frontiers in Genetics

METHODS
published: 22 May 2018
doi: 10.3389/fgene.2018.00177

Testing Calibration of Cox Survival Models at Extremes of Event Risk

David M. Soave[1,2]* *and Lisa J. Strug*[1,2,3]*

[1] Program in Genetics and Genome Biology, Research Institute, The Hospital for Sick Children, Toronto, ON, Canada, [2] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada, [3] The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada

Risk prediction models can translate genetic association findings for clinical decision-making. Most models are evaluated on their ability to discriminate, and the calibration of risk-prediction models is largely overlooked in applications. Models that demonstrate good discrimination in training datasets, if not properly calibrated to produce unbiased estimates of risk, can perform poorly in new patient populations. Poorly calibrated models arise due to missing covariates, such as genetic interactions that may be unknown or not measured. We demonstrate that models omitting interactions can lead to increased bias in predicted risk for patients at the tails of the risk distribution; i.e., those patients who are most likely to be affected by clinical decision making. We propose a new calibration test for Cox risk-prediction models that aggregates martingale residuals for subjects from extreme high and low risk groups with a test statistic maximum chosen by varying which risk groups are included in the extremes. To estimate the empirical significance of our test statistic, we simulate from a Gaussian distribution using the covariance matrix for the grouped sums of martingale residuals. Simulation shows the new test maintains control of type 1 error with improved power over a conventional goodness-of-fit test when risk prediction deviates at the tails of the risk distribution. We apply our method in the development of a prediction model for risk of cystic fibrosis-related diabetes. Our study highlights the importance of assessing calibration and discrimination in predictive modeling, and provides a complementary tool in the assessment of risk model calibration.

Keywords: calibration tests, cox proportional hazards model, extreme risk, goodness-of-fit, prediction

OPEN ACCESS

Edited by:
Mariza De Andrade,
Mayo Clinic, United States

Reviewed by:
Gengsheng Qin,
Georgia State University, United States
Jun Yin,
Mayo Clinic, United States

***Correspondence:**
David M. Soave
david.soave@mail.utoronto.ca
Lisa J. Strug
lisa.strug@utoronto.ca

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *13 March 2018*
Accepted: *30 April 2018*
Published: *22 May 2018*

Citation:
Soave DM and Strug LJ (2018) Testing Calibration of Cox Survival Models at Extremes of Event Risk. Front. Genet. 9:177. doi: 10.3389/fgene.2018.00177

1. INTRODUCTION

Genome-wide association studies have been very successful in identifying genetic contributors to disease (Welter et al., 2014). Following discovery and validation, it is desirable to determine whether these genetic findings translate to biomarkers that can be clinically useful (e.g., for disease prognosis).

Two important measures of predictive performance of a risk-prediction model are discrimination and calibration (Harrell, 2001; Moons et al., 2012). Discrimination measures how well model-estimated risks translate to patient outcomes, where patients grouped according to higher predicted risk should demonstrate higher event rates than patients in lower risk groups. Calibration is a measure of how closely the estimated and observed absolute risks agree, where miscalibrated models lead to biased estimates of risk. Both measures are important for model validation in both training (internal) and external datasets, however, calibration is rarely reported in risk prediction studies (Collins et al., 2014). Even if a new prediction model discriminates

well in a training dataset, if good calibration is not also achieved it can perform poorly in a new patient population. With the large cost and effort involved in obtaining external datasets for model validation, good model calibration should be demonstrated in the training set prior to collection of a second, independent sample.

The Cox proportional hazards (PH) model is a commonly used modeling technique for the analysis of time-to-event data. In the clinical setting, a Cox model can be used as a prediction tool to estimate an individual's relative (or absolute) risk of developing disease. Typically, a risk score is obtained as the linear predictor from the fitted Cox model. Patients can then be classified into risk groups to help inform clinical decisions. Methods to assess various aspects of the fit of a Cox model generally involve examination of plots of martingale residuals or their transforms (Schoenfeld, 1982; Barlow and Prentice, 1988; Therneau et al., 1990; Lin et al., 1993). Patterns in these plots can be challenging to identify in the presence of even moderate censoring, and thus, smoothers are typically applied as a visual aid. These smoothers can be useful in identifying trends, but give the impression of too little variation and therefore complementary formal testing is needed (Kalbfleisch and Prentice, 2002). Various calibration or goodness-of-fit (GOF) tests have been proposed in the Cox model setting, most of which can be characterized as variations of the Hosmer-Lemeshow GOF test for binary data (Hosmer et al., 1997). These methods assess the agreement between observed and expected risk across all risk levels, and therefore reflect a global assessment of lack of fit. Gronnesby and Borgan (1996) used counting process notation to derive a score (GB) test using the sums of martingale residuals across risk group deciles. The GB test is similar to the Hosmer-Lemeshow test since the martingale residuals correspond to the observed minus expected number of events for each subject. D'Agostino and Nam (2003) proposed a test comparing the average risk predictions with the observed Kaplan-Meier (K-M) failure probabilities across the deciles. This approach ignores censoring, however, leading to an incorrect variance estimate with increased instability for increased censoring (Crowson et al., 2016). Demler et al. (2015) proposed to use the robust Greenwood variance estimators of the K-M failure probabilities to improve performance of the testing procedure. While this approach maintains correct type 1 error control, it demonstrated comparable or lower power against the GB test under their simulation examples for model misspecification.

Clinical decisions about treatment and monitoring are most often made for patients at the extremes of the risk distribution (high or low). Therefore, accuracy of their predicted risks are a priority. While the available calibration tests for survival data have been shown to perform reasonably well as global tests, their power will be limited in detecting deviations in predicted risk at the extremes of the risk distribution where the proportion of subjects is small. Song et al. (2015) developed a method to test calibration of risk models at extremes of disease risk for binary outcomes. Their work was motivated by deviations between observed and expected risk near the tails of the risk distribution due to misspecification of either additive or multiplicative effects of the covariates on the risk. The Cox model also assumes that the effects of the covariates on the hazard rate (HR) are multiplicative, which may or may not be reasonable (Weinberg, 1986), and could result in bias in the expected hazard rate at the extremes of risk.

Genetic interaction (gene-gene and gene-environment) can contribute to complex traits. Many of these interactions remain unknown and are a challenge to model directly (Soave et al., 2015). Here we show that working models omitting relevant interactions are also likely to produce biased estimates of risk at the extreme tails of the population risk distribution.

Following the martingale theory used by Gronnesby and Borgan (1996), we propose a new calibration test for Cox models, that has improved power to detect biased risk estimates at the tails of the risk distribution. Our test aggregates martingale residuals for subjects from extreme high and low risk groups with a test statistic maximum chosen by varying which risk groups are included in the extremes. An estimate of the empirical significance of our test statistic is obtained by simulating from a Gaussian distribution using the covariance matrix for the grouped sums of martingale residuals. We describe and demonstrate how to implement our method using existing software. We conduct an extensive simulation study that shows the extreme risk (ER) test maintains good control of type 1 error and demonstrates improved power over the GB test when risk estimates are less accurate at the tails of the risk distribution. We consider scenarios where interaction effects are missing from the working model and the multiplicative risk assumption is violated. The ER test is complementary to existing global methods for examining risk model calibration.

2. MODEL AND TEST PROCEDURES

For simplicity, we consider fixed time covariates, and right censoring of event times. For an independent sample of size n, let each individual i have a px1 vector of fixed covariates, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^T$. Let T_i and C_i be the event/failure time and censoring time, respectively for individual i, and only the earlier of the two times is observed. Following the counting process notation of Andersen et al. (1993), we observe $Y_i(t)$ and $N_i(t)$ at each time t, where $Y_i(t) = I(T_i \geq t, C_i \geq t)$ is the at risk indicator and $N_i(t)$ counts the number of observed events for individual i until time t. We assume that an event can occur only once for each individual. Thus, for each individual i, we observe a follow-up time $x_i = min(T_i, C_i)$ and an indicator of whether an event occurred prior to censoring $\delta_i = I(T_i \leq C_i)$. Under the Cox PH model, the intensity process $h_i(t; \boldsymbol{z}_i)$ for $N_i(t)$ can be written as

$$h_i(t; \boldsymbol{z}_i) = h_0(t) exp(\boldsymbol{\beta}^T \boldsymbol{z}_i) Y_i(t), \quad (1)$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{\beta}$ is a px1 vector of regression parameters and $\boldsymbol{\beta}^T \boldsymbol{z}_i$ is the risk score for individual i.

2.1. Gronnesby and Borgan (GB) Test

The GB test is based on martingale residuals, which are estimated for each individual at time t as

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(u) exp(\widehat{\boldsymbol{\beta}}^T \boldsymbol{z}_i) d\widehat{\Lambda}_0(u), \quad i = 1, \ldots, n,$$

where

$$\widehat{\Lambda}_0(u) = \int_0^t \frac{dN_*(u)}{\sum_{l=1}^n Y_l(u) exp(\widehat{\boldsymbol{\beta}}^T \boldsymbol{z}_l)}$$

is the Breslow estimator (Breslow, 1972) of the baseline cumulative intensity process, and $N_*(u) = \sum_{i=1}^n N_i(t)$. We denote the estimated martingale at time $t = \infty$ as $\widehat{M}_i(\infty) = \widehat{M}_i$.

For the GB test, the data are divided into D groups based on their estimated risk score, $\widehat{r}_i = \widehat{\boldsymbol{\beta}}^T \boldsymbol{z}_i$, from the fitted Cox model of Equation (1). The martingale residuals are then summed within each group, $H_{J_d} = \sum_i K_{di} \widehat{M}_i$, where $K_{di} = I(\widehat{r}_i \epsilon J_d)$ is an indicator for whether the risk score of the ith observation is in the risk score interval for the dth group, $J_d, d = 1, \ldots, D$. If the model fit is good [i.e., model (Equation 1) holds], H_{J_d} should be close to zero for each group, and $\boldsymbol{H} = (H_{J_1}, \ldots, H_{J_{D-1}})^T$ converges to a mean zero multivariate Gaussian random vector (Gronnesby and Borgan, 1996). Therefore, the GB procedure uses the following test statistic:

$$T = (H_{J_1}, \ldots, H_{J_{D-1}}) \widehat{\Sigma}^{-1} (H_{J_1}, \ldots, H_{J_{D-1}})^T,$$

where $\widehat{\Sigma}$ is an estimate of the covariance matrix of $\boldsymbol{H}$. When model (Equation 1) holds, T is asymptotically distributed as χ^2_{D-1}. Note that one of the group-wise martingale sums is omitted for model identifiability since $\sum H_{J_d} = 0$.

May and Hosmer (1998) showed that the GB test is algebraically equivalent to a score test of $D - 1$ risk group indicator variables, $\boldsymbol{K}_i = (K_{1i}, \ldots, K_{(D-1)i})$, in the Cox model (Equation 1). Thus, the GB test is equivalent to the following two-stage procedure:

Stage 1.1. Obtain the estimated risk score $\widehat{r}_i$ for each subject from the Cox regression fit of model (Equation 1).

Stage 1.2. Divide the subjects into D groups based on the ordered risk score estimates, and specify group membership for each subject using the group indicator covariate vector, $\boldsymbol{K}_i$.

Stage 2. Test for association with the indicator vector $\boldsymbol{K}_i$ in the full model,

$$h_i(t; \boldsymbol{z}_i, \boldsymbol{K}_i) = h_0(t) exp(\boldsymbol{\beta}^T \boldsymbol{z}_i + \boldsymbol{\gamma}^T \boldsymbol{K}_i) Y_i(t). \quad (2)$$

In this framework, the GB procedure is a score test of the null hypothesis that $\boldsymbol{\gamma} = \boldsymbol{0}$. The GB test now simplifies to fitting two standard Cox models with existing software.

2.2. Extreme Risk (ER) Test

The GB test is a global test of the model fit across the entire distribution of estimated risk scores. However, it may be desirable to focus additional attention at the extreme tails of the risk distribution where patients are more likely to be affected (positively or negatively) by clinical decisions. We propose a modification of the GB test to improve the power to detect model bias in risk prediction at the extremes of the risk distribution. This work is motivated by a recently proposed calibration test at extremes of disease risk for binary risk models (Song et al., 2015) and by recognition that risk-prediction models omitting relevant genetic interactions will increase bias in risk estimates at the extremes (section 3).

Again, suppose the data are divided evenly into D groups based on the estimated risk scores, as described above for the GB test. For a given pair of thresholds, $c = (c_l, c_u)$, defining a set of "extreme" risk score groups, $R_c = (J_1, \ldots, J_{c_l}) \cup (J_{c_u}, \ldots, J_D)$, we propose the following test statistic

$$T_c = \sum_{d=1}^{D} (H_{J_d})^2 I(J_d \epsilon R_c).$$

This test statistic is observed to be the sum of the squared group martingales sums, over only those groups contained in the extreme risk set, R_c. We do not incorporate the covariance matrix $\widehat{\Sigma}$ in the definition of T_c but instead use it in a Monte Carlo simulation procedure as outlined below.

The motivation for the ER test arises from the departures detected at the tails of the risk distribution. However, specifying which groups should belong to the extreme risk set is arbitrary. The risk set should not be chosen by first looking at the data as this sort of adaptive procedure will lead to incorrect type 1 error control. Therefore, we propose taking our ER test statistic to be the maximum of a scaled version of T_c, over all possible risk group sets (Song et al., 2015), $T^{max} = max_c(\widetilde{T}_c / n_c)$, where n_c is the number of groups included in R_c and $\widetilde{T}_c$ is constructed using $\widetilde{\boldsymbol{H}}$, a scaled transformation of $\boldsymbol{H}$ such that each component has mean 0 and variance 1. In this way, R_c is chosen as a series of equally balanced groups beginning at both ends of the group list [i.e., $c = (c_1, c_D), (c_2, c_{D-1}), (c_3, c_{D-2}), \ldots$ etc.].

Under model (Equation 1), Gronnesby and Borgan (1996) derived explicit formulas for estimating the covariance matrix of $\boldsymbol{H}$, $\widehat{\Sigma}$. To achieve model identifiability, and estimate $\widehat{\Sigma}$, the GB test arbitrarily omits the martingale sum for group D, H_{J_D}, from $\boldsymbol{H}$. The ER test also requires estimation of $\widehat{\Sigma}$, however, the focus is on detecting departures from the null hypothesis in the tails of the distribution. Thus we redefine $\boldsymbol{H}$ by omitting $H_{J_{D/2}}$, when D is even, and $H_{J_{(D+1)/2}}$, when D is odd, resulting in direct estimation of the covariance $\widehat{\Sigma}$ for all groups except the median.

Next, $\boldsymbol{H}$ and $\widehat{\Sigma}$ are scaled to be $\widetilde{\boldsymbol{H}}$ and $\widetilde{\Sigma}$, such that $\widetilde{\Sigma}$ is a correlation matrix. Unfortunately, the distribution of the test statistic T^{max} is intractable. However, with $\widetilde{\Sigma}$ available, we can simulate realizations of $\widetilde{\boldsymbol{H}}$ (and correspondingly $\widetilde{T}_c$ from each of the defined risk group sets, R_c). Therefore, we propose the following steps to estimate the empirical p-value of the ER test, $P(T^{max} \geq t^{max})$, where t^{max} is the observed value of T^{max}, using simulations as follows.

1. Generate a new realization of $\widetilde{\boldsymbol{H}}$ from a mean zero multivariate Gaussian distribution with covariance matrix $\widetilde{\Sigma}$, and calculate a new value for the test statistic, t_s^{max}.
2. Repeat Step 1 R (Replicate) times, to create a simulated "null" distribution for T^{max}.
3. Estimate the p-value, $P(T^{max} \geq t^{max})$, empirically as the proportion of simulation replicates where the simulated t_s^{max} is greater than the observed t^{max}.

2.3. Implementing the ER Test Using Existing Software

Software packages generally estimate the regression coefficients, $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, for a Cox model (Equation 2) by maximizing the log-partial likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}^T \boldsymbol{z}_i + \boldsymbol{\gamma}^T \boldsymbol{K}_i - log\left(\sum_{l=1}^{n} exp(\boldsymbol{\beta}^T \boldsymbol{z}_l + \boldsymbol{\gamma}^T \boldsymbol{K}_l) Y_l(x_i) \right) \right].$$

May and Hosmer (1998) showed that the partial likelihood score vector for $\boldsymbol{\gamma}$ under $\boldsymbol{\gamma} = \boldsymbol{0}$ and $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}=\boldsymbol{0}}$ corresponds to the vector of risk group sums of martingale residuals, $\boldsymbol{H}$. That is,

$$\left. \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \right|_{\substack{\boldsymbol{\gamma}=\boldsymbol{0} \\ \boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}} = \boldsymbol{H}.$$

Therefore, we can extract an estimate of the covariance matrix of $\boldsymbol{H}$ from the observed information as

$$\widehat{\Sigma} = (\widetilde{\mathcal{J}}_{\gamma\gamma} - \widetilde{\mathcal{J}}_{\gamma\beta} \widetilde{\mathcal{J}}_{\beta\beta}^{-1} \widetilde{\mathcal{J}}_{\beta\gamma}),$$

where

$$\widetilde{\mathcal{J}} = \begin{pmatrix} \mathcal{J}_{\beta\beta} & \mathcal{J}_{\gamma\beta} \\ \mathcal{J}_{\beta\gamma} & \mathcal{J}_{\gamma\gamma} \end{pmatrix} \Bigg|_{\substack{\boldsymbol{\gamma}=\boldsymbol{0} \\ \boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}} \quad \text{and} \quad \mathcal{J}_{\beta\gamma} = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T}.$$

Although not directly available as output from the coxph() function in the "survival" R software package (R Core Team, 2016), $\widehat{\Sigma}$ can be obtained as follows (Web Appendix A in the Supplementary Materials provides example R code for this procedure).

1. Fit a Cox model corresponding to model (Equation 1) using coxph() to obtain estimates of the coefficients, $\widehat{\boldsymbol{\beta}}$.
2. Substitute these fixed estimates for $\boldsymbol{\beta}$ in a Cox model corresponding to model (Equation 2) while also specifying $\boldsymbol{\gamma} = 0$; coefficients can be fixed in a coxph() fit by specifying *"iter.max=0"*.
3. Use the vcov() function to return the inverse of the observed information, $\widetilde{I}$, and obtain $\widehat{\Sigma}$ by taking the inverse of the submatrix with rows and columns corresponding to $\boldsymbol{\gamma}$.

Obtaining $\widehat{\Sigma}$ in this way allows one to avoid explicit specification of the formulas for $\widehat{\Sigma}$ in Gronnesby and Borgan (1996). We can now simulate values of the grouped martingale sums, under the assumption of a correct model (Equation 1), and implement the ER test according to section 2.2.

2.4. Grouping Strategy-Choosing *D* and Dealing With Sparse Vents Within Risk Groups

Typically, a sample is stratified into 10 risk groups for the GB test. This convention is consistent with implementation of the Hosmer-Lemeshow test for binary data (Hosmer et al., 1997) and has been shown to yield good statistical properties for the GB test with samples sizes of 500 (May and Hosmer, 2004). We considered sample sizes of 5,000 and 1,500 in our simulation study, assuming that genetic markers individually contribute small effects to clinical outcomes. For implementation of both the GB and ER tests we use $D = 11$. The odd number of groups ensures that when the median group is omitted (as the reference group) from the simulation algorithm of the ER test, there is always a balanced number of upper and lower risk groups included in T_c, for all values of the threshold, c.

A second convention for application of the Hosmer-Lemeshow test is the "no less than 5" events rule, directing that successive groups be collapsed based on a minimum of five expected events per group. Examination of this grouping convention for the GB test generally supports its application, although it may be conservative (May and Hosmer, 1998, 2004; Parzen and Lipsitz, 1999). For estimates of the expected number of events within the risk groups, we take the sum of the Cox-Snell residuals. The Cox-Snell residual corresponds to the second term in the martingale residual at $t = \infty$,

$$\int_0^{\infty} Y_i(u) exp(\widehat{\boldsymbol{\beta}}^T \boldsymbol{z}_i) d\widehat{\Lambda}_0(u), \quad i = 1, \ldots, n.$$

We compared our results with and without application of the 'no less than 5' events rule, and denote the corresponding GB and ER tests by GB_{adj} and ER_{adj}.

3. SIMULATIONS

We conducted a simulation study to evaluate the performance of the ER test and compared it with the conventional GB test. To emulate calibration testing for risk-prediction models using genetic and environmental factors, we simulated 5 or 10 single nucleotide polymorphism (SNP) genotypes (G = 0, 1, or 2), one evnironmental exposure variable (E = 0, or 1), and event times (t in years) for each subject according to the Weibull hazard

$$h_i(t; \boldsymbol{G}_i) = h_0(t) exp(g(\boldsymbol{\beta}, \boldsymbol{G}_i, E_i)), \tag{3}$$

where $h_0(t) = \lambda \alpha t^{\alpha - 1}$ is the baseline hazard, α is the shape parameter and λ is the scale parameter. The function $g(\cdot)$ specifies the model for the joint risk of the disease associated with the genotype-covariate vector $\boldsymbol{G}_i$, exposure variable E_i, and corresponding effect coefficients, $\boldsymbol{\beta}$. We will use the notation $g_{0\cdot}(\cdot)$ and $g_{A\cdot}(\cdot)$ in (Equation 3) to specify various forms of the null (working) and alternative (true) hazard models, respectively. Each G_{ij} was simulated under Hardy-Weinberg equilibrium from a Binomial distribution to reflect the number of minor alleles (0,1,2) with minor allele frequency (MAF) 30% at each SNP $j = 1, \ldots, p$, for each subject $i = 1, \ldots, n$. All event times greater than 10 years were treated as censored at 10 years (administrative censoring). In addition, event times were uniformly censored prior to 10 years at a 0 or 50% censoring rate for different scenarios (lost to follow-up censoring). We considered sample sizes of n = 5,000 and 1,500. Type 1 error and power were assessed at the 0.05 significance level using 10,000 and 1,000 simulation replicates, respectively. To estimate the p-value for each ER test statistic, we used

$R = 1,000$ replicate simulations, which provided sufficient precision for the 0.05 significance threshold. **Table 1** provides an outline of the simulation (and working) models used for power comparisons with the corresponding results figures. We chose to use small to moderate main effects for $\boldsymbol{G}$ that might be plausible for polygenic risk prediction models invovling complex traits. We also considered alternative effect sizes to those described in **Table 1** and throughout the simulations, and obtained qualitatively similar comparisons between the ER and GB tests (results not shown).

3.1. Type 1 Error Control of the ER Test

To assess the type 1 error of our ER test, we simulated data from (Equation 3) under the null model (Equation 1) using $g_{01}(\boldsymbol{\beta}, \boldsymbol{G}_i) = \sum_{j=1}^{p} \beta_{G_j} G_{ij}$, and then fit a corresponding Cox model of the same form. We used a fixed β_{G_j} across all SNPs of log(1.2) and log(1.15) for the models with $p = 5$ and 10 SNPs, respectively. We specified α at 1, 3 and 0.3, corresponding to a constant, increasing and decreasing baseline hazard, respectively. Under each scenario, λ was chosen such that the event rate prior to 10 years, in the absence of censoring, was 5%, 10% and 20%. For each simulation replicate we tested the Cox model for lack-of-fit using the ER and GB tests.

The proposed ER test maintained good control of type 1 error across all simulation scenarios for both the 5-SNP and 10-SNP models with constant, decreasing and increasing hazards (**Table 2**, Web Tables 1, 2 in the Supplementary Materials, respectively). Collapsing of groups based on the 'less than 5' expected events rule rarely occurred in the simulations with $n =$ 5,000. For the simulations with $n = 1{,}500$, collapsing occurred more frequently, however, type 1 error results were similar for ER_{adj} and GB_{adj} (Web Table 2 in the Supplementary Materials) compared to the unadjusted ER and GB, respectively.

Traditional application of the GB test compares the test statistic (T in section 2.1) to the χ^2_{D-1} distribution to obtain a p-value. Simulation based p-value estimates for GB similarly to the ER test method are also possible. We compared the type 1 error and power between the asymptotic and simulation based methods for GB and found the results to be very similar. Thus, we report only the results of GB based on the asymptotic p-value estimates.

3.2. Power of the ER Test Under Misspecified Models-Missing Interactions

To examine the power of the ER test to detect model misspecification due to missing interactions, we simulated data from (Equation 3) with the addition of an exposure, E, that interacted with one or more of the covariates, $g_{A1}(\boldsymbol{\beta}, \boldsymbol{G}_i, E_i) = \sum_{j=1}^{p} \beta_{G_j} G_{ij} + \beta_E E_i + \sum_{j=1}^{p} \beta_{G_j E} G_{ij} E_i$. We then fit a Cox model corresponding to (Equation 3) that included the main effect of E but omitted the interactions, $g_{02}(\boldsymbol{\beta}, \boldsymbol{G}_i, E_i) = \sum_{j=1}^{p} \beta_{G_j} G_{ij} + \beta_E E_i$, as well as one that completely ignored the presence of the exposure, $g_{01}(\boldsymbol{\beta}, \boldsymbol{G}_i)$. This may be a common occurrence as interacting exposures are often unknown to researchers. Each E_i was a binary (0,1) variable simulated from a Bernoulli distribution with frequency 0.3; smaller frequencies were also evaluated (0.2 and 0.1) but led to similar conclusions (results not shown). We used a fixed β_{G_j} of log(1.15) across all SNPs for the model with $p = 10$ SNPs. The main effect of the exposure, β_E, was also fixed at the same value as the β_{G_j}. We specified α at 1, corresponding to a constant baseline hazard. Under each scenario, the event rate prior to 10 years was 20% in the absence of censoring. To examine the empirical power, we varied a single exposure effect common to all interactions in the model, $\beta_{G_j E}$, across all scenarios.

We expected that data simulated under g_{A1} and fit using a Cox PH model with g_{01} or g_{02} would show increased deviation in observed vs. expected risk at the extremes of the subject risk distribution. **Figure 1** demonstrates the deviation between observed and predicted survival probabilities across the population risk distribution for the missing interaction scenario.

Empirical power results from the fitted models that included the main effect of E but omitted the interaction effect(s), g_{02}, are presented in **Figure 2** ($n = 5{,}000$) and Web Figure 1 in the Supplementary Materials ($n = 1{,}500$) for the data simulated with 1 and 10 interactions between $\boldsymbol{G}$ and E. For both scenarios, the power of the ER and GB tests increased as the size of the omitted interactions increased. The ER test appeared to show an increase in power over GB for both the single and multiple (10) SNP-interaction models with 0% or 50% lost to follow up in the samples, for much of the range of the interaction effect, $\beta_{G_j E}$. The GB test showed a slight advantage for some of the larger interaction effect sizes considered.

Results from the fitted models that completely omitted E (both main effect and interactions), g_{01}, are presented in **Figure 3** ($n = 5{,}000$) and Web Figure 2 in the Supplementary Materials ($n = 1{,}500$). For data simulated from models including one interaction between $\boldsymbol{G}$ and E, ER was more powerful than GB. For the data simulated under 10 interactions, the performance of the two tests was comparable. When one large interaction effect exists (say for $G_1 \cdot E$) but is omitted in the fitted model along with the main effect E, the estimated effect on risk due to $G_1 = 1$ will be larger (much larger in the case of $G_1 = 2$) than the estimated effect of the other SNPÕs. Depending on the minor allele frequency of G_1, individuals with either $G_1 = 0$ or $G_1 = 2$ will be over represented in either the lowest- or highest-risk groups, respectively, and contribute less stable martingale residuals to their risk group (larger in absolute value). On the other hand, when smaller interaction effects exist for each SNP, but are omitted, the bias that is introduced is likely to be spread across more risk groups, reducing the improvement observed for the ER test.

To demonstrate the usefulness of the ER test beyond models using only genotypes as predictors, we considered scenarios involving covariates from a standard Gaussian distribution. For this scenario, we simulated data under (Equation 3) using $g_{A3}(\boldsymbol{\beta}, \boldsymbol{Z}_i) = \beta_{Z_1} Z_{i1} + \beta_{Z_2} Z_{i2} + \beta_{Z_1 Z_2} Z_{i1} Z_{i2}$, where $Z_{i1}, Z_{i2} \sim \mathcal{N}(0, 1)$, and β_{Z_j} was fixed at log(1.15) for both covariates, and $\beta_{Z_1 Z_2}$ was varied. We then fit a Cox model (Equation 3) that omitted the interaction term, using $g_{03}(\boldsymbol{\beta}, \boldsymbol{Z}_i) = \beta_{Z_1} Z_{i1} + \beta_{Z_2} Z_{i2}$, to examine the empirical power to detect the model misspecification. Similar to the SNP risk model examples, we expected that data simulated under g_{A3} and fit using a Cox PH model with g_{03} would show increased deviation in observed vs. expected survival at the extremes of the subject risk distribution

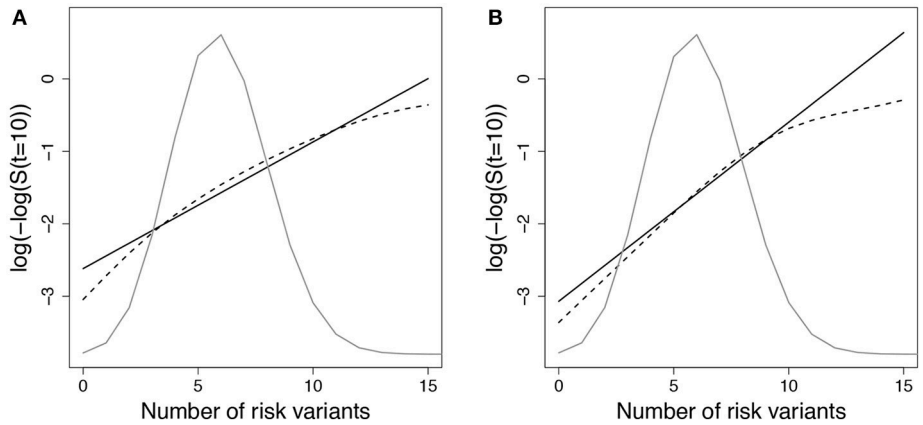

FIGURE 1 | Log-log survival probability [$t = 10$ years] versus cumulative number of risk variants under missing interaction models. Time to event data were simulated under model (Equation 3) using g_{A1} with 10 SNPs (MAF=0.3 for each) and **(A)** one interaction with E ($g_{A1}(\boldsymbol{\beta}, \boldsymbol{G}_i, E_i) = \sum_{j=1}^{10} 1.15G_{ij} + 1.15E_i + 10G_{i1}E_i$) or **(B)** ten interactions with E ($g_{A1}(\boldsymbol{\beta}, \boldsymbol{G}_i, E_i) = \sum_{j=1}^{10} 1.15G_{ij} + 1.15E_i + \sum_{j=1}^{10} 1.35G_{ij}E_i$). Cox models were fit using the null model $g_{02}(\boldsymbol{\beta}, \boldsymbol{G}_i, E_i) = \sum_{j=1}^{10} \beta_{G_j} G_{ij} + \beta_E E_i$ to estimate survival probabilities at time t=10 years ($\widehat{S}_i(t) = \widehat{S_0}(t)^{exp(\sum_{j=1}^{10} \hat{\beta}_{G_j} G_{ij} + \hat{\beta}_E E_i)}$), which were then averaged by risk group (defined by cumulative number of risk variants) (solid line - predicted survival). Observed survival probabilities (dashed line - observed survival) were obtained from a Cox model comparing patients in each risk group with the reference group. The reference group was the subset of individuals with 6 risk alleles; the mean number of risk alleles from the 10-SNP model. The distribution of subjects per number of risk alleles is shown as a density curve. Survival probabilities estimated for other time points ($t < 10$ years) yielded similar results.

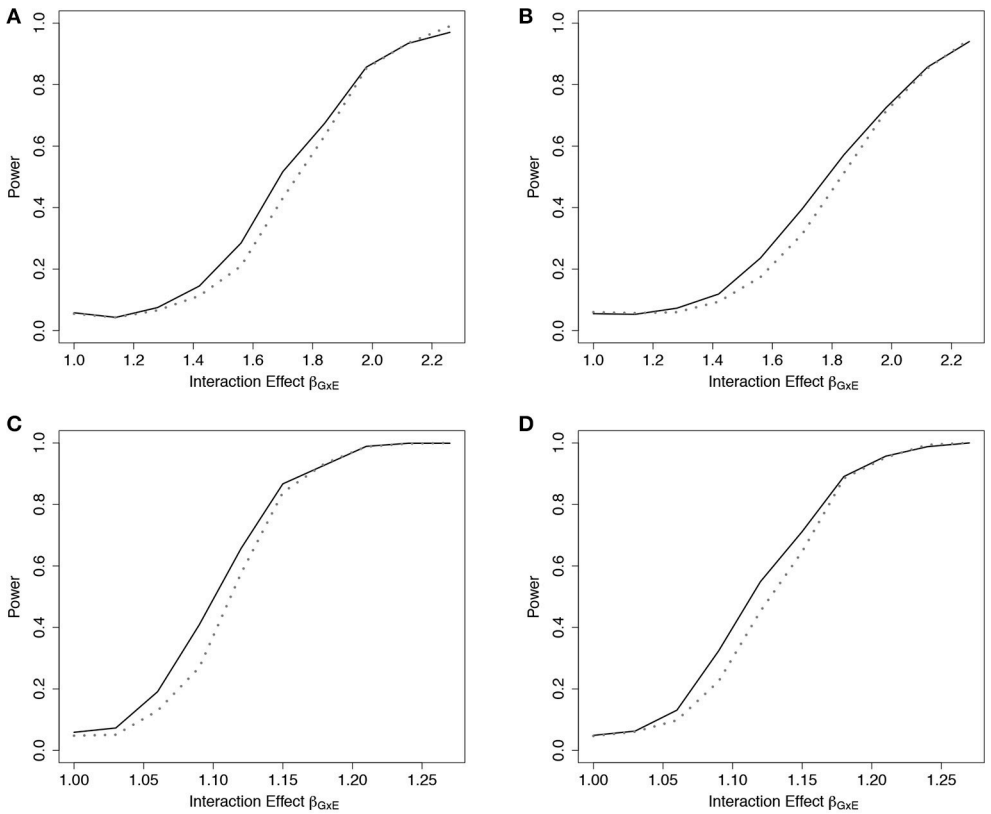

FIGURE 2 | Power of tests under missing interaction models g_{02}. Power of the ER (solid line) and GB (dotted line) tests when $\boldsymbol{G}$xE interaction terms are missing from the fitted Cox model (3), but exposure E is included, using g_{02}. $n = 5000$ for the 10-SNP model with one interaction **(A**,**B)** or ten interactions **(C**,**D)** in the underlying true model g_{A1}, and either 0% **(A**,**C)** or 50% **(B**,**D)** lost to follow-up censoring.

TABLE 1 | Outline of simulation study and results figures for power comparisons.

Generating model	Working model	Parameter values	Results tables	Conclusions
Missing Interactions (Categorical Covariates) - section 3.2				
$g_{A1}(\boldsymbol{\beta}, \boldsymbol{G}_i, E_i) = \sum_{j=1}^{p} \beta_{G_j} G_{ij}$	$g_{01}(\boldsymbol{\beta}, \boldsymbol{G}_i) = \sum_{j=1}^{p} \beta_{G_j} G_{ij}$	$n = 5{,}000$, $p = 10$	**Figure 2**	ER showed increased power over GB for most models considered, however, GB showed a slight advantage for some of the larger interaction effect sizes.
		$n = 1{,}500$, $p = 10$	Web Figure 1 in the Supplementary Materials	
$+\beta_E E_i + \sum_{j=1}^{p} \beta_{G_j E} G_{ij} E_i$	$g_{02}(\boldsymbol{\beta}, \boldsymbol{G}_i, E_i) = \sum_{j=1}^{p} \beta_{G_j} G_{ij} + \beta_E E_i$	$n = 5{,}000$, $p = 10$	**Figure 3**	For data simulated under one interaction, ER was more powerful than GB. For data simulated under 10 interactions, the two tests performed comparably.
		$n = 1{,}500$, $p = 10$	Web Figure 2 in the Supplementary Materials	
Missing Interactions (Continuous Covariates) - section 3.2				
$g_{A3}(\boldsymbol{\beta}, \boldsymbol{Z}_i) = \beta_{Z_1} Z_{i1} + \beta_{Z_2} Z_{i2} + \beta_{Z_1 Z_2} Z_{i1} Z_{i2}$	$g_{03}(\boldsymbol{\beta}, \boldsymbol{Z}_i) = \beta_{Z_1} Z_{i1} + \beta_{Z_2} Z_{i2}$	$n = 5{,}000$, $p = 10$	**Figure 4**	ER demonstrated a noticeable power increase over GB.
		$n = 1{,}500$, $p = 10$	Not Shown	
Additive Effects - section 3.3				
$g_{A4}(\boldsymbol{\beta}, \boldsymbol{G}_i) = log(1 + \sum_{j=1}^{p} \beta_{G_j} G_{ij})$	$g_{01}(\boldsymbol{\beta}, \boldsymbol{G}_i) = \sum_{j=1}^{p} \beta_{G_j} G_{ij}$	$n = 5{,}000$, $p = 5$	**Figure 5**	ER was more powerful than the GB to detect departures from the multiplicative model.
		$n = 5{,}000$, $p = 10$	Web Figure 5 in the Supplementary Materials	

Event times were simulated under the Weibull hazard model (Equation 3), $h_i(t; \boldsymbol{G}_i) = \lambda\alpha t^{\alpha-1} exp(g(\boldsymbol{\beta}, \boldsymbol{G}_i))$ with a constant baseline hazard, $\alpha = 1$, and λ chosen to provide a 20% event rate by time $t = 10$(years) in the absence of censoring. The function $g(\cdot)$ specifies the model for the joint risk of the disease associated with the genotype-covariate vector $\boldsymbol{G}_i$ and corresponding effect coefficients, $\boldsymbol{\beta}$. Simulation results were grouped into figures by sample size (n) and number (p) of covariates ($\boldsymbol{G}$ or $\boldsymbol{Z}$). See section 3 for complete simulation details.

TABLE 2 | Type-1 error of tests for constant baseline hazard.

Event rate	$p = 5$				$p = 10$			
	0% censoring		50% censoring		0% censoring		50% censoring	
	GB	ER	GB	ER	GB	ER	GB	ER
$n = 5{,}000$								
0.05	0.049	0.048	0.050	0.049	0.053	0.056	0.054	0.046
0.1	0.050	0.052	0.053	0.050	0.054	0.052	0.052	0.054
0.2	0.052	0.049	0.052	0.051	0.048	0.049	0.052	0.05
$n = 1{,}500$								
0.05	0.050	0.046	0.054	0.052	0.049	0.046	0.053	0.052
0.1	0.051	0.052	0.048	0.045	0.053	0.049	0.051	0.050
0.2	0.051	0.050	0.052	0.052	0.055	0.052	0.057	0.054

Event times simulated under the Weibull hazard model (Equation 3) with g_{01} for a constant baseline hazard, $\alpha = 1$, with p = 5 or 10 covariates (genotypes). Empirical size is presented for the Gronnesby and Borgen test (GB) and the proposed Extreme Risk test (ER). The empirical type 1 error was estimated from 10,000 simulated replicates at the nominal 5% level.

(Web Figure 3 in the Supplementary Materials). The results showed a noticeable power increase in ER over GB for this scenario (**Figure 4**).

As demonstrated in Web Figure 3 in the Supplementary Materials, predicted risk can be highly sensitive to quadratic effects (e.g., interaction terms) of normally distributed covariates. As an example, patients with large negative values for both Z_1 and Z_2 will be predicted to have low risk from a fitted model with positive estimated main effects. However, if a large positive interaction effect is omitted from the model, the actual observed risk will be much higher resulting from the product of the two negative values. This contributes to the more significant improvement of ER over GB with Gaussian covariates compared to categorical ones.

3.3. Power of the ER Test Under Misspecified Models-Additive Covariate Effects on the HR

Similar to Song et al. (2015), we examined the power of the ER test to detect model misspecification due to additive effects on the hazard. For this, we simulated data from (Equation 3) with $g_{A4}(\boldsymbol{\beta}, \boldsymbol{G}_i) = log(1 + \sum_{j=1}^{p} \beta_{G_j} G_{ij})$; it's easy to see that this corresponds to additive effects for a fixed baseline hazard, $h_0(t)$. We then fit the corresponding Cox model for (Equation 3) with g_{01}, which assumes a multiplicative effect of $\boldsymbol{G}$ on the hazard. We expected that the deviation from the assumed model of multiplicative effects on the HR would show increased deviation in observed vs. expected survival in the extremes of

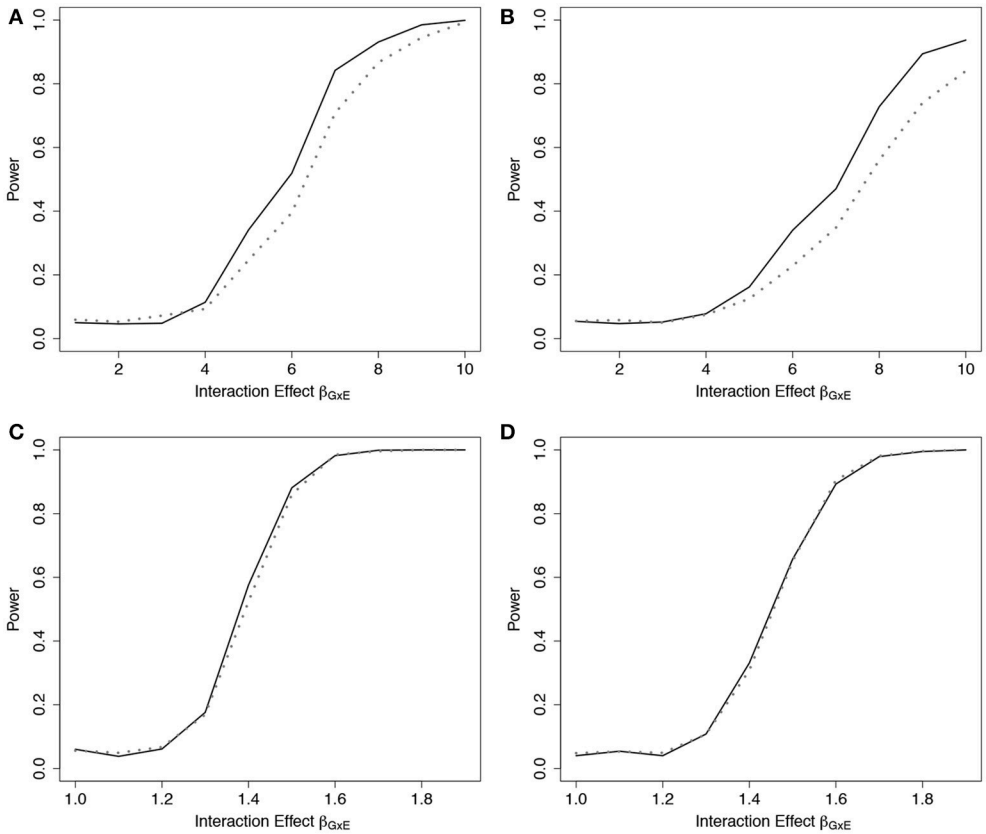

FIGURE 3 | Power of tests under missing interaction models g_{01}. Power of the ER (solid line) and GB (dotted line) tests when main effect of *E* and ***G***x*E* interaction terms are missing from the fitted Cox model (3), using g_{01}. $n = 5{,}000$ for the 10-SNP model with one interaction **(A**,**B)** or ten interactions **(C**,**D)** in the underlying true model g_{A1}, and either 0% **(A**,**C)** or 50% **(B**,**D)** lost to follow-up censoring.

the subject risk distribution (Web Figure 4 in the Supplementary Materials).

We simulated data with a fixed β_{G_j} across all SNPs so that the marginal HR for each SNP in the fitted Cox model would range between 1 and 1.2, and between 1 and 1.4, for models with $p = 10$ SNPs and 5 SNPs, respectively. We specified α at 1, corresponding to a constant baseline hazard. Under each scenario, we fixed the event rate prior to 10 years at 20 and 50% in the absence of censoring.

Simulation results demonstrated that the ER was more powerful than the GB to detect departures from the multiplicative model when the true model was additive under the 5-SNP and 10-SNP models (**Figure 5** and Web Figure 5 in the Supplementary Materials, respectively).

4. APPLICATION TO CYSTIC FIBROSIS-RELATED DIABETES

Cystic Fibrosis (CF) is a life-limiting recessive genetic disorder caused by mutations in the CF transmembrane conductance regulator (*CFTR*). CF affects multiple organs including the pancreas. As CF patients with severe *CFTR* mutations age, there is an increased risk of CF-related diabetes (CFRD), with a prevalence of 40% by the fourth decade of life (Moran et al., 2009). Uncontrolled CFRD is associated with muscle loss and declining lung function, which can be prevented by early detection and treatment. Models that predict CF individuals at high risk of developing CFRD could enable targeted surveillance programs with more frequent glucose monitoring so that CFRD is diagnosed early.

Multiple genetic factors, beyond *CFTR*, have been shown to contribute to CFRD including Type 2 diabetes susceptibility genes (Blackman et al., 2013) and CF-specific modifier genes (Li et al., 2014). Age-dependent predictive models (Heagerty et al., 2000) for CFRD based on genetic markers could be applied at birth to determine individuals at high risk, potentially benefiting the length and quality of life for individuals living with CF. However, if a CFRD risk model is well calibrated globally but poorly calibrated for the low or high risk groups, we question the utility of the model.

With CFRD event times based on data from the Canadian Cystic Fibrosis Gene Modifier Study we build and calibrate a predictive model for CFRD using a Cox PH model that includes as predictors six SNPs from six risk genes (coded additively) in addition to indicator variables for *CFTR* genotype

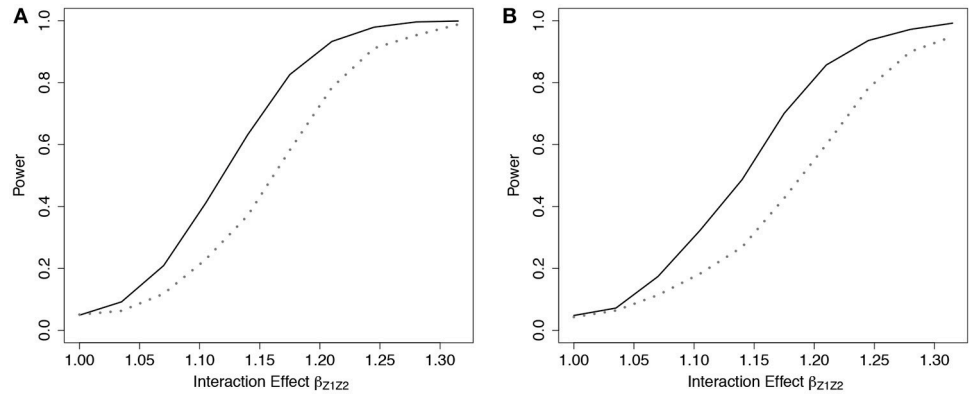

FIGURE 4 | Power of tests under missing interaction models with Gaussian covariates g_{03}. Power of the ER (solid line) and GB (dotted line) tests when interaction effect of Z_1Z_2 is missing from the fitted Cox model (3), using g_{03}. $n = 5{,}000$ for the 2-covariate model with one interaction in the underlying model g_{A3}, with either 0% **(A)** or 50% **(B)** lost to follow-up censoring. $Z_1, Z_2 \sim \mathcal{N}(0, 1)$.

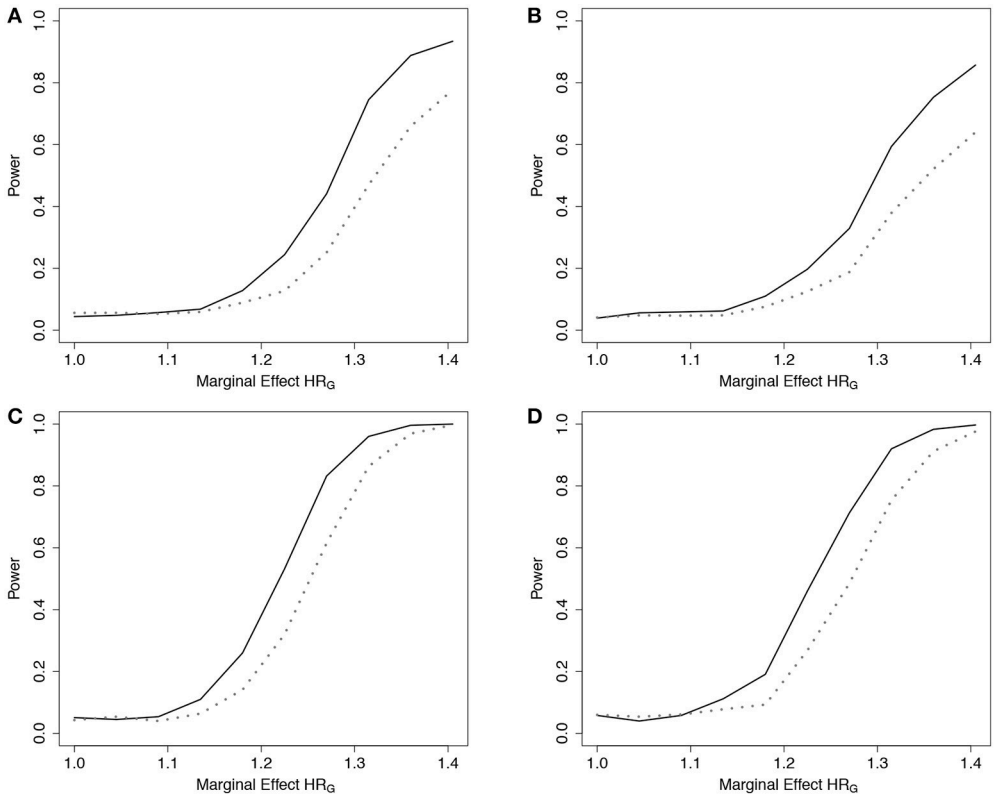

FIGURE 5 | Power of tests under additive effect models with 5 SNPs. Power of the ER (solid line) and GB (dotted line) tests for detecting departures from the multiplicative model with 5 SNPs. Data simulated under additive effects on the HR, g_{A4}, and fit using the multiplicative effects Cox model (3), with g_{01}. $n = 5{,}000$ for the 5-SNP model with event rate 20% **(A**,**B)** or 50% **(C**,**D)**, and either 0% **(A**,**C)** or 50% **(B**,**D)** lost to follow-up censoring. $G_j \sim Bin(2, 0.3)$.

severity and sex. Five of the genes (*SLC26A9*, *TCF7L2*, *CDKAL1*, *CDKN2A/B*, *IGF2BP2*) were previously identified in Blackman et al. (2013) and the sixth gene, *PRSS1*, encodes the enzyme cationic trypsinogen which is a biomarker of CFRD at birth (Soave et al., 2014).

Analysis of 1,330 unrelated CF patients with complete information on the eight covariates is presented. Details of data collection, CFRD diagnosis, genotyping, and quality control procedures are reported elsewhere (Sun et al., 2012; Soave et al., 2014). Of the 1,330 included in the analysis, 203 patients had a CFRD diagnosis and the median age in years at last study visit (or diabetes) was 16.2 (21.6). To illustrate the fit of the model across the distribution of risks, we plot the observed vs. expected average absolute risk for each of 11 risk groups stratified according to

their ordered expected risks (**Figure 6**). Ideally the points should cluster around the identity line, and if the model fits the data well no point should display a large deviation. However, for the groups with larger expected average risk, the deviations are large (**Figure 6**). For patients in the highest risk group, the model appears to over-estimate risk on average. The *p*-value for the ER test for this model is 0.047, whereas, the global GB test calculates a *p*-value = 0.2. This discrepancy is not surprising since ER is designed to have greater power than GB when the bias is greatest in the tails of the risk distribution.

The observed bias in risk-prediction for these patients, therefore, gives reason to reconsider the current model. A number of model fitting issues could contribute to the lack-of-fit in the tails, as we have demonstrated in this paper. Additional analyses involving interaction effects, scaling of covariates, and appropriateness of the multiplicative effects assumption should be considered.

5. DISCUSSION

The Cox PH model for time-to-event data is straightforward to implement and does not require specification of a distribution for survival times. However, the Cox model does make several strong assumptions that may not be appropriate. Evaluation of a given model as a prediction tool requires assessment of both discrimination and calibration. Since the time and cost to obtain a second independent sample for model validation may be great, calibration assessment in a training sample can provide valuable information. Unfortunately, calibration is rarely reported (Collins et al., 2014).

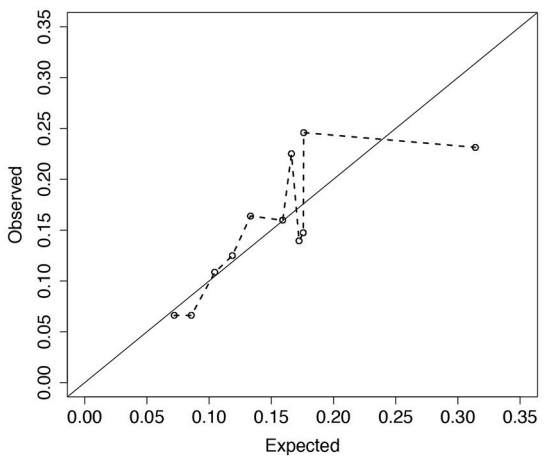

FIGURE 6 | Observed vs. expected average number of CFRD events across 11 risk groups. Risk groups were defined according to the Cox PH model for CFRD onset risk with 8 covariates, $\boldsymbol{z}_i$ (see section 4 for details). Expected number of events for each subject was calculated as the estimated cumulative hazard from the fitted Cox model with 8 covariates, $\widehat{\Lambda}_i(x_i) = -log(\widehat{S}_i(x_i))$, where $\widehat{S}_i(x_i) = \widehat{S}_0(x_i)^{exp(\boldsymbol{z}_i^T \widehat{\boldsymbol{\beta}})}$. Observed number of events for each subject was calculated similarly using a Cox model with the risk-group indicator variables ($\boldsymbol{K}_i$) replacing the 8 covariates. Larger deviations from the diagonal line correspond to larger risk-group sums of martingales. Calibration test results for ER and GB were $p = 0.047$ and 0.21, respectively.

The GOF test of Gronnesby and Borgan (1996) has been proposed as an omnibus test for global lack-of-fit assessment in Cox models. While this and other GOF methods have been shown to perform reasonably well as global tests, their power might be limited to detect bias of predicted risk at the extremes of the risk distribution where clinical decisions are generally made. Here, we proposed a new ER calibration test designed to examine the accuracy of risk predictions for patients at extreme (high or low) risk to be used alongside existing methods. Due to its construction, the distribution of the ER test statistic is intractable, and therefore we demonstrated a simulation method to estimate empirical significance that can be easily implemented using existing software.

Model misspecification can result in poorly calibrated risk estimates that are not detectable from standard GOF tests. Prediction tools that do not account for important interactions are likely to produce biased estimates of risk at the extreme tails of the population risk distribution, and these deviations should be detected more effectively through the proposed ER test. The simulation examples in section 3.2 indicate that the ER test could have increased power over the GB test to detect model misspecification when there are missing interactions. Although additive models may be a more natural starting point because they correspond to simple independent effects on the underlying risk factors (Weinberg, 1986), they are rarely implemented due to less convenient statistical properties. Similar to the logistic regression model (Song et al., 2015), model misspecification of the Cox model due to the assumption of multiplicative effects can also create bias at the extremes of the risk distribution. We observed that an incorrect multiplicative effects assumption in the Cox model can lead to dramatically underestimated survival probabilities for both extreme high and low risk groups when the underlying effects are additive (Web Figure 4 in the Supplementary Materials). In section 3.3, we showed that the ER test has advantages over the GB test in detecting the resulting bias.

We recognize that among competing calibration tests for time-to-event data, no single test will be most powerful for detecting lack-of-fit in all situations. Certainly, for model misspecification that results in bias over much of the range of estimated risks, the GB test should demonstrate greater power over a "max" test that simultaneously considers multiple subgroups of the data, such as the ER test. In section 3.2, we observed that the GB test slightly outperformed the ER test when very large interaction effect(s) were omitted from the working model but all main effects were included. This is likely because the larger interaction effects, when omitted, create deviations in predicted vs. observed risk over more of the risk distribution range, not just in the tails. In light of this, we recommend that our test should be thought of as a complementary tool when examining calibration of a risk model. Such an assessment should include visual inspection of a calibration plot similar to **Figure 6** that can also be useful in explaining situations where the ER and GB tests give contradictory results.

Based on the convention for collapsing groups with fewer than 5 expected events, we found that collapsing was rarely required for the sample sizes of $n = 5{,}000$. For samples of $n = 1{,}500$ observations, collapsing occurred, however, it had little

impact on type 1 error control (**Table 2** and Web Table 3 in the Supplementary Materials). The power of the ER and GB tests compared with and without collapsing was also quite similar under most simulation models considered. Not surprisingly, for simulation models that resulted in frequent collapsing to fewer than 5 groups, there was a large decrease in power after applying the collapsing rule. Thus, for both type 1 error and power considerations, we recommend applying the ER test without collapsing, provided the sample size is no smaller than considered here, $n = 1{,}500$.

Our development of the ER test here focuses on detecting lack-of-fit for the purpose of internal model validation in a training dataset. Often, we need to evaluate the performance of a model in a new external cohort, where we might examine predicted survival probabilities. The implementation of both the GB and ER tests, however, only assesses accuracy of the linear predictor (estimated model coefficients) and does not incorporate information about the baseline hazard. As a result, these tests would be insensitive, in an external dataset, to detect any systematic bias (high or low) of predicted survival probabilities that require the baseline hazard estimated in the training dataset.

As the research community amasses new information about the molecular basis for disease, progress toward personalized medicine is being realized, revolutionizing disease treatment and preventative care. Accurate assessment of individual risk, incorporating both genetic and environmental factors plays a critical role in this initiative. Calibration tests such as the ER will become integral to risk-model determination to safeguard against biased risk-estimates for extreme risk patients, potentially most affected by clinical decisions.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of The Hospital for Sick Children, Research Ethics Board. The protocol was approved by the Research Ethics Board. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

DS and LS developed the method and wrote the manuscript. DS performed all analyses.

ACKNOWLEDGMENTS

The authors thank Professor Jerry F. Lawless and Dr. Lei Sun for helpful suggestions and critical reading of the original version of the paper. This work was funded by the Canadian Institutes of Health Research (CIHR; MOP-258916 to LS); the Natural Sciences and Engineering Research Council of Canada (NSERC; 371399-2009 to LS); Cystic Fibrosis Canada #2626 (to LS). DS is a trainee of the CIHR STAGE (Strategic Training in Advanced Genetic Epidemiology) training program at the University of Toronto and is a recipient of the SickKids Restracomp Studentship Award and the Ontario Graduate Scholarship (OGS).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00177/full#supplementary-material

REFERENCES

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Methods Based on Counting Processes*. New York, NY: Springer-Verlag.

Barlow, W. E., and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika* 75, 65–74. doi: 10.1093/biomet/75.1.65

Blackman, S. M., Commander, C. W., Watson, C., Arcara, K. M., Strug, L. J., Stonebraker, J. R., et al. (2013). Genetic modifiers of cystic fibrosis-related diabetes. *Diabetes* 62, 3627–3635. doi: 10.2337/db13-0510

Breslow, N. (1972). Discussion on regression models and life-tables(by dr cox). *J. Roy. Statist. Soc. Ser. B* 34, 216–217.

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., et al. (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med. Res. Methodol.* 14:40. doi: 10.1186/1471-2288-14-40

Crowson, C. S., Atkinson, E. J., and Therneau, T. M. (2016). Assessing calibration of prognostic risk scores. *Stat. Methods Med. Res.* 25, 1692–1706. doi: 10.1177/0962280213497434

D'Agostino, R. B., and Nam, B.-H. (2003). "Evaluation of the performance of survival analysis models: Discrimination and calibration measures," in *Handbook of Statistics*, Vol. 23 (Elsevier), 1–25.

Demler, O. V., Paynter, N. P., and Cook, N. R. (2015). Tests of calibration and goodness-of-fit in the survival setting. *Stat. Med.* 34, 1659–1680. doi: 10.1002/sim.6428

Gronnesby, J. K., and Borgan, O. (1996). A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Anal.* 2, 315–328. doi: 10.1007/BF00127305

Harrell, F. E. (2001). *Regression Modeling Strategies : With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer series in statistics. New York, NY: Springer.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics* 56, 337–344. doi: 10.1111/j.0006-341X.2000.00337.x

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Stat. Med.* 16, 965–980. doi: 10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical nalysis of Failure Time Data, 2nd Edn*. Wiley Series in Probability and statistics. Hoboken, NJ: J. Wiley.

Li, W., Soave, D., Miller, M. R., Keenan, K., Lin, F., Gong, J., et al. (2014). Unraveling the complex genetic model for cystic fibrosis: pleiotropic effects of modifier genes on early cystic fibrosis-related morbidities. *Hum. Genet.* 133, 151–161. doi: 10.1007/s00439-013-1363-7

Lin, D. Y., Wei, L. J., and Ying, Z. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika* 80, 557–572. doi: 10.1093/biomet/80.3.557

May, S., and Hosmer, D. W. (1998). A simplified method of calculating an overall goodness-of-fit test for the cox proportional hazards model. *Lifetime Data Anal.* 4, 109–120. doi: 10.1023/A:1009612305785

May, S., and Hosmer, D. W. (2004). A cautionary note on the use of the gronnesby and borgan goodness-of-fit test for the cox proportional hazards model. *Lifetime Data Anal.* 10, 283–291. doi: 10.1023/B:LIDA.0000036393.29224.1d

Moons, K. G., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., et al. (2012). Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 98, 683–690. doi: 10.1136/heartjnl-2011-301246

Moran, A., Dunitz, J., Nathan, B., Saeed, A., Holme, B., and Thomas, W. (2009). Cystic fibrosis-related diabetes: current trends in prevalence, incidence, and mortality. *Diab. Care* 32, 1626–1631. doi: 10.2337/dc09-0586

Parzen, M., and Lipsitz, S. R. (1999). A global goodness-of-fit statistic for cox regression models. *Biometrics* 55, 580–584. doi: 10.1111/j.0006-341X.1999.00580.x

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* 69, 239–241. doi: 10.1093/biomet/69.1.239

Soave, D., Corvol, H., Panjwani, N., Gong, J., Li, W., Boelle, P. Y., et al. (2015). A joint location-scale test improves power to detect associated snps, gene sets, and pathways. *Am. J. Hum. Genet.* 97, 125–138. doi: 10.1016/j.ajhg.2015.05.015

Soave, D., Miller, M. R., Keenan, K., Li, W., Gong, J., Ip, W., et al. (2014). Evidence for a causal relationship between early exocrine pancreatic disease and cystic fibrosis-related diabetes: a mendelian randomization study. *Diabetes* 63, 2114–2119. doi: 10.2337/db13-1464

Song, M., Kraft, P., Joshi, A. D., Barrdahl, M., and Chatterjee, N. (2015). Testing calibration of risk models at extremes of disease risk. *Biostatistics* 16, 143–154. doi: 10.1093/biostatistics/kxu034

Sun, L., Rommens, J. M., Corvol, H., Li, W., Li, X., Chiang, T. A., et al. (2012). Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nat. Genet.* 44, 562–569. doi: 10.1038/ng.2221

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* 77, 147–160. doi: 10.1093/biomet/77.1.147

Weinberg, C. R. (1986). Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. *Am. J. Epidemiol.* 123, 162–173. doi: 10.1093/oxfordjournals.aje.a114211

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JY and handling Editor declared their shared affiliation.