Published in final edited form as:

Comput Stat Data Anal. 2016 March 1; 95: 192–206. doi:10.1016/j.csda.2015.10.004.

Exploratory Failure Time Analysis in Large Scale Genomics

Cheng Cheng¹

Department of Biostatistics, St. Jude Children's Research Hospital 262 Danny Thomas Place,
Memphis, TN 38105-2794, USA

Abstract

In large scale genomic analyses dealing with detecting genotype-phenotype associations, such as genome wide association studies (GWAS), it is desirable to have numerically and statistically robust procedures to test the stochastic independence null hypothesis against certain alternatives. Motivated by a special case in a GWAS, a novel test procedure called correlation profile test (CPT) is developed for testing genomic associations with failure-time phenotypes subject to right censoring and competing risks. Performance and operating characteristics of CPT are investigated and compared to existing approaches, by a simulation study and on a real dataset. Compared to popular choices of semiparametric and nonparametric methods, CPT has three advantages: it is numerically more robust because it solely relies on sample moments; it is more robust against the violation of the proportional hazards condition; and it is more flexible in handling various failure and censoring scenarios. CPT is a general approach to testing the null hypothesis of stochastic independence between a failure event point process and any random variable; thus it is widely applicable beyond genomic studies.

Keywords

Censored failure time data; Exploratory analysis; Failure event point process; Stochastically monotone dependence; Correlation Profile Test; Hybrid permutation test; Large scale genomic analysis; GWAS; genotype-phenotype association

1. Introduction

The advance of high throughput genotyping assays have allowed biomedical researchers to obtain genome-wide the genotypes of hundreds of thousands single nucleotide polymorphisms (SNPs) in a large number of individuals, and have provided unprecedented opportunities to identify new genetic variations underlying various phenotypes of interest, by genome-wide association study (GWAS). Common methods to test genotype-phenotype association include parametric and nonparametric models and tests, such as linear regression, Pearson's Chi-square test, Kruskal-Wallis test, Cochran-Armitage trend test, etc., depending on the experimental design and the data type of the phenotype. Phenotypes of

¹Tel: +1 901 595 2935; Fax: +1 901 595 8843. cheng.cheng@stjude.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

failure times subject to right censoring are often analyzed in cancer clinical GWAS. For this kind of phenotype semi-parametric hazard rate regression models (Cox, 1972; Kalbfleisch and Prentice, 2002, Chapter 4–6; Fine and Gray, 1999) or accelerated failure time models (Buckley and James, 1979; Wei, Ying and Lin, 1990; Lin and Geyer, 1992; Lin, Wei and Ying, 1998) are possible choices. Motivated by a phenomenon encountered in a GWAS described below, this paper presents a novel approach called *Correlation Profile Test* (CPT).

A motivating example

In a GWAS of long-term treatment outcome of childhood leukemia with relapse as a cause-specific end point (detailed description in Section 4), we encountered a SNP for which all failures (relapses) and competing events occurred only in the individuals with the wild type genotype AA. The total sample size $n = 707$ (8 had genotype value missing), with 679 wild type (AA), 11 heterozygous (AB), and 9 homozygous mutant (BB) individuals. The three genotypes AA, AB, BB were coded into ordinal values as 0, 1, 2 respectively. The 8 individuals with missing genotype values had no events, and were removed from analysis. The total number of failures (relapse) was 70; 24 individuals had competing events, and 605 were censored. Individuals who had events are all wild type (AA). Traditional Cox regression with genotype as the explanatory variable could not produce a reliable estimate of the coefficient because iterations for the maximum pseudolikelihood estimation did not converge properly; whereas Fine and Gray (1999) regression modeling of the hazard rate function of the time to relapse subdistribution produced a highly significant result: coefficient estimate -11.36 , standard error 0.33 , two-sided P value 0 . On the other hand Gray's (1988) test comparing cumulative incidence functions of relapse across the three genotypes is not significant: $P = 0.35$.

The goal of GWAS is not to model precisely the hazard rate of the failure-time phenotype as a function of SNPs, but rather to identify SNPs on which the phenotype is stochastically dependent. Currently this goal can be partially fulfilled by testing the coefficients in a hazard rate regression model, or a nonparametric test derived from hazard function estimators. Despite representing an extreme case, the above example does prompt the need in GWAS-type applications for a more general stochastic dependence test procedure that is robust in computation and maintains performance when the assumptions required for hazard rate regression modeling fail to hold. On this account, completely nonparametric omnibus tests (Fleming and Harrington, 1991, Chapter 7; Gray, 1988) are preferable. On the other hand in vast majority of GWAS applications, genotypes are regarded as ordinal variables, and it is often desirable to test the independence (“no association”) null hypothesis against the “additive”, “dominate” or “recessive” genetic effect alternatives, which is a stochastically monotone alternative (detailed in Section 2.1). It would be then compelling to develop test procedures specifically targeting the stochastically monotone alternative hypotheses. Additionally, there are often applications in which the phenotype measures repeated failures (e.g., multiple episodes of certain toxicity caused by chemotherapy), or the censoring mechanisms are mixed or informative (detailed in Section 5). The existing nonparametric procedures do not fit naturally in these scenarios.

Correlation Profile Test (CPT) is developed to address the above issues. Briefly, each subject's survival course is regarded as a point process on $[0, \infty)$ which increases by 1 at each failure time – the “event point process” (Kalbfleisch and Prentice, 2002, Chapter 5). The survival data in a sample of size n consist of n observed point processes, some of which are only partially observed up to the censoring time. Suppose a covariate variable (a SNP) is also observed on each subject in the sample and it is desired to test the null hypothesis that failure time and the covariate are stochastically independent. CPT aggregates the correlations between the covariate and the event point process at a finite number of properly placed time points into a test statistic. Statistical significance is determined by a novel procedure called *hybrid permutation test* that combines the asymptotic null distribution of the test statistic and permutations. Analytical insights and a simulation study show that compared to the popular methods CPT maintains nominal significance levels well, and outperforms the existing methods when the condition of proportional hazards does not hold. In exact or near proportional-hazards scenarios, CPT suffers some efficiency loss compared to existing methods that are sensitive to differences in hazard rates. Moreover, CPT is a very flexible procedure that can be easily extended to various failure time scenarios more complex than traditional survival data, by properly modifying the construction of the event point process.

Definitions, construction of the CPT statistic, hybrid permutation test, and stratified CPT are described in Section 2. A simulation study and some discussions are given in Section 3. A genomic application, along with a comparison of CPT to other methods on a real dataset, is presented and discussed in Section 4. Extensions of CPT to several non-traditional failure time scenarios are described and discussed in Section 5. Additional discussions and concluding remarks are made in Section 6.

2 Correlation Profile Test

2.1 The Correlation Profile Test (CPT) Statistic

First consider the classical survival case; extensions are discussed in Section 5. Suppose each individual in the population possesses a time to failure (event time) $T_E \sim F_E$ and a censoring time $T_C \sim F_C$, with T_E and T_C independent of each other. Only observable are the time $Y = \min(T_E, T_C)$ and the event indicator $\delta = I(T_E \leq T_C)$. For ease of development consider equivalently the event point process $\mathcal{N} := \{N(t) = I(T_E \leq t), t \in [0, \infty)\}$, which is observable till the time Y if $\delta = 0$. Let $X \sim F_X$ be an explanatory (factor) variable, which can be continuous, ordinal, or binary. The observable is the triplet (Y, δ, X) . Precisely speaking, exploratory failure time analysis deals with testing the null hypothesis

$$H_0: \mathcal{N} \text{ is independent of } X,$$

versus various alternatives. Different types of alternatives can be targeted by various constructions of the CPT statistic. The type of alternatives of interest most often in practice is the stochastically monotone alternative when X is continuous or ordinal: $\mathcal{N}$ is *stochastically increase* in X if for any $x_1 < x_2$ the conditional distribution of $N(t)$ given $X = x_2$ is stochastically greater than or equal to the conditional distribution of $N(t)$ given $X = x_1$,

for all $t \in [0, \infty)$, and strict stochastic order holds for at least one pair of distinct X values. This type of stochastic monotonicity includes the dominate and recessive effects in genomic applications: for example the conditional distributions of $N(t)$ given $X = 0$ and $X = 1$ are stochastically equal, and is stochastically greater given $X = 2$ — this defines the recessive pattern. A CPT statistic targeting stochastically monotone alternative hypotheses can be constructed through the *correlation profile function* $\rho(t) := \text{Corr}(N(t), X)$, $t \in [0, \infty)$.

A random sample of size n consists of (Y_i, δ_i, X_i) , $i = 1, \dots, n$, i.i.d. random vectors. For the i th individual, define the observed event point process as (NA denotes unobserved/missing value)

$$N_i(t) := \begin{cases} 0 & \text{if } \delta_i=1 \text{ and } t < Y_i \\ 1 & \text{if } \delta_i=1 \text{ and } t \geq Y_i \\ 0 & \text{if } \delta_i=0 \text{ and } t \leq Y_i \\ NA & \text{if } \delta_i=0 \text{ and } t > Y_i, \end{cases} \quad (1)$$

$t \in [0, \infty)$. Let $\mathcal{J}_E(t) := \{i : N_i(t) = 1, i = 1, \dots, n\}$ be the index of individuals who have failed before or at time t , let $\mathcal{J}_R(t) := \{i : N_i(t) = 0, i = 1, \dots, n\}$ be the index of individuals who are still at-risk at time t , let $\mathcal{J}(t) := \mathcal{J}_E(t) \cup \mathcal{J}_R(t)$, and let $|\mathcal{J}(t)|$ be the number of elements in $\mathcal{J}(t)$. Then define

$\bar{N}(t) := |\mathcal{J}(t)|^{-1} \sum_{i \in \mathcal{J}(t)} N_i(t)$, $\bar{X}(t) := |\mathcal{J}(t)|^{-1} \sum_{i \in \mathcal{J}(t)} X_i$, and define the *sample correlation profile function*

$$\tilde{\rho}_n(t) := \frac{\sum_{i \in \mathcal{J}(t)} (N_i(t) - \bar{N}(t))(X_i - \bar{X}(t))}{\sqrt{\sum_{i \in \mathcal{J}(t)} (N_i(t) - \bar{N}(t))^2 \sum_{i \in \mathcal{J}(t)} (X_i - \bar{X}(t))^2}}. \quad (2)$$

Fix a relatively small integer J (say $J = 9$), and let t_j be the $\frac{j}{J+1}$ th sample quantile of all observed failure times ($j = 1, \dots, J$). The sequence $\{t_j, j = 1, \dots, J\}$ will be referred as time-sampling points. The CPT statistic is defined as

$$S_n := \frac{1}{J} \sum_{j=1}^J \tilde{\rho}_n(t_j). \quad (3)$$

Note the time sampling points $\{t_j, j = 1, \dots, J\}$ are placed adaptively by taking the sample quantiles of the unconditional (marginal) empirical distribution of the failure times (Y_i 's with $\delta_i = 1$, for $i = 1, \dots, n$), not involving X . If beyond some $t_{J'}$ with $J' < J$ there is no longer any individual at risk, then the above sum just runs through $j = 1, \dots, J'$. In practice the summation can be stopped even earlier, at a J' such that the number of individuals still at risk is n' (say $n' = 3$).

A broader spectrum of stochastic monotone relationship can be tested by replacing X_i for $i \in \mathcal{J}(t)$ in the sample correlation profile function with its (average) rank R_i , and replacing

$\bar{X}(t)$ with $\bar{R}(t) := |\mathcal{J}(t)|^{-1} \sum_{i \in \mathcal{J}(t)} R_i$. Notably, if there are consistent positive

(negative) sample correlations at the vast majority of the time-sampling points t_j ($j = 1, \dots, J$), S_n tends to have a large magnitude; this test statistic is thus sensitive to the stochastic monotone alternatives. The asymptotic null distribution of S_n and a test procedure called *hybrid permutation test* will be derived in Section 2.2.

Some insights into this Pearson's correlation approach can be seen by examining the covariance $\text{Cov}(N(t), X)$. For this purpose we take the constructive definition of a finite point process (Daley and Vere-Jones 1988, P.121). Let D be a Bernoulli random variable indicating if an individual ever fail ($D = 1$), and assume D follows the Bernoulli($\pi(X)$) distribution given X . Let $F_E(t|X) := \Pr(T_E \leq t | D = 1, X)$, $t \in [0, \infty)$ be the conditional cdf of the failure time T_E given X and $D = 1$, $I(\cdot)$ be the indicator function, and IR be the real line. Then $N(t) = DI(T_E \leq t)$, $E[N(t)] = \Pr(DI(T_E \leq t) = 1) = \int_{IR} \pi(x)F_E(t|x)dF_X(x)$, and $E[N(t)X] = \int_{IR} x\pi(x)F_E(t|x)dF_X(x)$; together giving

$$\text{Cov}(N(t), X) = \int_{IR} \pi(x)F_E(t|x)(x - \mu_X)dF_X(x). \quad (4)$$

Notably this covariance (consequently $\rho(t)$) is large in magnitude if $\text{Cov}(\pi(X)F_E(t|X), X)$ is large in magnitude, and CPT is sensitive to this type of dependence structure. Two special cases are: (a) D is independent of X ; hence the overall failure probability π is free of x . CPT will be sensitive if the conditional cdf $F_E(t|X)$ is highly correlated with X for every t . (b) T_E is independent of X ; CPT will be sensitive if $\pi(X)$ is highly correlated with X . In this case

using a nonlinear monotone transformation on X (such as $\frac{\exp(X)}{1+\exp(X)}$) may help the power. To contrast with other failure time methods, it is instructive to consider the conditional sub-distribution function of the time to failure defined by the conditional probability that the individual fails before or at time t given X , namely, $F_1(t|X) := \Pr(N(t) = 1 | X) = \Pr(T_E \leq t, D = 1 | X) = \pi(X)F_E(t|X)$. This sub-distribution has the hazard rate function

$$\lambda_1(t|X) = \pi(X) \frac{f_E(t|X)}{1 - \pi(X)F_E(t|X)}, \text{ with } f_E(t|X) = dF_E(t|X)/dt. \text{ Case (b) implies that}$$

$$\lambda_1(t|X) = \pi(X) \frac{f_E(t)}{1 - \pi(X)F_E(t)}, \text{ which is very similar to the proportional hazards form}$$

$\lambda^*(t|X) = \pi(X) \frac{f_E(t)}{1 - F_E(t)}$ with the baseline hazard rate function $\frac{f_E(t)}{1 - F_E(t)}$. In fact $\lambda^*(t|X)$ is an upper bound of $\lambda_1(t|X)$ for every $t \in [0, \infty)$. Therefore this type of association (alternative hypothesis) favors the procedures mainly based on semi-parametric models of the hazard rate function, and CPT may not be as efficient. Notably in the traditional survival analysis setting it is assumed implicitly that $\pi(X) \equiv 1$ (every subject dies eventually); by the point process formulation adopted herein it can be seen that hazard rate regression can be more sensitive than CPT if the failure probability depends on but the time to failure is free of the

covariate X . Case (a) on the other hand gives $\lambda_1(t|X) = \frac{\pi f_E(t|X)}{1 - \pi F_E(t|X)}$, which strongly deviates from the proportional hazards structure. Hence in this case hazard rate function regression methods may not be as efficient as CPT. These qualitative properties are specifically demonstrated by the simulation results (Tables 1 and 2, Section 3): CPT is

sensitive to case (a), but not to case (b) particularly when it is difficult to capture $\pi(X)$ by a linear model. Notably, the conditional sub-distribution function $F_1(\cdot|X)$ is essentially the leading term of a mixture cure model; so in case (b) incorporation of certain cure model approach may help increase the power (for example Laska and Meisner 1992 for discrete X); this is a point worthy of future investigation.

Presence of a competing risk—In certain applications, in addition to censoring, a competing risk is an end-point event whose occurrence prevents further observation (follow up) on the individual (Gray 1988). In such case the event indicator is typically coded as

$$\delta_i = \begin{cases} 0 & \text{censored} \\ 1 & \text{failure of interest} \\ 2 & \text{competing event.} \end{cases}$$

The regular censoring indicated by $\delta = 0$ represents the “administrative censor” (Gray 1988). Because the interest here is not to estimate the cumulative incidence of each failure type, for the purpose of testing stochastic dependence the competing event $\delta = 2$ can be regarded merely as an additional censoring mechanism. The modification to the observed event point process is simply to replace in the condition $\delta_i = 0$ by $\delta_i = 1$ in Equation (1).

2.2 Hybrid permutation test

For either the classical survival or the multiple failure case (see Section 5), the asymptotic distribution of the test statistic S_n under the independence null hypothesis is established by the following theorem; the proof is given in Appendix.

Theorem—Assume that (a) the proportion of at-risk individuals $\lambda_n(t_j)$ in computing $\tilde{\rho}_n(t_j)$ is bounded almost surely from below by some $p_0 > 0$ for each n and each $j = 1, \dots, J' - J$, and (b) the second moment of X exists, $E(X^2) < \infty$. Under the independence null hypothesis, the CPT statistic S_n is asymptotically normal with zero mean and asymptotic variance τ_n^2 , i.e.,

$$\frac{S_n}{\tau_n} \Rightarrow N(0, 1), \quad n \rightarrow \infty,$$

where $\tau_n = \tau_n(F_E, F_C, F_X, J)$ ($n = 1, 2, \dots$) is a sequence depending on the underlying event time distributions and J . Moreover, $\tau_n^2 = O(n^{-1})$, $n \rightarrow \infty$.

Utilizing this result, the basic idea of the hybrid permutation test is to estimate the asymptotic standard deviation τ_n by permutation and then plug it into a z-test: Randomly permute the covariate X_i ($i = 1, \dots, n$) values; compute S_n according to Equations (2) and (3) using the permuted data; repeat B (a few hundred) times to obtain simulated observations of S_n under H_0 : $S'_{n,1}, \dots, S'_{n,B}$. The statistical significance of the observed test statistic $S_{n,obs}$ is given by, for a two-sided test, $P = 2 \{1 - \Phi(|S_{n,obs}/\hat{\tau}_n|)\}$, where $\hat{\tau}_n$ is the sample standard deviation of the simulated observations $\{S'_{n,1}, \dots, S'_{n,B}\}$, and $\Phi(\cdot)$ is the $N(0, 1)$ cdf.

Although the asymptotic mean is zero, there can be finite sample bias; thus it is recommended to use $(S_{n,obs} - \hat{\mu}_n)/\hat{\tau}_n$ instead, with $\hat{\mu}_n$ the sample mean of $S'_{n,1}, \dots, S'_{n,B}$.

In a large-scale genomics application a massive number of such test is performed. A slightly conservative test to help reduce the number of false positive hits is to replace $\Phi(\cdot)$ by the cdf of the Student t distribution with $B - 1$ degrees of freedom in the P value calculation. When B is on the order of 10^2 this modification does not make much difference at the customary 0.05 or 0.01 level, but can be appreciably different from $N(0, 1)$ at the extreme tails (say 10^{-6} and beyond).

Back to the motivating example described in Introduction, the CPT P value of the test using S_n in Equation (3) is 0.2956 from hybrid permutation (200 rounds) and $N(0, 1)$ approximation, giving the same conclusion as Gray's test.

2.3 A more general CPT

The CPT presented above targets the stochastic monotone alternative hypothesis. Although in genomic applications monotone alternatives are the easiest to interpret biologically, another class of alternatives that might be of interested is that the genetic effect switches direction during the survival course. For example, the event point process $\mathcal{N}$ is stochastically increasing in X in a time interval $[0, T]$ (T can be either deterministic or random), then becomes stochastically decreasing in X in (T, ∞) . A modification on the CPT test statistic in Equation (3) to target this type of alternative is to take the average of the absolute values of the correlation coefficients:

$$S_{n;Beta} := \frac{1}{J} \sum_{j=1}^J |\tilde{\rho}_n(t_j)|. \quad (5)$$

Heuristically, because under the null hypothesis each $\tilde{\rho}_n(t_j)$ follows a unimodal symmetric distribution on the interval $[-1, 1]$, the distribution of $S_{n;Beta}$ which is an average of the absolute values of several such random variables, can be approximated by a unimodal distribution on the interval $[0, 1]$ under the null hypothesis. Recently Ji *et al.* (2005) use mixtures of Beta distributions to model a set of sample correlations. Borrowing their idea, a hybrid permutation test can be constructed based on a $Beta(a, b)$ distribution approximation to the null distribution of $S_{n;Beta}$. Let $\hat{\mu}_n$ and $\hat{\tau}_n$ be the sample mean and standard deviation of the simulated observations of $S_{n;Beta}$ under the null hypothesis obtained by permutation runs as described earlier. Method of moments estimators of the Beta distribution parameters are $\hat{a} = \hat{\mu}_n \hat{c}$ and $\hat{b} = \hat{c} - \hat{a}$, where $\hat{c} = \hat{\mu}_n(1 - \hat{\mu}_n)/\hat{\tau}_n - 1$. In practice a boundary control should be added by setting $\hat{a} = \max(\hat{\mu}_n \hat{c}, \varepsilon)$ and $\hat{b} = \max(\hat{c} - \hat{a}, \varepsilon)$, with ε a small positive fraction (e.g. 10^{-5}). This approach will be shown to work well by the simulation study in Section 3.

2.4 Stratified CPT

Often in a study cohort there are well-defined strata (subgroups) across which the failure time may follow different distributions. In this case it is necessary to adjust the test of the stochastic independence null hypothesis for the stratum effect; one approach is to perform a stratified test.

A simple stratified test can be constructed as follows. Use a common number J , generate separately in each stratum the set of time-sampling points $\{t_j, j = 1, \dots, J\}$ and compute the CPT statistic as described in Equation (3). Let S_k be the CPT statistic for the k th stratum ($k = 1, \dots, K$). Let $\hat{\mu}_k$ and $\hat{\tau}_k$ be the permutation-estimated asymptotic null mean and standard deviation of S_k as described above, where permutations are performed within each stratum separately and independently. A stratified CPT test statistic is

$$S = \frac{\sum_{k=1}^K (S_k - \hat{\mu}_k)}{\sqrt{\sum_{k=1}^K \hat{\tau}_k^2}}. \quad (6)$$

Under the independence null hypothesis, S approximately follows $N(0, 1)$; thus a P value can be computed as $P = 2(1 - \Phi(|S|))$. More conservatively, $\Phi(\cdot)$ can be replaced by the cdf of the Student t distribution with $BK - K$ degrees of freedom, with B the common number of permutations used to compute each S_k . Notably, this test will have good power if there is high homogeneity, that is, the directions of the $\mathcal{N}$ - X stochastic monotone relationship are the same in (almost) all strata, so that most of $(S_k - \hat{\mu}_k)$ ($k = 1, \dots, K$) have the same sign and large magnitude.

Alternatively a stratified test statistic sensitive to stochastic monotonicity regardless direction in each stratum is

$$S_q = \sum_{k=1}^K \left(\frac{S_k - \hat{\mu}_k}{\hat{\tau}_k} \right)^2, \quad (7)$$

which follows approximately the Chi-square distribution with K degrees of freedom under the null hypothesis.

A test along the line of meta-analysis can be considered as well, using Fisher's transformation. Let P_k be the CPT P value computed from the hybrid permutation test on stratum k , and define $S_p = -\sum_{k=1}^K \log(P_k)$. Then under the null hypothesis S_p approximately follows the $Gamma(K, 1)$ distribution. This approach can be applied to perform stratified test using the $S_{n;Beta}$ statistic defined in Equation (5).

Here for simplicity a common number J across all strata is used for the time sampling points in the construction. In practice it is conceivable that one can simply use the order statistics of all observed failure times in the stratum as time-sampling points for S_k (or in unstratified test for S_n in Equation (3)); the asymptotic normality may still hold. If the minimum size of the set $\mathcal{J}_R(t)$ is set to n' (say = 3) for all strata, then depending on the censoring pattern in different strata, the truncation point J' may vary from stratum to stratum. Adaptive determination of J (or J' and n') in stratified or unstratified tests remains an open problem at this point. The simulation study and the real data example indicate that setting $J = 9$ works well in a wide spectrum of scenarios.

3 A Simulation Study

Because CPT is a completely nonparametric, moment-based test, it is expected that under the exact proportional hazard model, Cox regression model will perform better. On the other hand CPT does not assume any model form for the failure time, one thus expects it to perform better than the hazard rate function regression models when there is a deviation from the proportional hazards model. Some analytical insights along this line have been discussed following Equation (4). In this simulation study the performances of unstratified CPT (Equation (3)) and the more general CPT (Equation (5); referred as *CPT.Beta* henceforth) are compared with those of Cox regression, Fine-Gray regression, and Gray's test. Four scenarios of different combinations of the failure time models and a single ordinal covariate X (taking values 0, 1, 2) are considered. With large-scale genomics applications in mind, type-I error level and power of the assessed procedures are estimated at the 0.005 nominal significance levels in all cases. One thousand (1,000) simulations are run to estimate power under alternative hypotheses and 10,000 simulations are run to estimate the type-I error probabilities. All simulations are done using R. Cox regression, Fine-Gray regression and Gray's test are computed using the R functions `coxph`, `crr` and `cuminc`, respectively. P values of CPT and *CPT.Beta* are computed by the hybrid permutation procedure using 200 permutations and $N(0, 1)$ and Beta approximations respectively, as described in Section 2. In all cases the number of time-sampling points is set $J = 9$.

3.1 Monotone (additive) effect, near to and deviation from proportional hazards – Case 1

The ordinal covariate X takes values 0, 1, 2 with respective probability 0.1, 0.4, 0.5. The censoring time $T_C \sim \text{Exp}(0.2)$. An observation of the failure event point process is simulated by two steps. First conditional on X a Bernoulli observation D is drawn according to the probability $\pi(X) := \Pr(\text{Fail}|X) = 0.2 \exp(-\theta(X - 2))$, $\theta \geq 0$. Next if $D = 1$ then the event point process $N(\cdot)$ has a single jump at the time T_E such that given X and $D = 1$, $[T_E|(X, D = 1)] \sim \text{Lognormal}(\beta X, 1)$, $\beta \geq 0$; otherwise $N(\cdot)$ remains constantly 0. An observation of the competing risk point process $N_R(\cdot)$ is generated in the same way with different parameters: $D_R \sim \text{Bernoulli}(0.1)$, $N_R(\cdot)$ has a single jump of 1 at T_R with $[T_R|D_R = 1] \sim U(0, 7)$ if $D_R = 1$; otherwise $N_R(t) = 0$ for all $t \geq 0$.

In this case lower values of X are less frequent in the population, yet individuals with lower values of X are more likely to fail if $\theta > 0$, and tend to fail more quickly if $\beta > 0$. The null hypothesis corresponds to $\theta = \beta = 0$. Observed power and significance level are given for sample sizes $n = 200$ and 700 in Table 1. All procedures can hold the nominal significance levels fairly well although *CPT.Beta* is slightly liberal. FG and Gray's tests are slightly liberal only at the smaller sample size $n = 200$. When $\beta > 0$, a case deviating from proportional hazards, CPT uniformly outperforms the other procedures. When $\theta > 0$ and $\beta = 0$ (a near proportional-hazards case), CPT suffers some efficiency loss. At $(\theta, \beta) = (0.5, 0)$ and $n = 400$, the power of CPT and *CPT.Beta* is 0.57 and 0.63 respectively, which is a level comparable to the other methods at $n = 200$. The performance pattern observed here can be understood by the discussion following Equation (4) in Section 2: When $\theta > 0$ and $\beta = 0$, the power comes from only $\text{Cov}(\pi(X), X)$ which is relatively low for this model; yet at the same time this type of alternatives has the hazard rate function in the form favoring the methods

based on hazard rate function regression modeling. Notably all procedures have rather low power for small effect sizes ($\theta = 0.25$, $\beta = 0$). When $\beta > 0$ a substantial amount of power comes from $\text{Cov}(F_E(t|X), X)$ which is fairly high for the Lognormal model; and substantial deviations from the proportional hazards structure make the hazard rate regression methods less efficient.

3.2 Monotone (additive) effect, near to and deviation from proportional hazards – Case 2

In the second case, X takes values 0, 1, 2 with respective probability 0.98, 0.015, 0.005; all the other parameters are the same as in Case 1. Note in this case higher values of X is extremely rare in the population, yet individuals with higher X are less likely to fail (if $\theta > 0$) and tend to have longer time to failure if $\beta > 0$. This mimics the situation of the motivating example in Introduction. As shown in Table 2, it is clear that, after balancing the trade-off between type-I error probability and power, CPT and CPT.Beta performs better than the other methods. Notably, the same as in Case 1, all procedures have low power for small effect sizes ($\theta = 0.25$, $\beta = 0$).

3.3 Recessive effect – Case 3

Here a recessive effect relatively difficult to detect is considered. The distribution of X , T_C and the competing risk point process are the same as in Case 2. Given X , $\Pr(D = 1|X) = 0.6 \exp(\theta(I(X = 2) - 1))$, $\theta > 0$; and $[T_E|X, D = 1] \sim \text{Lognormal}(\beta(1 - I(X = 2)), 1)$, $\beta > 0$. In this model individuals with the very rare homozygous mutant genotype ($X = 2$) have the highest overall probability of failure and tend to fail earlier than those with the other genotypes; wild type and heterozygous individuals have the same overall probability of failure and the same failure time distribution. In order to have a non-trivial probability to observe the homozygous mutant genotype, a larger sample size $n = 1500$ is used in the simulation.

It is seen in Table 3 that among all five procedures only CPT can hold the nominal significance level well. Although theoretically the stochastic monotonicity as defined in Section 2 includes the recessive effect pattern, the CPT approach nonetheless has limited power in this case. The same pattern as in Case 1 can be observed here in Table 3; and when $\beta > 0$ the CPT tests have similar power compared to Gray's test. FG and Gray's procedures can have larger power than CPT, at the cost of much inflated type-I error probability. It should be noted that all procedures compared here have rather low power (0.43) in this difficult scenario. Balancing the trade off between power and the ability of holding the nominal significance level, CPT would perform the best here, followed by Cox regression.

3.4 Switching effect by delayed failures – Case 4

Here a case of delayed failures is considered. The delay causes a switch in the direction of the genetic effect on (association with) the failure event point process: before the delay $\mathcal{N}$ is stochastically increasing in X , and switches to be decreasing in X after the delay. X takes values 0, 1, 2 with respective probability 0.55, 0.43, 0.02. Under the null hypothesis, $\Pr(D) = 0.2$ and $T_E \sim \text{Lognormal}(0, 1)$. Under the alternative hypothesis, $\Pr(D = 1|X = 1) = \Pr(D = 1|X = 2) = 0.1$ and $\Pr(D = 1|X = 0) = 0.28$; T_E is generated as $T_E = T_d + T$, with $[T|D = 1, X] \sim \text{Lognormal}(\beta X, 1)$ ($\beta > 0$) and an independent random delay $[T_d|D = 1, X] \sim \text{Gamma}(10,$

$1)I(X = 0)$. In order to observe the delayed effect it is necessary to have prolonged follow up time; thus the censoring time is set $T_C \sim \text{Exp}(1/12)$, and for the competing risk point process $\Pr(D_R) = 0.02$ and $[T_R|D_R = 1] \sim U(0, 12)$. This model defines a fairly complex non-monotone relationship between the event point process $\mathcal{N}$ and X under the alternative hypothesis: The wild type ($X = 0$) individuals have the highest overall probability of failure and tend to fail late (if $\beta < 0$) and after a random delay.

As seen in Table 4, in this non-monotone scenario both CPT.Beta and Gray's test are slightly liberal at the 0.005 significance level; it is clear however CPT.Beta is very sensitive to this type of alternatives and the other methods have little power to detect the stochastic dependence.

4 An Example in Large Scale Genotype-Phenotype Association Studies

To illustrate the practical use of CPT and contrast it with the hazard rate regression approach, a dataset of germline SNPs on Chromosome 9 and treatment outcome of childhood acute lymphoblastic leukemia (ALL) patients is analyzed. The subjects came from three clinical study protocols, P9906 (a Children's Oncology Group legacy trial), TOT-XIIIB and TOT-XV (two St. Jude Children's Research Hospital front line trials), comprising of five treatment strata: P9906, TOT-XIIIB standard/high risk (T13B-SH), TOT-XIIIB low risk (T13B-L), TOT-XV-standard/high risk (T15-SH) and TOT-XV low risk (T15-L). P9906 is a study on high risk patients. These three protocols used very different risk classification criteria and treatment regimen for each risk level; thus it is necessary to stratify the analysis with the five protocol-risk defined treatment strata. Generally speaking high/standard risk patients received more intense chemotherapy (but very different between the three protocols), and are more likely to relapse (Figure 1). The goal here is to identify SNPs that are stochastically monotonically associated with the failure event (ALL relapse) point process consistently across the protocol-risk defined treatment strata, to identify germline mutations that may put patients at higher risk of relapse. SNPs demonstrating the stochastic monotonicity pattern are clinically and biologically most meaningful for this particular study. Thus a stratified CPT with directional statistics as defined in Equation (6) is appropriate.

Germline SNP genotype data were generated by Affymetrix 100K, 500K, and SNP6.0 GeneChips. Focusing on Chromosome 9 only, 21,909 SNPs are included in the analysis after quality control. The phenotype is time to relapse since the initiation of treatment of the primary ALL. Stochastic monotonicity of the failure event (ALL relapse) point process in each SNP genotype coded as 0=AA, 1=AB, 2=BB is tested by two methods: Fine-Gray (1999) hazard regression model and stratified CPT (Equation (6)) with five different treatment strata (P9906, T13B-SH, T15-SH, T13-BL, T15-L). Death in remission and second cancer are regarded as competing events, and failure to achieve complete remission is considered as a failure at time zero. The profile information criteria for massive multiple tests I_p (Cheng, Pounds, Boyett, Pei, Kuo, and Roussel 2004) is applied to determine the statistical significance (P value) threshold in each analysis.

Both analyses yield a large number of germline SNPs for their stochastic association with ALL relapse. Interestingly, the findings include SNPs near two genes, ABL1 and MLLT3, well known for their structural alterations in leukemia (tumor) DNA resulting in hard-to-cure ALL subtypes. Findings here suggest that germline point mutations near these genes may also affect the treatment outcome. Despite a substantial overlap in the “statistically significant” SNPs by each method, many top hits by stratified CPT are not captured by Fine-Gray regression, and vice versa (Table 5).

It is instructive to examine a few SNPs in detail. The SNP A-2066957 is another incidence the same as the motivating example described in Introduction – the minor allele is infrequent and all events occurred in the wild-type individuals.

SNP A-1998030 in the ABL1 gene is captured by stratified CPT but not by Fine-Gray regression. The B allele is the minor allele. Notably, despite that the cumulative incidence functions in each treatment stratum are not well separated by the genotype (A–E, Figure 2), it is evident that consistently across the five strata individuals having the B allele tend to relapse earlier (F–J, Figure 2), resulting in the event point process monotonically increasing in the number of B allele (K–O, Figure 2). Despite the relatively small effect size of the B allele, the accumulation of consistent evidence across all strata drives the stratified CPT to a high significance level. Additionally, by visual examination the proportional hazards assumption does not seem to hold (A–E, Figure 2).

SNP A-1996168 downstream the MLLT3 gene is captured by Fine-Gray regression but not by stratified CPT. The A allele is the minor allele. It is evident that the statistical significance is driven by the large differences in cumulative incidence functions in two strata, P9906 and T15-SH (A and C, Figure 3). These differences are also reflected well in the sample correlation profile functions (K and M, Figure 3), but the lack of consistent evidence in the other strata lowers the significance of the stratified CPT (L, N, and O Figure 3). However the unstratified CPT nonetheless reaches a high significance level (Table 5); and stratified CPT with S_q defined in Equation (7) showed high significance (P value 0.00051). Other SNPs in Table 5 show patterns similar to what is described above. Another SNP downstream of the MLLT3 gene is captured by CPT but not by Fine-Gray regression. The two SNPs (A-1996168, A-1838092) are more than 190,000 base pairs away and are not in high linkage disequilibrium (correlation 0.0857).

These results essentially recapture what has been observed in the simulation study although stratified CPT is applied. Additionally it can be observed that CPT is more sensitive to the effect of infrequent genotypes involving the minor allele on the time to failure (Figure 2).

Although the primary objective of this application is to discover SNPs consistently associated with ALL relapse across different protocol-treatment strata, it is instructive to examine the findings by the more general CPT statistic $S_{n;Beta}$ defined in Equation (5) and compare with the other tests. The stratified version of $S_{n;Beta}$ based on combining stratum-specific P values (see Section 2.4) is applied. The test captures another SNP, A-2237842, downstream of the MLLT3 gene, and in very weak linkage with SNP A-1996868 in Figure 3 (correlation coefficient 0.2048). Figure 4 shows this SNP’s interesting possible interaction

effect with the treatment protocols: The count of the variant (B) allele is positively associated with relapse in the P9906 and T15 protocols (in both low and standard/high risk strata) but negatively associated with relapse in the T13B protocol. Stratified CPT with the statistic S_q showed the highest statistical significance among all five procedures. Table 6 also contains a few examples that the SNP has substantial stochastically monotone association with relapse consistently across all strata, leading to high significance for all five procedures.

5 Extensions

The CPT procedure can be extended to more complex situations by properly modifying the way of counting failures. This is fulfilled by redefining for each case the observed event point process $N_i(\cdot)$ of each individual in the sample, whereas the construction of the test statistic as defined by Equations (2), (3) and (5), and the test procedures remain unchanged.

Recurrent failures with or without competing risk

Unlike in the classical survival scenario where the failure of interest is an endpoint, it is now assumed that an individual can experience multiple times of a specific type of failure of interest, and occurrence of censoring or a competing event prevents the further observation (follow up) on the individual. The event point process now takes the form $\mathcal{N} := \{N(t) = \text{number of failures before or at } t, t \in [0, \infty)\}$, which is observable till the time of censoring or a competing event. It is assumed here that the competing event is not recurrent, and follow up on the individual stops if a competing event occurs on the individual. In a sample the data of individual i ($i = 1, \dots, n$) consist of (Y_{ij}, δ_{ij}) ($j = 1, \dots, K_i$), where Y_{ij} 's are increasingly ordered event times $Y_{i1} < \dots < Y_{iK_i}$ and $\delta_{ij} = 1$ for $j = 1, \dots, K_i - 1$ if $K_i > 1$, indicating repeated failures of interest. For any $K_i - 1$ ($i = 1, \dots, n$)

$$\delta_{iK_i} = \begin{cases} 0 & \text{censored} \\ 1 & \text{failure of interest} \\ 2 & \text{competing event(if present)}. \end{cases}$$

A fundamental difference here from the classical survival analysis setting is that an individual is still at-risk after the failure of interest has occurred. Thus arguably there is no more information on the event point process beyond the last time point Y_{iK_i} even if the last event is a failure of interest (i.e., $\delta_{iK_i} = 1$). Therefore the observed event point process should be redefined as (assuming $Y_{i1} < \dots < Y_{iK_i}$):

$$N_i(t) := \begin{cases} \sum_{j=1}^{K_i} \mathcal{I}(\delta_{ij}=1) \mathcal{I}(t \geq Y_{ij}) & \text{if } t \leq Y_{iK_i} \\ NA & \text{otherwise} \end{cases}$$

$t \in [0, \infty)$, $i = 1, \dots, n$. Note this is not a generalization of Equation (1) where the value of the event point process remains 1 beyond the failure time. Alternatively, it may be desirable to impute the unobserved part of the event point process when the individual has had at least one failure of interest, the “last observation carried forward” criteria gives the following definition which includes Equation (1) as a special case:

$$N_i(t) := \begin{cases} \sum_{j=1}^{K_i} \mathcal{J}(\delta_{ij}=1) \mathcal{J}(t \geq Y_{ij}) & \text{if } t \leq Y_{iK_i} \\ N_i(Y_{iK_i}) & \text{if } t > Y_{iK_i} \text{ and } N_i(Y_{iK_i}) > 0 \\ NA & \text{if } t > Y_{iK_i} \text{ and } N_i(Y_{iK_i}) = 0 \end{cases}$$

Note this definition reduces to Equation (1) for $K_i = 1$.

In computing $\tilde{\rho}_n(t)$ in Equation (2), one simply redefines $\mathcal{J}_E(t) := \{i : N_i(t) > 0, i = 1, \dots, n\}$.

Informative censoring

In certain applications where the relevant follow-up time is confined in a finite interval, the administrative censoring represents the fact that either the individual never failed in the entire relevant follow-up interval, or in recurrent failure case, never failed between the last episode of failure and the end of follow up. An example is the occurrences of certain acute toxicity during a (relatively) long period of chemotherapy; the study objective is to assess the association of the acute toxicity (failure) event process with some patient factor X during the chemotherapy. Thus adverse events beyond the chemotherapy period are not of interest; the follow up stops if the patient finishes the therapy (administratively censored, $\delta_i = 0$), or some competing event occurs during therapy forcing withdraw of treatment ($\delta_i = 2$). In this scenario the administrative censoring is in fact informative as it provides a complete observation of the event point process. It is desirable then to utilize this information in the test statistic. This can be fulfilled by again properly modifying the observed event point process $N_i(\cdot)$ as follows.

The same as done earlier, it is assumed that the competing event is not recurrent, and follow up on the individual stops if a competing event occurs on the individual before the administrative censoring point. If the failure of interest is an end point (non-recurrent) and there is no competing risk, then simply define $N_i(t) = I(\delta_i = 1)I(t \leq Y_i)$, $t \in [0, \infty)$. In presence of competing risk, define

$$N_i(t) := \begin{cases} I(\delta_i=1)I(t \leq Y_i) & \text{if } \delta_i \neq 2 \\ 0 & \text{if } t \leq Y_i \text{ and } \delta_i=2 \\ NA & \text{if } t > Y_i \text{ and } \delta_i=2. \end{cases}$$

For recurrent failures with or without a competing risk, define

$$N_i(t) := \begin{cases} \sum_{j=1}^{K_i} I(\delta_{ij}=1)I(t \geq Y_{ij}) & \text{if } \delta_{iK_i}=0 \\ \sum_{j=1}^{K_i} I(\delta_{ij}=1)I(t \geq Y_{ij}) & \text{if } t \leq Y_{iK_i} \text{ and } \delta_{iK_i} \neq 0 \\ NA & \text{if } t > Y_{iK_i} \text{ and } \delta_{iK_i} \neq 0. \end{cases}$$

Here again the individual is still at-risk after experiencing an episode of failure of interest as long as the administrative censoring point has not been reached. It may be desirable to impute the unobserved part of the event point process when the individual has had at least

one episode of failure of interest, then the “last observation carried forward” strategy gives the following alternative definition:

$$N_i(t) := \begin{cases} \sum_{j=1}^{K_i} I(\delta_{ij}=1)I(t \geq Y_{ij}) & \text{if } \delta_{iK_i}=0 \\ \sum_{j=1}^{K_i} I(\delta_{ij}=1)I(t \geq Y_{ij}) & \text{if } t \leq Y_{iK_i} \text{ and } \delta_{iK_i} \neq 0 \\ N_i(Y_{iK_i}) & \text{if } t > Y_{iK_i}, N_i(Y_{iK_i}) > 0, \text{ and } \delta_{iK_i} \neq 0 \\ NA & \text{if } t > Y_{iK_i}, N_i(Y_{iK_i}) = 0, \text{ and } \delta_{iK_i} \neq 0. \end{cases}$$

Notably, this setting is similar to the problem of nonignorable dropouts caused by censoring due to competing risks (Rotnitzky, Scharfstein, Su, Robins, 2001; Scharfstein, Rotnitzky, Robins, 1999), but the analysis objective here is exploratory.

6 Discussion and Concluding Remarks

Motivated by a GWAS application, a novel approach (CPT) to testing the stochastic independence null hypothesis involving failure times subject to right censoring is developed. This approach has its unique strengths and limitations compared to the existing hazard rate regression and nonparametric test methods. In human genetic studies dealing with failure time phenotypes, it is always of the most interest to detect the additive (strictly stochastically monotone), dominate, or recessive genetic effects which are the most biologically interpretable, regardless the proportional hazards condition. These genetic effect patterns are covered by the stochastic monotone alternative hypothesis described in Section 2. The simulation results indicate that, for monotone (additive) and recessive effect patterns the CPT approach uniformly has higher or similar power compared to existing methods when the proportional hazards condition does not hold. The more general CPT.Beta procedure clearly outperforms the existing methods when the stochastic dependence is non-monotone and complex. In the scenarios where the proportional hazards condition holds exactly or approximately, the CPT approach is less efficient. The loss of power can be compensated by larger sample sizes typical in GWAS applications. Because the nature of the genetic effect and the stochastic dependence structure is unknown *a priori*, the CPT approach effectively complements the existing failure time analysis methods.

The CPT approach is quite flexible and can be adapted to several different failure time scenarios with simple modifications to the counting schema. Modifications to the hazard rate regression models to accommodate such scenarios and recurrent failures do not appear straightforward. Lastly, CPT is numerically more robust because it is solely based on sample moments and does not rely on any convergence of numerical iterations to solve an estimation equation. This feature is particularly desirable in large-screening applications such as GWAS of failure time phenotypes.

The CPT approach employs a non-traditional and very flexible method, the hybrid permutation test, to determine statistical significance. A major feature of this method is that it combines the existing parametric analytical approximation (derived from asymptotic theory or by a heuristic) with simulation (permutation) to construct a much more computationally efficient test procedure compared to the naive permutation test. In GWAS applications high significance levels typically beyond 10^{-4} are desired; at least 10,000 naive

permutations have to be performed in order to differentiate such significance level from 0; whereas a hybrid permutation test requires only a few hundred (typically 200) permutations. The reduction in computational effort is at least 50 times, a reduction highly desirable in GWAS type applications. Note the estimation of the parameters (τ_n and μ_n for example) only involves computing a few sample moments from a few hundred data points; thus with the contemporary statistical software such as R, the computation effort for this part is negligible. The simulation study in Section 3 indicates that the hybrid permutation tests considered herein can hold the significance level fairly well at the relatively stringent 0.005 significance level, and in the case it fails (Case 2: Table 2), the other methods are even worse. A formal investigation and further development of hybrid permutation tests are certainly worthy of future efforts, although these are beyond the scope of this paper.

Notably CPT is a general approach to the inference of the association between a failure event point process and any random variable; thus it is widely applicable beyond genomic studies.

Acknowledgments

This research is supported in part by the NIH grants U01-GM92666 and CA21765, and the American Lebanese and Syrian Associated Charities (ALSAC). Thanks are due Dr. Xueyuan Cao, Dr. Wenjian Yang, Dr. Mary Relling and Dr. William Evans, all at St. Jude Children's Research Hospital, for technical assistance and inspiring collaborations. Many thanks are due two reviewers, an associate editor and a co-editor, for their constructive critiques and suggestions that substantially helped improve this paper.

References

- Buckley J, James I. Linear regression with censored data. *Biometrika*. 1979; 66:429–436.
- Cheng C, Pounds S, Boyett JM, Pei D, Kuo M-L, Roussel MF. Statistical significance threshold criteria for analysis of microarray gene expression data. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3 Article 36.
- Cox DR. Regression models and life tables (with discussion). *Journal of Royal Statistical Society Series B*. 1972; 34:187–220.
- Daley, D.; Vere-Jones, D. *An Introduction to the Theory of Point Processes*. New York: Springer-Verlag; 1988.
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*. 1999; 94:496–509.
- Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*. 1988; 16:1141–1154.
- Laska EM, Meisner MJ. Nonparametric estimation and testing in a cure model. *Biometrics*. 1992; 48:1223–1234. [PubMed: 1290799]
- Ji Y, Wu C, Liu P, Wang J, Coombes KR. Applications of beta-mixture models in bioinformatics. *Bioinformatics*. 2005; 21:2118–2122. [PubMed: 15713737]
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. Second Edition. New Jersey: John Wiley & Sons, Inc.; 2002.
- Lin DY, Grayer CJ. Computational methods for semiparametric linear regression with censored data. *Journal of Computational and Graphical Statistics*. 1992; 1:77–90.
- Lin DY, Wei LJ, Ying Z. Accelerated failure time models for counting processes. *Biometrika*. 1998; 85:605–618.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of The American Statistical Association*. 1999; 94:1096–1146.

- Rotnitzky A, Scharfstein DO, Su T-L, Robins JM. Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*. 2001; 57:103–113. [PubMed: 11252584]
- Wei LJ, Ying Z, Lin DY. Linear regression analysis of censored survival data based on rank tests. *Biometrika*. 1990; 77:845–852.

References

- Aly E-E, Csörgő M, Horváth L. Strong approximations of the quantile process of the product-limit estimator. *Journal of Multivariate Analysis*. 1985; 16:185–210.
- Major P, Rejtő L. Strong embedding of the estimator of the distribution function under random censorship. *Annals of Statistics*. 1988; 16:1113–1132.

Appendix: Proof of the theorem

The proof is given for the classical survival case with the sample event point process defined in Equation (1). Proof for the multiple-failure case can be done in the same way.

It suffices to show that under the H_0 for each fixed j , $\rho_n(t_j)$ is Asymptotically Normal (AN) with zero mean and some asymptotic variance of order $O(n^{-1})$.

1° First consider the sample failure time quantiles t_j at $\frac{j}{J+1}$. Once condition (a) holds, Aly, Csörgő, and Horváth (1985) and Major and Rejtő (1988) imply that as $n \rightarrow \infty$ $t_j \rightarrow \xi$ with probability 1 for some true failure time quantile ξ .

2° Let $\rho(\xi) = \text{Corr}(N(\xi), X)$, which equals to 0 under H_0 . Consider the sample correlation $\rho_n(\xi)$. Standard large sample arguments show that $\rho_n(\xi)$ is AN with mean $\rho(\xi)$ and some asymptotic variance $\sigma_n^2 = O(n^{-1})$.

3° Next consider $\rho_n(t_j) - \rho(\xi) = \rho_n(\xi) - \rho(\xi) + \rho_n(t_j) - \rho_n(\xi)$. By 2° above, $\rho_n(\xi) - \rho(\xi)$ is AN with mean zero and asymptotic variance $\sigma_n^2 = O(n^{-1})$. Let $\mathcal{E}_n := \rho_n(t_j) - \rho_n(\xi)$.

The asymptotic normality of $S_n = \sum_{j=1}^J \tilde{\rho}_n(t_j)$ follows with desired asymptotic variance if $\mathcal{E}_n \rightarrow 0$ in probability as $n \rightarrow \infty$. Next show that this condition actually holds.

4° Show that $\mathcal{E} \xrightarrow{p} 0$ as $n \rightarrow \infty$ by analyzing the increment of the process $\rho_n(\cdot)$ between t_j and ξ . Note that $\rho_n(\cdot)$ is a right-continuous function with a jump at a time point t where one or more $N_i(t)$ changes value either from 0 to 1 or from 0 to NA. By (1), as n gets sufficiently large, the true quantile ξ estimated by t_j is between t_{j-1} and t_{j+1} . So between ξ and t_j , $\rho_n(\cdot)$ has a jump at t_j because one of the $N_i(\cdot)$'s changes from 0 to 1; without loss of generality, say it is N_1 . For simplicity assume that no additional missing value occurs between t_j and the corresponding true quantile ξ . A jump in $\rho_n(\cdot)$ due to missing values can be handled in the same fashion. The effective sample size at t_j and ξ is then $n\lambda_n(\xi)$ plus the number of individuals who failed before time t_j . For notational simplicity λ_n is dropped in the following calculations; this will not materially affect the derivation because of the bounded-away-from-zero assumption on λ_n . The change between $\bar{N}(\xi)$ and $\bar{N}(t_j)$ is $1/n$, and without loss of

generality, assume $N_1(t_j) = N_1(\xi) + 1 = 1$ and $N_i(\xi) = N_i(t_j)$ for $i = 2, \dots, n$. Now at the time point t_j the numerator of $\rho_n(t_j)$ is

$$\begin{aligned} & \sum_{i=2}^n (N_i(\xi) - \bar{N}(\xi) - \frac{1}{n})(X_i - \bar{X}) + (N_1(\xi) + 1 - \bar{N}(\xi) - \frac{1}{n})(X_1 - \bar{X}) \\ &= \sum_{i=2}^n (N_i(\xi) - \bar{N}(\xi))(X_i - \bar{X}) - \frac{1}{n} \sum_{i=2}^n (X_i - \bar{X}) + (N_1(\xi) - \bar{N}(\xi))(X_1 - \bar{X}) + \frac{n-1}{n}(X_1 - \bar{X}) \\ &= \sum_{i=1}^n (N_i(\xi) - \bar{N}(\xi))(X_i - \bar{X}) + (X_1 - \bar{X}) \end{aligned}$$

and in the denominator

$$\begin{aligned} & \sum_{i=1}^n (N_i(t_j) - \bar{N}(t_j))^2 \\ &= \sum_{i=2}^n (N_i(\xi) - \bar{N}(\xi) - \frac{1}{n})^2 + (N_1(\xi) + 1 - \bar{N}(\xi) - \frac{1}{n})^2 \\ &= \sum_{i=2}^n (N_i(\xi) - \bar{N}(\xi))^2 - \frac{2}{n} \sum_{i=2}^n (N_i(\xi) - \bar{N}(\xi)) + \frac{n-1}{n^2} + (N_1(\xi) - \bar{N}(\xi))^2 \\ &\quad + \frac{2(n-1)}{n} (N_1(\xi) - \bar{N}(\xi)) + \left(\frac{n-1}{n}\right)^2 \\ &= \sum_{i=1}^n (N_i(\xi) - \bar{N}(\xi))^2 + \frac{n-1}{n^2} + 2(N_1(\xi) - \bar{N}(\xi)) + \left(\frac{n-1}{n}\right)^2 \\ &= \sum_{i=1}^n (N_i(\xi) - \bar{N}(\xi))^2 - 2\bar{N}(\xi) + \frac{n-1}{n}, \end{aligned}$$

recognizing that $N_1(\xi) = 0$ in the last step. Let $S_{NX}(t)$ be the numerator of $\rho_n(t)$, let

$SS_N(t) = \sum_{i=1}^n (N_i(t) - \bar{N}(t))^2$, and let $SS_X = \sum_{i=1}^n (X_i - \bar{X})^2$. The the above calculation

shows that $S_{NX}(t_j) = S_{NX}(\xi) + (X_1 - \bar{X})$ and $SS_N(t_j) = SS_N(\xi) - 2\bar{N}(\xi) + \frac{n-1}{n}$. Therefore (dropping ' (ξ) ' in the notation in latter steps),

$$\begin{aligned}
\mathcal{E}_n &:= \tilde{\rho}_n(t_j) - \tilde{\rho}(\xi) \\
&= \frac{S_{NX}(\xi) + (X_1 - \bar{X})}{\sqrt{(SS_N(\xi) - 2\bar{N}(\xi) + (n-1)/n)SS_X}} - \frac{S_{NX}(\xi)}{\sqrt{SS_N(\xi)SS_X}} \\
&= \frac{\sqrt{SS_N}(S_{NX} + (X_1 - \bar{X})) - S_{NX}\sqrt{SS_N - 2\bar{N} + (n-1)/n}}{\sqrt{SS_NSS_X(SS_N - 2\bar{N} + (n-1)/n)}} \\
&= \frac{S_{NX}\left(\sqrt{SS_N} - \sqrt{SS_N - 2\bar{N} + (n-1)/n}\right) + \sqrt{SS_N}(X_1 - \bar{X})}{\sqrt{SS_NSS_X(SS_N - 2\bar{N} + (n-1)/n)}} \\
&= \frac{\frac{1}{n}S_{NX}\left(\sqrt{\frac{1}{n}SS_N} - \sqrt{\frac{1}{n}SS_N - 2\frac{1}{n}\bar{N} + (n-1)/n^2}\right) + \frac{1}{n}\sqrt{\frac{1}{n}SS_N}(X_1 - \bar{X})}{\sqrt{\frac{1}{n}SS_N\frac{1}{n}SS_X\left(\frac{1}{n}SS_N - 2\frac{1}{n}\bar{N} + (n-1)/n^2\right)}}.
\end{aligned}$$

By the Strong Law of Large Numbers, almost surely, as $n \rightarrow \infty$,

$$\frac{1}{n}S_{NX} \rightarrow \text{Cov}(N(\xi), X), \frac{1}{n}SS_N \rightarrow \text{Var}(N(\xi)) := \sigma^2, \bar{N} \rightarrow N(\xi) := \mu \text{ and } \frac{1}{n}SS_X \rightarrow \text{Var}(X) := \sigma_X^2,$$

and hence almost surely $\mathcal{E}_n/D_n \rightarrow 1$, where

$$D_n = \frac{\text{Cov}(N(\xi), X) \left(\sqrt{\sigma^2} - \sqrt{\sigma^2 - 2\mu/n + (n-1)/n^2} \right)}{\sqrt{\sigma^2\sigma_X^2(\sigma^2 - 2\mu/n + (n-1)/n^2)}} + \frac{n^{-1}\sigma(X_1 - \bar{X})}{\sqrt{\sigma^2\sigma_X^2(\sigma^2 - 2\mu/n + (n-1)/n^2)}}.$$

By the finite second moment condition on X , for arbitrarily small $\delta > 0$,

$$\Pr(|n^{-1}\sigma(X_1 - \bar{X})| > \delta) \leq \frac{\text{Var}(X_1 - \bar{X})\sigma^2}{n^2\delta^2} = O(n^{-2});$$

thus the the second term of D_n is $O_p(n^{-1})$. Let

$$B_n = \frac{\sqrt{\sigma^2} - \sqrt{\sigma^2 - 2\mu/n + (n-1)/n^2}}{\sqrt{\sigma^2\sigma_X^2(\sigma^2 - 2\mu/n + (n-1)/n^2)}}.$$

Then the numerator of $n^{1/2} B_n$ is

$$\Delta_n = (n\sigma^2)^{1/2} - (n\sigma^2 - 2\mu + (n-1)/n)^{1/2}.$$

Taylor series expansion of $f(x) = x^{1/2}$ at $x_0 = n\sigma^2$ with the Lagrange remainder and setting $x = n\sigma^2 - 2\mu + (n-1)/n$ give that the second term

$$\begin{aligned}
& (n\sigma^2 - 2\mu + (n-1)/n)^{1/2} \\
&= (n\sigma^2)^{1/2} + \frac{1}{2}(n\sigma^2)^{-1/2}(-2\mu + (n-1)/n) - \frac{1}{8}\eta_n^{-3/2}(-2\mu + (n-1)/n)^2 \\
&= (n\sigma^2)^{1/2} + \frac{1}{2}(n\sigma^2)^{-1/2}(-2\mu + (n-1)/n) + O(n^{-3/2}),
\end{aligned}$$

where $\eta_n = n\sigma^2 + \theta(-2\mu + (n-1)/n)$ and $0 < \theta < 1$. Therefore

$$\Delta_n = -\frac{1}{2}n^{-1/2}\sigma^{-1}(-2\mu + (n-1)/n) + O(n^{-3/2});$$

consequently $n^{1/2}B_n = O(n^{-1/2})$; thus $D_n = O_p(n^{-1})$, and consequently $\mathcal{E}_n = O_p(n^{-1})$, completing the proof.

- A novel procedure called Correlation Profile Test (CPT) is proposed.
- A novel permutation test, the hybrid permutation test, is proposed.
- CPT fits genomics applications better than common survival analysis methods.
- CPT is general and can have wide applications beyond genomics.

Figure 1. Cumulative incidence function estimates (Kalbfleisch and Prentice 2002, Chapter 8) of ALL relapse by protocol and treatment group

Figure 2.

Details of SNP A-1998030. The left column contains cumulative incidence estimates by genotype (AA solid line, AB dashed line, BB dotted line) in each stratum. The middle column contains scatter plot showing time to event or last follow up by genotype in each stratum; relapsed individuals are represented by bullets, others by gray circles. The right column contains stratum-specific sample correlation profile function [Eq. (2), Section 2] linearly interpolated between two consecutive event times (solid line) and the zero-reference line (dashed line) in each stratum.

Figure 3.

Details of SNP A-1996168. This figure is organized in the same way as Figure 2.

Figure 4.

Details of SNP A-2237842. This figure is organized in the same way as Figure 2.

Table 1

Case 1: Estimated significance level and power of CPT, CPT.Beta, Fine-Gray (FG) regression, Cox regression and Gray's test at nominal significance level 0.005; numbers in each pair correspond to sample size $n = 200, 700$ respectively.

	CPT	CPT.Beta	FG	Gray	Cox
Null	0.004, 0.005	0.007, 0.007	0.006, 0.005	0.007, 0.005	0.005, 0.005
$\theta = 0, \beta = 1.2$	0.44, 0.97	0.48, 0.98	0.28, 0.86	0.19, 0.77	0.21, 0.83
$\theta = 0.25, \beta = 0$	0.05, 0.13	0.06, 0.15	0.07, 0.23	0.05, 0.16	0.06, 0.23
$\theta = 0.25, \beta = 1.2$	0.80, 1.00	0.84, 1.00	0.63, 0.99	0.57, 0.99	0.62, 0.99
$\theta = 0.5, \beta = 0$	0.28, 0.85	0.33, 0.89	0.48, 0.96	0.37, 0.96	0.45, 0.96
$\theta = 0.5, \beta = 1.2$	0.97, 1.00	0.98, 1.00	0.94, 1.00	0.92, 1.00	0.94, 1.00

Table 2

Case 2: Estimated significance level and power of CPT, CPT.Beta, Fine-Gray (FG) regression, Cox regression and Gray's test at nominal significance level 0.005; numbers in each pair correspond to sample size $n = 700, 1500$ respectively.

	CPT	CPT.Beta	FG	Gray	Cox
Null	0.0128, 0.0092	0.0163, 0.0115	0.1601, 0.0242	0.0283, 0.0129	0.0132, 0.0084
$\theta = 0.5, \beta = 0$	0.007, 0.079	0.019, 0.11	0.055, 0.052	0.007, 0.055	0.000, 0.023
$\theta = 0.5, \beta = 1.2$	0.053, 0.64	0.12, 0.74	0.17, 0.39	0.009, 0.40	0.000, 0.14
$\theta = 0.8, \beta = 0$	0.45, 0.82	0.53, 0.87	0.32, 0.90	0.44, 0.94	0.30, 0.94
$\theta = 0.8, \beta = 1.2$	0.85, 1.00	0.91, 1.00	0.77, 0.99	0.78, 1.00	0.48, 0.99

Table 3

Case 3: Estimated significance level and power of CPT, CPT.Beta, Fine-Gray (FG) regression, Cox regression and Gray's test at nominal significance level 0.005; sample size $n = 1500$.

	CPT	CPT.Beta	FG	Gray	Cox
NULL	0.0058	0.0096	0.0095	0.0139	0.0062
$\theta = 0, \beta = 1.2$	0.11	0.14	0.044	0.15	0.089
$\theta = 0.8, \beta = 0$	0.071	0.078	0.14	0.17	0.12
$\theta = 0.8, \beta = 1.2$	0.38	0.43	0.25	0.43	0.22

Table 4

Case 4: Estimated significance level and power of CPT.Beta, Fine-Gray (FG) regression, Cox regression and Gray's test at nominal significance level 0.005; sample size $n = 700$.

	CPT.Beta	FG	Gray	Cox
NULL	0.0061	0.0045	0.0062	0.0044
$\beta = 0$	0.78	0.014	0.012	0.015
$\beta = -1.2$	0.84	0.013	0.022	0.018
$\beta = -2.4$	0.86	0.014	0.022	0.025

Hypothesis test P values: Upper half shows several top hits by stratified CPT not captured by Fine-Gray regression, lower half shows several opposite SNPs.

Table 5

Affy SNP ID	Stratified CPT	Unstratified CPT	Fine-Gray regression	Stratified Gray's test	Annotation
A-1847412	5.56E-05	0.021615	0.0089	0.2024	C9orf88 intron
A-2181937	0.000301	0.032986	0.0091	0.0576	TUSC1 downstream
A-2084786	0.000350	0.059541	0.0619	0.2577	CCRK downstream
A-2089381	0.000458	0.005826	0.0630	0.0565	C9orf82 downstream
A-1838092	0.000464	0.001871	0.0244	0.2473	MLLT3 downstream
A-1998030	0.001470	0.006650	0.0477	0.1044	ABL1 in gene
A-2066957	0.086142	0.0264	0.0000	0.0314	MPDZ downstream
A-1956981	0.092463	1.10E-05	1.10E-07	0.0215	HSDL2 in gene
A-1997856	0.738998	7.12E-07	8.64E-07	0.0232	DBC1 downstream
A-2157502	0.007280	7.86E-05	1.40E-06	0.0106	TLR4 downstream
A-1996168	0.131733	0.003425	0.003088	0.0319	MLLT3 downstream

Table 6

Hypothesis test P values of several biologically interesting SNPs captured by $S_{n;Beta}$ in comparison with the other tests.

Affy SNP ID	Stratified CPT ($S_{n;Beta}$)	Stratified CPT (S_g)	Stratified CPT S	Fine-Gray regression	Stratified Gray's test	Annotation
A-2237842	0.000124	1.76E-11	0.2259	0.0697	0.0150	MLLT3 downstream
A-2040281	0.000177	1.02E-12	0.0376	0.0029	0.0170	NOTCH1 in gene
A-2289668	0.000286	<1E-16	0.0156	0.0078	0.00019	GLIS3 in gene
A-2024771	8.63E-6	3.33E-16	2.44E-5	1.56E-5	2.52E-5	STOM in gene
A-1902372	1.40E-6	<1E-16	2.45E-6	1.04E-6	1.65E-5	GSN in gene
A-1996410	7.11E-7	<1E-16	0.00058	0.0078	0.0058	LRRN6C in gene