CHAPTER 1

INTRODUCTION TO CANCER SURVIVAL ANALYSIS
1.0 Introduction

The field of survival analysis emerged in the 20th century and experienced tremendous growth during the latter half of the century (Fleming and Lin, 2000). Survival analysis has become one of the major fields of modern bio-statistics. Information on survival has long been recognized as an important component in maintaining cancer control activities (WHO/IARC, 1979). An important and indirect contribution to patient care and to health care planning is the monitoring of population based survival rates (Young et al., 1984). A lot of work on the estimation of survival probability (Berkson and Gage, 1950; Kaplan and Meier, 1958; Axtell et al., 1976; Hakulinen, 1982; Ganesh, 1995; Strauss and Shavelle, 1998) and modeling based on survival times and covariates (Cox, 1972; Prentice et al., 1978; Kay, 1984; Hakulinen et al., 1987; Collett, 1991) had been carried out in the last few decades. Statistical methodology stood to gain by these and there is a lot of scope for advancement. In practice, especially from India, only a few conventional methods are being routinely applied to cancer survival data even when inappropriate. This dissertation makes a modest attempt to overcome the lacunae and applies different statistical models to the same data to choose the most appropriate one under the given conditions.

1.1 Cancer Registry

The potential source of reliable data on cancer research all over the world has been the "cancer registry". It is an organization for the systematic collection, storage, analysis, interpretation and reporting of data on subjects with cancer. Therefore, the cancer registry
is an essential part of any rational programme on cancer control (Muir et al., 1985). The process of such data collection is called ‘cancer registration’.

Hospital Cancer Registry (HCR)

This is one of the two distinct types of cancer registries. It is concerned with the recording of information on cancer patients seen in a particular hospital. The emphasis is on clinical care and hospital administration. This permits collection of “high resolution data” on all desired aspects of the subject that are required for research purpose by direct interview of patients or their by-standers. Lifetime follow up of cancer cases following treatment is one of the prime objectives of such registries.

Population Based Cancer Registry (PBCR)

This is concerned with the recording of all new or incident cancer cases occurring in a defined population, most frequently a geographic area. The emphasis is on epidemiology and public health. The indices generated like the “incidence rate” and the “risk”, relate to the underlying population giving rise to the cancer cases. The availability of information on individual cancer cases will be limited.

Cancer registration in India

In India, the first PBCR was established in Mumbai (formerly Bombay) in 1963 with the objective of estimating the incidence and pattern of cancer in Mumbai (Yeole et al., 1998). This was followed by establishing three satellite registries in the vicinity in Maharashtra. In 1981, the Indian Council of Medical Research (ICMR), the premier body for medical research of Government of India, launched the National Cancer Registry Programme (NCRP) (National Cancer Registry Programme, 1992) to obtain a reliable
estimate of cancer incidence for the entire country. At present, there are six PBCRs (Bangalore, Bhopal, Chennai (formerly Madras), Delhi and Mumbai cities and Barshi in rural Maharashtra) in this network apart from seven others outside it generating data on cancer incidence in India (Gajalakshmi et al., 2001).

The Cancer Institute (W.I.A), Chennai, was founded in 1955 to provide comprehensive cancer care for all sections of the society. It is recognized as the Regional Cancer Centre by Government of India since 1975, with the state of art facilities for cancer diagnosis, treatment and research. It is the seat of both a HCR and the Chennai PBCR. The HCR is in existence since 1955. Data on identity, socio economic and other factors related to the patient, disease and treatment are collected by direct interview of the patients. Follow up of cancer cases treated at the Cancer Institute (W.I.A) is inherent in the HCR operations. It has been predominately done by active method. Various means have been evolved during the years to enhance the response rate for follow up. All treated cases are followed up till their death (Gajalakshmi and Shanta. 1995). The Chennai PBCR, also known as “Madras Metropolitan Tumour Registry”, was established at Cancer Institute (W.I.A) in 1981 under the NCRP. The systematic and continuous collection of data on the incident cancer cases among the residents of the city of Chennai commenced from 1st January 1982. Since cancer is not a notifiable disease in India, registration of cancer cases is done by active method. In this way, the registry staff visit all the government hospitals and cooperating private hospitals, nursing homes, pathology laboratories, imaging centres and hospices, to collect data from residents of Chennai city only (duration of stay in Chennai city being at least a year at the time of first diagnosis of
cancer) by in-person interview and/or from medical records. The residential criterion was devised to exclude registering cases from a floating population (Shanta et al., 1994). The age adjusted incidence rate of cancer in Chennai, standardized to the world population (Parkin et al., 1997), was 104 per 100,000 among males and 118 per 100,000 among females during 1998. Thus, it is estimated that one in nine males and one in eight females have a cumulative lifetime risk of acquiring cancer in Chennai (Shanta et al., 2001). Follow up of cancer cases of selected cancer sites registered in Chennai PBCR commenced in 1985, to collect information on the vital status (whether alive or dead) of the cases. The major source of data on mortality is the Vital Statistics Division of the Corporation of Chennai. Data on all the deaths occurring in Chennai are collected and matched with the cancer cases registered in Chennai PBCR. Cancer cases whose death information is not available by this method, are followed up to know their vital status (Gajalakshmi and Shanta, 1995).

Thus establishment of the HCR and Chennai PBCR at Cancer Institute (W.I.A). Chennai, India, has provided the infrastructure for high quality research and the follow up activities undertaken by these registries are conducive for a successful conduct of survival studies. Data on cases arising from both HCR and Chennai PBCR have been utilized in this dissertation.

1.2 Survival analysis in cancer research

In order to describe completely the experience of cancer in a population, it is necessary to know not only its incidence and mortality, but also the survival of cancer patients. There are three main sources of information about cancer survival: the
randomised controlled clinical trial which represents the 'gold standard' for the evaluation of forms of treatment for cancer; the hospital based study which aims to provide information about the outcome of treatment in particular settings; population based survival which reflects a broader range of cancer control activities (Black et al., 1998). The field of survival analysis offers a wide range of valuable statistical methods and models for use on data that arise in cancer research from any of the above sources.

Survival analysis in cancer research dates back to many decades. One of the basic tools in the description of mortality experience of a population, the life table, was developed as early as 1693 by E. Halley in England. Only in the past three decades, much progress has been achieved in all spheres of cancer survival analysis encompassing estimation and testing of survival probabilities and elicitng of prognostic factors for survival in different framework. It is common practice to use the term “survival rate” to describe the probability of survival even though we seek to estimate an “individual probability” rather than a “rate”. Hence the terms, “survival rate”, “survival probability” or simply “survival” are used interchangeably in this dissertation.

Pre requisites of cancer survival analysis

Follow up

A careful, adequate and complete follow up is the primary goal in the conduct of an end result or a survival study. Lengthy periods of time may be required until an event of interest say, death of a patient or recurrence of the disease, occurs and maintenance of surveillance on patients may be extremely difficult. The most common event studied in a survival study on cancer is the “death” of the patient. Reliable mortality statistics in
developing countries are difficult to obtain due to incompleteness in death registration owing to lack of infrastructure in health information system. The data so collected in longitudinal studies in this environment may be complex and may not lend themselves to analysis by some standard follow up methods (Colton, 1974; Ganesh, 1995). Much of the methods are applicable under standard conditions only. Hence, simple intuitive approaches will lead to serious misinterpretations of the results. There is a need for tested methods to be evolved in effective tracing of the patients under the given conditions so that the results can be generalised to a larger population, especially in a developing environment like India (Mathew, 1996).

Methods of follow up

The follow up of patients in a cohort is an arduous task. It is done broadly in two ways. One is by passive method and the other by using active method. In passive method, the data on mortality of all cases under study are routinely received by the organization either by law or an understanding from the vital statistics division. By this procedure, those for whom, if no information of death is received, may be considered to be “alive” until that point of time. The main requirement for this method to work efficiently is that there should be a high quality of registration of mortality data and unique data linkage possibilities. This method is often followed in developed countries where the data linkage procedures are carried out using computer network connecting various spheres of social activity. Besides this, the high quality of mortality registration both in terms of completeness and quality and the availability of unique identification number for every citizen makes it possible to link data on emigration and immigration of the population.
This makes the follow up of cases almost complete. On the contrary, the situation in most developing countries is not similar. Incompleteness and unavailability of data are often encountered. Hence, follow up of cases has to be predominately done by active method. This essentially requires extra man power to specially undertake this task of follow up of cases. The different ways by which this is accomplished are by scrutiny and collection of mortality data from the records maintained at the vital statistics division, perusal of medical records in hospitals periodically, follow up surveys called “medical follow up” through the patients’ own doctors, scanning the population register (city directory), health registers of national health service, health insurance register, electoral lists, driving license registers, postal/telephone enquiries and visits to the houses of the cases or persons known to them. These are cumbersome and require elaborate man power and resources (Black and Swaminathan, 1998).

For getting the hospital based cancer survival as well as the prognostic factors for survival, the Cancer Institute (WIA), has evolved an efficient follow up system that is inherent in its HCR operations, for all cases given treatment there. Apart from the regular hospital visits by the patients for check up, postal enquiries are the most effective means of getting follow up information. For obtaining the population based cancer survival rates, Chennai PBCR was the first in India to start collecting follow up data on selected cancers. The methods of follow up data collection were as follows: matching of mortality data from death certificates, house visits, telephone/postal enquiries and repeated scrutiny of medical records in respective hospitals (Gajalakshmi and Shanta, 1995; Swaminathan et al., 1998).

7
The process of a survival study comprises (i) a well defined case, (ii) a starting point and a closing date of follow up and (iii) a well defined outcome of interest (Swaminathan et al., 1998; dos Santos Silva 1999).

Case

An unambiguous definition of a “case or subject” of interest with respect to place, time and person along with other factors should be determined before the start of the study. For cancer patients, this should include the site of cancer and may be the method of diagnosis of cancer. A valid example for cases could be patients who were “diagnosed as carcinoma of stomach between 1st January 1999 to 31st December 1999 at the Cancer Institute (WIA) on the basis of histological verification”.

Starting point

This often refers to the point of time at which the subject enters into the study. In certain situations, all subjects could be observed from the same point of time. In medical application, often, the subjects enter into the study at various points of time. Thus the starting point is usually within a period of time. The valid starting points used frequently in cancer studies are the dates of diagnosis of cancer or admission for treatment or start of treatment, depending upon the availability of information and the research setting. This starting point of the study is also referred to as the ‘index date’.

Closing date

This is a unique point of time which indicates the end of the study period for those still under observation or termination of the study. The choice of this date could be arbitrary, say the beginning or end or middle of a calendar year. It is usually imposed
upon by the researcher keeping in mind the minimum follow up time required for the latest subject entering into the study and the availability of data on follow up. It is not always practical to wait for the occurrence of outcome of interest in all the subjects in the study. The date of closure of study corresponds to the end of the period of observation.

**Outcome of interest or end point**

This is a well defined event whose occurrence is what we are looking for among the subjects. It could be anything say any death or death due to a disease or occurrence of new disease or recurrence of the same disease or progression of the disease. This is also referred to as an ‘end point’ or ‘status indicator’. It is important that the end point is a binary variable and that each subject can have one and only one end point.

**Survival time**

It is the person time elapsed from the starting point to the date of occurrence of the outcome or closing date or loss to follow up, whichever is earlier. The interpretation of the results of a survival study greatly depends upon the length of time, each case was followed up.

1.3 **Basic statistical functions in survival analysis**

The random variable “$T$” measures survival time; $T \geq 0$, a continuous variable. The actual survival time “$t$”, of an individual is the value of the variable “$T$”. “$T$” has the probability density function, $f(t)$, $t \geq 0$ and distribution function $F(t)$, so that

$$F(t) = P[T \leq t]$$
$$f(t) = F'(t) \text{ and } F(t) = \int_0^t f(u) d(u)$$

The survivor function $S(t)$, is defined as the probability of surviving beyond $t$, so that,
\[ S(t) = P[T \geq t] = 1 - F(t) \]
\[ S'(t) = -f(t), \quad S(t) = \int_0^t f(u)du \]

The hazard rate, \( h(t) \), measures the "risk" or "proneness" to, say death, at time "\( t \)" given survival up to time '\( t \)'.

\[ h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t / T \geq t]}{\Delta t} \]

All of these are equivalent ways of defining a specific survival pattern uniquely and are inter-related.

\[ h(t) = f(t) / S(t), \quad f(t) = F'(t) = -S'(t), \quad h(t) = -S'(t) / S(t) \]

If anyone of these is given, the other two can be derived. Their interrelationships are discussed and derived by Mike and Stanley (1982), Johnson and Johnson (1980) and Kalbfleisch and Prentice (1980).

Censoring

In a life time experiment or study, for example, in a medical setting, it may not be feasible to continue experimentation or follow up until all items under study have failed or experienced the event of interest (say death). If the study is terminated before all the subjects have died, or the subjects are withdrawn wilfully or getting dropped from the study, then for those that are still alive at the time of termination or withdrawal, only a lower bound on life time is available. This is not to conclude that no information is available on them, but the information on them is partial. This unique feature in lifetime data analysis which occurs when exact lifetimes are known for only a portion of the individuals in the study and known to exceed certain values in the remainder is called
"censoring". There are four ways in which censoring of an individual can occur during follow up (Chiang 1968).

(i) **by death due to a cause other than the one of interest so that the chance of dying from the specific cause cannot be determined directly.** The date of death corresponds to the “end of follow up”.

(ii) **by the individual lost to follow up.** The date at which the individual is lost to follow up corresponds to the end of the period of observation. The losses may or may not be independent of the outcome of interest. The available information on this date provides the “status indicator”.

(iii) **by withdrawal from follow up due to closure of study and these subjects are still “alive at the closing date”**.

(iv) **by ‘drop outs’, especially in clinical trials, which occur due to the side effects of treatment resulting in discontinuation of treatment and hence withdrawn from the study or incomplete treatment on the part of the subject.**

These are examples of ‘right censoring’ which is the most common form of censoring in medical studies. In this, we know that the event has not occurred during follow up but we are unable to follow up the patient further. Less common is ‘left
censoring’ where we know the event has occurred prior to the time of observation but we do not know exactly when. ‘Interval censoring’ occurs when we know that the event has occurred between two time points but don’t know the exact date. In this dissertation, we concentrate only on right censoring.

In a follow up study of cancer cases, when the withdrawal of cases is attained due to the termination of study (or closing date), it is solely technical and is a function of ‘when’ the data is analysed and not the survival status. This type of censoring is known as “non informative censoring”. The crucial assumption made is that, conditional on the values of any explanatory variable (like age, gender, tumour stage etc.), censoring is “unrelated” to the prognosis or the outcome of the disease studied. Here, the follow up is complete but without experiencing the outcome event.

When the withdrawal of cases is attained due to the cases being not traceable up to the closing date and getting lost to follow up before that, we encounter with “informative censoring” type. It now becomes a function of ‘what’ would have been the survival status and ‘how’ it should be dealt with (ie) withdrawal from follow up is associated with prognosis. Here the follow up is incomplete.

Follow up of subjects in a survival study

The follow up of three subjects A, B and C in a survival study are illustrated in Figure 1.
FIGURE 1. ILLUSTRATION OF FOLLOW-UP OF SUBJECTS IN A SURVIVAL STUDY

Subject        Calendar Time
                y_0-------------------y_1-------------------y_2-------------------y_3-------------------y_4

[----Period of registration----]

[-------------------Period of Follow up------------------]

A               i---------------------------------------------------d
B               i-----------------------------------I
C               i-----------------------------------W

Subject        Follow-up Time
                t_0-------------------t_1-------------------t_2-------------------t_3-------------------t_4

[-------------------Duration of Follow up------------------]

A               i---------------------------------------------------l
B               i-----------------------------------0
C               i-----------------------------------0
It is divided into two parts. The first shows follow up in terms of calendar time (in years, \( y_0, y_1, ..., y_s \)) and the second part shows the same information in terms of the duration of follow up (in units \( t', t_0, t_1, ..., t_d \)). The period of registration of subjects is less than the follow up period. In the first part of the figure, subject \( A \) is diagnosed with cancer with the date of diagnosis shown as \( 't' \) and dies between \( y_2 \) and \( y_4 \), shown as \( 'd' \), within the possible follow up period. In the second part, this is shown as three units of time \( 'r' \) between diagnosis and death. Subject \( B \) is diagnosed at the beginning of \( y_1 \), but is lost to follow up \((l)\) between \( y_2 \) and \( y_3 \), after duration of follow up of 1.5 units of \( 'r' \). Finally, subject \( C \) is diagnosed between \( y_1 \) and \( y_2 \) and is still alive at the end of follow up period \( y_4 \) and is said to be withdrawn \((w)\) alive after 2.5 units of follow up time \( 'r' \). Thus if death is what we are interested, only subject \( A \)'s death is known while we are not aware about the deaths of \( B \) and \( C \). But we can use the information that they did not die during the period. Here \( B \) has incomplete follow up while \( C \) has completed follow up without dying. This is key to formal survival analysis (Black and Swaminathan, 1998).

1.4 Purpose of this dissertation

Reports on survival from cancer are enormous, especially from the developed countries. However, only in the last few years, there is an appreciable increase in the number of publications on cancer survival analysis from India and other less developed regions in the world. This was made possible due to setting up of HCRs and PBCRs in these regions, realization on the part of the registry personnel on the need for follow up of registered cancer cases and international collaborative initiatives (Sankaranarayanan et al., 1998). The choice of methods and models utilized in the analysis of a vast majority of
these survival studies has rather been based on convention rather than any valid statistical considerations concerning the particular data sets and ignoring the standard assumptions that these methods require.

This dissertation is aimed at (i) a modest review of statistical methods and models in cancer survival analysis in different framework (ii) emphasize on the statistical considerations and assumptions to form the basis in the choice of methods in cancer survival analysis and (iii) their application in real time data sets from both HCR and PBCR settings in the developing countries.

1.5 Organization of chapters in this dissertation

Chapter 1 deals with the introduction to cancer registries and cancer survival analysis as summarized above. The review of literature on the various aspects of survival analysis is dealt with, in each of chapters 2 to 5.

Chapter 2 deals with the "estimation of survival probability" in the presence of informative and non-informative censoring using non-parametric statistical methods.

Chapter 3 deals with "regression models in eliciting the risk or prognostic factors for survival from cancer". The models discussed are Cox's Proportional Hazard (PH) and Accelerated Failure Time (AFT) models.

Chapter 4 deals with "multistate survival analysis" in competing risk and multistate framework. The underlying model for analysis is the Cox PH model. The data is also envisaged as arising from a matched case control study and analysed using conditional logistic regression (CLR).
Chapter 5 deals with "relative survival analysis". This includes estimation and testing of the relative survival estimates using different methods and age standardization of the relative survival rates.

Chapter 6 gives an account on "current status of research in survival analysis" encompassing frailty models and tree structured survival analysis.

Chapter 7 summarizes the conclusions drawn from chapters 2 to 5 and briefly mentions the highlights and limitations of this dissertation.