Predicting Risk of Direct-to-Customer Drug Prescription using K-Mean Clustering Technique

Francisca N. Ogwueleka
Computer Science Department
Federal University Wukari, Taraba State - Nigeria

Timothy Moses
Computer Science Department
Federal University Wukari, Taraba State - Nigeria

ABSTRACT
Exploration of patients medical record have necessitated the need for tracking customers with adverse drug reactions so as to deliver an analysis on risk involved and which cause of action proved effective. This paper acquires customer insights on adverse reactions experienced by taking anti-malaria drugs without medical diagnosis in north eastern part of Nigeria. The data collected were computed using k-means clustering algorithm implemented on Excel Visual Basic for Applications (VBA) Macro. Cluster values generated from the program was plotted. The graph predicts age of customers who are at risk of buying drugs without proper diagnosis and medical examination. Result obtained shows that about 38.47% of working population are at risk of direct-drug prescription. The result implies that, there is tendency of low productivity and inefficiency among 38.47% of the working force.

General Terms
Cluster analysis; segmentation techniques; data mining; centroid; k-means; Macro.

Keywords
Direct-to-customer drug, prediction, risk, customer insight, adverse reactions, direct drug prescription..

1. INTRODUCTION
Due to advancement in computer technology today and how data are been explored, it is important for business organizations to understand their customers better. Any business organization that uses data driven analytical strategies have the opportunity to enjoy competitive advantage and will acquire better knowledge of their customers (Gledon, 2008). Such organization will be able to use models to predict diverse risk, fraud, or the likelihood of response to a particular disease. It is possible to apply segmentation techniques in predicting risk based on experience acquired, but such segmentation schema will definitely be restricted to the use of few variables at best. To perform a true multivariate segmentation so as to identify different segments in a population, cluster analysis must be used. Hence, this research work will use cluster analysis for acquiring customer insight and predicting risk in medical industry (Gledon, 2008).

K-means clustering technique is a method of clustering that involves portioning of set of data into a number of clusters with each of these data belonging to the cluster that has the nearest mean (Kardi, 2007). K-means clustering algorithm finds a partition with squared error among the empirical mean of cluster so that points in the cluster is minimized (Kardi, 2007). Data mining techniques using k-means clustering can be used to predict fraudulent activities in medical industry. It can help reduce the impact of fraud and abuse in the healthcare system and will support policy change decisions. Using cluster analysis helps to answer questions like; (1) who are the customers, (2) how can I identify and track chronic customers who experience adverse drug reactions? (3) how do I reduce number of hospital admissions and claims? (4) how can I compare and contrast treatments given to a customers?

Customer insight can be observed or acquired through an in-depth understanding of what the customer want. To understand a customer will require a data collection method of questionnaire administration, interview or personal observation. Methods of predicting risk in medical industry are many. In medical section, the risk varies from the other. Risks are predicted in medical claims, fraud and abuse, prescription of drugs and to some extent, insurance risk.

Purchase of drugs directly from drug vendors or chemist, which is prompted by advertisements in newspapers, magazines, television and industry-funded websites increases risk of patients in medical industry (Brownfield, 2004). Our media houses today commercialize prescription of drugs and what you see on television today are likely to attract customers of the latest medication aid featuring happy and attractive people. Television and radio adverts do not provide consumers with well-known risks, companies that produce these drugs will not do same so as to sell their products and the regulating body we have, do not control the so-called seeking adverts which sell diseases rather than drugs. Observation also shows that when a customer requests a specific drug, doctors, who are supposed to prescribe drugs base on diagnosis and treatment may prescribe it even if the drug is not in the patient's health condition.

Customers who take drugs prescribed by health personnel sometimes complain of being dizzy, having general body weakness, eye disturbance or even itching. But these symptoms are most times not listed in the label insert of the drug taken as possible side effect, why then should a customer be feeling so odd? Unfortunately, there are no answers to such questions (Krista, 2012). It is generally true that clinical trials on new drugs are meant to show that a drug is safe and effective, and though these trials are generally very thorough and reliable, they could not possibly foresee all of the potential effects the drug could have on individuals with different underlying conditions and medical histories. Most times, the most serious adverse reactions are not known until after the medicine has been on the market for quite some time. Drugs that have undergone clinical trials are most times not produced to study how they react with one another in the human body; a consideration that is necessary since people age and their medicine cabinets begin to overflow (Krista,

2012). It was observed from U.S. Food and Drug Administration (FDA) report that most adverse drug reactions can be traced to common brain-related side effect categories. For example Ambien CR has been traced to memory loss (amnesia) if taken outside medical examination. Propecia can cause inhibition of libido side effects; Mirapex is traced to compulsive behaviour side effects.

To predict risk in drug prescription, there is need to collate data through administration of questionnaire to customers who took drugs base on prescription by their health personnel but at later time discovered serious side effect of such drugs. Data collected for a particular drug prescription from different customers are segmented and possible risk are discovered by matching different variables such as age, gender, diet and customer's genetic system. This work uses cluster analysis to group customers with similar needs and response, identify and track high-risk patients due to drugs prescribed by health personnel, predict age group affected by adverse drug effects and whether this work force (age group) reduces the nation's productivity.

2. REVIEW OF RELATED LITERATURE

Medical decisions supported through concrete data have existed for centuries. John Snow; known as father of modern epidemiology, was able to discover source of cholera through the use of maps with forms of bar graphs in 1854 (Tufte, 1997). He traced the spread of cholera through water supply. Snow was able to achieve this by counting number of deaths and plotting address of those affected with cholera on the map as black bars. After careful analysis, Snow discovered that most of the deaths clustered towards a particular water pump in London. Cao described how data mining techniques has helped to monitor trends in clinical trials of cancer vaccines. They explained how data mining techniques has helped medical experts to discover patterns and irregular behaviour rather than looking at a set of data (Cao, 2008). Wong et al (2005) introduced "What's Strange About Recent Events (WSARE)", which is an algorithm that helps unfold disease outbreaks in their early stages. The algorithm is based on association rule and Bayesian network. Application of this algorithm has resulted in accurate prediction of disease outbreak. Thangavel (2006) used k-means clustering techniques to determine the level of risk in a cervical cancer patient and showed that this technique produces a better result when compared to existing techniques used (Thangavel, 2006). Doctors can now recommend biopsy for patient with cervical cancer through attributes obtained from the techniques.

Harleen (2006) explained how to use rule induction with illustration of its application to medical industry. The process involves extracting useful "if–then" rules from a set of data. Harleen used this method to predict concentration of blood alcohol. Shantakumar (2009) described how intelligent and effective data mining and neural network can be used to predict heart attack. They explained how significant patterns were extracted from heart disease data warehouse and how data was clustered using K-means clustering algorithm. Margaret (2006) also explained how data mining has helped to predict length of stay of patients with spinal cord injury. In their submission, they are most interested in knowing a subset of total patient population most especially those with spinal cord injury. Wynee (2006) described how you can explore mining in diabetic patients' database with 200,000 screening records, obtained between 1992 and 1996. There are 60 fields in each record. An analysis discovers that patient's identification number, sex, date of birth, race, date of screening and duration of diabetics were among the fields captured. Due to the clumsiness of the data captured, there was need for data cleaning before mining is done, the cleaning process done made it possible to map attributes in different formats, know the encoding schemes used and to be sure if the attributes are kept in a cleaned database (Wynne, 2006). This method has helped to generate a standardized format schema, which allow file to be transformed accordingly (Wynne, 2006). Tatonetti (2006) developed an algorithm that was able to access the US Food and Drug Administration database and discover true adverse reaction of drugs. The model was able to compare interactions between pairs of drugs associated with blood pressure which have led to deadly heart condition. A model that matches people who were alike in terms of the kind of drug they took was developed. The model was observed to see whether there are more people who react to a particular drug to those that did not react to the same drug. If this is achieved, it then implies that the drug is indeed the culprit. This approach was used to predict health records of patients at Standford Hospital and Clinics and was able to understand the biological effects of drugs on human. The system has helped physicians to know exactly what drug should be prescribed to patient and have also help to discover biological pathways among drugs with similar reactions (Tatonetti, 2012).

2.1 Existing System and its Limitation

The system mostly practiced in medical industry as regards new drugs is that, every new drug undergoes extensive laboratory test. Most times, animals are used before these drugs are tested on humans. The drugs go through variety of clinical trials among closely monitored patients so as to observe what positive and negative reactions occur on these patients. After thorough and reliable clinical trials are performed on few individuals, the drugs are certified for use and adverse effects based on few individuals tested are written on the drug labels, which are packaged and sent to the market for consumers' consumption (Tatonetti, 2012).

A careful analysis of the existing system shows that direct drug prescription is always dangerous and have list of negative effects. Though clinical trials are reliable to prove that a drug is safe and effective, these trials would not be able to foresee all potential effects such drug may have on very large number of customers with diverse health conditions hence, putting customers at risk of taking such drugs. It was also observed that most serious adverse reactions are not known until the medicine has been in the market for quite some times. List of side effects listed on drug labels are most times a small portion of what consumers of such drugs usually experienced. Most consumers of direct-to-customer prescription of drugs have no knowledge of the drugs they take, not because they do not wish to but because there is no concrete data analysis that shows what adverse effects these drugs have on individuals that have taken them in the past. This also put consumers of direct-to-customer prescription of drugs at high risk.

3. METHODOLOGY

Questionnaires were administered to customers in North East region of Nigeria, who take anti-malaria drugs without medical diagnosis so as to gain insights on what adverse effects they have experienced. Adverse effects recorded are

itching, nausea, vomiting, blistered rash, general weakness, eye disturbance, headache, abdominal upset, dizziness, diarrhea, drowsiness and palpitation. These data are cleaned by separating un-completed questionnaire forms from those that gave quality response. The data from completed questionnaire forms were structured and integrated into one flat table to help in segmentation process. This process was achieved by grouping of customers based on drugs they have taken in the past and adverse reactions experienced. Because we are constraints by non-availability of a data mart in medical industry that keeps adverse reactions from drugs taken by customers, our techniques used data from questionnaires collected. These data were run in MS Excel with Excel VBA Macro developed, which generates cluster values. A square Euclidean distance on pair of cluster values was computed to find the closest centroid. Iterative process continued until all objects are grouped based on minimum distance. Prediction was made based on grouping of objects. This paper identified possible side effects of drug prescription on high-risk patient; predict which age group is greatly affected and whether age group affected reduces the nation's productivity. The proposed model in figure 1 gives a general view of how the system behaves in meeting its target.

To achieve the objectives of this research paper, the following five steps were followed.

Step One:- Related literatures as it regards antimalaria drugs, prescriptions and adverse effects were studied.

Step Two:- Interviews were conducted with doctors and pharmacists to know what drugs are used to cure malaria parasites and what side effects each of these drugs may cause if taken outside medical instructions.

Step Three:- Questionnaires were administered based on information gathered. Responses were documented. The sample population for this study is 2866, which covers all valid questionnaires received. The data collection method applied for this research is a close ended questionnaire method, where customers are provided with an option to questions asked.

Step Four:- k-means clustering program was run on Microsoft Excel environment, the data analyzed was clustered into three (3), five (5) and ten (10) similar chunks (centroids) respectively. K-means clustering is an algorithm that partitions set of observations into a number of clusters, where each data belongs to the cluster with the nearest mean. The algorithm finds a partition with squared error among the empirical mean of cluster so that points in the cluster is minimized.

Description

Given a set of observations A = (a_1, a_2, …, a_n), with each observation having d-dimensional real vector, k-means algorithm partitions the set of n observations into k number of clusters ($k \leq n$) S = {s_1, s_2, …, s_k} so that the within-cluster sum of squares (WCSS) is minimized as defined in equation (1)

$$\arg\min \sum_{i=1}^{k} \sum_{aj \in si} \|ai - \mu i\|^2 \qquad (1)$$

Where μ_i is the mean of points in S_i (Kardi, 2007).

The algorithm uses an iterative refinement technique. Let p1(1), …, pk(1) be an initial set of k means, the algorithm iterate between two steps; first is the assignment step, which assigns each data to the cluster with the closest mean as shown in equation (2)

$$S_i^{(t)} = \{a_m : \|a_m - p_i^{(t)}\| \leq \|a_m - p_j^{(t)}\| \forall 1 \leq j \geq k\} \quad (2)$$

Note that each a_m goes into one $S_i^{(t)}$, even if it can go into two of them (Kardi, 2007)

The second step is the update step, which calculates the new mean to be the centroid of the observations in the cluster (Kardi, 2007).

$$p_i^{(t+1)} = \frac{1}{S_i^{(t)}} \sum_{a_j \in S_i^{(t)}} a_j \qquad (3)$$

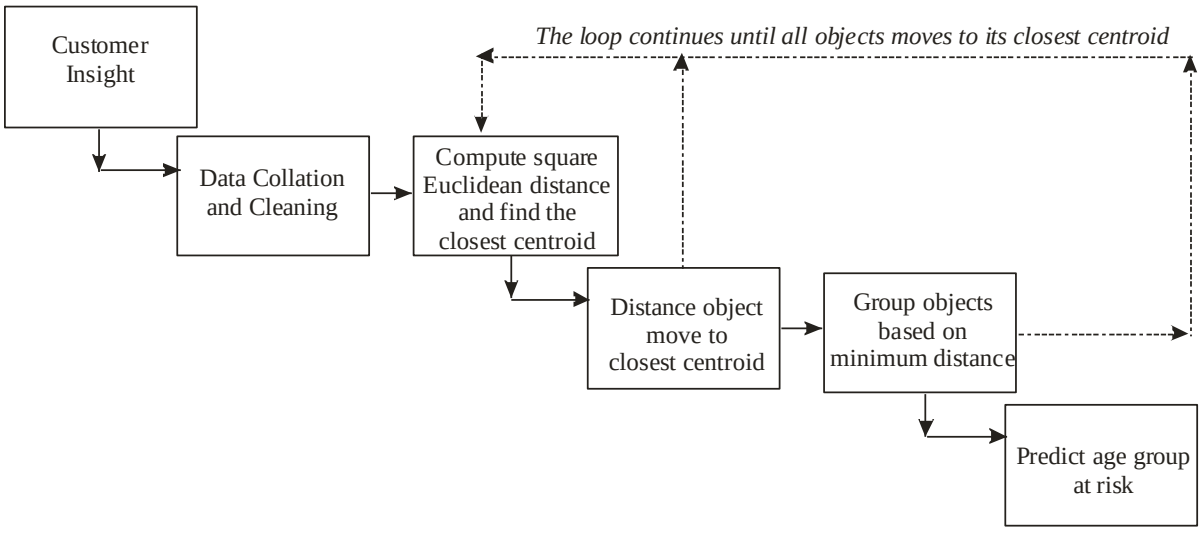

Figure 1: Proposed new model for predicting risk of direct-to-customers drug prescription

3.1 k-means algorithm for this system

Step 1: Start
Step 2: Select Data Range
Step 3: Select cluster centroids
Step 4: DO
Step 5: For j = least(record) To highest(record)
Step 6: ld = 0 //ld means lowest distance
Step 7: For i = least(centroid) To highest(centroids)
Step 8: m = 0, n = 0
Step 9: For k = least(record(j).dimension) To highest(record(j).dimension)
Step 10: n = record(j).dimension(k) – centroid(i).dimension(k)
Step 11: n = n^2, m = m + n //sum of distance space
Step 12: next k
Step 13: m = sqrt(m) //square root of distance
Step 14: if i = least(centroid) OR (m < ld) Then ld = m
Step 15: record(j).distance(i) = ld
Step 16: record(j).cluster = i
Step 17: endif
Step 18: next i
Step 19: next j

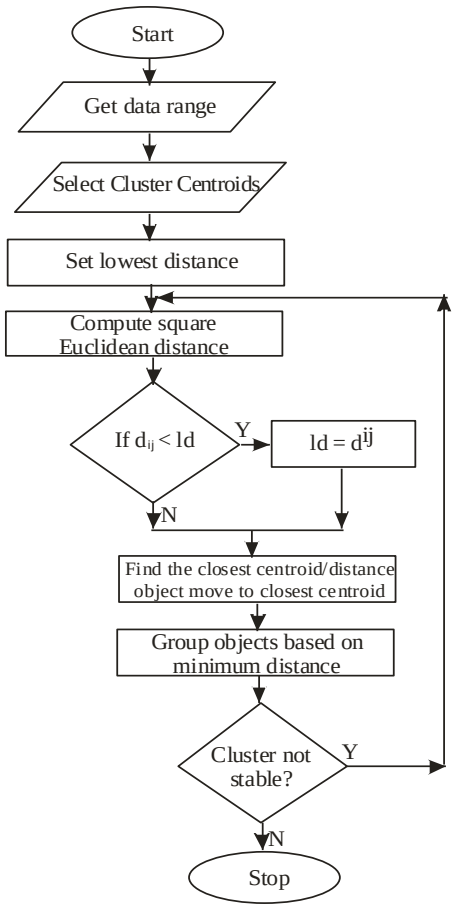

Figure 2: k-means algorithm for the system

Step Five:- Customers' age group was plotted against cluster values generate from our macro. The plotted graph shows age group at risk of direct drug prescription

4. RESULT AND DISCUSSION

4.1 Data Analysis

The data collected were structured and integrated into one flat table to help in segmentation process. The data mining table to be used was stored in Microsoft Excel. The characteristics features of customers (demographic data) are shown in table 1 to 7.

Table 1: Age of customer/patient

Age	Frequency	Percentage (%)
18 – 30	987	34.44
31 – 43	927	32.34
44 – 55	652	22.75
56 – above	300	10.47
Total	**2866**	**100**

From table 1, 987 respondents representing 34.44% are within the age of 18 and 30 while 927 (32.34%) are within the age bracket 31-43. 652 respondents representing 22.75% are within 44 – 55 while 300 (10.47%) represents respondents with age 56 and above. This shows that a high percentage of working class and youths are victims of direct-to-customer prescription of drugs.

Table 2: Response base on customer qualification

Qualification	Frequency	Percentage (%)
FSLC	15	0.52
SSCE	237	8.27
HND	807	28.16
BSc	1174	40.96
MSc	580	20.24
Ph.D	8	0.28
None	45	1.57
Total	**2866**	**100**

Table 2 shows customers' qualification. 15 customers have First School Leaving Certificate, which represents 0.52%. SSCE holders are 237 representing 8.27%. These two groups of customers filled their questionnaires with the help of an older person assisting them. 807 customers, representing 28.16% are HND holders. 1174 respondents representing 40.96% are BSc holders, while 580, 8 and 54, representing 20.24%, 0.28% and 1.57% are MSc, Ph.D holders and customers with no qualification respectively. The result from table 2 shows that high percentages of our respondents are educated.

Table 3: Analysis on anti-malaria drugs frequently purchased by customers

Anti-malaria drug purchase	Frequency	Percentage (%)
Chloroquine	511	17.83
Artesunate	336	11.72
Sulphadoxine/pyrimethamine combination	707	24.67
Quinine	35	1.22
Halofantrine (Helfan)	28	0.98
Artemether	1193	41.63
Amodiaquine	56	1.95
Total	**2866**	**100**

Data collected in Table 3 shows that most customers buy Artemether (41.63), which represent 41.63%, followed by sulphadoxine/pyrimethamine combination (707), representing 24.67%.

Table 4: Adverse drug reactions on respondents

Reactions	Frequency	Percentage (%)
Itching	260	9.07
Nausea	256	8.93
Vomiting	206	7.19
Blistered Rash	224	7.82
General Weakness	257	8.97
Eye disturbance	239	8.34
Headache	243	8.48
Abdominal upset	240	8.37
Dizziness	255	8.90
Diarrhea	251	8.76
Drowsiness	233	8.13
Palpitation	202	7.05
Total	**2866**	**100**

Table 4 shows that, about 9.07% (260) of respondents experience itching when they take anti-malaria drugs. This may be as a result of the quinine content in most of the anti-malaria drugs they take. While vomiting and nausea are common malaria symptoms, we observed that about 8.34% (239) complain of eye disturbance. This implies that if drugs are not taken base on diagnosis, there is likelihood of a patient going blind.

Table 5: Frequency of anti-malaria drug use

Anti-malaria drug use	Frequency	Percentage (%)
Every day	978	34.13
Anytime I feel feverish	1008	35.17
Recommended in the hospital	880	30.70
Total	**2866**	**100**

From Table 5, 978 customers, representing 34.13% take anti-malaria drugs everyday while 1008 respondents representing 35.17% take anti-malaria drugs whenever they feel feverish. Only 880 (30.70%) customers go to hospitals.

Where do customers get their drugs? Table 6 gives a summary of where most customers get their anti-malaria drugs. This also goes a long way to tell whether customers are at risk of adverse drug reactions.

Table 6: Sources of purchase of anti-malaria drugs

Sources	Frequency	Percentage (%)
Hospital	479	16.71
Chemist	1315	45.88
Pharmaceutical shops	726	25.33
Traders	346	12.07
Total	**2866**	**100**

Table 6 shows that most customers buy drugs from chemists (1315), representing 45.88%. This result shows that customers do not go for proper diagnosis before they purchase drugs rather; choose to buy from chemist probably, because of its proximity to their house. Customers that buy drugs from pharmaceutical shops were most times, not asked by the Pharmacist what the drug is meant for, he/she only sells base on what the customer want. This might also result in high risk of customers since Table 6 shows that 726 customers, representing 25.33% buy from pharmaceutical shops.

It has been observed that most customers take drugs base on recommendation from either friends or relatives. Some customers believe in "if it has worked for you, then I will also buy the drug". Table 7 shows an analysis of the person that recommends drugs to customers.

Table 7: Respondents responses on who recommend drugs

Recommendation	Frequency	Percentage (%)
Doctor	398	13.89
Nurse	121	4.22
Self	1405	49.02
Pharmacist	942	32.87
Total	**2866**	**100**

Table 7 shows 1405 customers representing 49.02% as high percentage of customers that buy drugs base on self-medication. These categories of customers buy drugs because it has worked for somebody close to them or such drugs were advertised in television or radio.

4.2 Result of Analysis

The data collected were structured and integrated into one flat table to help in segmentation process. The data collected were run in MS Excel with an Excel VBA Macro developed in Visual Basic. Table 8 - 10, shows cluster results obtained after running the Excel VBA macro.

Table 8: Cluster values for number of clusters = 3

Reactions	Centroid1	Centroid2	Centroid3
Itching	8	5.555556	6.190476
Nausea	4.6	5.66667	7.57143
Vomiting	4.7	4.55556	5.61905
blistered rash	6.3	4.333333333	5.80952381
General weakness	6.4	6.111111111	6.571428571
Eye disturbance	7.2	3.111111111	6.619047619
Headache	5.5	7.111111111	5.904761905
Abdominal upset	5.2	5.222222222	6.714285714
Dizziness	5.5	5.777777778	7.047619048
Diarrhea	6.3	5.444444444	6.619047619
Drowsiness	4.7	5.111111111	6.666666667
Palpitation	3.2	4.888888889	6

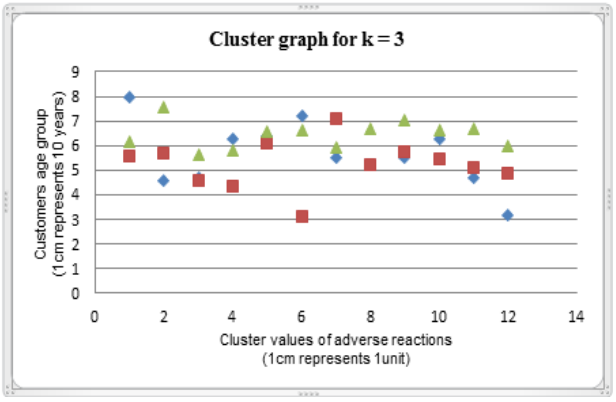

Figure 3: Cluster graph with number of clusters = 3

Cluster graph in Figure 3 indicates that customers between ages 40 – 70years experience more adverse reactions. Other age groups have little or no adverse reactions from direct drug prescription.

Table 9: Cluster values for number of clusters = 5

Reactions	Centroid1	Centroid2	Centroid3	Centroid4	Centroid5
Itching	12	5.16667	5	6.1875	6.58333
Nausea	4.33333	5	8	6.4375	7.16667
Vomiting	5.66667	5.33333	6	4.9375	5
Blistered rash	5.666666667	4.5	5.333333333	5.75	6
General weakness	7	6	6.666666667	6.375	6.5
Eye disturbance	6.666666667	2.166666667	4	6.8125	7.083333333
Headache	5.333333333	6.5	3.333333333	6.1875	6.583333333
Abdominal upset	5	5.166666667	8	6.0625	6.083333333
Dizziness	6.333333333	5.166666667	7	5.875	7.5
Diarrhea	7	5.5	8	6.1875	6.1666667
Drowsiness	4	6.1666667	7.6666667	4.1875	7.8333333
Palpitation	1	3.166666667	6.666666667	6.5	4.666666667

Table 10: Cluster values for number of clusters = 10

Reactions	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Itching	11.5	5	4	6.2	6.090909091	12	6	5	6	13
Nausea	4	4.8	8	6.4	7.090909091	8	6.5	8	8	5
Vomiting	5.5	5.4	6	5	4.909090909	6	4.5	6	6	6
Blistered rash	6	3.8	5	5.8	6	6	6.5	6	5	5
General weakness	7	6	7	6.333333333	6.545454545	6	6.5	7	6	7
Eye disturbance	7	2.2	5	6.866666667	7.181818182	6	4	4	3	6
Headache	5	5.4	3	5.6	6.454545455	8	13.5	3	4	6
Abdominal upset	4.5	5.2	8	6.066666667	5.818181818	9	5.5	8	8	6
Dizziness	6	5.2	6	5.8	7.454545455	8	6	7	8	7
Diarrhea	6	6	9	6.06666667	6	8	5.5	6	9	9
Drowsiness	2.5	5.8	8	4.133333333	7.727272727	9	6.5	9	6	7
Palpitation	0.5	2.6	7	6.533333333	4.545454545	6	6	6	7	2

Centroid is represented by the alphabet C in the above table

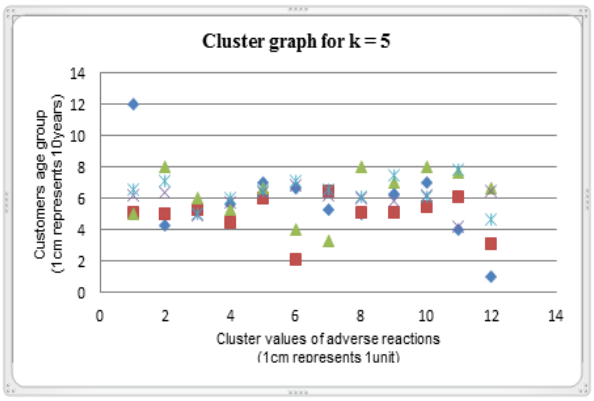

Figure 4: Cluster graph with number of clusters = 5

Figure 4 shows that adverse reactions are concentrated around customers with ages 40 – 80years. This graph has similar behavior with cluster graph in figure 3. This indicates that, as number of clusters used increases, customers adverse reactions tend towards same age bracket.

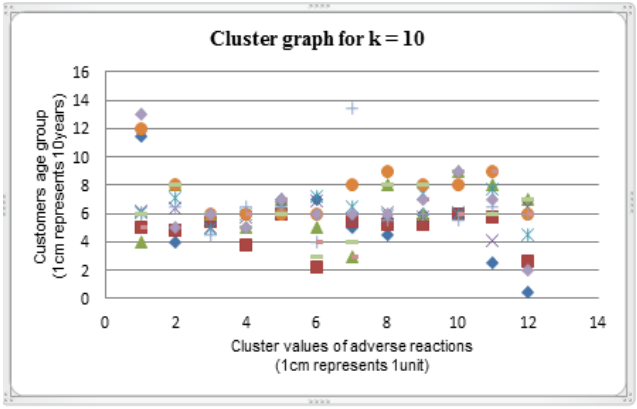

Figure 5: Cluster graph with number of clusters = 10

Figure 5 also shows that adverse reactions are concentrated around customers with ages 40 – 80years. This graph has similar behavior with cluster graph in figure 3 and 4.

4.3 Discussion

Our results shows that increase in the number of clusters produces slight difference on the cluster graph with most cluster points concentrated around customers with ages 40 – 80years. Considering age of our work force according to Trading Economic (2015), data has shown that working age is between 15 – 64years. The result obtained in this paper indicates that about 38.47% of our work force experience adverse reactions due to direct drug prescription. The result implies that, there is tendency of low productivity and inefficiency among 38.47% of the working force. Since medicine is a safety critical context where decision activities must be supported by explanations, it is therefore, important that adverse drug reactions experienced by customers be clinically validated and direct to customers drug prescription totally discouraged.

5. CONCLUSION

In this paper, we provided a simple and qualitative analysis of risks involved in taking direct-drug prescription common in the north eastern part of Nigeria. The result obtained shows that victims of adverse reactions on direct prescription of anti-malaria drugs are between ages 40 – 80years. About 38.47% of these victims are of the working age hence, the tendency of low productivity and ineffectiveness among these populations. This technique can also be applied in any class of drug that has caused adverse reactions on customer who do not undergo proper medical diagnosis and examination. For instance, we can take customers that use body lotion cream. It has been observed that some particular cream causes change in skin of customers; while some customers become light once they apply such cream, other become dark. While some cream makes the skin soft, on other customers, the skin becomes hard and to some extent broken skin. These reactions can be clustered to be sure of the customers that are at risk of applying these creams and whether they should be advised to stop the cream or should undergo medical examination first before applying such cream.

The technique can also be used in production industry to check effects of chemicals inhaled by workers during production and what age group is mostly affected. It will be suitable for business organization to determine the risk of producing particular product/goods considering the age group of targeted customers.

6. REFERENCES

[1] Brownfield, E. D., Bernhardt, J. M., Phan, J. L., Williams, M. V., & Parker, R. M. (2004). Direct-to-customer drug advertisements on network television: an exploration of quantity, frequency and placement. *J-health communication*, 9, 491-497.

[2] Cao, X., Maloney, K. B., & Brusic, V. (2008). Data mining of cancer vaccine trials: a bird's eye view. *Immune Research*, 4, 1-11.

[3] Gledon, C, & Wayne, T. (2008). Understanding your customer: Segmentation techniques for gaining customer insight and predicting risk in Telecom Industry. *AT&T Corporation* (pp 1-14), Cary North Carolina: SAS Institute, Inc.

[4] Harleen, K, & Siri, K. W. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science,* 2, 194-198.

[5] Kardi, T. (2007). K-Means clustering tutorial. Retrieved from http://people.revoledu.com/kardi/tutorial/kMean/

[6] Krista, W. J. (2012). Dangers of popular drugs used to cure malaria. Retrieved from http://www.thehealthierlife.co.uk/natural-health-articles *Natural Health Article*.

[7] Margaret, R. K., Kevin, C. D., & Ida, A. (2002). Data mining in healthcare information systems: case study of veterans' administration spinal cord injury population. *IEEE Computer Society* (pp 1-9), Hawaii: Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03).

[8] Shantakumar, B. P., & Kumaraswamy, Y. S. (2009). Intelligent and effective heart attack prediction system using data mining and Artificial Neural Network. *European Journal of Scientific Research,* 3, 642-656.

[9] Tatonetti, P. N., Patrick, P. Y., Roxana, D., & Russ, B. A. (2012). Data-driven prediction of drug effects and interactions. *Clinical data analysis, science translational medicine,* 4, 1-14.

[10] Thangavel, K., Jaganathan, P. P., & Easmi, P. O. (2006). Data mining approach to cervical cancer patients analysis using clustering technique. *Asian Journal of Information Technology,* 5, 413-417.

[11] Trading Economic (2015) Age dependency ratio (% of working-age population) in Nigeria. Retrieved from http://www.tradingeconomics.com/nigeria/age-dependency-ratio-percent-of-working-age-population-wb-data.html

[12] Tufte, E. (1997). Visual explanations, images and quantities, evidence and narrative. *Journal of the American Statistics Association,* 1, 1-2.

[13] Wynne, H., Mong, L. L., Bing, L., & Tok, W. L. (2006). Exploration mining in diabetic patients databases: findings and conclusion. Association for Computing Machinery (pp 430-435), New York: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining.

[14] Wong, W. K., Moore, A., Cooper, G., & Wagner, M. (2005). What's Strange About Recent Events (WSARE): An algorithm for early detection of disease outbreaks. *Journal of Machine Learning Research,* 6, 5-8.