DETERMINANTS OF CHILDHOOD LEUKEMIA: A RISK ASSESSMENT MODEL FOR SRI LANKA

T. Priyadharshan[1], T. Rupasinghe[1] and P. Abeysinghe[2]

[1]Department of Industrial Management, University of Kelaniya, Sri Lanka.
Email: priyadharshant@gmail.com, thashika@kln.ac.lk
[2] National Cancer Institute, Maharagama, Sri Lanka. Email: prasadabeysinghe@hotmail.com

ABSTRACT

Leukemia is the most frequent occurrence type of cancer for children. Although it is etiologic is largely unknown, leukemia is believed to result from an interaction between genetic and environmental factors. The purpose of this study is to generate a risk assessment model for the Sri Lanka. This study will help the medical professionals to assess the risk levels of a child with various combinations of risk factors on a defined period of time for earlier or more frequent screening and counselling of behavioral changes to decrease risk. This model was developed from data collected at Sri Lankan National Cancer Institute (NCI) and through focus group interviews of patients, relatives of the patients, non-patients and oncologists. The rigorous data collection was followed up by Principal component analysis (PCA) to identify the most influencing risk factors and the risk assessment model was developed using Logistic regression. This model calculates the individualized leukemic risk through the information of the relative risk and the maximum exposure limits, considering the influential factors identified from the PCA. The study models the risk assessment and reveals the significant risk factors of the pediatric leukemia patients in Sri Lanka which can be applicable for other developing countries in the Asian continent as well.

Keywords: Childhood leukemia, Logistic regression, Principal Component Analysis, Risk assessment

1. INTRODUCTION

Childhood leukaemia cancer is considered as a deadly disease which is divided into four main categories named as acute lymphoblastic leukemia **(**ALL), acute myeloid leukemia **(**AML), chronic lymphocytic leukemia (CLL), chronic myeloid leukemia (CML) [1] and at the same time it is considered as non-communicable disease as well. Early detection of any kind of leukaemia will be beneficial for the children as well as the nation. Statistical risk assessment models will helps the clinicians to identify the individuals at higher risk to get the leukaemia. Earlier detection both by screening and early clinical diagnosis represents an important component of childhood leukaemia cancer control which will helps the patients to decrease risk, so predict the childhood leukaemia cancer earlier through these statistical models will help the children at high risk of cancers in the general population.

In here we presented the developed model for the childhood leukaemia for Sri Lankan context with considering the risk factors such as gender, exposure to benzene, age and exposure to radiation [2]. Further in model development several advanced Statistical methods were used, including multivariate methods such as principal component analysis [3, 4] and categorical data modeling such as Logistic Regression [5, 6] were used and the model will evaluate the risk level of a child which is explained through probability terms for subsequent time interval, who is free of the disease at an initial age, and will probably develop that disease in the coming years.

2. METHODOLOGY

The research investigates the determinants of childhood leukemia cancer in Sri Lanka. Literature survey was used to find out the approaches, gaps, influencing factors which are causing the childhood leukemia cancer and applicability of existing models in the literature. Subsequently, the research questions were formulated. The research was conducted as a study on Sri Lankan context with the support from existing literature available for other countries.

After identification of the influencing factors, these were validated through Delphi survey and focus group discussion to evaluate, whether there are any connections between influencing factors identified and childhood leukemia in Sri Lanka.

Then **principal component analysis** is used for eliminating unnecessary risk factors which were focusing the same dimension. With the use of **logistic regression** the relationship with the

dependent variable (level of risk) and the independent variables (such as age, gender, race, benzene exposure, ionizing radiation) were modeled to evaluate the chances of getting childhood leukemia cancer by the determined risk level in probability terms.

Qualitative data is gathered through literature reviews and primary data such as interviews with stakeholders and focus group discussions.

Quantitative data is gathered through secondary data such as "Annual health bulletins published by ministry of health, Sri Lanka", the systematic review of literature was also used to identify the back ground, approaches, needs and situations of the healthcare (cancer and other diseases) risk assessment, National cancer institute registry, interviews and questionnaires. Furthermore the questionnaires, and Delphi surveys were used to collect primary data and the participants included experts such as senior oncologists and healthcare professionals in the area of pediatric leukemia.

Since the study focused on child hood leukemic cancer patients in Sri Lanka, children with or without leukemia cancer in the boundaries of the country belonging to the targeted population. Sampling frame divided into non-leukemic cancer patients, and leukemic cancer patients. Under non-leukemic cancer patients, children who do not have leukemia cancer from different places from Sri Lanka, different gender, different ethnicity and different age group were selected. Under the leukemic cancer patients, the same characteristics were taken into consideration to select children who had leukemia.

Through stratified sampling and probability sampling methods, biasness of data were reduced. Furthermore, by separating the population into strata/groups as leukemia cancer patients, non-leukemia cancer patients and then following up simple/systematic random sampling the biasness was further reduced. If stratified sampling was applied, this may provide a chance of selecting irrelevant children for data collection who are unaware about the scenario.

Since this research had a data collection which had a portion of qualitative research, the quota sampling which is a non -probability sampling method was applied, when quotas were selected, for each quota we have considered different risk factors such as their age (age limit is up to 14 years), gender, food habits, exposure to radiation, exposure to benzene from different areas, and different ethnicity etc.

The data was collected from the Sri Lankan National Cancer Institute (NCI), interviews and questionnaires. Sampling Units contain leukemia cancer patients, Relatives/friends of leukemia cancer patients, oncologists, and nurses. The developed model was validated internally, the data set was divided into two sets and using the chi square goodness of fit test it was further validated.

3. RESULTS

The analyses of the survey results identifies the influencing factors which were participating in causing leukemia are discussed below. The collected data were analyzed using software package IBM SPSS version 20. Likert scale score evaluation was used to identify the importance of different influencing factors contributing overall cause of childhood leukemia cancer. From the literature review the risk factors of causing childhood leukemia was identified; they are the exposure to electromagnetic field, exposure to benzene, exposure to pesticides, exposure to radiation, maternal alcoholic consumption, aged parents, age, gender, smoking (mentioning the second hand smoking) [2] and genetic factors [7].

For the study, we have considered 50 leukemic children and 50 non- leukemic children, within the sample 33% of were females and 67% were males were taken in analysis, 48 % of the participants were not exposed to benzene exposure and 52% were exposed to benzene exposure, 51 % of the participants were not exposed to radiation exposure and 49% were exposed to radiation exposure, 53% has no first degree relatives and 47% have leukemia cancer relatives, 65% data not belongs to second hand smoking and 35% belong to second hand smoking, 48% do not have genetic abnormality and 52% have genetic abnormality, In here 42% children having young parents and 58% children having aged parents. While highest age participating age is age 2, and the least participation age were age 10 and age 14. If we take less than age 2 as category 1, age 2 to 10 as category 2 and age 11 to 14 as category 3, the valid percentage of each categories are 10%, 75% and 15% respectively.

Figure 1 shows the results of the hypothesis testing for the identified risk factors of the childhood leukemia cancer. According to Table first degree relatives and leukemia are independent, Second hand Smoking and leukemia are related, Genetic abnormalities and leukemia are independent, exposure to benzene and leukemia are related, gender and leukemia are

related, exposure to radiation and leukemia are related and aged parents and leukemia are related.

	Null Hypothesis	Test	Sig	Decision
1	First degree relatives and leukemia are independent.	chi square test of independence	0.548	accept the (H_0) null hypothesis
2	Second hand Smoking and leukemia are independent	chi square test of independence	0.000	reject the (H_0) null hypothesis
3	Genetic abnormalities and leukemia are independent	chi square test of independence	0.423	accept the (H_0) null hypothesis
4	Exposure to benzene and leukemia are independent	chi square test of independence	0.000	reject the (H_0) null hypothesis
5	Gender and leukemia are independent	chi square test of independence	0.001	reject the (H_0) null hypothesis
6	Exposure to radiation and leukemia are independent	chi square test of independence	0.000	reject the (H_0) null hypothesis
7	Aged parents and leukemia are independent	chi square test of independence	0.005	reject the (H_0) null hypothesis

Figure 1: Results for hypothesis of chi square test of independence

Before conducting principal component analysis (PCA) the dataset needs to be validated for the test named KMO (Kaiser-Mayer-Olkin) test, the test score was 0.502 and which exceed the minimum requirement of 0.500 to conduct the Principal component analysis and the Bartlett's Tests' level of significance value is 0.000 which is less than the level of significance 0.001 shows that the correlation matrix is not an identity matrix. According to the component matrix for the PCA factors which have greater than 0.572 eigenvalues which were selected as the most influencing factors for the getting leukemia; they are: Gender, Exposure to benzene, Age, Exposure to radiation.

When logistic regression was performed using SPSS software for risk factors and leukemia the following results were received.

The Initial Log Likelihood (-2 Log Likelihood or -2LL) value is 138.629 on step 0, before any variables have been added to the model and the model child-square value 28.504 has a significance value of 0.000, less than 0.05, so we conclude that there is a significant relationship between the dependent variable and the set of independent variables. The Nagelkerke R^2 has a value of 0.465, so it interprets the relationship as strong. The Hosmer and Lemeshow goodness-of-fit statistic for the model has a value of 6.501 and the significance value is 0.591, greater than 0.05 which has the desirable outcome of non-significance.

The coefficient estimation of individual variables and the intercept coefficient estimation are mentioned in Table 1, in the column denoted as **"B"** and the **"Exp (B)"** column respectively and it contains the odds ratio for each independent variable**.** The standard errors (S.E) and B coefficients were not excessively large, so there is no evidence of a numeric problem with this analysis (Refer to Table 1).

So the logistic regression model will be shown below in eq. (01).

$$P_{(getting\ leukemia)} = 1/\ (1+ e^{-(\infty + 0.738\ \textbf{Gender} - 0.077\ \textbf{Age} + 1.175\ \textbf{Exposure to benzene} + 1.634\ \textbf{Exposure to radiation})}) \quad (01)$$

For validation purpose, the data set was split into two sets using **SPSS**, after that again the logistic regression for each data set was executed. For data set 1 the chi square value of the model is 20.822 and sig .006, Nagelkerke R^2 value is 0.454, The overall accuracy of the model is 95 %, for data set 2 the chi square value of the model is 20.974 and sig .001, Nagelkerke R^2 value is 0.421, the overall accuracy of the model is 94%. Based on this validation analyses, we can conclude that this results are generalizable.

Table 1: Summary of variables in the equation- logistic regression

		B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Step 1[a]	**Gender**	.738	.476	2.405	1	.121	2.092
	Age	-.077	.067	1.301	1	.254	.926
	Exposure to benzene	1.175	.488	5.814	1	.016	3.240
	Exposure to	1.634	.484	11.382	1	.001	5.125
	Constant	-1.421	.542	6.885	1	.009	.241

4. CONCLUSION

The main objective of the research was to develop a risk assessment model for Sri Lankan childhood leukemia which will assess the risk level with the mostly influencing risk factors.

The limitation of this research were: only one cancer treatment center (National Cancer Institute, Maharagama) was considered to collect data and other treatment centers were not used.

Because of the insignificant relationship between the dependent variable and some independent variables, those variables were not considered in model development. Inadequate sample may be a reason for some variables not being identified as non-significant in the model development.

Another limitation was the close observation of patients was not done, incase if any patients were reluctant to state the risk factors clearly that may have created an impact on the results.

Therefore, the researchers had to put-in extra efforts to extract the information required. There are significant opportunities to strengthen this model in the future to generate integrated results to develop better risk assessment models for childhood leukemic cancer and any other variants of cancer.

5. REFERENCES

[1] R. W. Miller, J. L. Young, and B. Novakovic, *"Childhood cancer"*, Volume 75, Issue Supplement S1, pages 395–405, 1995.

[2] M. Belson, B. Kingsley, and A. Holmes, *"Risk factors for acute leukemia in children: A review-environmental health perspectives"*, Vol. 115, No. 1, pp. 138-145, 2007.

[3] I. Jolliffe, *"Wiley StatsRef: Statistics reference Online"*, DOI:10.1002/9781118445112.stat06472, 2014.

[4] K. A. Schultz, K. K. Ness, J. Whitton, C. Recklitis, B. Zebrack, L. L. Robinson, L. Zelter, and A. C. Mertens, *"Behavioral and social outcomes in adolescent survivors of childhood cancer: A report from the childhood cancer survivor study"*, J Clin Oncol 2007; 25:3649-56, 2007.

[5] M. H. Gail, R. M. Pfeiffer, *"On criteria for evaluating models of absolute risk"*, Biostatistics, 6.2, pp.227-239, 2005.

[6] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, and C. Schairer, *"Projecting individualized probabilities of developing breast cancer for white females who are being examined annually"*, J Natl Cancer Inst 1989; 81:1879–86, 1989.

[7] M. Dordelman, M. Schrappe, A. Reter, M. Zimmerman, N. Graf, and G. Schott, *"Dow's syndrome in child hood acute lymphoblastic leukemia: clinical characteristic and treatment outcomes in four consecutive BFM trials, Berlin-Franlum-Munster Group"*, Leukemia 12(5):645-51, May 1998.

[8] National Cancer Institute Statistics [online]: http://www.ncisl.lk/statistic.php