RESEARCH COMMUNICATION

Prediction of Cancer Cases for a Hospital in Nepal: A Statistical Modelling

B Sathian[1*], CR Bhatt[2,3], S Jayadevan[4], J Ninan[5], NS Baboo[6], G Sandeep[7]

Abstract

Objective: The aim of this study was to predict the number and trends of cancer cases for radiotherapy up to the year 2015 in Manipal Teaching Hospital, Pokhara, Nepal. Methods: A retrospective study was carried out on data retrieved from the radiotherapy treatment records of patients treated at Manipal Teaching Hospital between 28 September 2000 and 31 December 2008. Different statistical programmes were used for statistical modelling and prediction. Using curve-fitting methods, Linear, Logarithmic, Inverse, Quadratic, Cubic, Compound, Power, Exponential, and Growth models were tested. Results: Including constant term, none of the models were best fitted. However, excluding the constant term, the cubic model was best fitted; R^2=0. 95, p=0.001 for total cancer cases, R^2=0. 94, p=0.001 for female cancer cases and R^2=0. 95, p=0.001 for male cancer cases. The cancer cases estimated using cubic model showed a steady increase in the total frequency of cancers (including male and female cancer cases) following the year 2010. The three most common cancers reported were head and neck 24.2% (CI 21.6 - 27.0), lung 20.9% (CI 18.4 -23.6), cervix 15.9% (CI 13.7-18.3) respectively. Conclusion: The cancer cases in need of radiotherapy will increase in future years. The curve fitting method could be an effective exploratory modelling technique for predicting cancer frequency and trends over the years.

Keywords: cancer cases - curve fitting method - prediction- radiotherapy - statistical modelling - Nepal

Asian Pacific J Cancer Prev, **11**, 441-445

Introduction

Cancer is a global public health problem and radiotherapy is one of the treatment modalities of the disease. Of 12 million new cancer cases estimated globally in the year 2007, 6.7 million cases were in developing countries alone (Garcia et al., 2007). Furthermore, there is an assumption that the developing countries will have an increase of 70% new cancers cases by the year 2020 (Jones, 1999). The population based cancer registry is important in evaluating the burden of the disease (Shanmugaratnam, 1991). The cancer registries cover only 16 % of the global population and the data from developing countries are scarce (Parkin et al., 2006).

The estimation of future cancer cases and their incidence trends could be important for understanding need of resource allocation and mobilization in the cancer care sector. One approach to predict the incidence of cancer is to create the appropriate statistical models by analyzing the available data of the disease. Statistical modelling has been used for estimating diseases (Mukerji, 1989; Anderson et al., 1991) including different cancer incidences (Boyle et al., 1987; Chu et al., 1996). Nepal does not have a population based cancer registry; however, an initial seven-hospital based cancer registry has been published (Pradhananga et al., 2009). Estimation of the cancer burden for the upcoming years using the available data would be valuable also in the Nepalese context. The aim of this study was to predict the number and trend of cancer cases for radiotherapy up to the year 2015 at Manipal Teaching Hospital, Pokhara, Nepal.

Materials and Methods

This study utilized the data of the cancer cases treated with external radiotherapy between 28 September 2000 and 31 December 2008 at the Department of Radiotherapy and Oncology, Manipal Teaching Hospital (MTH), Pokhara, Nepal. MTH is a 750 bedded tertiary care hospital affiliated to Manipal College of Medical Sciences located in the western development region of Nepal (Bhatt et al., 2009). Of 1011 total cancer cases, 10 lacked relevant data, and therefore only 1001 cases were analysed in this study. The cancer cases were categorized into groups; brain, head and neck, breast, lung, cervix, gastrointestinal tract (GIT), hepatobiliary, urogenital, and

*[1]Department of Community Medicine, Manipal College of Medical Sciences, [2]Department of Radiology, Gandaki Medical College, Pokhara, Nepal, [3]Department of Plant and Environmental Sciences, Norwegian University of Life Sciences, Aas, Norway, [4]Research Division, Gulf Medical University, Ajman, United Arab Emirates, [5]Department of Radiation Therapy, Tarini Cancer Hospital and Research Institute, Alwar, Rajasthan, India, [6]Department of Physiology Medicine, Manipal College of Medical Sciences, [7]Department of Radiotherapy, Manipal Teaching Hospital, Pokhara, Nepal *For Correspondence : brijeshstat@gmail.com*

gynaecological cancers. Head and neck cancers included cancers of oral cavity, tongue, oropharynx, larynx, buccal mucosa, maxilla, paranasal sinuses, parotid glands and nasopharyngeal carcinomas. Gynaecological cancers included ovarian, uterine and vulvar/vaginal cancers. Cervix cancer was categorized as a separate category due to its high incidence. Cancers not falling into above mentioned categories were put into 'others'. Approval for the study was obtained from the institutional research ethical committee. Data were analysed using Excel 2003, R 2.8.0, SPSS 16.0 and EPI Info 3.5.1 windows versions.

Curve fitting, also known as regression analysis, was used to find the "best fit" curve for a series of data points in this study. The curve fit often produces an equation that can be used to find points anywhere along the curve. Curve fitting method was chosen to fit Linear, Logarithmic, Inverse, Quadratic, Cubic, Compound, Power, Exponential, Growth models. F-test was used for selecting the best fitting curve for hypothesis testing. P-value <0.05 (two sided) was taken statistically significant. Similarly, R^2 value > 0.80 was taken significant, where R^2 is the correlation of the contribution of years (independent variables) in predicting cancer cases (dependent variables).

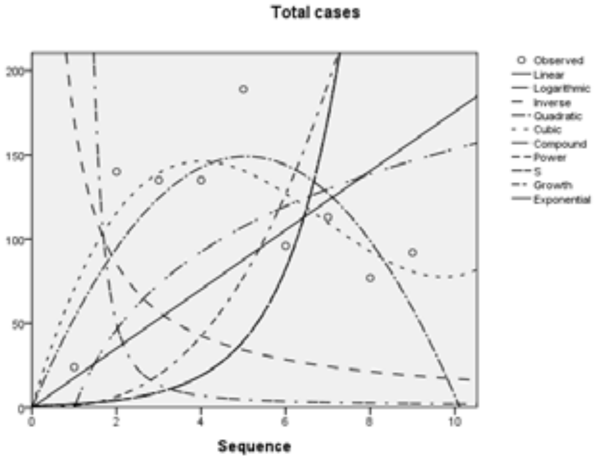
Results

Of 1001 cases, gender distribution showed a female preponderance; 53.6% (CI 50.5, 56.8) female vs. 46.4% (CI 43.2, 49.5) male. The gender difference is statistically significant (p=0.04). The patients' mean age was 58.2 years with SD 15.3 (range 1.5 - 94 years). The observed (treated) and estimated cancer cases showed an increasing trend. The estimated cancer cases then showed a decreasing trend until 2009, and the trend started to increase steadily afterwards (Table 1 and Graph 2).

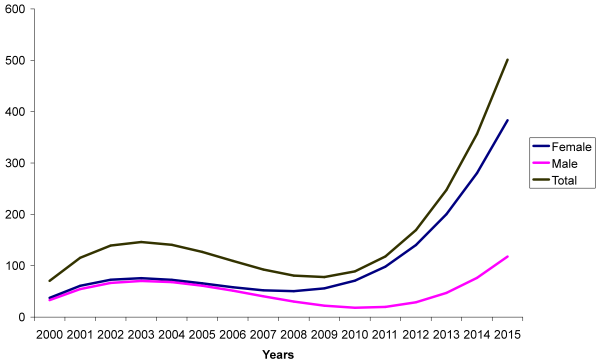
The data were modelled using curve-fitting method. Using the method, Linear, Logarithmic, Inverse, Quadratic, Cubic, Compound, Power, Exponential, and Growth function curves were explored. When constant term was included, the p values were >0.05 in all the models, and none of the models were best fitted. However, when applied without constant term in equation, almost all models were fitted.

Table 1. Observed and Estimated Number of Total Cancer Cases from 2000-2015

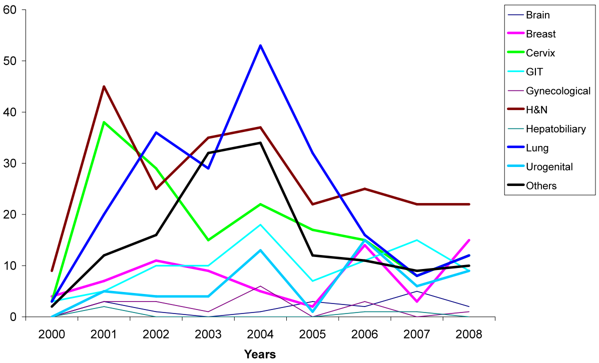
Year	Observed Cases	Estimated Cases	Confidence Intervals Lower limit	Confidence Intervals Upper limit
2000	24	71	0	159
2001	140	116	22	209
2002	135	139	47	232
2003	135	146	56	236
2004	189	141	50	231
2005	96	127	34	220
2006	113	110	17	203
2007	77	93	0	185
2008	92	81	0	190
2009	-	78	0	252
2010	-	89	0	383
2011	-	118	0	590
2012	-	169	0	882
2013	-	247	0	1268
2014	-	357	0	1760
2015	-	501	0	2370



Graph 1. Fitted Curves for Observed Cancer Cases (X-axis shows years; 1=2000, 2=2001, 3=2002, 4=2003 and so on, Y-axis shows number of cancer cases)



Graph 2. Year Wise estimates of Cancer Cases



Graph 3. Observed Number of Cancer Cases according to Anatomical Site and Year

While exploring the estimation of total annual frequency of cancers, all other curves (except inverse curve) were fitted to the data when the constant term is excluded (p-values < 0.05). The fitted curves were Linear (R^2=0.69, p=0.003), Logarithmic (R^2=0.78, p=0.001), Quadratic (R^2=0.93, p=0.001), Cubic (R^2=0.95, p=0.001), Compound (R^2=0.80, p=0.001), Power (R^2=0.85, p=0.001), Exponential (R^2=0.80, p=0.001), Growth models (R^2=0.80, p=0.001). If the p value is < 0.05, the curve with higher R^2 value is the best fitted curve. Therefore, cubic model gave the best-fitted curve in our case (Graph 1). The cubic model is a third degree polynomial, represented by the equation $y = m_0 + m_1 * x + m_2 * x^2 + m3 * x^3$, where m_0 is the constant term and m_1, m_2, m_3 are coefficient terms (Beyer, 1976; Chambers, 1983). The cubic model equations below (1, 2 and 3) contain X and Y, which are the corresponding year and frequency of cancers respectively. m_1, m_2, m_3 calculated from the observed data.

The equation for the cubic model for total number of cancers is

$$Y= 84.750X - 14.952X^2 + 0.726X^3 \text{ --------- } (1)$$

Using the equation (1), cancer cases were estimated. Table 1 shows the observed and estimated number of cancer cases up to the year 2008, and estimated number of cancer cases up to the year 2015. The observed number of cancer cases in 2008 was 92 while estimated number of cancer cases was 81, so model is fairly good to project the data. The observed and estimated number of cancer cases is almost equal up to the year 2008, however, an obvious variation in frequency of cases was also seen (Table 1). The estimated number of cancer cases during the year 2010-15 showed a decline in 2009 and later a steady increase after 2010 (Graph 2).

Similarly, the estimation of male cancer cases was done. All the curves (except inverse curve) were fitted to the data when the constant term is excluded (p-values < 0.05). The fitted curves were Linear (R^2=0.64, p=0.005), Logarithmic (R^2=0.74, p=0.001), Quadratic (R^2=0.93, p=0.001), Cubic (R^2=0.95, p=0.001), Compound (R^2=0.79, p=0.001), Power (R^2=0.85, p=0.001), Exponential (R^2=0.79, p=0.001), Growth models (R^2=0.79, p=0.001). Therefore, cubic was the best-fitted model.

The equation for cubic model for the cancers among males

$$Y= 39.293X - 6.561X^2 + 0.285X^3 \text{ -------- } (2)$$

From equation 2, the observed and estimated number of male cancer cases was derived (Table 2). The observed and estimated number of cancers up to the year 2008 was closely matched. The number of projected cancer cases showed a decline in 2009 and an increase from 2010.

Similarly, the estimation of female cancer cases was also done. All curves, except inverse curve, were fitted to the data when the constant term is excluded (p-values < 0.05).

The fitted curves were Linear (R^2=0.711, p=0.002), Logarithmic (R^2=0.792, p=0.001), Quadratic (R^2=0.922, p=0.001), Cubic (R^2=0.944, p=0.001), Compound (R^2=0.804, p=0.001), Power (R^2=0.852, p=0.001), Exponential (R^2=0.804, p=0.001), Growth models

Table 2. Observed and Estimated Number of Male Cancer Cases from 2000-2015

Year	Observed	Estimated	Confidence Intervals	
	Cases	Cases	Lower limit	Upper limit
2000	10	33	0	76
2001	62	55	9	100
2002	61	67	21	112
2003	74	70	26	115
2004	93	68	24	112
2005	46	61	16	107
2006	47	51	6	97
2007	34	41	0	86
2008	37	30	0	84
2009	-	22	0	107
2010	-	18	0	162
2011	-	20	0	251
2012	-	29	0	378
2013	-	47	0	548
2014	-	76	0	765
2015	-	118	0	1035

Table 3. Observed and Estimated Number of Female Cancer Cases from 2000-2015

Years	Observed	Expected	Confidence Intervals	
	Cases	Cases	Lower limit	Upper limit
2000	14	38	0	86
2001	78	61	9	113
2002	74	73	22	124
2003	61	76	26	126
2004	96	73	22	123
2005	50	66	14	117
2006	66	58	6	110
2007	43	52	1	103
2008	55	51	0	111
2009	-	56	0	152
2010	-	71	0	234
2011	-	98	0	360
2012	-	140	0	536
2013	-	200	0	766
2014	-	280	0	1059
2015	-	383	0	1420

Table 4. Distribution of Cancer Cases by Anatomical Regions and Years (%)

Year	Brain	Breast	Cervix	GIT	Gyn	H&N	Hepatobiliary	Lung	Urogenital	Others
2000	0.0	5.7	1.9	3.4	0.0	3.7	0.0	1.4	0.0	1.4
2001	17.6	10.0	23.9	5.7	17.6	18.6	50.0	9.6	8.8	8.7
2002	5.9	15.7	18.2	11.4	17.6	10.3	0.0	17.2	7.0	11.6
2003	0.0	12.9	9.4	11.4	5.9	14.5	0.0	13.9	7.0	23.2
2004	5.9	7.1	13.8	20.5	35.3	15.3	0.0	25.4	22.8	24.6
2005	17.6	2.9	10.7	8.0	0.0	9.1	0.0	15.3	1.8	8.7
2006	11.8	20.0	9.4	12.5	17.6	10.3	25.0	7.7	26.3	8.0
2007	29.4	4.3	5.0	17.0	0.0	9.1	25.0	3.8	10.5	6.5
2008	11.8	21.4	7.5	10.2	5.9	9.1	0.0	5.7	15.8	7.2

Table 5. Distribution of Cancer Cases by Anatomical Region

Case	Frequency	Percentage	Confidence Interval
Brain	17	1.7	1.0 - 2.8
Breast	70	7.0	5.5 - 8.8
Cervix	159	15.9	13.7 - 18.3
GIT	88	8.8	7.1 - 10.8
Gynecological	17	1.7	1.0 - 2.8
Head & neck	242	24.2	21.6 - 27.0
Hepatobiliary	4	0.4	0.1 - 1.1
Lung	209	20.9	18.4 - 23.6
Urogenital	57	5.7	4.4 - 7.4
Others	138	13.8	11.7 - 16.1
Total	**1001**	**100.0**	

(R^2=0.804, p=0.001). The cubic model is the best-fitted model.

The equation for cubic model for cancers among female is

$$Y= 45.458\ X - 8.39X^2 + 0.44X^3 \text{ ------ } (3)$$

Using the equation (3), observed and estimated number of female cancer cases were derived (Table 3). The observed and estimated numbers of cases up to the year 2008 were matched.

The observed annual cancer cases according to the anatomical regions showed observed trends of the diseases (Table 4). Likewise, the frequency of different observed cancers, according to the respective anatomical regions, was also summarized (Table 5) which showed head and neck cancer being the most common followed by the cancers of lung and cervix.

Discussion

Making future prediction of cancer incidences in Nepal is difficult due to the lack of population based cancer registry. In this study, we have used the data of cancers reported for radiotherapy at MTH to estimate the future annual incidences of the disease up to the year 2015. Due to the limitation of the study design, this estimate may not exactly predict the way the population based future estimates would have. However, we hope that it could provide useful information about the possible cancer incidences at MTH indicating the burden of the disease particularly in the catchment region. Such estimation can also help in planning future resources for cancer treatment in the hospital or in the region. Seven major Nepalese hospitals, including MTH, participated in the nation's first hospital-based cancer registry, and MTH alone treated 3.7% of total registered patients (n=162) in 2005 (Pradhananga et al., 2009). This figure included the patients treated with all available modalities (radiotherapy, surgery and chemotherapy) in combination or alone. It is interesting to mention here that India has been an alternative destination of health care among Nepalese patients (Chanda, 2002), therefore, there remains a possibility of under-estimation of cancer patients in the Nepalese cancer registry system.

The statistical estimations have been done using different approaches in the different situations. A regression equation based empirical approach can be applied when incidence rates and mortality data are available (Yang, 2005). Similar methods have been employed in predicting the cancer incidence or mortality for the countries that had a limited coverage of cancer registration system (Parkin et al., 1988; Jensen et al., 1990; Parkin et al., 1999). The decision regarding the selection of a suitable prediction approach is governed by the relative performance of the models for monitoring and prediction, and the phenomenon under study should be adequately interpreted (Nobre et al., 2001). Kamo et al. (2007) used the mathematical estimation and regression curves for the estimation of cancer incidence in Japan. The disease incidence trends are difficult to interpret as they could be related to use of screening and diagnostic practices as well as changes in exposure to risk factors. Jemal et al. (2008) used two different statistical methods for two different geographic sets of aggregate data to describe cancer trends in the US; a single linear regression model to describe short-term trends of the disease for two-thirds of geographic areas in the country, and a joinpoint model to describe long-term trends combined in a subset of these geographic areas covering approximately one-tenth of the US population. In our study, cubic model was chosen for estimating cancer trend because we had only a small cancer data set to explore. For our data, cubic model was best fitted with scattered observations as compared to with other models including linear model. Thus, our selection criteria are not so rigid, and this could make our estimation under or over.

Using the curve fitting method, we estimated the number and trend of cancer cases to receive radiotherapy at MTH from the year 2002 to 2015. Cubic model provided closely fitted curves for estimated and observed cancer cases (Graph 1). While building model, the extremities (maximums and minimums) play a great role. If the points are scattered more, the curve tries to adjust with maximum number of observed points. Therefore, it might give over-and under-estimation inevitably, but that is not the case in all the situations. A sudden annual decrease and increase in the trend is normal, as the curve cannot exactly connect these data points because of its shape. For adjusting the over-and under-estimation, the model gave wide confidence intervals in case of some years (Table 1). Devesa et al. (1995) suggested that the graph should be designed and studied carefully. Furthermore, it should provide sufficient overview so that conclusions can be drawn from it without overemphasising. In our study, the future annual estimated cancer-case curves (Graph 2) shows an increasing trend of the disease following the year 2010. Such an increase might be convincing as cancer incidence in developing countries is expected to rise principally due to the possible decline of mortality from infectious diseases, population growth and increasing life expectancy (Magrath, 2004).

Our study hereby establishes the applicability of statistical modelling in predicting the cancer incidence in the Nepalese context.

This study demonstrates the predictive value of statistical modelling to understand the burden of cancer cases in need of radiotherapy. Present prediction at MTH estimates an increasing trend of cancer burden in future years. Though cancer predictions do not match exactly the actual cancer incidences, the prediction is often helpful to understand and assess variation and trends of the disease over the years.

Acknowledgement

We would like to thank Manipal Teaching Hospital, Pokhara, Nepal for giving us permission to use the hospital data in this study. Our sincere gratitude also goes to Mr. Suresh Awasthi, Department of Optometry and Visual Sciences, Buskerud University College, Kongsberg, Norway for his invaluable inputs while writing the manuscript of this study.

References

Anderson RM, May RM (1991). Infectious disease of humans: dynamics and control. Oxford: Oxford University Press, 768 pp.

Beyer WH (1976). Standard mathematical tables. Cleveland: CRC Press.

Bhatt CR, Sharan K, Ninan J, et al (2009). Cancer treatment by radiotherapy in western

Nepal: A Hospital-based study. *Asian Pac J Cancer Prev*, **10**, 205-8.

Boyle P, Robertson C (1987). Statistical modelling of lung cancer and laryngeal cancer incidence in Scotland, 1960-1979. *Am J Epidemiol*, **125**, 731-44

Chambers JM, Cleveland WS, Kleiner B, et al (1983). Graphical methods for data analysis. Boston: Duxbury Press.

Chanda R (2002). Trade in health services. *Bull World Health Organ*, **80**, 158-63.

Chu KC, Tarone RE, Kessler LG (1996). Recent trends in U.S. breast cancer incidence, survival, and mortality rates. *J Natl Cancer Inst*, **88**, 1571-9.

Devesa SS, Donaldson J and Fears T (1995). Graphical presentation of trends in rates. *Am J Epidemiol*, **141**, 300-4.

Garcia M, Jemal A, Ward EM, et al (2007). Global cancer facts & figures American Cancer Society, Atlanta, GA. Accessed on 22. 12.09 via http://www.cancer.org/downloads/STT/Global_Cancer_Facts_and_Figures_2007_rev.pdf

Jemal A, Thun MJ, Ries LAG, et al. (2008). Annual report to the nation on the status of cancer, 1975-2005, featuring trends in lung cancer, tobacco use, and tobacco control. *J Natl Cancer Inst*, **100**, 1672-94.

Jensen OM, Esteve J, Møller H, et al (1990). Cancer in the European ommunity and its member states. *Eur J Cancer*, **26**, 1167-256.

Jones SB (1999). Cancer in the developing world: a call to action. *BMJ*, **319**, 505-8.

Kamo K, Kaneko S, Satoh K, et al (2007). A Mathematical estimation of true cancer incidence using data from population-based cancer registries. *Jpn J Clin Oncol*, **37**, 150-5.

Magrath I (2004). The international network for cancer treatment and research: helping poorer nations confront a growing problem. *Cancer Futures*, **3**, 55-8.

Mukerji S (1989). Dynamics of population and family welfare (Eds. Srinivasan K and Pathak KB). New Delhi: Himalaya Publishing House, pp. 300-14.

Nobre FF, Monteriro ABS, Telles PR, et al (2001). Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statist Med*, **20**, 3051-69.

Parkin DM, Laara E, Muir CS (1988). Estimates of the worldwide frequency of sixteen major cancers in 1980. *Int J Cancer*, **41**, 184-97.

Parkin DM, Pisani P, Ferlay J (1999). Estimates of the worldwide incidence of 25 major cancers in 1990. *Int J Cancer*, **80**, 827-41.

Parkin DM, Fernández LM (2006). Use of Statistics to Assess the Global Burden of Breast Cancer. *Breast J*, **12(Suppl 1)**, S70-80.

Pradhananga KK, Baral M, Shrestha BM (2009). Multi-institution hospital-based cancer incidence data for Nepal-an initial report. *Asian Pac J Cancer Prev*, **10**, 259-62.

Shanmugaratnam (1991). Cancer registration: Principles and methods IARC scientific publication No. 95. Jensen OM et al., (Eds) Chap 1.

Yang L (2005). Estimating Cancer Burden in China, Academic dissertation. Accessed on 26. 03.2010 via http://acta.uta.fi/pdf/951-44-6475-3.pdf