Published in final edited form as:
Ann Appl Stat. 2014 ; 8(2): 1232–1255.

GENE-LEVEL PHARMACOGENETIC ANALYSIS ON SURVIVAL OUTCOMES USING GENE-TRAIT SIMILARITY REGRESSION

Jung-Ying Tzeng[¶,**,*,†], **Wenbin Lu**[¶,*,‡], and **Fang-Chi Hsu**[||,§]

[¶]North Carolina State University

[||]Wake Forest University

[**]National Cheng-Kung University

Abstract

Gene/pathway-based methods are drawing significant attention due to their usefulness in detecting rare and common variants that affect disease susceptibility. The biological mechanism of drug responses indicates that a gene-based analysis has even greater potential in pharmacogenetics. Motivated by a study from the Vitamin Intervention for Stroke Prevention (VISP) trial, we develop a gene-trait similarity regression for survival analysis to assess the effect of a gene or pathway on time-to-event outcomes. The similarity regression has a general framework that covers a range of survival models, such as the proportional hazards model and the proportional odds model. The inference procedure developed under the proportional hazards model is robust against model misspecification. We derive the equivalence between the similarity survival regression and a random effects model, which further unifies the current variance-component based methods. We demonstrate the effectiveness of the proposed method through simulation studies. In addition, we apply the method to the VISP trial data to identify the genes that exhibit an association with the risk of a recurrent stroke. *TCN2* gene was found to be associated with the recurrent stroke risk in the low-dose arm. This gene may impact recurrent stroke risk in response to cofactor therapy.

Keywords and phrases

association study; gene/pathway; pharmacogenetics; similarity regression; survival data; proportional odds model; proportional hazard model

1. Introduction

Genetic variations play a significant role in drug responses. A gene that participates in a particular physiological mechanism might influence the response to a specific therapeutic

[*]Equal contribution

[†]Supported by National Institutes of Health grants R01 MH084022 and P01 CA142538
[‡]Supported by National Institutes of Health grants R01 CA140632 and P01 CA142538
[§]Supported by National Institutes of Health grant U01 HG005160.

Department of Statistics, North Carolina State University, Raleigh, NC, USA
jytzeng@ncsu.edu
lu@stat.ncsu.edu
Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA
fhsu@wakehealth.edu

agent that targets the mechanism. Identifying these influential genes may help to clarify if an individual might benefit from or be harmed by the therapy. Understanding the genetic diversity of drug responses can help to identify medications that maximize treatment effectiveness and minimize the risk of adverse effects for individuals. Such an understanding will also lead to improved risk stratification, prevention, and treatment strategies for human diseases. Pharmacogenetics studies how an adverse reaction or positive response to pharmaceutical treatment is affected by an individual's genetic makeup and has the potential to deliver both public health and economic benefits rapidly. With the recent advancements in high-throughput technologies, it is becoming common for pharmacogenetic researches to systematically investigate genetic markers across the genome. Nevertheless, appropriate and efficient analysis of the data remains a challenge.

Gene- or pathway-based analyses can assess pharmacogenetic effects more effectively than single-marker based analyses (Goldstein, Tate and Sisodiya, 2003; Goldstein, 2005). First, there often exist obvious candidate genes and pathways that metabolize the drug and carry variants that are relevant to the drug responses. Responses to therapies usually involve complex relationships between gene variants within the same molecular pathway or functional gene set. When applied to pharmacogenetic studies, gene- or pathway-based methods might identify multiple variants of subtle effects that are missed by single-marker based methods. Second, pharmacogenetic studies typically enroll only a moderate number of patients, which limits the power of the association detection. Gene-based analyses have been shown to yield higher power than standard single-marker and haplotype analyses. This type of analysis can particularly facilitate studies on rare-event drug responses, such as adverse reactions, where it could take many years to collect a sufficient number of samples to obtain adequate power for standard analyses. In gene-based analyses, the association signals are aggregated across variants, and the total number of tests is reduced; the amplification of the association signals and the alleviation of the multiple testing burdens result in improved power.

Our study was motivated by the need for a gene-based analysis of the time-to-event data of the Vitamin Intervention for Stroke Prevention (VISP) trial (Toole et al., 2004; Hsu et al., 2011). Our goal is to assess the association between the recurrent stroke risk and the 9 candidate genes involved in the homocystein (Hcy) metabolic pathway (see Data section for more details). In our preliminary analysis, we used the Cox proportional hazards (PH) model (Cox, 1972) to perform single-SNP screening on 69 SNPs across 9 genes from 969 individuals. There were no SNPs past the significance threshold after accounting for multiple testing. However, the top 6 hits, thresholding at unadjusted p-values <0.05, were concentrated in two genes. Specifically, 4 SNPs are from *TCN2* (i.e., rs1544468, rs731991, rs2301955, and rs2301957 have Wald's test p-values of 0.0065, 0.0072, 0.0346, and 0.0346, respectively) and 2 SNPs are from *CTH* (i.e., rs648743 and rs663465 each have a Wald's test p-value of 0.0115). The Kaplan-Meier curves of these 6 SNPs are shown in Figure 1 and indicate the potential for different risk patterns among different variants at these loci. The clustering within the two genes suggests that it would be more efficient to combine the individual signal strengths and model the joint effect of multiple loci in a gene.

We perform the gene-based analysis using a gene-trait similarity regression inspired by Haseman-Elston regression from linkage analysis (Elston et al., 2000; Haseman and Elston, 1972) and haplotype similarity tests for regional association (Beckmann et al., 2005; Qian and Thomas, 2001; Tzeng et al., 2003). First, we quantify the genetic and trait similarities for each pair of individuals. The genetic similarity is determined using identity by state (IBS) methods. The trait similarity is obtained from the covariance of the transformed survival time conditional on the covariates. We then regress the trait similarity on the genetic similarity and test the regression coefficient to detect the genetic association. There are several gene-based approaches for censored time-to-event phenotypes in the literature, including Goeman et al. (2005) and Lin and colleagues (Cai, Tonini and Lin, 2011; Lin et al., 2011). In these approaches, the multimarker effects were modeled under the Cox PH model using linear random effects (Goeman et al., 2005) or a nonpara-metric function induced by a kernel machine (Cai, Tonini and Lin, 2011; Lin et al., 2011). The global effect of a gene was detected by testing for the corresponding genetic variance component. These approaches were found to be superior in identifying pathways or genes that are associated with survival.

For many years, similarity-based methods have been successfully used to evaluate gene-based associations in quantitative and binary traits (Beck-mann et al., 2005; Lin and Schaid, 2009; Qian and Thomas, 2001; Tzeng et al., 2003; Wessel and Schork, 2006). Our work makes such approaches available for survival phenotypes. In addition, our similarity regression covers a variety of risk models, including the commonly used PH model and the proportional odds (PO) model. Furthermore, we show that the coefficient of the similarity regression obtained for survival phenotypes can be re-expressed as a variance component of a certain working random effects model. Such results facilitate the derivation of the test statistic and unify the similarity model and previous variance-component methods (Goeman et al., 2005; Cai, Tonini and Lin, 2011; Lin et al., 2011). Specifically, under the Cox PH model, our test statistic is equivalent to the test statistic defined by a kernel machine approach (Lin et al., 2011). We also show that the test statistic can be robust to model misspecification. Specifically, the proposed test gives the correct type I error even if the true risk model is misspecified. However, the correct specification of the true risk model generally leads to a test with better power. Finally, we demonstrate the utility of the similarity regression by identifying the important *TCN2* gene in the VISP study. The significance of *TCN2* to stroke risk has been reported by other association studies (Giusti et al., 2010; Low et al., 2011) and has been supported by molecular biology evidence (Afman et al., 2003; von Castel-Dunwoody et al., 2005). Our findings further suggest potential interactions between *TCN2* and B12 supplementation. This new information furthers the possibility that *TCN2* could be utilized to predict recurrent strokes, identify at-risk individuals and identify therapeutic targets for ischemic stroke.

2. Data

The VISP study was a prospective, double-blind, randomized clinical trial (Toole et al., 2004). The trial was designed to study if high doses of folic acid, vitamin B6, and vitamin B12 reduce the risk of a recurrent stroke as compared to low doses of these vitamins. The trial enrolled patients who were 35 or older, had a non-disabling cerebral infarction within

120 days of randomization, and had Hcy levels in the top quartile of the U.S. population. Subjects were randomly assigned to receive daily doses of either a high-dose formulation (containing 25 mg vitamin B_6, 0.4 mg vitamin B_{12}, and 2.5 mg folic acid) or a low-dose formulation (containing 200 μg vitamin B_6, 6 μg vitamin B_{12}, and 20 μg folic acid). Patient recruitment began in August 1997 and was completed in December 2001. A total of 3680 participants were enrolled at 56 clinical sites across the United States, Canada, and Scotland. The patients were followed for a maximum of two years, and the average follow-up duration was 1.7 years. In the VISP genetic study, 2206 participants provided informed consent and blood samples. The SNP genotypes of 9 genes related to the enzymes and cofactors in the Hcy metabolic pathway were collected: *BHMT1, BHMT2, CBS, CTH, MTHFR, MTR, MTRR, TCN1*, and *TCN2* (Hsu et al., 2011). In a previous study, Hsu et al. (2011) conducted single-SNP analyses on targeted loci (e.g., Hcy-associated variants) to examine the genetic association with the recurrent stroke risk. In the low-dose arm, the authors found that *TCN2* SNP rs731991 under a recessive mode was associated with the risk of a recurrent stroke with an unadjusted logrank test p-value of 0.009. The associations for the remaining SNPs within the 9 genes in the low dose arm were not studied. We extend this previous analysis to all 9 genes using a gene-based approach. After quality control screening of the data (e.g., removing loci with >99% missing proportion or HardyWeinberg disequilibrium under additive mode and removing individuals with missing genotypes), the analysis included 969 individuals in the low-dose arm with 69 recessively-coded SNPs.

3. Gene-Trait Similarity Regression for Survival Traits

3.1. The Model

For individual i ($i = 1, 2, \ldots, n$), let T_i denote the survival time of interest and C_i denote the censoring time. We observe $\tilde{T}_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = I(T_i \le C_i)$. In addition, let X_i denote the $K \times 1$ vector of covariates and G_{mi} denote the allele count vector of marker m for person i, where the length of G_{mi}, ℓ_m, is the number of distinct alleles at marker m, $m = 1, 2, \ldots, M$. For example, for a tri-allelic locus m, $\mathrm{G}_{mi} = (1, 0, 1)^T$ if person i has genotype 'A_1A_3' and $(0, 2, 0)^T$ if person i has genotype 'A_2A_2'.

For each pair of individuals i and j, we measure the genetic similarity, S_{ij}, of the targeted gene and the trait similarity, Z_{ij}. The genetic similarity is quantified using the weighted IBS sum across the M markers in the gene, i.e., $S_{ij}=\sum_{m=1}^{M} w_m S_{ij}^{m}$, where $S_{ij}^{m}=2$ if $|\mathrm{G}_{m,i} - \mathrm{G}_{m,j}|$ is a zero vector, $S_{ij}^{m}=1$ if $|\mathrm{G}_{m,i} - \mathrm{G}_{m,j}|$ contains exactly two 1s (and if $\ell_m > 2$, the remaining entries are 0.), and $S_{ij}^{m}=0$ otherwise. The weights, w_m, are specified to up-weight or down-weight a variant based on certain features. Examples include weights that are based on allele frequencies, the degree of evolutionary conservation, or the functionality of the variations (Wessel and Schork, 2006; Schaid, 2010a; Schaid, 2010b; Price et al., 2010). We can use the minor allele frequency of marker m, denoted as q_m, to up-weight similarities that are contributed by rare variants. Specifically, one can set a moderate weight, such as $w_m=q_m^{-3/4}$ (Pongpanich et al., 2012) or $w_m=q_m^{-1}$ (Tzeng et al., 2011), to promote similarity attributed by rare alleles, or use a more extreme weight, such as $w_m = (1 - q_m)^{24}$ (Wu et al., 2011), to target rare variants only.

The trait similarity, Z_{ij}, is quantified as follows. First, we define $Y_i = H(T_i)$, where $H(\cdot)$ is an (unspecified) monotonic increasing transformation function, such as the logarithm transformation $Y_i = \log(T_i)$. Assume that the conditional mean of Y_i given the covariates and genes is $E\,(Y_i|X_i, g_i)=\theta+X_i^T\gamma+g_i$, where θ is the intercept, g_i is the multi-locus genetic effect of person i, and γ is the K-dimensional covariate effect. Further, define $\mu_i^0=\theta+X_i^T\gamma$. The trait similarity is defined as the product of the paired residuals adjusting for the covariate effects, i.e., $Z_{ij}=(Y_i-\mu_i^0)\,(Y_j-\mu_j^0)$. The expected value of the trait similarity is the covariance between the transformed survival times of subjects i and j.

The gene-trait similarity regression has the form

$$E\,(Z_{ij}|X_i, X_j)=b\times S_{ij},\ \ i\neq j. \quad (3.1)$$

Just as in Tzeng et al. (2009) and Tzeng et al. (2011), the regression has a zero intercept and does not have the covariate term X_iX_j because the baseline and covariate effects have been adjusted when defining Z_{ij}. This argument will become more obvious from the viewpoint of variance components in the following subsection. Under model (3.1), the overall association of a gene can be evaluated by testing the null hypothesis: $b = 0$.

3.2. Score Test for the Gene-Level Effect

We derive the score test statistic based on the equivalence between the similarity regression and a mixed model. This equivalence is demonstrated as follows. Consider a working mixed model for the transformed survival time:

$$Y_i=H(T_i)=X_i^T\gamma+g_i+\theta+\varepsilon_i, \quad (3.2)$$

where $(g_1, \ldots, g_n)^T \sim N\,(0,\ \tau S)$ with $S=\{S_{ij}\}_{i,j=1}^n$, i.e., the covariance between g_i and g_j depends on the genetic similarity between subjects i and j, and $\varepsilon_i^* \equiv \alpha+\varepsilon_i$, $i = 1, \ldots, n$ are independently and identically distributed with a known distribution that is independent of X_i and g_i. Given X_i and g_i, model (3.2) specifies a general class of linear transformation models (Cheng, Wei and Ying, 1995), which contains many popular survival models as special cases. For example, when ε_i^* follows the standard extreme value distribution, the linear transformation model becomes the PH model (Cox, 1972). When ε_i^* follows the standard logistic distribution, the linear transformation model becomes the PO model (Bennett, 1983).

Under (3.2), the conditional expectation of the trait similarity between individuals i and j ($i \neq j$) is

$$\begin{aligned} E(Z_{ij}|X_i, X_j)=&\text{cov}\,(Y_i, Y_j|X_i, X_j)=\text{cov}(g_i+\varepsilon_i^*, g_j+\varepsilon_i^*)\\ =&\text{cov}(g_i, g_j)\\ =&\tau\times S_{ij}. \end{aligned}$$

Therefore, we have $b = \tau$, i.e., the regression coefficient in the similarity regression (3.1) is the genetic variance component in the mixed model (3.2). This motivates us to develop a score test for the variance component in the working model. As shown in the Appendix, the score test statistics for $\tau = 0$ can be written as

$$Q_n=\frac{1}{n}(\hat{r}_1,\ldots,\hat{r}_n)S(\hat{r}_1,\ldots,\hat{r}_n)^T,$$

where

$$\hat{r}_i=\int_0^\infty\hat{\omega}_i(t)dM_i(t;\hat{\gamma},\hat{H})=\delta_i\frac{\dot{\lambda}\{\hat{H}(\tilde{T}_i)-\hat{\gamma}^TX_i\}}{\lambda\{\hat{H}(\tilde{T}_i)-\hat{\gamma}^TX_i\}}-\lambda\{\hat{H}(\tilde{T}_i)-\hat{\gamma}^TX_i\},$$

$$M_i(t;\gamma,H)=\delta_iI(\tilde{T}_i\le t)-\int_0^tI(\tilde{T}_i\ge s)d\Lambda\{H(s)-\gamma^TX_i\},$$

$\hat{\omega}_i(t) = \dot{\lambda}\{\hat{H}(t) - \hat{\gamma}^T X_i\}/\lambda\{\hat{H}(t) - \hat{\gamma}^T X_i\}$, and S is as defined after equation (3.2). Here, $\lambda(\cdot)$ and $\Lambda(\cdot)$ are the hazard and cumulative hazard functions of ε_i^*, respectively, $\dot{\lambda}(\cdot)$ is the first derivative of $\lambda(\cdot)$, and $\hat{\gamma}$ and $\hat{H}(\cdot)$ are the estimates of γ and $H(\cdot)$, respectively, in model (3.2) under the null hypothesis: $\tau = 0$. For example, if the PH model is imposed, i.e., $\lambda(u) = \dot{\lambda}(u) = e^u$, the estimators $\hat{\gamma}$ and $\hat{\Gamma}(\cdot) \equiv e^{\hat{H}(\cdot)}$ can be taken as the maximum partial likelihood estimator and Breslow's estimator, respectively. Under this case, $\hat{\omega}_i(t) \equiv 1$ and $\hat{r}_i = \delta_i - \hat{\Gamma}(\tilde{T}_i) \exp(-\hat{\gamma}^T X_i)$, i.e., the martingale residual for the null model. If the PO model is used, i.e., $\lambda(u) = e^u/(1 + e^u)$ and $\dot{\lambda}(u) = e^u/(1 + e^u)^2$, we have $\hat{\omega}_i(t) = 1/[1 + \exp\{\hat{H}(t) - \hat{\gamma}^T X_i\}]$. In general, γ and $H(\cdot)$ can be estimated using the martingale-based estimating equations (Chen, Jing and Ying, 2002) or the nonparametric maximum likelihood estimation method (Zeng and Lin, 2006) for the semiparametric linear transformation model. In the Appendix, we show that under the null hypothesis the test statistic, Q_n, asymptotically follows a weighted χ^2 distribution where the weights can be estimated consistently. The p-values can then be calculated numerically using a resampling method or moment-matching approximations (Pearson, 1959; Duchesne and Lafaye, 2010).

We note that the proposed score test can be robust to the misspecification of the true survival model. As an illustration, consider the test derived under the PH model. Based on the results for the robust inference of the PH model (Lin and Wei, 1989), when the PH model is misspecified, the maximum partial likelihood estimator, $\hat{\gamma}$, and Breslow's estimator, $\hat{\Gamma}(\cdot)$, do not converge to the true parameters but converge to some deterministic values γ^* and $\Gamma^*(\cdot)$ under certain regularity conditions. The corresponding $r_i^*=\delta_i-\Gamma^*(\tilde{T}_i)\exp(-X_i^T\gamma^*)$ is not a martingale residual but has a mean of 0 under the null hypothesis. Therefore, it can be shown that Q_n converges in distribution to a weighted χ^2 distribution under the null hypothesis even with model misspecification. As shown in our simulation studies, the proposed score test derived under the PH model still gives correct type-I error when the true model is the PO model. However, the power of the test depends on the assumed working model. In general, the test derived under the true risk model may have better power.

4. Results of the Analysis of the VISP Trial Data

We now return to the VISP trial to evaluate the association between the recurrent stroke risk and the 9 candidate genes studied in Hsu et al. (2011). In our analysis, we conducted a gene-based screening on the 9 genes using the low-dose samples. After removing loci with >1% of missingness and subjects with missing genotypes, there were 969 individuals with 69 polymorphic SNPs under recessive coding. Of the 969 individuals, 86 experienced a recurrent stroke (i.e., 91.1% censoring). We used the proposed similarity regression (referred to as SimReg) with inverse allele frequency weights $w_m = q_m^{-3/4}$, i.e., the weight recommended in Pongpanich et al. (2012) when analyzing a mixture of common and rare variants. We calculated the p-values of the SimReg statistics using the resampling method. Specifically, we computed the nonzero eigenvalues, ξ°_1, ···, ξ°_d, of Σ° as defined in the Appendix. We generated 10^4 sets of $\left(\chi^2_{1,1}, \cdots, \chi^2_{1,d}\right)$. Each set consisted of *d* independently and identically distributed χ^2_1 random variables. For each set, we calculated the value $\sum_{k=1}^{d} \hat{\xi}_k \chi^2_{1,k}$, and the 10^4 values formed an empirical null distribution of the SimReg statistics. The SimReg p-value was the proportion of the generated null statistics that were greater than the observed statistic. We performed SimReg analyses under the PH model (referred to as SimReg-PH) and the PO model (referred to as SimReg-PO). The performances of the SimReg methods were benchmarked against three approaches: (a) the single SNP minimum p-value method using the Cox PH model (referred to as minP), (b) the multi-SNP method using the global test for survival under the PH model (Goeman et al., 2005) as implemented in the R-package globaltest (referred to as Global), and (c) the multi-SNP method using kernel machine (Lin et al., 2011) as implemented in the R-package KMTest.surv (referred to as KM) with 10^4 resamplings. Although the SimReg-PH test statistic is identical to the KM test statistic, the results may be slightly different due to the different resampling methods adopted to obtain the p-values. Specifically, in the KM method, the score statistic was perturbed by multiplying i.i.d. normal random variables to achieve the same limiting distribution of the test statistic. In SimReg-PH, the weights in the limiting weighted χ^2 distribution, i.e., ξ_1, ···, ξ_d, were consistently estimated and then the samples were directly generated from the estimated weighted χ^2 distribution based on a large number of i.i.d. χ^2_1 random variables. The different resampling approaches also lead to different computational burden. For example, using a 3.6 GHz Xeon Processor with 60GB RAM with 10^4 resamplings, the system run-time of SimReg-PH was < 1/6 of KM in the VISP analysis, and the time difference became greater with a larger number of resamplings. For the minP method, we fitted the standard PH model to each SNP in a gene, took the smallest p-value, and calculated the adjusted p-value of a gene to correct for the multiple SNPs using 1–(1–minimum raw p-value)$^{K_{eff}}$. The effective number of independent tests, K_{eff}, was estimated using the method of Moskvina and Schmidt (2008) and accounts for the correlations in recessive coding of loci. As studied by Hsu et al. (2011), all analyses were considered under the recessive mode and were adjusted for age, sex, and race.

The p-values for each of the methods are shown in Table 1, and the p-values are compared with the Bonferroni threshold adjusted for the 9 gene analyses, i.e., 0.05/9 = 0.0056. SimReg-PH detected a significant association between the recurrent stroke risk and *TCN2* (i.e., p-value = 0.0040), which strengthens the observation of differential survival between different variants from the single SNP analysis. Gene *CTH* had the second smallest p-value (0.0073), which did not pass the Bonferroni threshold but was near the cutoff. These results also agree with the findings in the single SNP analysis. The results of SimReg-PO are similar to those of SimReg-PH except that the p-values are slightly larger, i.e., p-value = 0.0052 for *TCN2* and 0.0075 for *CTH*. On the other hand, the KM, Global and minP methods did not yield any significant findings. However, the smallest p-values of the 9 genes were obtained for *TCN2* (i.e., p-values of *TCN2* are 0.0075 for KM, 0.0457 for Global and 0.0704 for minP). As expected, the results of KM were very similar to those of SimReg-PH, except that the p-value of *TCN2* was slightly above the 0.0056 threshold. The smallest p-values for the minP method were from the *TCN2* SNP rs731991 with a raw Wald's test p-value of 0.0065. However, neither the raw minimum p-value (0.0065) nor the adjusted p-value (0.0704) survived the significance threshold corrected for multiple testing (0.0056). All methods also had *CTH* as the gene with the second smallest p-value (i.e., p-values 0.0078 for KM, 0.0518 for Global, and 0.00918 for minP).

Next, we assessed the prediction performance of Cox PH models built with and without *TCN2* using the procedure described in Li and Luan (2005). Specifically, we randomly divided the samples into a training set (*n* = 646) and a testing set (*n* = 323). Based on the training set, we fitted the Cox PH regression with two models: Model 1 included only the baseline covariates (age, sex and race), i.e., no genetic information, and Model 2 included the baseline covariates plus the top 7 principal components (PCs) of *TCN2* SNPs that explained 95% of the variations. The PCs were used instead of the 15 original genotypes because of the high linkage disequilibrium among them. Based on the fits of the PH model from the training set, we obtained the risk scores of every subject under each model and computed the medians of the risk scores. We also computed the risk scores for the testing set using the estimated coefficients from the training set. Next, we divided the subjects into 2 risk groups: high-risk and low-risk. Individuals with a risk score higher/lower than the median risk scores obtained from the training data comprised the high-risk/low-risk group. Finally, we plotted the Kaplan-Meier curves for the 2 risk groups in the training data and the two risk groups in the testing data separately and obtained the p-values of the corresponding log-rank tests. The results are given in Figure 2. As expected, the p-values for both models under the training set were very significant. However, only Model 2 is significant (p-value is 0.048) under the testing set. This result implies that *TCN2* gives a more accurate prediction of the risk for recurrent stroke.

Vitamin supplements have been identified as a potential treatment for vascular diseases. The beneficial effects of vitamin supplements on stroke recurrence are not yet fully understood. In VISP, vitamin supplements did not show an effect on the recurrent stroke risk during the 2 years of followup. However, we found that genetic variants, such as SNPs in *TCN2*, were associated with the recurrent stroke risk in the low-dose arm. This finding is consistent with the literature. *TCN2* was previously found to be associated with ischemic stroke risk (Low et al., 2011; Giusti et al., 2010) and premature ischemic stroke risk (Giusti et al. 2010). It has been reported that *TCN2* interferes with the intracellular availability of vitamin B12 (von Castel-Dunwoody et al. 2005). The gene is associated with plasma homocysteine levels and affects the proportion of vitamin B12 bound to transcobalamin (Afman et al., 2003). It is suspected that SNPs on the genes coding for enzymes involved in the methionine metabolism have been suspected to beare associated with hyperhomocysteinaemia, which can result in occurrence of stroke (Giusti et al., 2010). Our significant findings in the low-dose arm suggest that there may be an interaction between *TCN2* and B12 supplementation; a finding that warrants further studies. The findings lead to a hypothesis that there may be one specific combination of genotypes of *TCN2* that is more efficient at transporting B12 and thus impacts the effectiveness of cofactor therapy on recurrent stroke risk. A functional study is being planned to localize possibly independent regions of association and determine their function.

Besides *TCN2*, *CTH* is marginally associated with recurrent stroke risk in the low-dose arm. It encodes cystathionine gamma-lyase, which is an enzyme that converts cystathionine to cysteine in the trans-sulfuration pathway (Wang et al., 2004). It may be a determinant of plasma Hcy concentrations, which may increase the risk of recurrent stroke because of arterial disease. It is worth further study as well.

5. Simulation Studies

We performed simulations to assess the validity and effectiveness of the proposed SimReg methods based on the 15 SNPs in *TCN2* of the 969 VISP low-arm samples. The rarest minor allele frequency (MAF) is approximately 3%. We generated genotypes of *n* individuals by randomly sampling with replacement from the 15-SNP genotypes of the 969 samples. For individual *i*, we generated the covariate, *Xi*, from *N* (0, 1). We generated the survival time, T_i, based on the genetic and covariate information under two models: the PH model and the PO model. Specifically, for the PH model, $\log(T_i) = -(X_i + \sum_\ell \gamma_\ell \mathcal{G}_{\ell i}) + \varepsilon_i^*$, where ε_i^* follows the standard extreme value distribution; for the PO model, $\log\{\exp(T_i) - 1\} = -(X_i + \sum_\ell \gamma_\ell \mathcal{G}_{\ell i}) + \varepsilon_i^*$, where ε_i^* follows the standard logistic distribution. The value of $\mathcal{G}_{\ell i}$ is determined by the genotypes at the causal locus $\ell$and the mode of inheritance. For example, if *A* is the causal allele at locus ℓ, then $\mathcal{G}_{\ell i}$ = 2, 1, and 0 for genotypes *AA*, *Aa*, and *aa*, respectively, under an additive mode. Under a dominant mode, $\mathcal{G}_{\cdot i}$ = 1, 1, and 0, respectively. Under a recessive mode, $\mathcal{G}_{\cdot i}$ = 1, 0, and 0, respectively. For type I error analysis, no SNPs were set to be causal, i.e., $\gamma_\ell$was set to be 0 for all ℓ. For power analysis, we selected 3 SNPs with different MAFs and LD patterns as causal loci from the 15 SNPs in *TCN2* and referred to them as SNP R, SNP U, and SNP C. The MAFs are 0.036 (rare) for SNP R, 0.132 (uncommon) for SNP U, and 0.419 (common) for SNP C. The average R^2s between a causal locus and the remaining loci are 0.002 (low) for SNP R, 0.003 (low) for SNP U, and 0.216 (high) for SNP C. The specific values of γ_ℓs are given in Table 2 for each scenario under different inheritance modes and censoring rates. The values were set to consider 1, 2, and 3 causal loci in the gene and to consider causal loci that have either the same or different effect sizes (e.g., rarer variants with larger effect sizes). All of the power scenarios assumed linear additive effects of the causal loci, which favors the linear random effects model (e.g., Global) and can be used to examine the utility of using the non-linear IBS function to capture the multi-marker effects.

We generated the censoring time, C_i, from Unif(0, c), where c is uniquely chosen for each of 3 censoring rates: 15%, 40%, and 90%. Specifically, we set c = 6.7, 2.0 and 0.2 for censoring rates of 15%, 40%, and 90%, respectively. The sample sizes, n, were 500 for the 15% and 40% censoring rates under the additive and dominant modes. For the 90% censoring rate under the additive and dominant modes and all censoring rates under the recessive mode, n = 1000. Each scenario was analyzed using SimReg (PH and/or PO), minP, Global and KM. SimReg-PH and KM have identical test statistics but used different resampling approaches to obtain p-values. Because both resampling approaches are asymptotically equivalent, we expect minor differences in finite sample performance between SimReg-PH and KM. In all analyses, the causal loci were excluded.

5.1. Results of Type I Error Analyses

We first examined the performance of the proposed SimReg-PH model. Table 3 displays the type I error rates of different methods when the survival times were generated from the PH model. The results were based on 10^5 replications except that KM was based on 5×10^4 replications due to computational cost. In each replication, the p-values of SimReg-PH were obtained from 5×10^5 resamplings. We report the type I error rates evaluated at the nominal levels of 5×10^{-2}, 5×10^{-3}, and 5×10^{-4}. The type I error rates obtained by SimReg-PH remained around the nominal levels. However, the deviations were larger for $\alpha = 5 \times 10^{-4}$, mainly due to fewer resampled statistics observed on the extreme tail. In particular, for the low censoring proportion (i.e., 15%), the type I error rates were slightly inflated, while for the high censoring proportion (i.e., 90%), the type I error rates became a little conservative when $\alpha = 5\times 10^{-4}$. Nevertheless, the overall results suggest that the SimReg-PH test maintained an appropriate size, which confirms the validity of the derived null distribution of the test statistic, Q_n. As expected, the type I error rates obtained by KM were very similar to SimReg-PH. The type I error rates obtained by the Global test are overly conservative; similar behavior has been reported in the literature (Zhong and Chen, 2011). The minP method had correct type I error rates under additive and dominant modes. However, the method had inflated type I error rates under the recessive mode, and the inflation was more severe with smaller α. Under the recessive mode, the Bonferroni corrected p-values obtained by replacing K_{eff} with the total number of SNPs also yielded inflated type I error rates. Specifically, the empirical type I error rates for 15%, 40%, and 90% censoring are (0.0627, 0.0687, 0.0608) for $\alpha = 5 \times 10^{-2}$, (0.0145, 0.0152, 0.0163) for $\alpha = 5\times10^{-3}$, and (0.0042, 0.0041, 0.0053) for $\alpha = 5\times10^{-4}$, respectively. The anti-conservation appears to be somewhat related to the rare recessive loci; when the rare loci are excluded from the analysis, the empirical type I error rate became closer to the nominal level (data not shown). However, such an exclusion strategy might give uninformative results because the relevant signals were excluded from the analysis.

Table 4 displays the type I error rates when the survival times were generated from the PO model. Both SimReg-PH and SimReg-PO were implemented. The results were based on 5000 replications, and the type I error rates were calculated at the nominal level of 0.05. The type I error rates for both SimReg-PO and SimReg-PH were close to 0.05 independent of the inheritance mode and the censoring proportions. These results show the validity of the SimReg-PO tests and the robustness of the SimReg-PH tests. Though not performed, KM is expected to have the same robustness as SimReg-PH. As seen previously, the Global test had conservative type I error rates, but the magnitude of conservation is less than that seen in Table 3. The minP method yielded slightly inflated type I error rates for the additive and dominant modes with low censoring proportion (i.e., 15%). As before, it yielded inflated type I error rates for the recessive mode.

5.2. Results of Power Analyses

The power analyses were performed using the settings specified in Table 2. The results were based on 100 replications under each scenario. Figure 3 shows the power when the survival times were generated from the PH model. We first consider the additive mode. When one large-effect causal locus has low MAF and low LD with the other markers (e.g., Scenarios 1, 5 and 11), minP tends to have the highest power independent of the censoring proportion. The good performance of minP is not unexpected in these scenarios because the overall association of the gene was driven by a single large-effect locus, for which the majority of the other SNPs did not carry much information. As a result, there is no power gain when borrowing information from other SNPs, which is what SimReg-PH does. However, the power gain of minP over other methods generally diminishes as the number of causal loci increases (e.g., Scenarios 5 to 10). In scenarios where the marker set is not dominated by a single causal locus of low MAF and low LD, SimReg-PH showed comparable or higher power. As expected, KM had near identical power as SimReg-PH in all scenarios. In most scenarios, the Global test produces the least amount of power largely due to the over-conservative test size. The overall performance under the dominant mode has a similar trend to that of the additive mode. However, the power of SimReg-PH is comparable to or better power than the power of minP in more cases under the dominant mode than under the additive mode. For the recessive mode, SimReg-PH appears to have better power than the Global test. We did not perform a power analysis for minP because the type I error rates were inflated.

Figure 4 shows the power performance when the survival times were generated from the PO model. The power obtained using SimReg-PO is similar to or better than the power obtained using SimReg-PH, indicating that efficiency is gained when the correct model is used. The power gain of SimReg-PO is more substantial when the censoring proportion is low to medium, and the power is comparable to or better power than the minP power in some of those difficult cases, such as Scenario 1.

6. Discussion

In this work, we extended the similarity regression (SimReg) approaches, which have shown effectiveness in modeling marker-set effects on binary and continuous outcomes, to survival models to facilitate the assessment of gene or pathway effects on drug responses. The genetic effect is evaluated by assessing the association between the IBS status of a pair of individuals and the covariance of their survival times. We derived the equivalence between the similarity survival regression and a random effects model. The equivalence facilitates the derivation of the score test statistics and unifies the current variance-component based methods. Specifically, the KM approach (Lin et al., 2011) is equivalent to SimReg-PH when the same kernel function is used to quantify the genetic similarity, S_{ij}. The Global test (Goeman et al. 2005) can be viewed as a special case of SimReg-PH with $S_{ij}=\sum_m \mathrm{G}_{m,i}^T \mathrm{G}_{m,j}$ (i.e., the linear kernel). However, the results of Global and SimReg-PH with linear kernels may be different because different approaches were used to derive the asymptotic distributions of the test statistics. Compared to these existing gene-based approaches, our proposed method has the generality to incorporate a variety of risk models in the class of linear transformation models, and we explicitly constructed the SimReg tests under the PH model and the PO model in this work. We also proposed a resampling approach to obtain the p-values that improves the computational efficiency. Finally, we showed that the derived inference procedure is robust against the misspecification of the risk model, which is an attractive feature because the underlying risk model is often unknown. Through simulations, we showed that the power of the SimReg method is comparable to or higher than the power of the minP and Global methods across various scenarios. We also verified that the SimReg-PH test statistics remain valid even if the risk model is misspecified.

In the data application on the VISP study, we illustrated how SimReg can be used to search for genes or pathways that are associated with a time-to-event outcome and confirmed previous findings using this gene-based approach with statistical significance. Although we focused our method development and demonstrated its utility based on pharmacogenetics studies, the proposed method is applicable to other genetic clinical researches or observational studies with time-to-event outcomes. For a pharmacogenetic study with sample sizes 1000, such as the VISP trial, 1000 runs of the SimReg analyses took ≤1 hour to complete on an Intel Xeon 3.33 GHz machine with 12 Gb RAM using one processing core. We expect a gene-based whole genome analysis on ~20K genes should be completed in a day using a comparable computing facility. We implemented the proposed methods in R and made it available at the authors' websites. We are incorporating the software into the SimReg R package.

Motivated by the data application where the risk variant acted recessively, we further investigated the behavior of the gene-based approaches under different modes of inheritance. We found that all of the studied gene-based methods performed appropriately under the additive and dominant modes. However, caution should be used when performing the minimal p-value approaches under the recessive mode; the minimal p-value approaches had severely inflated type I error rates. The inflation might be related to the extremely rare recessive loci. However, excluding those rare recessive loci is suboptimal because the important signals can be artificially removed and lead to power loss. In contrast, the global test and the similarity regression are not vulnerable to such a situation and appear to be more suitable options given their reasonable performance under the recessive mode.

This work focused on assessing the genetic main effect on drug response in an effort to understand how individual variation affects drug efficacy and toxicity. In pharmacogenetics and personalized medicine, one major topic is to study if the genetic effects are modified by treatments and how the effects differ across treatment options. As observed in the VISP genetic studies, the effect of *TCN2* on recurrent stroke risk is restricted to the low-dose treatment. An analysis stratified by treatments allows for the evaluation of such heterogeneous effects between different treatment groups, but its efficiency can be further improved by incorporating the gene-treatment interaction in the regression model. Such an extension is not straightforward because the calculation of the score test requires the variance component estimates for the genetic main effect under a mixed effects survival model. We are developing further extensions of SimReg to incorporate interaction effects.

Acknowledgments

The authors thank Dr. Michèle M. Sale and Dr. Bradford B. Worrall for providing the data. They thank Drs. Chen-Hsin Chen for his insightful discussion and helpful suggestions to improve the work. They also thank Dr. Shannon Holloway for her constructive input and proofreading to improve the manuscript and assistance in creating the software package.

References

Afman LA, Lievers KJ, Kluijtmans LA, Trijbels FJ, Blom HJ. Gene-gene interaction between the cystathionine beta-synthase 31 base pair variable number of tandem repeats and the methylenetetrahydrofolate reductase 677C>T polymorphism on homocysteine levels and risk for neural tube defects. Molecular Genetics and Metabolism. 2003; 78:211–215. [PubMed: 12649066]

Beckmann L, Thomas D, Fischer C, Chang-Claude J. Haplotype sharing analysis using mantel statistics. Human Heredity. 2005; 59:6778.

Bennett S. Analysis of survival data by the proportional odds model. Statistics in Medicine. 1983; 2:273–277. [PubMed: 6648142]

Cai T, Tonini G, Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics. 2011; 67:975–986. [PubMed: 21281275]

Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. Biometrika. 2002; 89:659–668.

Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. Biometrika. 1995; 82:835–845.

Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B. 1972; 34:187–220.

Duchesne P, Lafaye De Micheaux P. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. Computational Statistics & Data Analysis. 2010; 54:858–862.

Elston RC, Buxbaum S, Jacobs KB, Olson JM. Haseman and Elston revisited. Genetic Epidemiology. 2000; 19:117.

Giusti B, Saracinim C, Bolli P, Magi A, Martinelli I, Peyvandi F, Rasura M, Volpe M, Lotta LA, Rubattu S, Mannucci PM, Abbate R. Early-onset ischaemic stroke: analysis of 58 polymorphisms in 17 genes involved in methionine metabolism. Thrombosis and Haemostasis. 2010; 104:231–242. [PubMed: 20458436]

Goeman JJ, Oosting J, Cleton-Jansen AM, et al. Testing association of a pathway with survival using gene expression data. Bioinformatics. 2005; 21:1950–1957. [PubMed: 15657105]

Goldstein DB. The genetics of human drug response. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005; 360:1571–1572.

Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. Nature Reviews Genetics. 2003; 4:937–947.

Haseman J, Elston R. The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics. 1972; 2:319.

Hsu FC, Sides EG, Mychaleckyj JC, et al. A Transcobalamin 2 gene variant associated with post-stroke homocysteine modifies recurrent stroke risk. Neurology. 2011; 77:1543–1550. [PubMed: 21975197]

Li H, Luan Y. Boosting proportional hazards models using smoothing spline, with application to high-dimensional microarray data. Biostatistics. 2005; 21:2403–2409.

Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. Journal of American Statistical Association. 1989; 84:1074–1078.

Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. Genetic Epidemiology. 2009; 33:183–197. [PubMed: 18814307]

Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, Lin X. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genetic Epidemiology. 2011; 35:620–631. [PubMed: 21818772]

Low HQ, Chen CP, Kasiman K, Thalamuthu A, Ng SS, Foo JN, Chang HM, Wong MC, Tai ES, Liu J. A comprehensive association analysis of homocysteine metabolic pathway genes in Singaporean Chinese with ischemic stroke. PLoS One. 2011; 6:e24757. [PubMed: 21935458]

Moskvina V, Schmidt KM. On multiple-testing correction in genome-wide association studies. Genetic Epidemiology. 2008; 32:567–573. [PubMed: 18425821]

Pearson ES. Note on an approximation to the distribution of non-central χ^2. Biometrika. 1959; 46:364.

Pongpanich M, Neely M, Tzeng JY. On the aggregation of multi-marker information for marker-set and sequencing data analysis: genotype collapsing vs. similarity collapsing. Frontiers in Statistical Genetics and Methodology. 2012; 2:110.

Price AL, Kryukov GV, de Bakker PI, et al. Pooled association tests for rare variants in exon-resequencing studies. American Journal of Human Genetics. 2010; 86:832–838. [PubMed: 20471002]

Qian D, Thomas D. Genome scan of complex traits by haplotype sharing correlation. Genetic Epidemiology. 2001; 21(Suppl 1):S582–S587. [PubMed: 11793742]

Schaid DJ. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Human Heredity. 2010a; 70:109–131. [PubMed: 20610906]

Schaid DJ. Genomic similarity and kernel methods II: methods for genomic information. Human Heredity. 2010b; 70:132–140. [PubMed: 20606458]

Toole JF, Malinow MR, Chambless LE, et al. Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction, and death: the Vitamin Intervention for Stroke Prevention (VISP) randomized controlled trial. Journal of American Medical Association. 2004; 291:565–575.

Tzeng JY, Devlin D, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. The American Journal of Human Genetics. 2003; 72:891–902.

Tzeng JY, Zhang D, Chang SM, et al. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009; 65:822–832. [PubMed: 19210740]

Tzeng JY, Zhang D, Pongpanich M, et al. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. American Journal of Human Genetics. 2011; 89:277–288. [PubMed: 21835306]

von Castel-Dunwoody KM, Kauwell GP, Shelnutt KP, Vaughn JD, Griffin ER, Maneval DR, Theriaque DW, Bailey LB. Transcobalamin 776C→G polymorphism negatively affects vitamin B12 metabolism. The American Journal of Clinical Nutrition. 2005; 81:1436–1441. [PubMed: 15941899]

Wang J, Huff Am, Spence JD, Hegele RA. Single nucleotide polymorphism in CTH associated with variation in plasma homocysteine concentration. Clin Genet. 2004; 65:483–486. [PubMed: 15151507]

Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. American Journal of Human Genetics. 2006; 79:792–806. [PubMed: 17033957]

Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics. 2011; 89:82–93. [PubMed: 21737059]

Zeng D, Lin DY. Efficient estimation of semiparametric transformation models for counting processes. Biometrika. 2006; 93:627–640.

Zhong PS, Chen SX. Tests for High Dimensional Regression Coefficients with Factorial Designs. Journal of American Statistical Association. 2011; 106:260–274.

APPENDIX SECTION

Derivation of the score statistic Q_n

Given the working mixed model: $H(T_i)=\gamma^T X_i+g_i+\varepsilon_i^*$, the log likelihood function of the observed data can be written as

$$l_n(\gamma,H,\tau)=\log\int\cdots\int\prod_{i=1}^{n}[\lambda\{H(\tilde{T}_i)-\gamma^T X_i-g_i\}]^{\delta_i}e^{-\Lambda\{H(\tilde{T}_i)-\gamma^T X_i-g_i\}}f_G(g_1,\ldots,g_n)dg_1\ldots dg_n,$$

where λ and Λ are the specified hazard and cumulative hazard functions of ε_i^*, $f_G(g_1, \ldots, g_n)$ is the joint density of $g_1, \ldots, g_n$, i.e., a multivariate normal density with mean 0 and variance-covariance matrix τS. Consider the variable transformation $(g_1^*,\ldots,g_n^*)'=\tau^{-1/2}S^{-1/2}(g_1,\ldots,g_n)'$, where $S = S^{1/2}S^{1/2}$. Then, $(g_1^*,\ldots,g_n^*)'$ follows a standard multivariate normal distribution. The result leads to

$$l_n(\gamma,H,\tau)=\log\int\cdots\int\prod_{i=1}^{n}[\lambda\{H(\tilde{T}_i)-\gamma^T X_i-\tau^{1/2}S^{1/2}g_i^*\}]^{\delta_i}e^{-\Lambda\{H(\tilde{T}_i)-\gamma^T X_i-\tau^{1/2}S^{1/2}g_i^*\}}\times f_G^*(g_1^*,\ldots,g_n^*)dg_1^*\ldots dg_n^*,$$

where f_G^* is the density for the standard multivariate normal distribution. After some algebra, we have

$$\frac{1}{n}\frac{\partial l_n(\gamma,H,\tau)}{\partial\tau}\Big|_{\tau=0}=\frac{1}{2n}(r_1,\ldots,r_n)S(r_1,\ldots,r_n)^T+\frac{1}{2n}\sum_{i=1}^{n}\left[\frac{\ddot{\lambda}(e_i)\lambda(e_i)-\{\dot{\lambda}\,(e_i)\}^2}{\lambda^2(e_i)}-\dot{\lambda}\,(e_i)\right]s_i^T s_i$$

where $e_i = H(\tilde{T_i})-\gamma^T X_i$, $\ddot{\lambda(\cdot)}$ is the second derivative of $\lambda(\cdot)$, s_i is the ith row of the matrix $S^{1/2}$ and $r_i=\int_0^\infty\dot{\lambda}\,\{H(t)-\gamma^T X_i\}/\lambda\{H(t)-\gamma^T X_i\}dM_i(t;\gamma,H)$. The equality in the above equation is obtained by first taking the derivative of $l_n(\gamma, H, \tau)$ with respect to τ and then deriving its limit as $\tau \rightarrow 0$ using L'Hôpital's rule. Note that the first term on the right-hand side of the above equation is nonnegative, and the second term converges in probability to a constant as n goes to infinity. In addition, under the null hypothesis, r_i's have expectation of 0 at the true values of γ and H because they are martingale integrations. Therefore, if the null hypothesis is correct, the first term in the summation should be close to 0. This result motivates us to consider a score test and reject the null hypothesis when the score $(1/n)$ $\partial l_n(\gamma^\circ, H^\circ, \tau)/\partial\tau\,|_{\tau=0}$ is bigger than some value, where γ° and H° are the estimates of γ and H, respectively, under the null model. It is asymptotically equivalent to consider the test statistic $Q_n = n^{-1}(\hat{r}_1, \ldots, \hat{r}_n)S(\hat{r}_1, \ldots, \hat{r}_n)^T$ and reject the null hypothesis when $Q_n > c_\alpha$, where c_α is the critical value for a level-α test.

Null Distribution of the score statistic Q_n

Here, we consider the estimators $\hat{\gamma}$ and $\hat{H}(\cdot)$ obtained via the martingale-based estimating equations for the standard linear transformation model (Chen, Jin and Ying, 2002) under the null hypothesis. Note that $Q_n=(n^{-1/2}\sum_{i=1}^n \hat{r}_i s_i)^T (n^{-1/2}\sum_{i=1}^n \hat{r}_i s_i)$. Let γ_0 and H_0 denote the true values of γ and H, respectively, in the null model. Based on the derivations given in Chen, Jin and Ying (2002), we have the following asymptotic representations:

$$\sqrt{n}(\hat{\gamma}-\gamma_0)=-A^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\infty \{X_i-\mu_X(t)\}dM_i(t;\gamma_0, H_0)+o_p(1),$$

$$\sqrt{n}\{\hat{H}(t)-H_0(t)\}=-b_1(t)^T A^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\infty \{X_i-\mu_X(t)\}dM_i(t;\gamma_0, H_0)+\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \phi(t,s)dM_i(s;\gamma_0, H_0)+o_p(1).$$

The definitions of A, $\mu_X(\cdot)$, $b_1(\cdot)$ and $\phi(\cdot, \cdot)$ can be found in Chen, Jin and Ying (2002). By Taylor expansion, we can show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{r}_i s_i \rightarrow_d N(0, \sum),$$

as $n \to \infty$, where $\sum=E(\psi_i\psi_i^T)$ and

$$\begin{aligned}\psi_i=&\left[\delta_i \frac{\dot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}}{\lambda\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}}-\lambda\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}\right] s_i\\ &-(B_1-B_2)A^{-1}\int_0^\infty \{X_i-\mu_X(t)\}dM_i(t;\gamma_0, H_0)\\ &-\int_0^\infty b_2(t)dM_i(t;\gamma_0, H_0).\end{aligned}$$

Here

$$\begin{aligned}B_1=&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n s_i X_i^T \left[\dot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\} -\delta_i \frac{\ddot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}\lambda\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}-(\dot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\})^2}{(\lambda\{H_0(\tilde{T}_i)-\gamma_0^T X_i\})^2}\right],\\ B_2=&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n s_i b_1(\tilde{T}_i)^T \left[\dot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\} -\delta_i \frac{\ddot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}\lambda\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}-(\dot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\})^2}{(\lambda\{H_0(\tilde{T}_i)-\gamma_0^T X_i\})^2}\right],\\ b_2(t)=&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n s_i I(\tilde{T}_i \geq t)\phi(\tilde{T}_i, t) \left[\dot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\} -\delta_i \frac{\ddot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}\lambda\{H_0(\tilde{T}_i)-\gamma_0^T X_i\}-(\dot{\lambda}\{H_0(\tilde{T}_i)-\gamma_0^T X_i\})^2}{(\lambda\{H_0(\tilde{T}_i)-\gamma_0^T X_i\})^2}\right].\end{aligned}$$

Therefore, Q_n converges in distribution to a weighted χ^2 distribution: $\sum_{k=1}^{d} \xi_k \chi_{1,k}^2$, where $\chi_{1,1}^2, \cdots, \chi_{1,d}^2$ are d independently and identically distributed χ^2 random variables with degree freedom of 1, and $\xi_1, \cdots, \xi_d$ are the d nonzero eigenvalues of the matrix Σ. To obtain the critical value, c_α, of the limiting weighted χ^2 distribution, we use a numerical method. Specifically, we first obtain a consistent estimator, $\hat{\Sigma}$, of Σ using the usual plug-in method and compute the nonzero eigenvalues $\hat{\xi_1}, \cdots, \hat{\xi_d}$ of the matrix $\hat{\Sigma}$. Next we generate a large set (e.g., 10000 sets) of independent and identically distributed random variables $\chi_{1,1}^2, \cdots, \chi_{1,d}^2$. For each set of χ^2 random variables, we compute $\sum_{k=1}^{d} \hat{\xi}_k \chi_{1,k}^2$. We can then estimate c_α by the upper α-quantile of $\sum_{k=1}^{d} \hat{\xi}_k \chi_{1,k}^2$'s and the associated p-value of the score test can be computed accordingly.

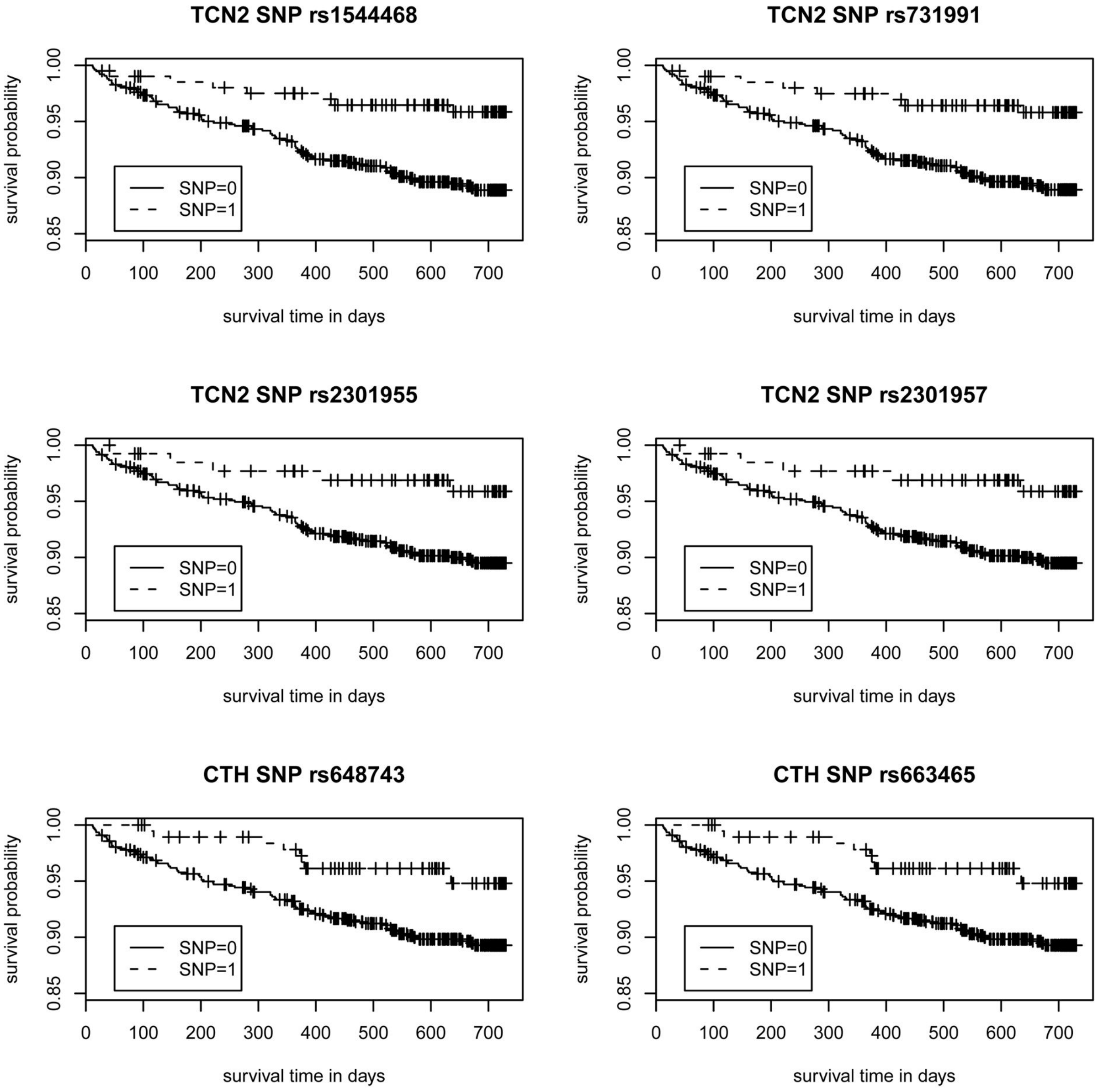

Figure 1. The Kaplan-Meier survival curves for the top 6 SNPs identified from the single SNP association analysis with risk of recurrent stroke in the VISP study

The single SNP association analysis on the 969 subjects under a recessive mode using Cox proportional hazards model showed that 6 out of 69 SNPs within the 9 candidate genes were potentially associated with risk of recurrent stroke (i.e., unadjusted p-values <0.05) in the low-dose arm. The Kaplan-Meier curves of the four *TCN2* SNPs (p-values for rs1544468, rs731991, rs2301955, and rs2301957 are 0.0065, 0.0072, 0.0346, and 0.0346, respectively) and two *CTH* SNPs (p-values for rs648743 and rs663465 are 0.0115.) are included. Note that SNP = 1 if homozygous of a risk allele, SNP = 0 otherwise. The numbers of individuals of SNP=1 are 208, 206, 137, 137, 194, and 194 for rs1544468, rs731991, rs2301955, rs2301957, rs648743, and rs663465, respectively.

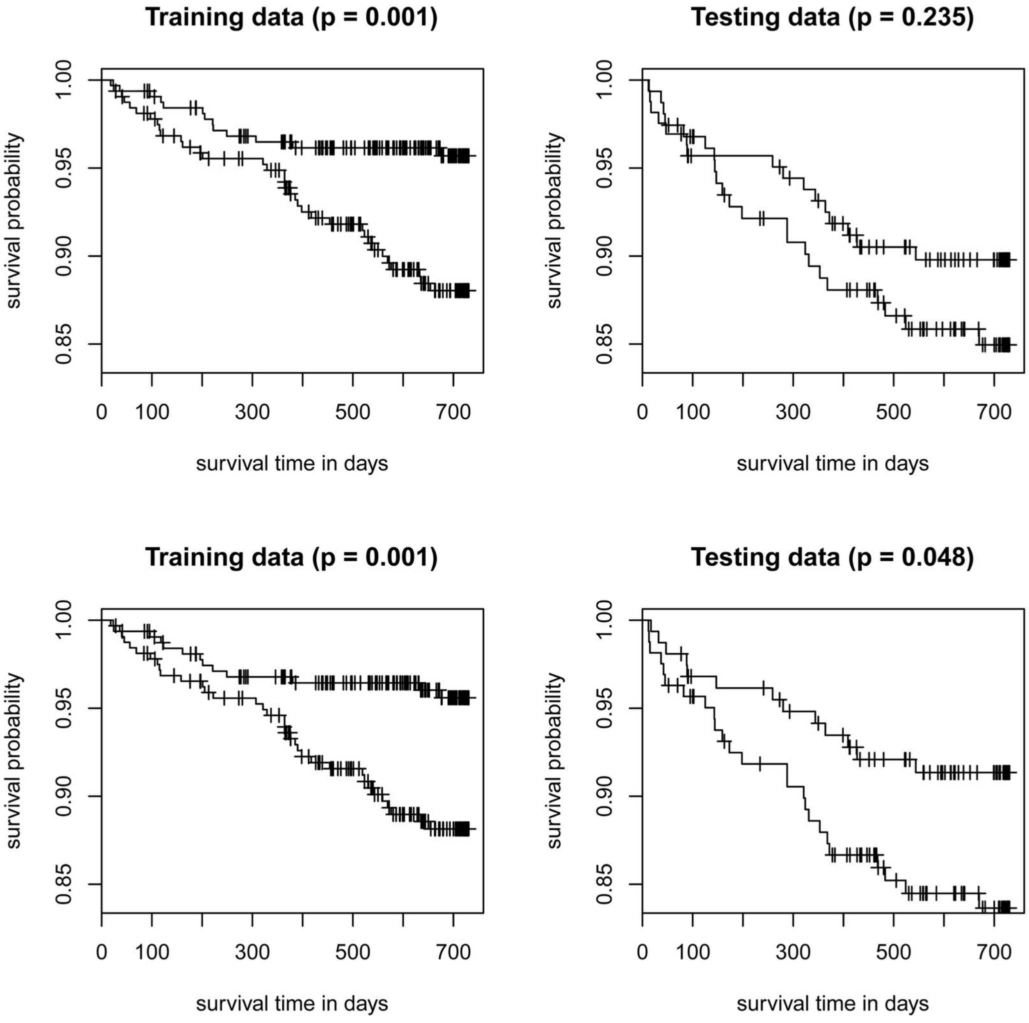

Figure 2. The Kaplan-Meier survival curves for the *TCN2* gene

The VISP subjects were randomly divided into a training dataset (n = 646) and a testing dataset (n = 323). Each plot shows the Kaplan-Meier curves of the high-risk group vs. the low-risk group in the training dataset (left panel) or in the testing dataset (right panel). A subject is in the high-risk (low-risk) group if his/her risk score, calculated from one of the PH models described below, is higher (lower) than the median risk score in the training dataset. Two PH models were considered: Model 1 (top row) included only baseline covariates and Model 2 (bottom row) included baseline covariates and the TCN2 gene. The p-values in the parentheses are for the log-rank tests comparing the corresponding two Kaplan-Meier curves.

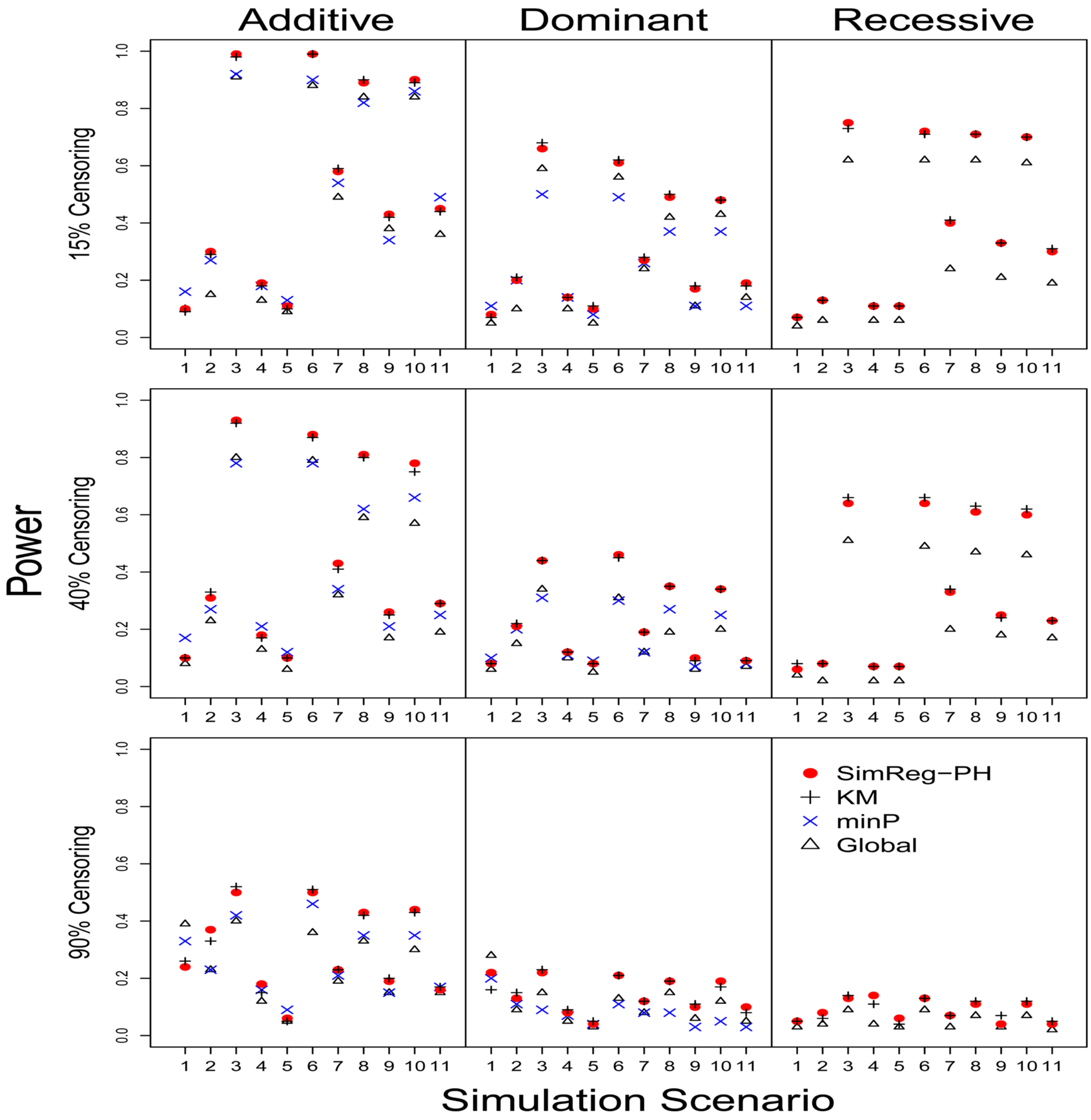

Figure 3. Power when the survival times are generated from the PH model

The 11 scenarios are defined in Table 2. Each row shows the results under different censoring proportions, 15%, 40% and 90% from top down. The Y axis is the power. Powers were obtained based on 100 replications using different methods under a PH model assumption.

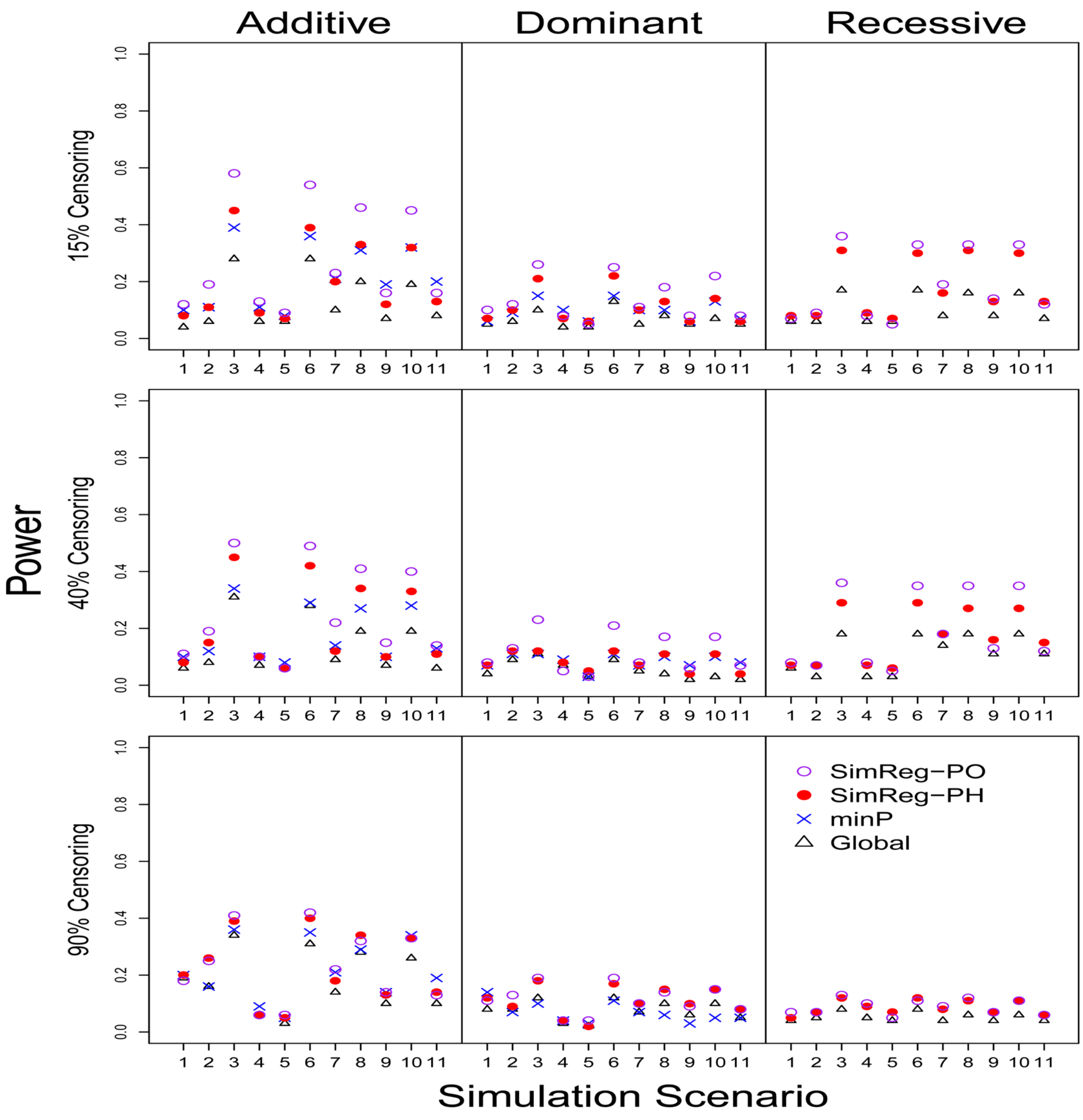

Figure 4. Power when the survival times are generated from the PO model

The 11 scenarios are defined in Table 2. Each row shows the results under different censoring proportions, 15%, 40% and 90% from top down. The Y axis is the power. Powers were obtained based on 100 replications using different methods under a PH model assumption for minP, Global and SimReg-PH, and under a PO model for SimReg-PO.

Table 1

Results of the VISP genetic study.

	BHMT1	BHMT2	CBS	CTH	MTHFR	MTR	MTRR	TCN1	TCN2
Numbers of SNPs	5	3	6	10	7	20	5	3	15
minP	0.3399	0.5968	0.3354	0.0918	0.9105	0.8933	0.6183	0.9764	0.0704
(K_{eff})	(3.99)	(2.83)	(4.41)	(8.35)	(4.64)	(10.64)	(4.78)	(2.77)	(11.28)
Global	0.6457	0.7391	0.2669	0.0518	0.7819	0.9154	0.7363	0.9689	0.0457
KM	0.4863	0.6142	0.2386	0.0078	0.7094	0.8289	0.7289	1.0000	0.0075
SimReg-PH	0.5794	0.6402	0.1889	0.0073	0.6833	0.8835	0.5988	0.9845	0.0040
SimReg-PO	0.6136	0.6942	0.2877	0.0075	0.7807	0.9011	0.6422	0.9922	0.0052

The adjusted p-values for the gene was obtained using $1 - (1 - \text{minimum raw p-value})^{Keff}$. Significance are concluded by comparing the p-values with the Bonferroni threshold 0.05/9 = 0.0056, which accounts for the 9 gene analysis

Table 2

Effect sizes for power analyses.

		Additive & Dominant	Recessive
		15%* ***n*=500** **40%** ***n*=500** **90%** ***n*=1000**	**15%** ***n*=1000** **40%** ***n*=1000** **90%** ***n*=1000**
Scenario	**Causal SNPs (MAF)**	**Effect size (γ_R, γ_U, γ_C)**	**Effect size (γ_R, γ_U, γ_C)**
1	R (0.036)	(1.5, 0.0, 0.0)	(4.0, 0.0, 0.0)
2	U (0.132)	(0.0, 1.0, 0.0)	(0.0, 3.0, 0.0)
3	C (0.419)	(0.0, 0.0, 0.3)	(0.0, 0.0, 0.3)
4	R,U	(0.6, 0.6, 0.0)	(4.0, 4.0, 0.0)
5	R,U	(0.6, 0.4, 0.0)	(2.5, 2.0, 0.0)
6	R,C	(0.3, 0.0, 0.3)	(0.3, 0.0, 0.3)
7	R,C	(0.6, 0.0, 0.2)	(2.5, 0.0, 0.2)
8	U,C	(0.0, 0.3, 0.3)	(0.0, 0.3, 0.3)
9	U,C	(0.0, 0.4, 0.2)	(0.0, 2.0, 0.2)
10	R,U,C	(0.3, 0.3, 0.3)	(0.3, 0.3, 0.3)
11	R,U,C	(0.6, 0.4, 0.2)	(2.5, 2.0, 0.2)

*: censoring proportion.

Table 3

Type I error rates for survival time generated from the PH model.

	Additive			Dominant			Recessive		
Analyzed under PH model	**15% (n=500)**	**40% (n=500)**	**90% (n=1000)**	**15% (n=500)**	**40% (n=500)**	**90% (n=1000)**	**15% (n=1000)**	**40% (n=1000)**	**90% (n=1000)**
$\alpha = 5 \times 10^{-2}$ (Rates shown on the scale of 10^{-2})									
minP	4.89	4.85	4.44	4.92	4.84	4.16	7.63	7.46	7.25
Global	2.73	2.71	2.72	3.01	2.99	2.96	2.75	2.74	2.73
KM	5.07	5.11	4.73	5.10	5.17	4.81	5.12	4.97	4.78
SimReg-PH	5.11	5.10	4.79	5.15	5.11	4.83	5.21	5.04	5.01
$\alpha = 5 \times 10^{-3}$(Rates shown on the scale of 10^{-3})									
minP	6.52	6.22	5.50	6.13	5.44	4.39	17.09	17.62	19.21
Global	2.82	2.65	2.67	2.98	2.74	2.86	2.61	2.50	2.59
KM	5.18	4.86	4.02	5.26	4.86	4.28	5.40	5.54	4.96
SimReg-PH	5.18	4.91	4.15	5.22	5.03	4.38	5.84	5.20	5.96
$\alpha = 5 \times 10^{-4}$ (Rates shown on the scale of 10^{-4})									
minP	8.6	8.1	6.8	8.0	7.0	6.2	48.4	47.9	60.3
Global	2.9	3.9	2.7	4.0	3.2	2.4	3.1	2.9	2.8
KM	5.8	5.2	2.8	6.8	5.8	4.4	8.4	4.4	6.6
SimReg-PH	7.1	5.7	2.7	6.9	5.4	3.2	8.0	4.6	8.1

The type I error rates are shown on the scale of 10^2, 10^3, and 10^4 for nominal level at 0.05, 0.005, and 0.0005, respectively. The survival times were generated from the PH model and were analyzed using different approaches under the PH model. The results were based on 10^5 replications except that the results for KM were based on 5×10^4 replications.

Table 4

Type I error rates for survival time generated from the PO model.

	Additive			Dominant			Recessive		
	15% (n=500)	**40% (n=500)**	**90% (n=1000)**	**15% (n=500)**	**40% (n=500)**	**90% (n=1000)**	**15% (n=1000)**	**40% (n=1000)**	**90% (n=1000)**
minP	0.0584	0.0548	0.0444	0.0560	0.0492	0.0410	0.0798	0.0774	0.0740
Global	0.0358	0.0320	0.0256	0.0352	0.0332	0.0266	0.0342	0.0312	0.0250
SimReg-PH	0.0488	0.0496	0.0464	0.0492	0.0504	0.0478	0.0526	0.0518	0.0480
SimReg-PO	0.0446	0.0486	0.0468	0.0452	0.0496	0.0508	0.0544	0.0492	0.0490

The survival times were generated from the PO model and were analyzed using different approaches under the PO model or the PH model (to examine the impact of model misspecification). The results were based on 5000 replications.