancer. To find out the prevention factors, the authors first constructed a prevention factor data set with an extensive literature review and subsequently employed three association rule mining algorithms, Apriori, Predictive apriori and Tertius algorithms in order to discover most of the significant prevention factors against specific types of cancer. Experimental results illustrate that Apriori is the most useful association rule-mining algorithm to be used in the discovery of prevention factors.

In the current study, data mining concept is applied to oral cancer database to explore the oral cancer risk stratification in terms of patient's demographics. Our result shows that 62.04% of oral cancer patients are male whereas only 37.95% of patients are female. The mean age of diagnosis is 58.7 years and the age group of 50-

55 has more chance to build up malignant cells. Significant increase in the incidences is observed in the patients of lower socioeconomic status. 67.31% of patients are from lower socoeconomic strata whereas only 11.41% patients are from higher socioeconomic strata of society.

Conclusion

The research work focuses on applying data mining for oral cancer risk stratification in terms of age, gender and socioeconomic status. The presented result shows that male have more probability to develop oral cancer in comparison to female. Risk associated with oral cancer is low socio-economic-status and the age group which is suceptible to oral cancer is middle age because of recent change in lifestyle. These results provide sufficient evidence to steer.

Acknowledgments

The authors would like to place on record their sincere thanks to Dr. Vijay Sharma, MS, ENT, for his valuable time and help in understanding the occurrence and diagnosis of Oral Cancer. We are also grateful towards the management and staff of Indian School of Mines, for their continuous support and motivation.

Conflicts of Interest: None declared.

References

[1] Kantardzic M. (2003) *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons.

[2] Han J. and Kamber M. (2012) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, USA, 3rd ed.

[3] Shital C.S., Andrew K., Michael A. and Donnell O. (2006) *Computers in Biology and Medicine,* 36, 634-655.

[4] Lin W.T., Wang, S.T., Chiang T.C., Shi Y.X., Chen W.Y. and Chen H.M. (2010) *Expert Systems with Applications*, 37, 2733-2741.

[5] Hen L.E. and Lee S.P. (2008) *Journal of Computer Science*, 4 (10), 826.

[6] Fayyad U.M., Piatetsky-Shapiro G. and Smyth P. (1996) *American Association for Artificial Intelligence,* AAAI-AI Magazine, 37-54.

[7] Fayyad U.M., Piatetsky-Shapiro G. and Smyth P. (1996) *Advances in Knowledge Discovery and Data Mining,* AAAI Press/ MIT Press, 1-36.

[8] ACM SIGKDD (2006) *Data Mining Curriculum*, 04-30.

[9] Clifton C. (2010) *Encyclopædia Britannica: Definition of Data Mining*.

[10]Piatetsky-Shapiro G. (1991) *Knowledge Discovery in Databases,* AAAI/MIT Press, Cambridge, MA, 229-248.

[11]Agrawal R., Imielinski T. and Swami A. (1993) *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207-216.

[12]An J., Chen Y.P.P. and Chen H. (2005) *Information Systems*, 30, 333-348.

[13]Chen Y.P.P. and Chen F. (2008) *Targets*, 12(04), 383-389.

[14]Lau R.Y.K., Tang M., Wong O., Milliner S.W. and Chen Y.P.P. (2006) *International Journal of Intelligent Systems*, 21(01), 41-72.

[15]Ordonez C. (2006) *IEEE Transaction on Information Technology. Biomed*, 10(02), 334-343.

[16]Ordonez C. and Omiecinski E. (1999) *IEEE Advances in Digital Libraries Conference* (ADL'99), 38-49.

[17]Ordonez C., Santana C.A. and Braal L. (2000) *ACM DMKD Workshop*, 78-85.

[18]Agrawal R., Imielinski T. and Swami A. (1993) *Proceedings of the 1993 ACM SIGMOD International Conference on Management of data,* 207-216.

[19]Hipp J., Güntzer U. and Nakhaeizadeh G. (2000) *ACM SIGKDD Explorations Newsletter,* 2(1), 58-64.

[20]Parkin D.M., Bray F., Ferlay J. and Pisani P. (2001) *Int. J. Cancer*, 94, 153-156.

[21]Marmot M. (2005) *Lancet*, 365, 1099-1104.

[22]Mucci L. and Adami H.O. (2002) *Cancer Epidemiology,* New York: Oxford University Press, 115-136.

[23]Sharma N. and Om H. (2012) *International Journal of Advances in Engineering and Technology*, 4(2), 302-310.

[24]Witten I.H. and Frank E. (2005) *Data Mining: Practical Machine Learning Tool and Techniques*, 2nd ed., Elsevier.

[25]Yeole B., Ramanakumar A.V. and Sankaranarayanan R. (2003) *Cancer Causes & Control*, 14(10), 945-952.

[26]Misra V., Singh P.A., Lal N., Agarwal P. and Singh M. (2009) *Indian J. Community Med*., 34(4), 321-325.

[27]Warnakulasuriya S. (2009) *Evid. Based Dent.,* 10(1), 4-5.

[28]Conway D.I., Petticrew M., Marlborough H., Berthiller J., Hashibe M. and Macpherson L.M. (2008) *Int. J. Cancer,* 122 (12), 2811-2819.

[29]Hashibe M., Jacob B.J., Thomas G., Ramadas K., Mathew B., Sankaranarayanan R. and Zhang Z.F. (2003) *Oral Oncol.,* 39 (7), 664-671.

[30]Conway D.I., Brewster D.H., McKinney P.A., Stark J., McMahon A.D. and Macpherson L.M.D. (2007) *Br. J. Cancer,* 96(5), 818-820.

[31]Ramchandran N.B. (2012) *Int. J. Head and Neck Surgery*, 3(3), 143-146.

[32]Exarchos K.P., Rigas G., Goletsis Y. and Fotiadis D.I. (2012) *Data Mining for Biomarker Discovery*, 199-212.

[33]HariKumar R., Vasanthi N.S. and Balasubramani M. (2012) *International Journal of Soft Computing and Engineering (IJSCE)*, 2(3), 263-269.

[34]Kaladhar D.S.V.G.K., Chandana B. and BharathKumar P. (2011) *International Journal of Research and Reviews in Computer Science (IJRRCS),* 2(2), 340-343.

[35]Nahar J., Kevin S.T., Ali A.B.M.S. and Chen Y.P. (2009) *J. Med. Syst*., 35(3), 353-367.