Brazilian Journal of Probability and Statistics (2001), 15, pp. 201–220.
©Associação Brasileira de Estatística

SURVIVAL ANALYSIS: PARAMETRICS TO SEMIPARAMETRICS TO PHARMACOGENOMICS

Pranab K. Sen
Departments of Biostatistics and Statistics, University of North Carolina at Chapel Hill, USA. Email: pksen@bios.unc.edu

Summary
Survival analysis with genesis in biometry and reliability analysis evolved with statistical modeling and analysis of life time data and survival distributions. During the past three decades, the advent of clinical trials has expanded its domain to clinical epidemiology, biomedical sciences, and public health disciplines, encompassing a variety of regression and noncompliance models. In this evolutionary process, parametrics gave way to more robust nonparametreics and semiparametrics where counting processes have stolen the methodoloic limelights. The past few years have witnessed a phenomenal growth of research in genomic sciences. From drug research and developmental perspectives, in pharmacogenoimcs there is a genuine need to incorporate survival analysis (albeit in a somewhat different form) in genomic clinical trials (involving biomarkers). The present study aims to provide a general appraisal of survival analysis with due emphasis on clinical trials as well as pharmacogenomics.

Key words: Bioinformatics; clinical trials; counting processes; genetic markers; genome; martingales; MRL; noncompliance; pharmacogenetics; pharmacokinetics; regression models; reliability; surrogate endpoint.

1 Introduction

Life time data models typically involve nonnegative stochastic response variables (Y) along with some *explanatory* or *auxiliary* variables $\mathbf{Z}$ which may include nonstochastic design and stochastic concomitant variables (all of which need not be continuous or even count ones). In a simple univariate setup, for Y having a *distribution function* (d.f.) $F(y), y \in \mathbf{R}^+$, the *survival function* (s.f.) is defined as

$$S(y) = 1 - F(y) = P\{Y > y\},\ y \in \mathbf{R}^+. \tag{1.1}$$

Note that $S(0) = 1$ (if there is no probability mass at $y = 0$), $S(y)$ is nonincreasing in $y(\geq 0)$, and $S(\infty) = 0$. If F admits a *density* $f = F'$ almost everywhere (a.e.), then the *hazard (failure) rate* is defined as

$$h(y) = f(y)/S(y) = -(d/dy)\log S(y),\ y \in \mathbf{R}^+, \tag{1.2}$$

so that $h(y)$ is nonnegative a.e. Also, let

$$H(y) = \int_0^y h(x)dx,\ y \geq 0, \tag{1.3}$$

be the *cumulative hazard (integrated failure)* rate. Then $H(0) = 0,\ H(\infty) = -\log S(\infty) = \infty$, and $H(y)$ is monotone nondecreasing in y. Further,

$$S(y) = \exp\{-H(y)\},\ y \geq 0. \tag{1.4}$$

Therefore, $S(.), f(.)$ or $F(.)$, more commonly used in biometric survival analysis, can equivalently be characterized in terms of $H(.)$ or $h(.)$, more popularly in use in reliability analysis. It did not, however, take too long to reconcile the two approaches in a common vein: *reliability and survival analysis* (RSA), as would be pursued here.

In the presence of explanatory variables ($\mathbf{Z}$), along with Y, we need to work with the *conditional s.f.* $S(y|\mathbf{z}) = P\{Y > y|\mathbf{Z} = \mathbf{z}\}$, or equivalently, the conditional $h(y|\mathbf{z})$ or $H(y|\mathbf{z})$, given $\mathbf{Z} = \mathbf{z}$. We relate such a conditional function to the explananatory variable $\mathbf{Z}$ by an appropriate *regression model*. In view of the properties of the s.f., or $h(.)$, a conventional linear regression model may not be appropriate. In this context, a *hazard regression model* may have some practical convenience over $S(y|\mathbf{z})$, particularly in a semiparametric setup, as would be considered later.

In a parametric formulation, drawing inference on $S(.)$ or $F(.)$ can be made by using the dual combination of *maximum likelihood estimators* (MLE) and *likelihood ratio tests* (LRT). For conditional models (e.g., $S(y|\mathbf{Z})$) such parametrifc models may lose their simplicity and computational ease, and may not be *robust* to plausible *model departures*. On top of that, *censoring* (due to withdrawl, drop out, or noncompliance) is common in clinical trials and follow-up studies where survival analysis is especially useful. Often, the precise nature of censoring can not be parameterized, and even it could be, the compound event may lead to a rather complicated likelihood function. As such, the MLE and LRT may not only lose their natural appeal (and optimality) to a greater extent, but also there could be some computational roadblocks. Nonparametrics has greater appeal in this respect, and much of this evolved around the Kaplan-Meier (1958) *product-limit estimator* (PLE).

Led by the Cox (1972, 1975) ingeneous *partial likelihood* formulation, semiparametrics captured the prime attention of mathematicians as well as researchers in survival analysis, and martingale-attracted counting processes flooded statistical journals in the 1980's and 1990's; the rampage is

still on. Yet there is some concern on unrestricted adoption of semiparametrics in survival analysis, especially when there are multiple end-points or surrogate variables, time-varying covariates, and more notably, informative censoring.

In the evolving field of *bioinformatics* (designated as applications of mathematics, statistics, and information technology (including computer science) to study and analysis of very large biological, and in particular, genetic data sets), *genomics* is the most significant component. To emphasize more the basic role of stochastics in this setup, *biostochastics* has been annexed to the same complex (Sen 2001). With the nearly completed human genome project, genomic information is to be properly incorporated in drug discovery, drug development, and pharmaceutical research, as a whole. With assistance from *knowledgre discovery and data mining* (KDDM) protocols, at the present, *pharmacokinetics, pharmacogenetics* and *pharmacpgenomics* are going for extensive data mining with the objectives of identifying (a) disease promoting genes, (b) disease preventing genes, (c) personal risks, and (d) the familial factors. The task is by no means simple. At the present, pharmaceutical industry plans to collect and utilize *genetic marker* data from patients in contemplated clinical trials with the hope that combining clinical data and genetic marker data would yield deeper insights into both *efficacy* and *safety*. Little is known on how to plan a pharmacogenomic clinical trial to incorporate genetic effects; presently, KDDM tools are used for *genomic simulations*. In this challenging statistical task, survival analysis, because of its affinity to clinical trials, has the right potentiality of providing resolutions.

The present study aims to appraise nongenomic survival analysis with an eye on the genomic counterpart, along with the hope that it would cater for the needed methodologic developments in the near future.

2 Phase I: RSA evolution

In this phase, the evolution of parametrics and its transition to nonparametrics are highlighted by the reconciliation with reliability analysis and formulaion of the PLE. Parametrics evolved around the exponential law:

$$S_1(t) = e^{-\lambda t} \quad \text{or} \quad h_1(t) = \lambda, \forall t \geq 0, \tag{2.1}$$

where λ is a positive constant; this constancy of $h(t)$ is a charateristic property of the exponential s.f. which may not generally match applications. To accommodate nonconstant hazard functions, other parametric s.f.'s have been considered; they include the Wibull and gamma families, having monote hazards. The Weibull s.f. has the form

$$S_2(t) = e^{-\lambda t^{\gamma}} \quad \text{or} \quad h_2(t) = \lambda\gamma t^{\gamma-1}, \ t \geq 0, \tag{2.2}$$

where λ and γ are both positive quantities. Thus, $h_2(t)$ is increasing or decreasing in t according as γ is $>$ or <1; for $\gamma = 1$, we have the exponential case. A gamma density is given by

$$f_3(t) = \frac{\lambda^\gamma}{\Gamma(\gamma)} e^{-t\lambda} t^{\gamma-1}, t \geq 0, \lambda > 0, \gamma > 0, \tag{2.3}$$

so that the hazard rate is $\nearrow$ or $\searrow$ in $t \geq 0$ according as γ is $>$ or <1. The first spark of nonparametrics exploited this monotonicity property under increasing generality.

(1) *The IFR/DFR family*: $S(.)$ belongs to the increasing/decreasing failure rate family according as $h(t)$ is increasing/decreasing in $t \geq 0$.

(2) *The NBU/NWU family*: $S(.)$ belongs to the new better/worse than used family according as

$$S(u) \geq \leq S(t+u)/S(t), \ \forall u, t \geq 0. \tag{2.4}$$

We also define the *mean remaining life* (MRL) at time t by

$$\mu(t) = E[Y - t | Y > t] = \frac{1}{S(t)} \int_0^\infty S(t+u) du, \ t \geq 0, \tag{2.5}$$

where $\mu(0) = EY =$ mean life time.

(3) *The DMRL/IMRL family*, characterized by $\mu(t)$ being decreasing or increasing in $t \geq 0$.

Similarly, *NBUE/NWUE* family is characterized by $\mu(t) \leq (\geq) \mu(0), \ t \geq 0$. There are other families studied extensively in reliability analysis; in survival analysis, the MRL function $M = \{\mu(t), t \geq 0\}$ has a special role, and we shall refer to that later on. In the parametric case, the *log-normal* and *log-logistic* s.f.'s are also sometimes used.

In many (biometric) survival analysis setups, without prior information on monotone hazards, or preference for log-normal or log-logistic s.f.'s, we may prefer a nonparametric approach. Based on n independent and identically distributed (i.i.d.) nonnegative random variables (r.v.) $Y_1, \ldots, Y_n$, the empirical (sample) s.f. is defined as

$$S_n(t) = 1 - F_n(t) = \frac{1}{n} \sum_{i=1}^n I(Y_i > t), t \geq 0. \tag{2.6}$$

Although $S_n(.)$ is an optimal nonparametric estortor of $S(.)$, it is a step function with decrement n^{-1} at the n ordered sample points. Thus, even though $S(.)$ may be absolutely continuous, $S_n(.)$ may not be smooth enough to match such properties. As such, if our goal is to estimate $f(.), h(.)$ or some other functional of $f(.)$, the empirical s.f. $S_n(.)$ may not suit; we may need suitable nonparametric estimators of $f(.), h(.)$ etc., derived essentially

from $S_n(.)$ through *smoothing* to a desired extent. Among the popular smoothing approaches we may mention the following.

(i) *Kernel Smoothing* (including the *nearest neighbor* (NN) method). For a suitable kernel density $k(.) = \{k(u), u \in \mathbf{R}\}$, normalized to have zero mean, unit variance and some other restraints, and for a suitable sequence $\{b_n\}$ of positive numbers, converging to zero with $n \to \infty$, let us consider an estimator of the density $f(.)$ of the form

$$\hat{f}_n(t) = \frac{1}{b_n} \int_{\mathbf{R}} k(\frac{t-y}{b_n}) dS_n(y), \ t \in \mathbf{R}^+. \tag{2.7}$$

It follows then by some routine steps that

$$\begin{aligned} E[\hat{f}_n(t)] &= f(t) + \frac{1}{2} b_n^2 f^{(2)}(t) + O(b_n^3), \\ V[\hat{f}_n(t)] &= (nb_n)^{-1} f(t) \int_{\mathbf{R}} k^2(z) dz + O(n^{-1}), \end{aligned} \tag{2.8}$$

so that the mean squared error (MSE) of $\hat{f}_n(t)$ is given by

$$E[\hat{f}_n(t) - f(t)]^2 = \frac{f(t)}{nb_n} \int_{\mathbf{R}} k^2(z) dz + \frac{1}{4} b_n^4 [f^{(2)}(t)]^2 + O(b_n^5 + n^{-1}). \tag{2.9}$$

In that sense, a minimizer of the MSE leads to a natural choice of b_n as $O(n^{-1/5})$ with an order $n^{-4/5}$ for the MSE (instead of the usual order n^{-1} prevailing in a parametric model). The actual choice of b_n depends on $t, f(t)$ and the chosen $k(.)$. Integrated MSE is usually used to have an adaptive choice. In particular, if the kernel density is chosen to be uniform on $(-\frac{1}{2}, b_n, \frac{1}{2} b_n)$, we have the usual version of the NN method; other versions can be treated in a similar way. Having obtained the density estimate, the hazard rate can be estimated by plugging-in the density estimate. Instead of the kernel method, *bounded spline* methodology has also been developed for log-density and hazard functions; we may refer to Green and Silverman (1994).

(ii) *Differentiable Smoother*. Following Chaubey and Sen (1996), let us define a triangular array of nonnegative weight functions:

$$w_{n,k}(ty) = \{(ty)^k / k!\} / \{\sum_{r=0}^{n} (ty)^r / r!\}, k = 0, \ldots, n; t \geq 0, y \geq 0. \tag{2.10}$$

Let $\{\lambda_n, n \geq 1\}$ be a sequence of (possibly stochastic) positive numbers, such that λ_n increases (a.s.) but $n^{-1}\lambda_n$ decreases (a.s.) with n. For example, one may choose $\lambda_n = \max\{Y_i : i = 1, \ldots, n\}$. Consider then the estimator

$$\tilde{S}_n(t) = \sum_{k=0}^{n} w_{n,k}(t\lambda_n) S_n(k/\lambda_n), \ t \in \mathbf{R}^+. \tag{2.11}$$

Note that the $w_{n,k}(ty)$ are contonuously differentiable (with respect to t), so that $\tilde{S}_n(t)$ is also so, and hence, it yields a smooth estimator of $f(.)$ as well as $h(.)$. We refer to Chaubey and Sen (1996) for details.
In survival analysis, often, *noncompliance* due to various reasons crop-up. Among the different modes of noncompliance, *censoring* of various types are most commonly encountered. For example, a clinical trial may be planned for a prefixed (nonstochastic) duration of time, so that responses observed in that time interval are all recorded, while those beyond the termination point are all censored. This is termed *(right) Type I censoring* or right truncation; here the number of censored observations is random. The left censoring may be defined in a similar way. In some other cases, a study may be planned for a (stochastic) duration of time, just enough to produce a prefixed number of responses; that is termed *Type II censoring.* In either case, the s.f. $S(t)$ can be estimated only for the period under study, and this can be done with simple adjustments. A more complex type of censoring, known as *random censoring*, may arise due to random entry of the subjects in the study plan, and / or their withdrawl or dropiing out also conceived in a random pattern. In this setup, the censoring may or may not be (statistically) dependent on the design or the primary response categories, and in that sense, they are said to be *informative* or *noninformative* (random) censoring. The informative censoring may generally require more complicated statistical modeling (depending on the informative nature), and in statistical modeling in survival analysis, noninformative censoring schemes have received greater attention (possibly due to greater mathematical convenience). This body of statistical research evolves around the celebrated Kaplan and Meier (1958) PLE of $S(.)$. Let the failure times $Y_1, \ldots, Y_n$ be i.i.d. nonnegative r.v.'s with the s.f. $S(.)$, and let censoring times $C_1, \ldots, C_n$ be i.i.d.r.v.'s with s.f. $S_C(.)$; it is assumed that C_i and Y_j are independent for all $i, j = 1, \ldots, n$. The observable random elements are

$$X_i = Y_i \wedge C_i,\ I_i = I(Y_i = X_i),\ i = 1, \ldots, n. \tag{2.12}$$

Let us define for every $t \geq 0$,

$$\begin{aligned} N_n(t) &= \sum_{i=1}^{n} I(X_i > t), \\ \alpha_i(t) &= I(X_i \leq t, I_i = 1),\ i = 1, \ldots, n, \\ \tau_n &= \max\{X_i, i = 1, \ldots, n\}, \\ \bar{H}_n(t) &= n^{-1} N_n(t). \end{aligned} \tag{2.13}$$

Thus, $N_n(t)$ is the number at risk (amnong the n subjects) who have not failed or censored prior to time t, while $\alpha_i(t)$ is the indicator function for the ith subject's failure before t. Under the assumed independence of Y and C, the s.f. for X is $S^o(t) = S(t)S_C(t),\ t \geq 0$. Based on this basic

feature, $\bar{P}_n(t)$, the PL estimator of $S(t)$ is defined as follows.

$$\begin{aligned}\bar{P}_n(t) &= \prod_{i=1}^{n}\Big\{\frac{N_n(X_i)}{1+N_n(X_i)}\Big\}^{\alpha_i(t)} I(t \le \tau_n) \\ &= \prod_{i=1}^{n}\Big\{\frac{n\bar{H}_n(X_i)}{1+n\bar{H}_n(X_i)}\Big\}^{\alpha_i(t)} I(t \le \tau_n) \\ &= \prod_{\{i:X_i \le t\}}\Big\{\frac{n\bar{H}_n(X_i)}{n\bar{H}_n(X_i)+1}\Big\}^{I_i},\ t \ge 0. \qquad (2.14)\end{aligned}$$

Thus, $\bar{P}_n(t) = 0, \forall t > \tau_n$, and the nonincreasing, nonnegative stochastic function $\bar{P}_n(t), t \ge 0$ can be regarded as a functional of the counting process $\{N_n(t), t \ge 0\}$. If there is no censoring, all the I_i are equal to 1, so that $\bar{P}_n(.)$ reduces to $S_n(.)$, as it should. When censoring is not stochastically independent of Y, the product formula that $S^o(t) = S(t)S_C(t)$ may not hold, and hence the PLE may not be appropriate.

Like $S_n(.)$, the PL estimator $\bar{P}_n(.)$ is also a step function. Moreover, its decrement may not be all equal (to n^{-1}), and hence, it may not be convenient to use it directly for estimating the density or the hazard function in the presence of random censoring. For this reason, suitable smoothing tools are desired to have estimates of functionals of the density under random censoring. Following Bhattacharjee and Sen (1995), we rewrite the PL estimator as follows. Let m_n be the total number of failure points, and let these ordered failure points be denoted by $X^*_{n:i}, i = 1, \ldots, m_n$. Then, we have

$$\hat{S}_n(t) = \prod_{i \le j}\Big\{\frac{n-k_i}{n-k_i+1}\Big\},\ X^*_{n:j} \le t < X^*_{n:j+1},\ j = 0, \ldots, m_n, \qquad (2.15)$$

where $X^*_{n:0} = 0, X^*_{n:m_n+1} = \infty$ and $X_{n:j}$ has the rank k_j among all the $X_i,\ i = 1, \ldots, n$, for $j = 1, \ldots, m_n$. It is naturraly appealing to use the weights $\{w_{n,k}(t\lambda_n)\}$ in (2.10) to obtain a smooth version which would lead us to suitable estimates of $f(.)$ or $h(.)$ in the presence of random censoring. However, there are some technical difficulties, particularly, for large values of t, and hence, Chaubey and Sen (1998) suggested the use of Poisson weights which lead to better estimates for the tail. Thus, we consider the smooth estimator

$$\tilde{S}^*_n(t) = \sum_{k=1}^{N}\{\mathrm{e}^{-t\lambda_n}(t\lambda_n)^k/k!\}S^*_n(k/\lambda_n),\ t \ge 0. \qquad (2.16)$$

The estimator of $f(.)$ (and $h(.)$) can readily be obtained from $\tilde{S}^*_n(.)$ (and -log$\tilde{S}^*_n(.)$) by direct differentiation with respect to t.

Let us briefly comment on the MRL function, defined in (2.5). One obvious way to estimate $\mu(t)$ is to replace $S(t)$ by $S_n(t)$ or $\bar{P}_n(t)$; but then for $t > \tau_n$, the estimator will not be defined properly. Moreover, the estimated MRL will be a nonsmooth function. Hence, Chaubey and Sen (1999) recommended the use of Poisson weights based smooth estimator, as introduced above in (2.16). That results in a smooth estimator which extends beyond τ_n.

3 Phase II: comparative clinical trials

Two-sample life testing models are the precursor to comparative clinical trials where placebo vs. treatment comparisons are quite analogous in this respect. In this comparison, there is a subtle emphasis on the hypothesis testing aspect. Although parametric testing procedures are quite appealing, excepting for the exponential s.f., in other cases, an exact test may often be difficult to construct, specially under Type I or Type II censoring. Nonparametrics fare better in this respect. A major part of the developments in nonparametrics took place in the 1960's and 1970's, and the progress is still flowing on. Basically, for the log-survival times for the placebo and treatment group observations, a classical two-sample linear rank test can be used. Such a two-sample linear rank statistic can be expressed as a statistical functional of the two s.f.'s, and more simply, in terms of the ranks of the individual sample observations (in the combined sample) incorporating suitable scores. A two-sample linear rank statistic is a special case of (regression) linear rank statistics when the regression constants are binary. Within this class of tests, an optimal one can be chosen under Pitman (contiguous) alternatives. Such a test statistic is conditionally distribution-free under Type I and distribution-free under Type II censoring. These tests are globally robust, and do not require the *proportional hazard model* (PHM) assumption underlying the semiparametric counterparts. Under a PHM, the log-rank score test can be identified as an optimal one, though plausible departures from the PHM may drastically take away its (asymptotic) optimality properties. We may refer to Sen (1981, Ch.11) for a general treatise of related nonparametrics.

In clinical trials, as contrasted to reliability analysis, there are a few features that deserve special mention. First, *interim analysis* and *time-sequential test*. In a comparative trial, often, a new drug is considered as marketable only if it performs at least as well as an existing one, and without any significant side-effect. Thus, in many cases, one has a one-sided alternative that the treatment is better than the placebo. If that alternative is statistically accepted then medical ethics may prompt to switch the surviving subjects to the treatment group as soon as possible. With that motivation, instead of waiting till the trial is over and then drawing statistical conclusions, it might be more appealing to look into the accumulating data sets at regular time intervals, and see whether or not an *early termi-*

nation of the trial can be made in favor of the treatment group. In interim analysis, this is done on a discrete basis, while in a time-sequential setup, this can be done even on a continual basis. In either way, dealing with multiple statistical looks in a clinical trial setup, a repeated significance testing scheme underlies the statistical modeling and analysis part. In principle this setup differs from the conventional *group sequential setup* where the sequential process involves independent and homogeneous increments, while here, in a typical follow-up scheme, neither independence nor homogeneity of increments can be taken for granted. For that reason, *Progressively censoring schemes* (PCS) based on suitable martingale characterizations of rank statistics were advocated by Chatterjee and Sen (1973) to formulate the methodology for time-sequential rank based procedures which may as well be used for interim analysis. We refer to Sen (1999) for an overview of this specific aspect.

One nice feature of rank tests under random censoring is that under the homogeneity of the censoring s.f.'s for the two groups, rank tests based on the X_i values alone (i.e., disregarding the indicator variables I_i) have the similar properties as in the uncensored case, though there is some loss of power due to the sacrifice of this information. To clarify this point, we refer back to (2.12) and (2.13). Note that under random censoring the s.f. $S^o(t)$ is the product of $S(t)$ and $S_C(t)$, so that under the null hypothesis, the X_i are i.i.d.r.v.'s with the s.f. $S^o(.)$. Hence the rank permutation principle holds. As the null hypothesis distribution does not depend on $S^o(.)$, the significance level and other features remain unaltered. On the other hand, because of the discount factor $S_C(.)$ under the alternative, the noncentrality is smaller than in the uncensored case, resulting in power loss. If, however, our interest lies in the modeling of regression relationship (with the explanatory variables), then the situation is different, and will be presented in the next section.

4 Phase III: rampage of semiparametrics

Depicting the dependence of the conditional s.f. or hazard rate on the explanatory variables is a delicate task. If the explanatory variables are binary or assume only a finite number of realizations, then the problem could be reduced to a multisample model, and simple nonparametric procedures can be posed. If, in addition to such nonstochastic auxiliary variables, there are stochastic concomitant variates, then in a nonparametric approach, multivariate rank procedures can be used to generate conditionally distribution-free tests for the null hypothesis of no difference in the treatment vs. placebo survival distributions, given the concomitant variables. Such procedures have been elaborately studied in a more general setup of time-sequential tests in Sen (1981, ch.11). If we assume that the regression of the (log) survival time on the stochastic concomitant is linear, but the conditional s.f. is arbitrary, then rank order estimates of the

regression parameters may as well be obtained from such rank tests. It is also possible to use *nonparametric regression* tools in this context, but that may generally require a larger sample size.

In comparative clinical trials, often the main interest centers around the regression of the hazard function on the explanatory variables. In the classical two-sample setup, Lehmann (1953) considered a class of alternatives that initiated a line of research leading to the celebrated Cox (1972) proportional hazards model (PHM). During the past 30 years the PHM has flourished under the limelight of *semiparametric models*. Although semiparametric models have touched almost every corner of statistical science, it has its genesis in survival analysis, and it has an astounding impact in this field. It is quite appropriate to introduce the PHM in a simple setup and to relate it to the semiparametrics in a more general formulation, where martingales and counting processes have cropped up in a very orchestered way.

With the same setup as in (2.12) - (2.13), we introduce covariates $\mathbf{Z}_i, i = 1, \ldots, n$ (which for the timebeing we assume to be not time-dependent). At time $t-$, there are $N_n(t-)$ units at risk, so that the conditional likelihood for a failure of the ith unit, at time t, given a faiure (not censoring) at that time can be computed. In that specific setup, Cox (1972) assumed that

$$h(t|\mathbf{z}) = [h_o(t)]e^{\boldsymbol{\beta}'\mathbf{z}}, \tag{4.1}$$

where the baseline hazard function $h_o(t)$ is arbitrary (i.e., nonparametric) while the regression function (on $\mathbf{z}$) is parametric. Under (3.1), the partial likelihood can be written as

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Big\{ \frac{e^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum_{r\in R_i} e^{\boldsymbol{\beta}'\mathbf{z}_r}} \Big\}^{I_i}, \tag{4.2}$$

where $R_i = \{r(\leq n) : X_r > X_i\}$, and the other notations are introduced in (2.12). Therefore, the partial likelihood score function (vector) is

$$\mathbf{U}_n(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}} \log PL(\boldsymbol{\beta}) = \sum_{i=1}^{n} I_i \Big\{ \mathbf{Z}_i - \frac{\sum_{j\in R_i} \mathbf{Z}_j e^{\boldsymbol{\beta}'\mathbf{z}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}'\mathbf{z}_j}} \Big\}. \tag{4.3}$$

Suppose now that we intend to test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ against alternatives that $\boldsymbol{\beta} \neq \mathbf{0}$. Sen (1981) exploited a discrete time-parameter martingale propoerty of the score statistics (at the null point), and formulated a class of time-sequential tests. We may look into this picture from a more general perspective as follows.

Adopting the notations in (2.13), we define $W_i(t) = I(X_i \geq t)$ as the at-risk indicator process for $i \geq 1$, so that $N_n(t) = \sum_{i=1}^{n} W_i(t),\ t \geq 0$.

Then, if $\Lambda_i(t)$ denotes the (conditional) cumulative hazard function of Y_i (given $\mathbf{Z}_i$), we note that

$$M_i(t) = \alpha_i(t) - \int_0^t W_i(s) d\Lambda_i(s),\ t \geq 0, i = 1, \ldots, n, \tag{4.4}$$

are (independent) square integrable martingales over $[0, \infty)$ with predictable covariation

$$< M_i, M_i > (t) = \int_0^t W_i(s)[1 - \Delta\Lambda_i(s)] d\Lambda_i(s),\ t \geq 0, \tag{4.5}$$

where $\Delta\Lambda_i(t) = \Lambda_i(t) - \Lambda_i(t-)$. Considering the flow of the events upto the time-point t, considering a more general case where the covariates are possibly time-dependent (and denoting them by $\mathbf{Z}_i(t)$) we may consider the score statistic (vector) upto time-point t expressed as

$$\begin{aligned} \mathbf{U}_n(t; \boldsymbol{\beta}) &= \sum_{i=1}^n \int_0^t \{\mathbf{Z}_i(s) - \frac{\sum_{j=1}^n W_j(s)\mathbf{Z}_j(s) e^{\boldsymbol{\beta}' \mathbf{Z}_j(t)}}{\sum_{j=1}^n W_j(s) e^{\boldsymbol{\beta}' \mathbf{Z}_j(s)}}\} d\alpha_i(s) \\ &= \sum_{i=1}^n \int_0^t \{\mathbf{Z}_i(s) - \frac{\sum_{j=1}^n \mathbf{Z}_j(s) W_j(s) e^{\boldsymbol{\beta}' \mathbf{Z}_j(s)}}{\sum_{j=1}^n W_j(s) e^{\boldsymbol{\beta}' \mathbf{Z}_j(s)}}\} dM_i(s), \\ &\qquad t \in [0, \tau_n), \end{aligned} \tag{4.6}$$

and a continuous time-parameter martingale property can be ascribed to this vector process $\mathbf{U}_n(t; \boldsymbol{\beta}), t \geq 0$ (viz., Andersen et al. (1993) for a very elaborate and unified treatise). In the two-sample model (when the Z_i are binary), the above statistic reduces to a version of the so-called log-rank statistic, and is known to have some desirable properties. Some other two-sample tests considered earlier can also be viewed as special cases of a more general weighted log-rank statistic:

$$L_n(\gamma) = \int_0^{\tau_n} \gamma(t; \bar{P}_n) dU_n(t; 0), \tag{4.7}$$

where $\gamma(t, \bar{P}_n)$ is a bounded nonnegative weight function that is made to depend on the pooled sample Kaplan and Meier (1958) product limit estimator. Fleming and Harrington (1991) suggested the family of weights:

$$\gamma(t, \bar{P}_n; a, b) = \{\bar{P}_n(t-)\}^a \{1 - \bar{P}_n(t-)\}^b, t \geq 0. \tag{4.8}$$

Lin and Kosorok (1999) considered a function-indexed stochastic processes wherein they allowed $(a, b) \in [0, K] \times [0.K]$ for a suitable values of K. Within this family, they used Monte Carlo method to prescribe a specific

weight function. Actually they also considered the case of vector valued covariates under the PHM as sketched before. Their study fails to reveal any analytical properties of their simulation based test construction. There is a need to supplement their Monte Carlo findings with analytical results, though the latter may not emerge in closed forms.

The past two decades have witnessed a phenomenal growth of research literature on semiparametrics covering all sorts of variations of the original Cox (1972) models, and these are discussed in detail in Andersen et al. (1993). Some of the developments taking place after 1992 deserve some appraisals. Two significant things in this respect are (i) multiple endpoints and surrogate endpoints models which are commonly encountered in clinical trials and survival analysis, and (ii) *interim analysis* which is generally recommended for trial monitoring and medical ethics consideration. Let me summarize the following remarks from Sen (1996). Due to excessive cost for measurement or other limitations, sometimes, in clinical trials, the primary endpoint may not be recordable, and it is not uncommon to make use of a very closely related but presumably more accessible (and less expensive) variate, termed the surrogate endpoint. This substitution generally results in some loss of statistical information, and also requires some regularity conditions that may not hold universally. This is particularly significant when the statistical interface of the surrogate and primary endpoints is not that precisely known, as may be the case when simultaneous recording of both these endpoints may not be possioble. Prentice (1989) formulated some operational criteria for surrogate endpoints, and Pepe (1992) formulated some alternative approaches involving a validation sample where both the endpoints are observable. In the context ofincomplete multiresponse designs in clinical trials, Sen (1996) has reviewed the available tools with special emphasis on rank based procedures.

The situation with multiple endpoint clinical trials is admittedly much more complex. Even for the classical multivariate multisample problems, sans the normality assumption, exact statistical analysis stumbles into difficulties. In the setup of clinical trials, the design aspects invoke complications due to the follow-up nature, possibly surrogate endpoints, censoring and noncompliance of varuious kind. Nonparametrics workout better in this respect, though the simplicity of statistical analysis may have to be compromised to a certain extent (Sen 1996). With respect to semiparametrics, we may note that the popular Cox PHM may not be directly applicable to multivariate failure time data, as the very basic assumption of conditional independence of the endpoint variables, given the covariates may not hold in general. To eliminate this roadblock, some alternative semiparametric approaches have been explored in the recent past. They include the following:

(1) The *copula* or *Frailty* model,
(2) The *marginal regression* model, and
(3) The *matrix-valued counting processes* model.

The frailty model is essentially a semiparametric model with conditional

independence of the multiple endpoints, given the unobservable stochastic frailty variable. In copula models, some specific (parametric) copula functions are assumed to tie up the marginal hazards or survival functions in a semiparametric setup. A recent article by Oakes and Ritz (2000) explores the bivariate case relating to both parametric and semiparametric setups. Their simple treatment may not be very appropriate in multiple endpoint clinical trials where multiple explanatory variables and noncompliance factors may cause considerable complications (Clegg et al. 2000). In marginal regression models, only the marginal hazards are formulated in a semiparametric setup, while nothing is specified on the dependence structure of the endpoints. As such, first, for each marginal failure time, a conventional univariate semiparametric approach is advocated, and then, to account for their interdependence, the naive covariance matrix of the estimators is replaced by some robust covariance matrix estimator (Prentice and Cai 1992). For an additional review of such related material, we refer to Clegg et al. (2000). In passing, we may remark that in the above the main issue is not related to the effect that the covariates may have on the survival experience, but rather to the estimation of the survival function. Pedroso de Lima and Sen (1997) proposed a model, termed the matrix-valued counting process model, that allows for modeling of multivariate failure time data in a Cox regression type setup, incorporating first-order interactive intensities. In this way, their model differs from the marginal models or the frailty models. For each marginal failure time, we can define a counting process as in (2.13). In that way, considering the set of all failure times, we will have a multivariate case where for each subject, we will have a p-variate counting process, so that the resulting case is a collection of n multivariate counting processes, i.e., a matrix-valued counting process. In a separate communication, Pedroso de Lima and Sen (2001) have dealt with a treatise of such counting processes in time-sequential analysis arising in clinical trials and survival analysis. As such, we shall not deal with this material in detail here.

In the preceding section, we have discussed the statistical aspects of interim analysis in the univariate setup. The picture becomes immensely complicated in the case of multiple and / or surrogate endpoint trials. Such complications may not only arise due to the complexities of the multiple endpoints, their interdependent structure, and noncompliance factors, but also due to the fact that the underlying statistical models could be much less precisely defined and may have too many parameters involved. Even for the simplest parametric models involving multivariate normal distributions, analytical determination of the cut-off points at the successive stages becomes an unmanageable task, and in the semiparametrics, the complications are many fold larger. Simulation studies have therefore been advocated on a large scale basis for drawing statistical conclusions.

Let us conclude this section with some pertinent observations on semiparametrics in this general setup of survival analysis. The basic idea of partial likelihood (Cox 1975) has been one of the most significant develop-

ments that paved the way for statistical modeling and analysis in far more complex setups than conceived before. However, in this development, the PHM, or in general, the multiplicative intensity processes that permit the formulation of the partial lilkelihood function can not be taken for granted in all applications. In a simple two-sample setup, this amounts to saying that a s.f. lies above the other; but in a more general regression setup, this entails more severe restrictions depending on the nature of the regressors or explanatory variables. As has been pointed earlier, in multiple endpoints such a condition is even less likely to hold, while the copula/frailty models entail some structuring of the copula or frailty functions, and in that way lack of robustness to such model departures remains a vital concern. In the context of survival analysis, often, we have more emphasis on the regression function on the covariates, and in that case, a nonparametric regression approach merits greater attention. On the other hand, a complete nonparametric formulation may require a very large sample size which may not always suit the practical application. As a compromise, it might be better to use a semiparametric model wherein the parametric regression function is replaced by a generalized additive model. Nevertheless, for multiple or surrogate endpoints this is going to create additional complications. With a complete nonparametric model, one ends up with a functional parameter space, and hence, interim analysis methodology requires considerable modifications.

5 Pharmacogenomics and clinical trials

With the advent of genomic science and information technology, genomic information is emerging at a furious rate and in astounding detals; this needs to be properly incorporated in drug discovery, drug development, and pharmaceutical research, in general. In a typical pregenomic era setup, in a clinical trial, the basic objectives were to compare a placebo and treatment group (or more generally multiple treatment groups) with respect to their survival patterns, possibly in the presence of numerous explanatory or auxiliary variables. This makes it naturally appealing to import the usual survival analysis methodology and to extend it to suit the needed statistical modeling and analysis task. In the present complex of pharmacokinetics, pharmacogenetics and pharmacogenomics, the situation is far more complex, though there is a great need for clinical trials to meet the general objectives. At the current state of developments, it is largely contemplated and assessed to a certain extent that there are certain genes which are carriers of certain diseases, so that a mapping of the discsease genes can be of immense help in better understanding the disease etiology as well as its treatment by medical device. Likewise, there are genes which are associated with the role of preventing certain diseases. In view of this complex, when one looks into personal risks for acquiring certain disease(s), such genetic information can be very helpful. It has also been recognized

for a long period that familial factors are very important explanatory variables in the disease diagnosis as well as prognosis. The use of biological markers in clinical trials is a first significant step in this utilization of genetic information in clinical studies. In pharmaceutical research, genetic marker data from patients in contemplated clinical trials are planned for inclusion in the statistical modeling and analysis with the hope that it would enhance the scope of assessment of both efficacy and safety of drugs. This way, genomic survival analysis methodology is emerging as an essential tool in pharmaceutical research.

Stochastic evolutionary forces act on genomes, and on top of that, sampling schemes, as may be needed for collection and monitoring of relevant data sets, also involve stochastics. Because of considerable variation in human metabolism and absorption of pharmacologic products, a precise mathematical law or deterministic equation may not be usually appropriate in pharmacogenomic research; statistical genetics plays a basic role in this context. However, much of the biological developments are taking place at a fast pace, and allowing the usual time to absorb the fundamental outcomes in an interpretable and measurable manner, statistical reasoning is somewhat lagging behind the technological advances. This might have prompted the exploratory data analysis tools under the fancy name of KDDM to take control of statistical reasonings in this highly computational biological field.

Mapping a disease means localizing the gene(s) associated with the disease on a chromosome. Many of the common diseases are thought to be under the influence of multiple environmental and genetic factors, so that *gene-environment interaction* has emerged as one of the most active research areas. Since there are more than 30,000 genes and the pool of diseases and disorders is not small either, drug discovery and target evaluation is a delicate task, lacking complete scientific precision. *Microarrays* have been incorporated as a tool for high throughput screening of genes and potential targets. The microarray analysis, at the present time, is largely in the court of KDDM wherein detection of outliers and quantification of variability are done routinely. There is a genuine need for statistical methods to address some of the basic issues: multiple testing dilemmas, multiple genes and diseases interaction, and their environmental impacts. As such, the potential problems are formidable. No wonder that the current analysis is more of exploratory nature. However, the statisticians could come up with appropriate methodology to provide better support for data mining and carry out statistical analysis in a more reasonable manner.

What is the relevance of survival analysis in this complex? In what way the classical survival analysis is amenable in this respect? Clinical trials in the conventional sense and in pharmacogenomics have somewhat different objectives, and yet they share a common goal. Because of the contemplated emphasis on clinical trials in pharmacogenomics and the affinity of survival analysis to clinical trials, there is no doubt of the relevance of survival analysis in this complex, though there are other pertinent queries at

the present. Because of the leading role of statistical genetics in this complex, and the fact that many common stochastic processes have genesis in genetics models, it is not difficult to foresee the need for more complex type of stochastic processes in genomic sciences. Such complexities are likely to arise due to excessively high diemsional spatial domains, and also due to foreseen as well as unforeseen structural restrians underlying the complex; in that way, functional models (involving latent effects) are more likely to be appropriate. This could put semiparametrics on a more favorable stand. As such, some of the developments sketched in earlier sections are more likely to provide the desired lead of statistical methodology. In physical sciences, Brownian motions (bridges, excursions) and related Gaussian processes crop up in many ways, so do the diffusion processes. In survival analysis, it is well known that such Gaussian processes provide the basis for refined statistical analysis. As such, here in genomic survival analysis, such Gaussian processes are expected to play a similar role, and in that way, survival analysis tools could provide the desired lead. In the sameway as semiparametrics grew out of survival analysis, it is not unlikely that genomic survival analysis will also be centered around the classical ones with more genetic annexations. At the sametime, because of the computational biological undercurrents, computational aspects in statistical reasoning should not be overlooked. At the present, genomic simulations and clinical trials attempt to take into account genetic marker allele frequencies, linkage disequilibrium and their impact on clinical outcomes, in a rather empirical way based solely on simulated clinical populations. Even in that respect, survival analysis models may provide methodologic incentives for simulating clinical populations preserving the genetic undercurrents to a desired extent. It's a challenge for the statistical researchers which we are obliged to meet.

6 Concluding remarks

Basically, clinical trials are designed for studying treatments for serious/chronic/life-threatening diseases or disorders. They need actual medical trials on concerned population, and often, are exploratory in nature (particularly, to start with); see Phase I and II trials. Drug developers and pharmaceutical groups, as well as, health regulatory agencies (like the FDA in USA) generally focus on treatments to *relieve symptoms* which may not match the real treatment objectives, raising questions and concern on the *cost-benefit* aspects of the treatment. Bioethics and public advocates have voiced concern about clinical trial exploitation in Third World Countries, the cost-benefit factor being a primary issue in this context. WHO and public health authorities all over the World are therefore trying to identify *effective* as well as *affordable* regimens to suit the need of developing countries. We refer to Rothman and Michel (1994), Agostini (1995), Aspinall and Goodman (1995), Angell (1997) and the 1997 *Helsinki Declaration* of

the World Medical Association (WMA). These developments raise a basic question: How far medical ethics can be implemented in clinical trials with such diverse perspectives in mind? To what extent cost-benefit aspects can overturn the basic medical prerequisites of a clinical trial (particularly, in developing countries)? How much statistical reasoning can be imparted in this broader context?

There is another important issue [viz., Temple and Ellenberg (2000)] that deserves our attention. *Placebo-controlled trials* (PCT), as have been discussed in earlier sections, are extensively used in developing new pharmaceuticals. There are allegations that PCT are invariably unethical when known effective therapy is available for the condition being treated or studied, regardless of the condition or the consequences of deferring treatments. The WMA 1997 Helsinki declaration documents ethical principles for clinical investigations. In any medical study, every patient - including those of control group - if any, should be assured of the best proven diagnostic and therapeutic method. Based on this declaration, patients asked to participate in a PCT must be informed of the existence of any effective therapy, must be able to explore the consequences of deferring such therapy with the investigator, and must provide fully informed consent. This would provide justification for the PCT even when effective therapy exists. This declaration has also led to the formulation of *active-controlled equivalence trials* (ACET) which may show that a new therapy is superior (or not inferior) to an existing one - but may not have all the other characteristics of a PCT. In any case, there is a genuine need for development of innovative statistical methodology to address such ACET, and also the more complex scenario of pharmacodynamics and pharmacokinetics.

References

Agostini, A. (1995). Placebo and EU guidelines. *Lancet*, **310**, 1710.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer.

Angell, M. (1997). The ethics of clinical research in the third world. *New Eng. J. Med.*, **337**, 847–849.

Aspinall, R. L. and Goodman, N. W. (1995). Denial of effective treatment and poor quality of clinical information in placebo-controlled trials of ondansetron for postoperative nausea and vomiting: A review of published trials. *Brit. J. Med.*, **311**, 844–846.

Bhattacharjee, M. C. and Sen, P. K. (1995). Kolmogorov-Smirnov type tests for NB(W)UE alternatives under censoring schemes. *Inst.Math. Statist. Lect. - Monog. Sr.*, **27**, 25–38.

Chatterjee, S. K. and Sen, P. K. (1973). Nonparametric testing under progressive censoring. *Calcutta Stat. Asso. Bull.*, **22**, 13–50.

Chaubey, Y. P. and Sen, P. K. (1996). On smooth estimation of survival and density functions. *Statist. Decs.*, **14**, 1–19.

Chaubey, Y. P. and Sen, P. K. (1997). On smooth estimation of hazard and cumulative hazard functions. *Frontiers in Probability and Statistics* (eds. S. P. Mukherjee et al.) Narosa, New Delhi, pp. 92–100.

Chaubey, Y. P. and Sen, P. K. (1998). On smooth functional estimation under random censoring. *Frontiers in Reliability* (eds. A. P. Basu et al.) World Scientific, Singapore, pp. 83–97.

Chaubey, Y. P. and Sen, P. K. (1999). On smooth estimation of mean residual life. *J. Stat. Plan. Infer.*, 75, 223–236.

Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *J. Roy. Statist. Soc. A*, **148**, 82–117.

Clegg, L. X., Cai, J. and Sen, P. K. (1999). A marginal mixed baseline hazards model for multivariate failure time data. *Biometrics*,**55**, 805–812.

Clegg, L. X., Cai, J. and Sen, P. K. (2000). Modeling multivariate failure time data. *Handbook of Statistics, Vol. 18: Bioenvironmental and Public Health Statistics* (eds. P. K. Sen and C. R. Rao), Elsevier, Amsterdam, pp. 803–838.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. B*, **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman-Hall.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman-Hall.

Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457–481, 562–563.

Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Statist.*, **24**, 23–43.

Lin, C.-Y. and Kosorok, M. R. (1999). A general class of function-indexed nonparametric tests for survival analysis. *Ann. Statist.*, **27**, 1722–1744.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.*, **23**, 182–198.

Oakes, D. and Ritz, J. (2000). Regression in a bivariate copula model. *Biometrika*, **87**, 345–352.

Pedroso de Lima, A. C. and Sen, P. K. (1997). A matrix-valued counting process with first-order interactive intensities. *Ann. Appl. Probab.*, **7**, 494-507.

Pedroso de Lima, A. C. and Sen, P. K. (2001). Time-sequential inference and counting processes. (in preparation).

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*, **79**, 495–512.

Prentice, R. L. (1989). Surrogate end-points in clinical trials: Definition and operational criteria. *Statist. Med.*, **8**, 431–440.

Prentice, R. L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, **79**, 495–512.

Rothman, K. J. and Mitchels, K. B. (1994). The continued unethical use of placebo-controls. *New Eng. J. Med.,* **331**, 394–398.

Sen, P. K. (1981a). The Cox regression model, invariance principles for some induced quantile processes and some repeated significance test. *Ann. Statist.*, **9**, 109–121.

Sen, P. K. (1981b). *Sequential Nonparametrics: Invariance Principles and Statistical Inference*. New York: Wiley.

Sen, P. K. (1996). Design and analysis of experiments : Nonparametric methods with application to clinical trials. *Handbook of Statistics, Vol. 13: Design and Analysis of Experiments* (eds. S. Ghosh and C. R. Rao), Elsevier, Amsterdam, pp. 91–150.

Sen, P. K. (1999). Multiple comparisons in interim analysis. *J. Statist. Plan. Infer.*, **82**, 5–23.

Sen, P. K. (2001). Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics. *Lect. Notes Inst. Statist. Acad. Sinica, Taipei, Taiwan.*

Temple, R. and Ellenberg, S. S. (2000). Placebo-controled trials and active-controlled trials in the evaluation of new treatments, I: Ethical and scientific issues. *Ann. Inter. Med.,* **133**, 455–463.

World Medical Association Declaration of Helsinki. (1997). Recommendation guiding physicians in biomedical research involving human subjects. *J. Amer. Med. Assoc.,* **277**, 925–926.