Journal of Computer Science 7 (12): 1819-1823, 2011
ISSN 1549-3636

Childhood Cancer-a Hospital based study using Decision Tree Techniques

[1]K. Kalaivani and [2]R. Shanmugalakshmi
[1]Department of Master of Computer Applications
SNS College of Technology, Coimbatore 641 035, TamilNadu, India
[2]Department of Computer Science and Engineering
Government College of Technology, Coimbatore 641 004, TamilNadu, India

Abstract: Problem statement: Cancer is generally regarded as a disease of adults. But there being a higher proportion of childhood cancer (ALL-Acute Lymphoblastic Leukemia) in India. The incidence of childhood cancer has increased over the last 25 years, but the increase is much larger in females. The aim was to increase our understanding of the determinants of south Indian parental reactions and needs. This facilitates the development of the care and follow-up routines for families, paying attention to both individual risk and resilience factors and to ways in which limitations related to treatment centre and organizational characteristics could be compensated. **Approach:** Decision Trees may be used for classification, clustering, affinity, grouping, prediction or estimation and description. One of the useful medical applications in India is the management of Leukemia, as it accounts for about 33% of childhood malignancies. **Results:** Female survivors showed greater functional disability in comparison to male survivors-demonstrated by poorer overall health status. Family stress results from a perceived imbalance between the demands on the family and the resources available to meet such demands. **Conclusion:** The pattern and severity of health and functional outcomes differed significantly between survivors in diagnostic subgroups. Family impact was aggravated by patients' lasting sequelae and by parent perceived shortcomings of long-term follow-up. Female survivors were at greater risk for health related late effects.

Key words: Acute lymphoblastic, Lymphoblastic leukemia, data mining, decision trees, knowledge discovery, parent perceived shortcomings, female survivors, greater risk, health related, late effects, limitations related

INTRODUCTION

Data mining may be defined as "the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules" (Berry and Gordon 1997). Hence, it may be considered mining knowledge from large amounts of data since it involves knowledge extraction, as well as data/pattern analysis (Han and Kamber, 2001).

Decision Trees may be used for classification, clustering, prediction, or estimation. One of the useful medical applications in Coimbatore is the management of Leukemia as it accounts for about 33% of pediatric malignancies.

Childhood Acute Lymphoblastic Leukemia (also called acute lymphocytic leukemia or ALL) is a cancer of the blood and bone marrow. This type of cancer usually gets worse quickly if it is not treated. It is the most common type of cancer in children. There are different approaches in Data Mining, namely hypothesis testing where a database recording past behavior is used to verify or disprove preconceived notions, ideas and hunches concerning relationships in the data and knowledge discovery where no prior assumptions are made and the data is allowed to speak for itself. As for knowledge discovery, it may be directed or undirected. Directed knowledge discovery tries to explain or categorize some particular data field while undirected knowledge discovery aims at finding patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes.

13% of the annual deaths worldwide are cancer-related and 70% of these are in the low-and middle-income countries. In India, the leading causes of

Corresponding Author: K. Kalaivani, Department of Master of Computer Applications SNS College of Technology, Coimbatore-641 035, TamilNadu, India

cancer-related death are carcinoma of the cervix in women and carcinoma of the lung and lower airways in men (WHO, 2011; Enskar *et al*., 1997;).

The focus of the National Cancer Control Program of India has been on primary prevention, by promoting tobacco control and genital hygiene; secondary prevention by screening for cervical cancer, breast cancer and oropharyngeal cancer; and palliative care (Dinshaw *et al*. (2006).

MATERIALS AND METHODS

Data Miningaims at discovering knowledge out of data and presenting it in a form that is easily comprehensible to humans. One of the useful applications in Coimbatore is the Cancer management, especially the management of Acute Lymphoblastic Leukemia or ALL, which is the most common type of cancer in children. The bone marrow produces stem cells (immature cells) that develop into mature blood cells.

In ALL, too many stem cells develop into a type of white blood cell called lymphocytes. These lymphocytes may also be called lymphoblasts or leukemic cells. In ALL, the lymphocytes are not able to fight infection very well. Also, as the number of lymphocytes increases in the blood and bone marrow, there is less room for healthy white blood cells, red blood cells and platelets. This may lead to infection, anemia and easy bleeding.

The following tests and procedures may be used:

- Physical exam and history
- Complete Blood Count (CBC)
- Bone marrow aspiration and biopsy
- Cytogenetic analysis
- Immunophenotyping
- Blood chemistry studies
- Chest x-ray

In childhood ALL, risk groups are used instead of stages. Risk groups are described as:

Standard (low) risk: Includes children aged 1 to 9 years who have a white blood cell count of less than 50,000 μL^{1} at diagnosis.

High risk: Includes children younger than 1 year or older than 9 years and children who have a white blood cell count of 50,000 μL^{1} or more at diagnosis. It is important to know the risk group in order to plan treatment.

Data collection:
Sources of data:

- Coimbatore Cancer Foundation's database
- Real patients' cases from proxy or parents
- Medical reviews (geographical divisions and disease categories)
- Doctors, Professors and Biostatisticians from Hospitals

Methods of data collection:

- Data acquisition from the CCF database using digital media such as CDs and hard copies in the form of printouts
- Capturing data from the CCF network from various spots
- Collecting data from patients' parents or proxies
- Note taking
- Collecting Questionnaires from patients and parent responders
- Structured interviews with experts.

Data cleaning: Real world data, like data acquired from CCF, tend to be incomplete, noisy and inconsistent. Data cleaning routines attempt to fill on missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.

Missing values: In Table: 1 the valueof row 3 is missing(i.e.,) found to be non filled.Many methods were applied to solve this issue depending on the importance of the missing value and its relation to the search domain:

- Fill in the missing value manually
- Use a global constant to fill in the missing value

Noisy data: Noise is a random error or variance in a measured variable. Many techniques were used to smooth out the data and remove the noise.

Table 1: Missing Values

39,000	free	90	L1-L2	C-All
13,000	free	96	L2	C-All
1,230	free	94	L2	-
74,000	free	95	L2	C-All
16,100	free	75	Disseminated	C-All
143,000	free	90	L2	C-All
	25,000	free	90	L2 C-All

Clustering: Outliers were detected by clustering, where similar values are organized into groups, or clusters, values that fall outside of the set of clusters may be considered outliers. In Table: 2 the value of row 5 is too small which do not match with the other rows.

Combined computer and human inspection: Using clustering techniques and constructing groups of data sets, human can then sort through the patterns in the list to identify the actual garbage ones. This is much faster than having to manually search through the entire database.

Inconsistent data: There may be inconsistencies in the data recorded for some transactions. Some data inconsistency may be corrected manually using external references, for example errors made at data entry may be corrected by performing a study trace (the most used technique in our search, to guarantee the maximum data quality possible, by reducing prediction factors). Other inconsistency forms are due to data integration, where a given attribute can have different names in different databases. Redundancies may also exist.. In Table: 3 the value of row 5 is too large when compared with all other values

Data Integration: Data Mining often requires data integration, the merging of data from multiple data sources into one coherent data store. These sources include in our case NCI database (Boman *et al*. (2010) flat files and data entry values.

Table 2: Outliers

F	254000
M	256000
F	280000
M	281000
M	2000
M	282000
F	315000
M	317000
F	325000

Table 3:Data Inconsistency

Lolita	3.00	F
Abinav	3.00	M
Monika	10.0	M
Mithun	3.00	F
Misra	156	M
Singh	1.00	M
Thanseela	10.0	F
Powar	15.0	M
Aziz	17.0	M
Aathira	11.0	F
Vishalini	15.0	F

Equivalent real-world entities from multiple data sources must be matched up, for example, patient_id in one database must be matched up with patient_number in another database.

Careful integration of the data from multiple sources helped reducing and avoiding redundancies and inconsistencies in the resulting data set. This helped improving the accuracy and speed of the subsequent mining process.

Data selection: Selecting fields of data of special interest for the search domain is the best way to obtain results relevant to the search criteria. In this research Acute Lymphoblastic Leukemia clustering was the aim, so data concerning the diagnosis of ALL (Barrera *et al*. (2005) and data concerning the patients of ALL were carefully selected from the overall data sets and mining techniques were applied to these specific data groups in order to reduce the interesting patterns reached to the ones that represent an interest for the domain.

Data transformation: In Data Transformation, the data is transformed or consolidated into forms appropriate for mining.

Smoothing: This study is to remove the noise form data. Such techniques include binning, clustering and regression.

Aggregation: where summary or aggregation operations are applied to the data.
Generalization of the data: where low-level data are replaced by higher-level concepts through concept hierarchies.

Normalization: where the attribute data are scaled so as to fall within a small specified range.

Attribute construction: where new attributes are constructed and added from the given set of attributes to help the mining process.

Choosing the tool (methodology): As a data mining application, Clementine offers a strategic approach to find useful relationships in large data sets. In contrast to more traditional statistical methods, you do not necessarily need to know what you are looking for when you start. You can explore your data, fitting different models and investigating different relationships, until you find useful information.

Working in Clementine is working with data. In its simplest form, working with Clementine is a three-step

process. First, you read data into Clementine, then run the data through a series of manipulations and finally send the data to a destination. This sequence of operations is known as a data stream because the data flows record by record from the source through each manipulation and, finally, to the destination-either a model or type of data output. Most of your work in Clementine will involve creating and modifying data streams.

At each point in the data mining process, Clementine's Visual interface invites your specific business expertise. Modeling algorithms, such as prediction, classification, segmentation and association detection, ensures powerful and accurate models. Model results can easily be deployed and read into databases, SPSS and a wide variety of other applications. You can also use the add-on component, Clementine Solution Publisher, to deploy entire data streams that read data into a model and deploy results without a full version of Clementine. This brings important data closer to decision makers who need it.

The numerous features of Clementine's data mining workbench are integrated by a visual programming interface. You can use this interface to draw diagrams of data operations relevant to your business. Each operation is represented by an icon or node and the nodes are linked together in a stream representing the flow of data through each operation.

Data evaluation: After applying the data mining techniques, comes the job of identifying the obtained results, in form of interesting patterns representing knowledge depending on interestingness measures. These measures are essential for the efficient discovery of patterns of value to the given user. Such measures can be used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre-specified interestingness constraints.

RESULTS

Female survivors show greater functional disability in comparison to male survivors-demonstrated by poorer overall health status. Family stress results from a perceived imbalance between the demands on the family and the resources available to meet such demands. Male and Female survivors are differed regarding current health and functional status and regarding the extent of health care needs and unmet such needs in adult life. Female survivors were at greater risk for health related late effects (Patterson *et al*. (2004).

DISCUSSION

The general aim of this study is to increase knowledge concerning short and long-term consequences of having a child with cancer and possible factors associated with those consequences. This includes the investigation of parental distress in general, as well as specific risks and strength factors, including variations in organization of care due to treatment Centre type.

Novel areas of focus were parental resilience to distress and a design that enabled a comparison between parents from two different types of treatment centre, investigating whether variations in parental distress might be understandable in the light of factors related to centre characteristics. The aim was to increase our understanding of the determinants of parental reactions and needs. This facilitates the development of the care and follow-up routines for families, paying attention to both individual risk and resilience factors and to ways in which limitations related to treatment centre and organizational characteristics could be compensated (Kennedy and Leyland, (1999).

A majority of earlier studies of parental reactions to childhood cancer have typically focused on the incidence and severity of distress, e.g., psychological and psychiatric reactive symptoms among parents of children with cancer. Findings indicate that these parents experience extraordinary strain which, in turn, can increase their vulnerability for developing various serious psychological symptoms

Mothers' appraisals of the strain of illness-related demands and their confidence in their own ability to deal with these were related to distress, both concurrently. Both this and the strong relationship between distress scores at the two time points indicates that problems are likely to continue, regardless of changes in the illness situation, for those mothers who find most difficulty in dealing with the situation in the early stage. Early identification of and intervention to support such parents are indicated (McDougall and Tsonis, (2009).

The results of the multivariable analysis suggested that for fathers, unlike for mothers, the effects of higher levels of stressors combined with lower levels of family support were additive. These points to the importance of service providers' awareness of the ways in which both parents are affected by the illness and other events

in their lives and recognition that responses of different family members will vary.

Although theoretical models emphasize the importance of the individual's perceptions in the stress and coping process, there are problems in interpretation of analyses of variables obtained from a single self-report source. Efforts were made to minimize such problems statistically by controlling for social desirability response bias and investigation of multicollinearity of descriptor variables before multivariable analysis. However, studies that obtain data from multiple sources, including staffs who are closely involved with families, could help to increase our understanding.

Studies using a wider range of measures could provide a clearer picture of the extent of distress. Further longitudinal research focusing on multicenter samples and using a range of measures and informants, is needed. Such research also needs to extend the period of follow-up and to investigate the effects of targeted family-focused interventions to reduce distress.

CONCLUSION

The pattern and severity of health and functional outcomes differed significantly between survivors in diagnostic subgroups. Family impact was aggravated by patients' lasting sequelae and by parent perceived shortcomings of long-term follow-up. A substantial proportion of survivors (40%) reported unmet health care needs in adult life. Most unmet needs concerned illness education and psychosocial services.

The considerable inter-regional variation in incidence and mortality rates across South India suggests a possible deficiency in the ascertainment of cases and death notification, particularly in rural areas. This facilitates the development of the care and follow-up routines for families, paying attention to both individual risk and resilience factors and to ways in which limitations related to treatment centre and organizational characteristics could be compensated

ACKNOWLEDGEMENTS

The researchers wish to thank M/s SNS College of Technology, Coimbatore for providing infrastructure facility to do optimization work. The authors also wish to thank the management and professionals from the department of Oncology in Coimbatore hospitals for providing us with necessary details and permission to conduct surveys at their premises.

REFERENCES

Barrera, M., A.K. Shaw, K.N. Speechely, E. Maunsell and L. Pogany, 2005. Educational and social late effects of childhood cancer and related clinical, personal and familial characteristics. Cancer, 104: 1751-1760. DOI: 10.1002/cncr.21390

Berry, M.J.A. and G. Linoof, 1997. Data Mining Techniques: For Marketing, Sales and Customer Support. 1st Edn., Wiley, New York, ISBN: 0471179809, pp: 454.

Choudhury, P. 2007. Indian pediatrics and child survival. Indian Pediatr, 44: 567-568. PMID: 17827628

Enskar, K., M. Carlsson, M. Golsater, E. Hamrin and A. Kreuger, 1997. Life situation and problems as reported by children with cancer and their parents. J. Pediatric, Oncol. Nursing, 14: 18-26. DOI: 10.1016/S1043-4542(97)90061-8

Han, J. and Kamber, M., 2001. Data Mining: Concepts and Techniques. 1st Edn., Morgan Kaufmann, San Francisco, ISBN: 1558604898, pp: 550.

Kennedy, C.R. and K. Leyland, 1999. Comparison of screening instruments for disability and emotional/behavioral disorders with a generic measure of health-related quality of life in survivors of childhood brain tumors. Int. J. Cancer, 83: 106-111. DOI: 10.1002/(SICI)1097-0215(1999)83:12+<106::AID-IJC19>3.0.CO;2-T

Patterson, J.M., K.E. Holm and J.G. Gurney, 2004. The impact of childhood cancer on the family: A qualitative analysis of strains, resources and coping behaviors. Psycho-oncology, 13: 390-407. DOI: 10.1002/pon.761

WHO, 2011. 10 facts about cancer. World Health Organization.