Research Article

Statistics in Medicine

Received 23 May 2014, Accepted 6 January 2015 Published online 11 February 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6428

Tests of calibration and goodness-of-fit in the survival setting

Olga V. Demler,*† Nina P. Paynter and Nancy R. Cook

To access the calibration of a predictive model in a survival analysis setting, several authors have extended the Hosmer–Lemeshow goodness-of-fit test to survival data. Grønnesby and Borgan developed a test under the proportional hazards assumption, and Nam and D'Agostino developed a nonparametric test that is applicable in a more general survival setting for data with limited censoring. We analyze the performance of the two tests and show that the Grønnesby–Borgan test attains appropriate size in a variety of settings, whereas the Nam-D'Agostino method has a higher than nominal Type 1 error when there is more than trivial censoring. Both tests are sensitive to small cell sizes. We develop a modification of the Nam-D'Agostino test to allow for higher censoring rates. We show that this modified Nam-D'Agostino test has appropriate control of Type 1 error and comparable power to the Grønnesby–Borgan test and is applicable to settings other than proportional hazards. We also discuss the application to small cell sizes. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: calibration; survival analysis; goodness-of-fit

1. Introduction

Risk prediction models are a centerpiece for clinical decision making and prediction. Models such as Gail's model for 10-year risk of breast cancer [1] or Framingham 10-year coronary heart disease (CHD) risk model [2–4] are used clinically for treatment decisions. It is critically important to have valid and objective means of evaluating performance of risk prediction models.

Calibration is one of the most important model performance characteristics because a miscalibrated model produces invalid risk estimates [5] and can introduce errors into decision making. The Hosmer–Lemeshow (HL) goodness-of-fit test is often used as a test of calibration [6, 7]. Originally developed for the logistic regression model, it was extended for survival data by Nam and D'Agostino [7] (ND test) and Grønnesby and Borgan (GB test) [8]. The latter group used martingale theory to develop a goodness-of-fit test for the proportional hazards (PH) regression model. While both tests perform well in their proposed settings, recent reports [9, 10] suggest that ND test has incorrect size for moderate to high censoring rates (above 15%). While May and Hosmer [11] showed that the GB test has incorrect size in other settings.

In this paper, we show that the GB test achieves the target size in a wide range of simulations in the model development setting. We demonstrate in our practical example, however, that the GB test can be insensitive to miscalibration when applied in the out-of-sample calibration context. Therefore, the GB test can be viewed as a pure goodness-of-fit test only in the model development setting and cannot be used as a test for external calibration. We propose a new calibration test based on the ND approach that is applicable in a more general setting than the GB test and which is not affected by the limitations of the ND test.

2. Model formulation

For all individuals, we record time when an outcome (for example a 10-year CHD event) has occurred or the time of censoring, which is the end of study/lost to follow up time. Denote the end of study time

Division of Preventive Medicine, Brigham and Women's Hospital Harvard Medical School, 900 Commonwealth Ave., East Boston, MA 02215, U.S.A

*Correspondence to: Olga V. Demler, Division of Preventive Medicine, Brigham and Women's Hospital Harvard Medical School, 900 Commonwealth Ave., East Boston, MA 02215, U.S.A.

†E-mail: olgademler@gmail.com

as T, that is, for 10-year CHD outcome $T=10$. In addition to the time variable and the censored/event indicator, for each person we collect information on fixed covariates $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}$ measured at baseline. For this paper, we assume that event times are right-censored. To model this kind of data, we can use any technique developed for survival right-censored data. For example, when the proportionality of hazards assumption is true, one can use the Cox PH regression model, otherwise use a parametric or nonparametric regression model for censored survival data. For each person, we calculate the predicted probability of an event. For cross-sectional data (no time of event) or data with a fixed time to event, Hosmer and Lemeshow developed a goodness-of-fit test to evaluate model fit in 1964 using a binary regression model (logistic regression). Data are divided into subgroups based on the deciles or other groupings of predicted probabilities. The HL goodness-of-fit method tests whether the average of the predicted probabilities follows the observed event rate across the deciles. Grønnesby, Borgan and Nam and D'Agostino used two different approaches to extend this test for survival data. In all the aforementioned tests and in this paper, we assume that the model is a good fit for the data under the null.

2.1. Grønnesby and Borgan test

Grønnesby and Borgan developed a goodness-of-fit test for the Cox PH regression model using a counting process approach. For $t \leqslant T$, define $N_i(t) =$ the number of observed events in our data for person i in $[0,t]$. Here, we consider one event per person so that $N_i(t)$ is just an event indicator for person i by time t. Using standard counting process notation [8], $N_i(t)$ is generated by an intensity process $\lambda_i(t) = Y_i(t)h_i(t)$, where $h_i(t)$ is a hazard rate and $Y_i(t)$ is at-risk indicator for person i. Under the assumptions of the Cox model, $h_i(t)$ is modeled as follows:

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$$

where $\boldsymbol{x}_i$ is the vector of p covariates for person i. Therefore, $N_i(t)$ is generated by $\lambda_i(t) = h_0(t)\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) Y_i(t)$.

The sum of $N_i(t)$ in group g is just the observed count of events in group g by time t.

$$\sum_{i\in g} N_i(t) = \text{observed}(t)$$

The expected number of events can be calculated as a sum of cumulative intensities in group g up through time t:

$$\sum\nolimits_{i\in g}\int_0^t \lambda_i(s)ds = \sum\nolimits_{i\in g}\int_0^t h_0(s)\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) Y_i(s)ds = \text{expected(t)} \qquad (1)$$

The difference between the number observed and expected in group g will tell us how close our model fits the data. $H_g(t) = \text{observed(t)} - \text{expected(t)} =$

$$\sum\nolimits_{i\in g}\mathrm{N}_i(\mathrm{t}) - \sum\nolimits_{i\in g}\int_0^t h_0(s)\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) Y_i(s) = \sum\nolimits_{i\in g} M_i(t) = H_g(t) \qquad (2)$$

where $M_i(t) = N_i(t) - \int_0^t h_0(s)\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) Y_i(s)ds$ in formula (2) is a martingale residual. We will denote it as $\hat{M}_i(t)$ when we use estimates of $\boldsymbol{\beta}$ in $M_i(t)$. $\hat{H}_g(t)$ is a group-wise sum of approximate martingale residuals. Using martingale theory, Grønnesby and Borgan show how to calculate $\hat{\Sigma}(t)$ the variance–covariance matrix of $(H_1(t), \ldots, H_{G-1}(t))$ and prove that

$$\chi^2_{GB}(t) = \left(\widehat{H_1(t), \ldots, H_{G-1}(t)}\right)\hat{\Sigma}^{-1}(t)\left(\widehat{H_1(t), \ldots, H_{G-1}(t)}\right)^T \sim \chi^2_{G-1} \qquad (3)$$

May and Hosmer [12] proved that $\chi^2_{GB}(t)$ is algebraically equivalent to the well-known score test statistic, which is available in most standard software packages. If the score test is not available, the likelihood ratio test statistics and Wald test are asymptotically equivalent to the score test [13]. Although May and Hosmer showed equivalency at time $= \infty$, $\chi^2_{GB}(t)$ is equivalent to the score test for any fixed a priori time $t < \infty$ as long as we do not use data beyond time t. Note that the test is inherently conditional on the censoring distribution in the observed data through the at-risk indicator $Y_i(t)$. The martingales are based on the observed and expected numbers, given the censoring times rather than at the end of a common time interval.

2.2. *Nam–D'Agostino test*

Nam and D'Agostino [7] introduced a goodness-of-fit test for survival data. They suggested splitting the data into groups based on the risk scores (for example, into deciles). To test how well the model fits the data in each decile, they define a test statistic:

$$\chi^2_{ND}(t) = \sum\nolimits_{g=1}^{G} \frac{\left[KM_g(t) - \overline{p(t)}_g\right]^2 n_g}{\overline{p(t)}_g \left(1 - \overline{p(t)}_g\right)} \tag{4}$$

where $KM_g(t)$ is the Kaplan–Meier failure probability in the gth decile at time t, $\overline{p(t)}_g$ is the mean predicted probability of failure for subjects in gth decile using any survival modeling technique, and n_g is the number of observations in a group g. Under the null, $\chi^2_{ND}(t)$ is distributed as a chi-square random variable with G-1 degrees of freedom.

The ND test statistic in (1) can be rewritten in a more familiar 'observed'–'expected' form:

$$\begin{aligned}\chi^2_{ND}(t) &= \sum\nolimits_{g=1}^{G} \frac{\left[KM_g(t) - \overline{p(t)}_g\right]^2 n_g}{\overline{p(t)}_g \left(1 - \overline{p(t)}_g\right)} = \sum\nolimits_{g=1}^{G} \frac{\left[n_g KM_g(t) - n_g\overline{p(t)}_g\right]^2}{n_g\overline{p(t)}_g \left(1 - \overline{p(t)}_g\right)} \\ &= \sum\nolimits_{g=1}^{G} \frac{\left[\text{observed}(t) - \text{expected}(t)\right]^2}{n_g\overline{p(t)}_g \left(1 - \overline{p(t)}_g\right)}\end{aligned} \tag{5}$$

where $n_g KM_g(t)$ is an estimator of the mean observed number of events by time t in n_g trials had there been no censoring. The average number of events in each group is $n_g KM(t)_g$(= 'observed'), and $n_g\overline{p(t)}_g$ is an estimate of the 'expected' number of events if the model is correct. The ND formula now has the familiar form of the HL statistic adapted for survival data. Note that the numerator differs from the sum of the martingales used in the GB statistic because the observed and expected events in the ND test are computed as if there were no censoring up to time t.

2.3. *Comparison of Grønnesby–Borgan and Nam–D'Agostino approaches*

Time is one of the most fundamental features of survival analysis and is always present explicitly or implicitly in tests and procedures. For example, Hosmer *et. al.* [13] point out that if the groups in the GB test are related to the probability of an event (i.e., top decile contains observations with the highest probability of the event), then proportionality of hazards allows forming the groups based on the risk score alone or equivalently on predicted probabilities estimated at time t, where t is common for all subjects. The same is true for the ND test if the expected number of events is calculated using the Cox PH model. However, if predicted risk is not a monotonic function of a linear risk score, then groups should be formed explicitly on predicted probabilities at some common time t. Also, note that it does not make sense to form deciles based on the survival probability estimated at different follow-up times for different subjects, because if the baseline hazard is nonconstant, risk would be calculated over different periods in the same decile ([13] Section 6.5).

Additionally, note that both tests themselves use time t in (3) and (4). The ND and GB tests can be estimated for $t = T$ but can be evaluated at some $t < T$ as well, if goodness-of-fit is of interest at some intermediate time t. In the GB test, to compare the observed and expected number of events by time $t < T$, individuals should be artificially censored at time t. This would ignore any events after t and potentially lose information in estimating the regression parameters of the Cox model. The ND test, however, could use later information to estimate the hazard ratios (HRs) or other model form but then estimate the predicted probabilities at time t. Note, however, that if T extends much further than t, such as estimating 10-year risk using 30 years of follow-up, then proportionality of hazards and other model assumptions may not hold and risk estimates may not be accurate.

Censoring is another fundamental feature of survival analysis. The two tests are conceptually different in their treatment of censoring. First, note that $\sum_{i \in g} N_i(t)$ in GB formula (2) is the unadjusted raw observed count in group g. Second, $n_g KM_g(t)$ is the observed count in the numerator of ND (5). $KM_g(t)$ is a Kaplan–Meier consistent estimator of survival probability, which by definition is unaffected by censoring. Therefore, the ND test is based on observed counts that are scaled up to compensate for censoring, whereas GB is using actual observed counts. The corresponding expected numbers used in the GB test are also as of the individual's follow-up time, while they are at time t in the ND test.

Finally, the GB test utilizes the powerful and efficient counting process theory. However, it is limited to the Cox PH model. In contrast, the ND test does not make any specific modeling assumptions. Therefore, $\overline{p(t)}_g$ in (4) and (5) may be estimated by any survival modeling technique, including statistical learning methods [14].

3. Performance of the Nam-D'Agostino goodness-of-fit test

Paynter *et al.* (for the ND test) and Guffey and May (for both tests) reported incorrect size for several settings. In this paper, we demonstrate how the two goodness-of-fit tests perform in a typical cohort study. Cohort studies often estimate risk as of a specific follow-up time, that is, 10-year risk of CHD, which can have in a population a low event rate such as 0.05. Censoring will further reduce the observed number of events. It is important to understand if and by how much the performance will deteriorate because of the small number of events and to find the best way to resolve problems with the performance.

3.1. Simulations set up

The goal of our simulations is to mimic a cohort study with a specific disease outcome (i.e., 10-year CHD event or 5-year risk of breast cancer). Using population event rates of 5%, 10%, or 40%, we simulate data under the assumption of a PH Cox regression model with three baseline hazard functions: decreasing, constant, and increasing. Event times follow the Weibull distribution $F(t) = \exp\left(-t^{\alpha}\lambda \exp\left(\beta_1 x_1 + \beta_2 x_2\right)\right)$.

In this paper, x_1 is called 'an old predictor', and x_2 is called 'a new predictor'.

Parameters α and λ are calculated in such a way that 5%, 10%, or 40% of observations will have an event by year 10 in the absence of censoring. This event time distribution translates into the following form of baseline hazard:

$$h_0(t) = \alpha t^{\alpha-1}\lambda$$

where α is a shape parameter. We used $\alpha \in \{0.3, 1.0, 3.0\}$, corresponding to monotone decreasing, constant, and increasing baseline hazard functions with respect to time t. λ is a scale parameter. It is calculated to ensure event rates of 0.05, 0.1, and 0.4 by year 10. β_1 is chosen so that HR(x_1) was 8.0, and β_2 corresponds to a set of HRs of x_2: $\{1, 1.3, 3.0, 8.0, 15.0\}$. Event times greater than 10 years are censored (administrative censoring). In addition, censoring due to lost to follow-up is implemented by generating uniformly distributed censoring times for all observations and ensuring that 25% (in one setting) or 50% (in another setting) of observations have censored times less than 10. A sample size of $N = 5000$ was used to generate data in all presented simulations unless indicated otherwise.

3.2. Dealing with sparse deciles

Ideally, a well-calibrated model performs well in a variety of divisions into groups. Asymptotic properties of all survival estimators (as $KM_g(t)$ in formula (4)) rely on having a sufficiently large expected number of events in g-th decile. May and Hosmer [11] show that GB test becomes too liberal when there are too few events per decile. We observed similar behavior in our simulations. To ensure a sufficient number of events per decile, May and Hosmer recommend either decreasing the number of groups or ignoring deciles with too few number of events. The latter strategy results in effectively discarding part of the data. We used a collapsing strategy instead. We started with 10 deciles but collapsed small deciles with their closest neighbors, until all groups contained a predefined minimum number of events. This strategy always uses all the data while ensuring convergence of the estimators. It will often lead to division into a sufficient number of groups, which could also help encourage a chi-square distribution for our test statistic. We tested several collapsing rules: collapse if less than two, five, and 20 events. Per our extensive simulations (Section 4.1), the collapse if less than five rule performs better than its two other alternatives, so we limit our results to that scenario.

3.3. Simulations results

In their original paper, Nam and D'Agostino [7] introduced their formula (4) for Framingham-type data–a cohort study with little censoring. Cook and Ridker [15] used a similar approach for the Women's Health Study (WHS), which also has low censoring rate: over 8 years of follow-up, only 1.87% of women were censored in WHS. Figure 1 demonstrates that the ND test has appropriate size for such low censoring scenarios.

Oftentimes, in practice, higher censoring rates are observed in a study [16]. Crowson *et al.* [17] note that the ND test ignores censoring. In order to apply the ND test to a more general situation, we first consider how it performs for higher censoring rates. Figure 2 demonstrates that at 50% censoring, the test rejects the null hypothesis about 15% of the time, which is much more often than we would have expected at a 5% significance level.

A possible explanation of this deterioration can be found by looking at the variance in the denominator of (4):

$$\frac{\left[KM_g(t)-\overline{p(t)}_g\right]^2 n_g}{\overline{p(t)}_g\left(1-\overline{p(t)}_g\right)}$$

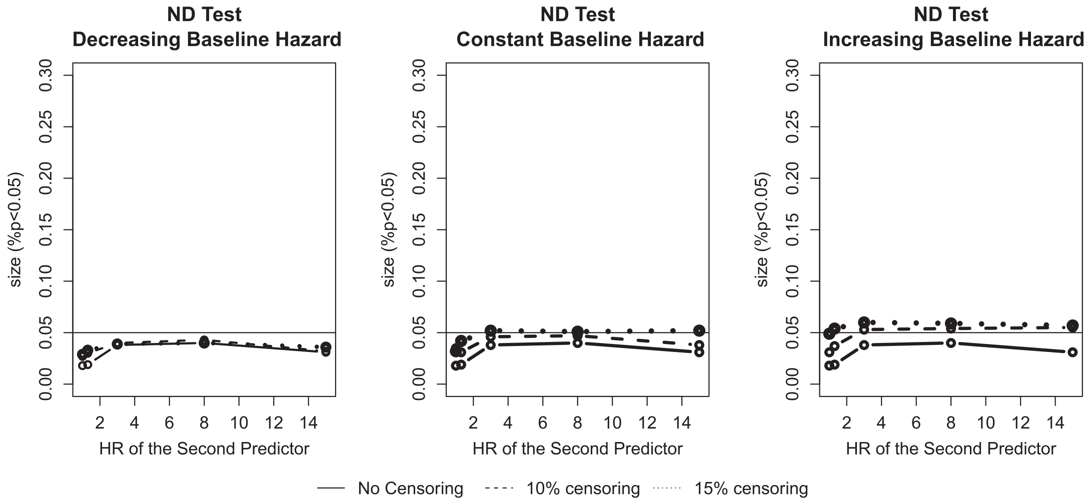

Figure 1. The size of the Nam–D'Agostino (ND) test for a low censoring rate for decreasing, constant, and increasing baseline hazards. The population event incidence is 10%. HR, hazard ratio.

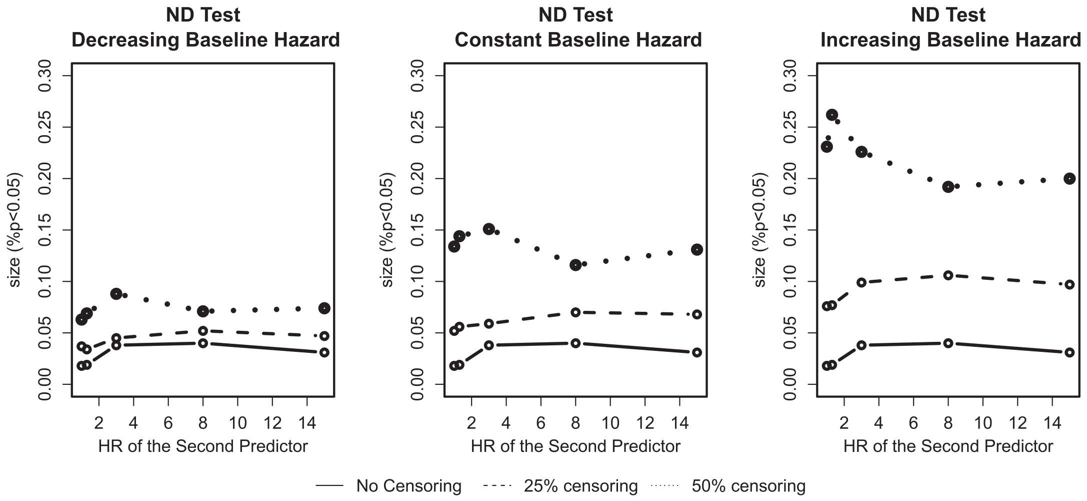

Figure 2. The size of the Nam–D'Agostino (ND) test for a high censoring rate for decreasing, constant, and increasing baseline hazards. The population event incidence is 10%. HR, hazard ratio.

It is a variance of a binomial probability estimator in n_g trials (size of gth decile). However, in the numerator, we have the estimator of the-survival probability in the presence of censoring. So, the more censoring there is, the less stable the numerator will be. Also, varying event times should be taken into account. Therefore, in the next section, we suggest using the Greenwood estimator of the variance of the probability of failure in the denominator.

4. New test using the Greenwood variance formula

The denominator of the ND test statistic (4) can be interpreted as the squared standard error of the binary proportion estimator $\bar{p}_g$. The binary proportion estimator was developed for binary data and does not account for censoring nor for time of event. It is appropriate to use it in the setting of a fixed event time, even though there the actual survival times are not taken into account. The Greenwood estimator is a consistent estimator of $Var(KMg(t))$ in a survival analysis setting and performed well in simulations [18]. We suggest using the Greenwood formula for the variance of failure probability $KM_g(t)$ in the denominator of ND (2). The Greenwood variance estimator $KM_g(t)$ can be written as follows [18]:

$$Var\left(KM_g(t)\right) = KM_g(t)^2 \sum\nolimits_{i|t_i \leqslant t} \frac{d_i}{n_i(n_i - d_i)},$$

where d_i and n_i are the number of failures and number at risk at time t_i The proposed Greenwood-Nam-D'Agostino (GND) statistic is

$$\chi^2_{GND}(t) = \sum\nolimits_{j=1}^{G} \frac{\left[KM_g(t) - \overline{p(t)}g\right]^2}{Var\left(KM_g(t)\right)} \sim \chi_{G-1} \tag{6}$$

In the Appendix, we prove that in the absence of censoring, GND test can be written as $\chi^2_{GND}(t) = \sum_{g=1}^{G} \frac{\left[observed_g - expected_g\right]^2}{n_g\hat{p}_g(1-\hat{p}_g)}$, where $\hat{p}_g = \frac{observed_g}{n_g}$. In this situation, it is exactly equal to the ND test, which is distributed as χ_{G-1} under the null. In the absence of censoring, the GND test can be viewed as a variation of the HL test [19] (Appendix).

5. Performance of Grønnesby–Borgan and Greenwood–Nam–D'Agostino goodness-of-fit tests

In this section, we run extensive simulations of a variety of scenarios to evaluate the performance of goodness-of-fit tests for survival data. We focus on two tests: the GB and the new GND tests. We include the GB test because it is one of the best tests to use in terms of performance [8] and ease of use [12] for PH survival data. The GND test is applicable in a wider variety of situations. In Figure 3, we plot the size of the tests: ND (4), GB (3), and GND(6) versus the hazard ratio of the second predictor variable x_2. One thousand simulations of sample size N=5000 were generated under the null; survival times were generated according to PH model with two normally distributed ($N(0,0.5)$) predictors (x_1 and x_2) and incidence rate of 0.1. In the top, middle, and lower rows, we present results for decreasing, constant, and increasing baseline hazard functions. We used 0%, 25%, and 50% censoring rates by generating uniform censoring times for zero, one-fourth, and one-half of the sample. Groups were defined by predicted probability deciles. Deciles with a small number of events were collapsed with the next decile, until each decile contained at least five events. We conclude that in these simulations, the GB and the new GND tests achieve the correct size under all three scenarios irrespective of censoring rate.

The performance of the original ND (4) test depends on the censoring rate and on the form of baseline hazard, with the increasing baseline hazard and 50% censoring rate being the most problematic. An increasing baseline hazard affects the observed censoring rate, because the increasing form of baseline hazard results in later survival times, so observations have more time to be censored, whereas for the decreasing hazard, more events are captured despite censoring. For example, 50% censoring reduced the 10% event rate to 9.09% for the decreasing hazard, whereas for the increasing hazard, the rate was only 6.42%. The censoring rate drives the performance of the ND (4) test.

Using a higher population event incidence (40%) resulted in similar plots. A new test for a lower incidence of 5% but with a doubled sample size also performed equally well.

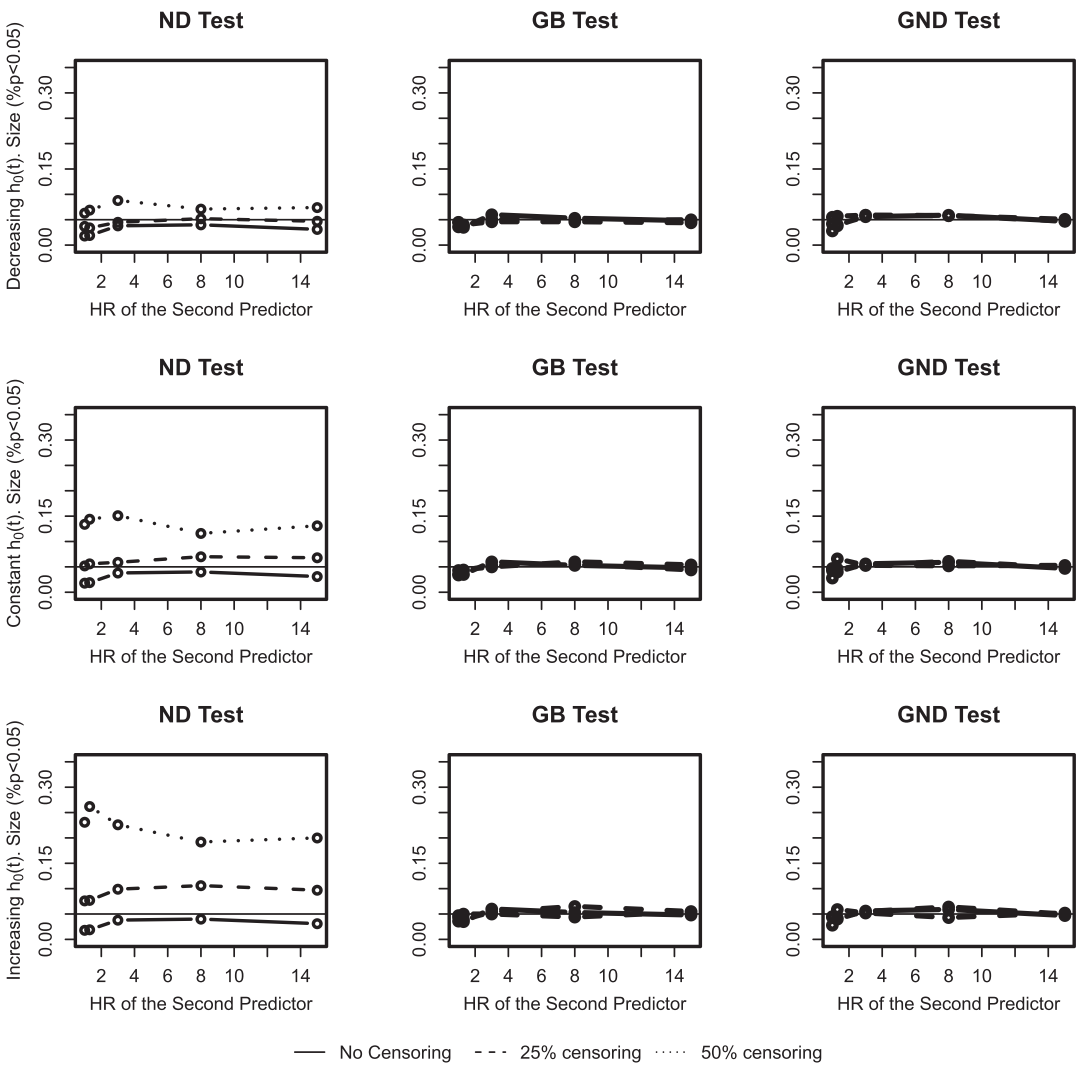

Figure 3. The size of the Nam–D'Agostino (ND), Grønnesby and Borgan (GB) and proposed Greenwood–Nam–D'Agostino (GND) tests (testing deciles under the null) for decreasing (top row), constant (center), and increasing (bottom row) baseline hazards. The population event incidence rate is 10%. Deciles with less than five events were collapsed with the next neighbor. HR, hazard ratio.

5.1. Performance for smaller sample size

Figure 4A demonstrates the performance of the two tests in samples of size 1000 with a 10% event rate. As in the previous analysis, we collapse deciles with less than five events. With such a small number of events (roughly 100 events per sample), the size of the tests (especially GND) varies more from simulation to simulation. As can be seen from Figure 4B, the GB test has a slightly higher size than the targeted 0.05 level, and the GND test's empirical size is a bit lower than 5%. So, in this situation, the GND test may be more useful than the GB test. When we use the 'collapse if less than two' rule, both tests' performance noticeably deteriorates. We find that as long as there are enough events per group, GB performs well in our simulations. However, as will be demonstrated in our practical example, there are other conditions when the GB test is less attractive (Section 5).

5.2. Power for the Grønnesby and Borgan and Greenwood–Nam–D'Agostino tests under various misspecified models

To examine the power of the two tests, we have to see how the tests perform when the model is mis-calibrated. We simulate data under the true model as described in Section 3, but we calculate predicted probabilities under various misspecified models. Ideally, a good goodness-of-fit test, particularly when

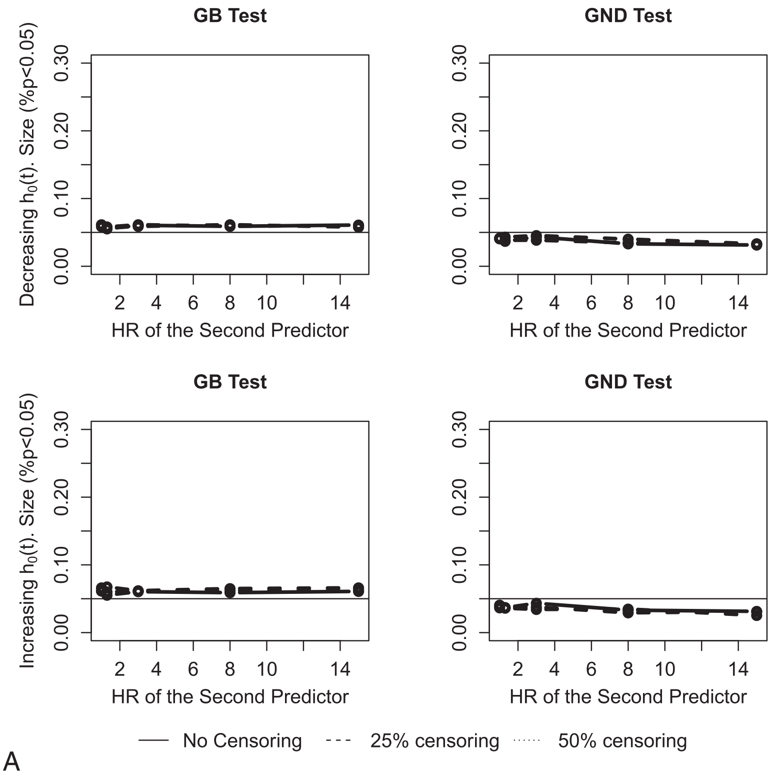

Figure 4A. The size of Grønnesby and Borgan (GB) and proposed Greenwood–Nam–D'Agostino (GND) tests with smaller sample size ($N = 1000$, $p = 0.1$, and at least five events per decile for decreasing (top row) and increasing (bottom row) baseline hazards. HR, hazard ratio.

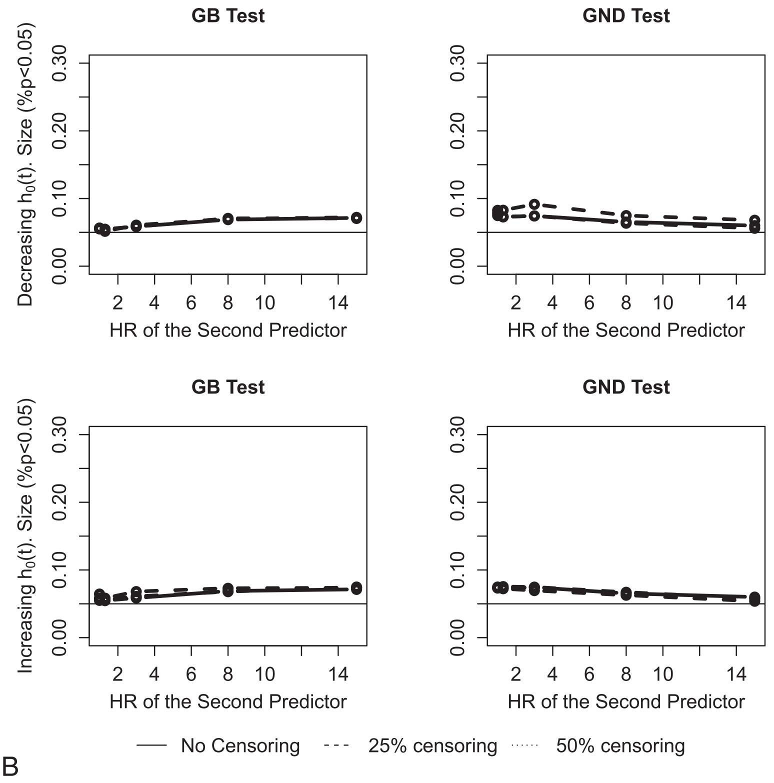

Figure 4B. The size of Grønnesby and Borgan (GB) tests and proposed Greenwood–Nam–D'Agostino (GND) with smaller sample size (N=1000, p=0.1, and at least two events per decile for decreasing (top row) and increasing (bottom row) baseline hazards. HR, hazard ratio.

it is intended to test calibration, ought to have enough power to detect that the model does not fit the data well. We considered three scenarios: (i) a missing quadratic term; (ii) a missing interaction; and (iii) a missing predictor.

First, in Figure 5A, the data were generated under the following model:

$$F_{true}(t) = \exp\left(-t^{\alpha}\lambda \exp\left(\beta_1 x_1 + \beta_2 x_1^2\right)\right) \tag{7}$$

We fit Cox regression using the following misspecified model to generate predicted probabilities:

$$F_{fitted}(t) = \exp\left(-t^{\alpha}\lambda \exp\left(\beta_1^* x_1\right)\right) \tag{8}$$

We plot in Figure 5A the power of the GB and GND tests for various HRs $\left(e^{\beta_2}\right)$ for the omitted quadratic term in model (7). The power for the GB test is excellent; the GND test has lower power, which is expected because of its minimal usage of parametrics. For an increasing baseline hazard and 50% censoring rate, the GND test has 55% power for HR = 2.5 and 73% for HR = 3.0 for the misspecified model (8) (Figure 5A).

In Figure 5B, we plot the power of the two tests for the true model:

$$F_{true}(t) = \exp\left(-t^{\alpha}\lambda \exp\left(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2\right)\right) \tag{9}$$

We omit the interaction term and fit the following misspecified Cox model:

$$F_{fitted}(t) = \exp\left(-t^{\alpha}\lambda \exp\left(\beta_1^* x_1 + \beta_2^* x_2\right)\right) \tag{10}$$

The performance of GND deteriorates somewhat more for this situation. For example, for increasing baseline hazard and 50% censoring rate, the GND test has only 18.3% power for HR = 3.0 and 68.7% for HR = 7.0 for the misspecified model (10) (Figure 5B).

Lastly, we omitted an important predictor. These data are generated under the model (11):

$$F_{true}(t) = \exp\left(-t^{\alpha}\lambda \exp\left(\beta_1 x_1 + \beta_2 x_2\right)\right) \tag{11}$$

But we used the following model to fit the data:

$$F_{true}(t) = \exp\left(-t^{\alpha}\lambda \exp\left(\beta_1^* x_1\right)\right) \tag{12}$$

The power plot is shown in Figure 5C. We observe that both tests completely fail to detect a missing predictor and the power plots in Figure 5C are similar to size plots. Unlike the power plots in Figure 5A and 5B, the GND and GB tests remain extremely low powered (close to 5%) across the range of HRs of the omitted predictor x_2, as has been seen previously [20, 21].

We conclude that the two tests have enough power to detect an omitted term in situations when there is some information about the omitted term in the model, that is, omitted quadratic term when the linear term is present and omitted interaction when main effects are present. However, it is impossible to detect omitted but completely unknown information.

6. Practical example

We applied the GB and GND tests to WHS data to illustrate the performance of the two tests in a real-life situation. The WHS is a large-scale nationwide 10-year cohort study of women beginning in 1992. A full description of the WHS data can be found elsewhere [22]. To calculate the risk of 10-year hard CHD, including myocardial infarction and death from coronary heart disease, we use the published Framingham Adult Treatment Panel III (ATP III) model [23]. This model was developed for women without diabetes, 30–79 years old without intermittent claudication at baseline. We removed from analysis observations for women with diabetes, or 80+ years old leading to a sample size of $N = 26,865$. The median follow-up was 10.2 years up through March 2004. A total of 213 women developed hard CHD by that time, and 36.6% of women were censored prior to year 10, most of censoring occurring after year 8.

ATP III predictors [23] are four log-transformed variables: age, total cholesterol, high-density lipoprotein cholesterol, and systolic blood pressure; two categorical variables: current smoking status and

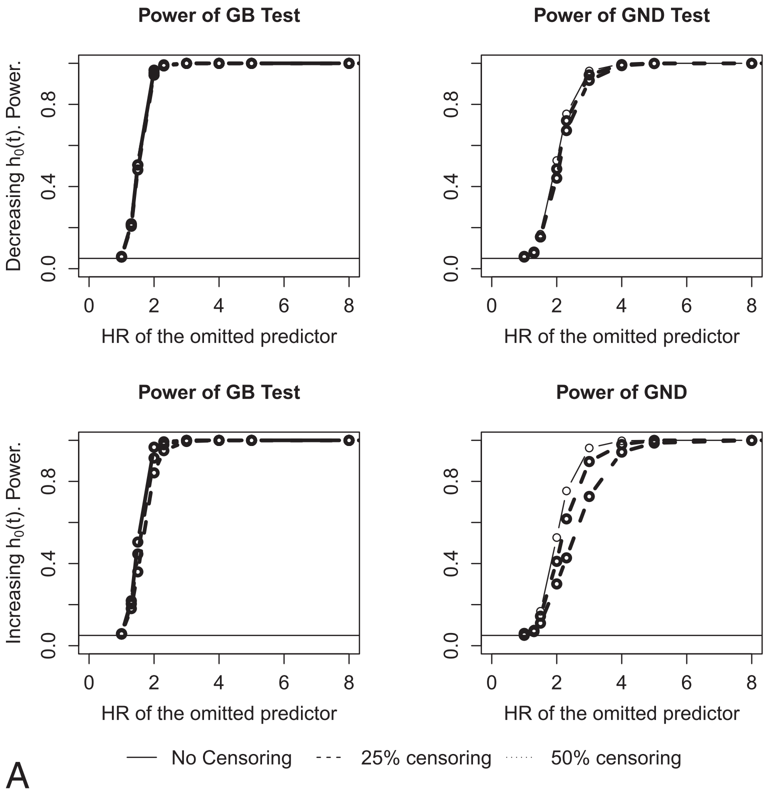

Figure 5A. Power of Grønnesby and Borgan (GB) and proposed Greenwood–Nam–D'Agostino (GND) tests when missing a quadratic term. $N = 5000$ and $p = 0.1$ for decreasing (top row) and increasing (bottom row) baseline hazards. (Models 7 and 7*). HR, hazard ratio.

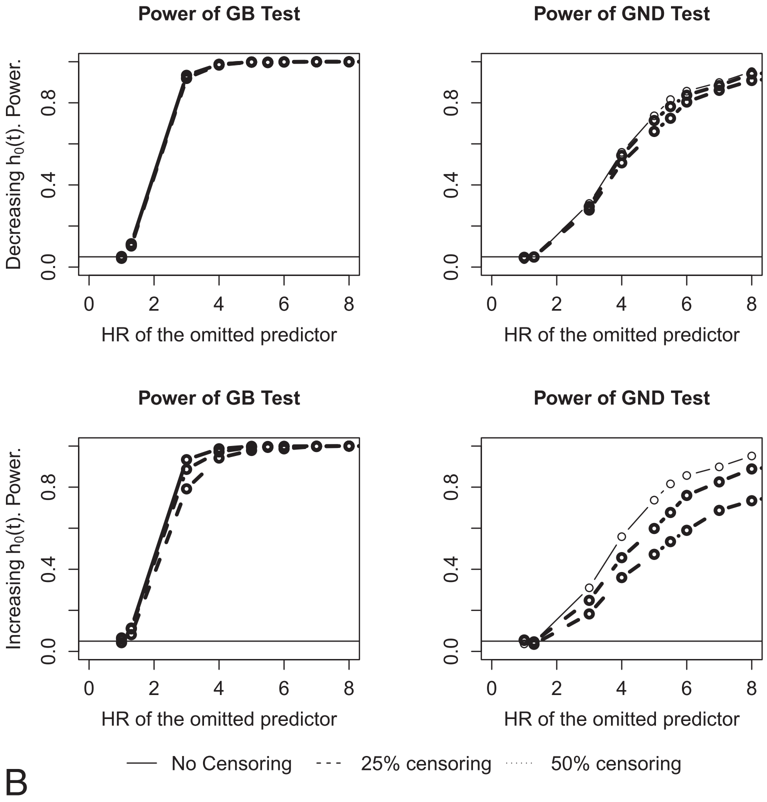

Figure 5B. Power of Grønnesby and Borgan (GB) tests and proposed Greenwood–Nam–D'Agostino (GND) tests when missing an interaction term. N = 5000 and p = 0.1 for decreasing (top row) and increasing (bottom row) baseline hazards. (Models 8 and 8*).

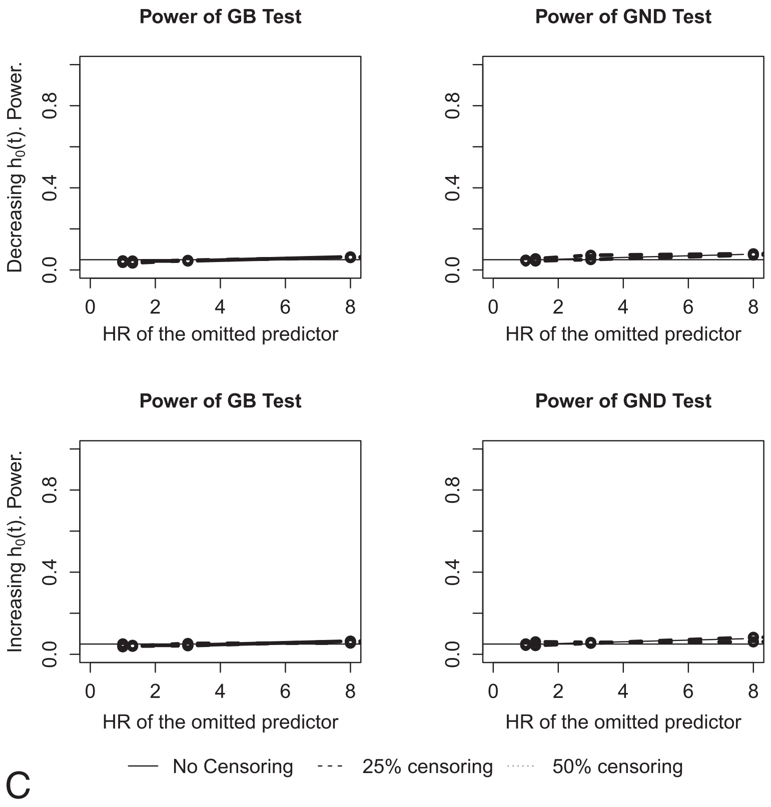

Figure 5C. Power of Grønnesby and Borgan (GB) tests and proposed Greenwood–Nam–D'Agostino (GND) tests when missing an important predictor. N = 5000 and p = 0.1 for decreasing (top row) and increasing (bottom row) baseline hazards. (Models 9 and 9*). HR, hazard ratio. HR, hazard ratio.

Table I. Published ATP III model for women: Cox regression coefficients and estimated means.

		Means	
Independent variable	Cox parameter coefficient	Framingham	WHS
Ln(age)	31.764001	3.92132	3.982996
Ln(total cholesterol)	22.465206	5.362898	5.335407
Ln(HDL cholesterol)	−1.187731	4.014637	3.951109
Ln(SBP)	2.552905	4.837649	4.809493
TRT for HTN (SBP > 120)	0.420251	0.142802	0.12172
Current smoker	13.07543	0.32362	0.116397
Ln(age)*Ln(total cholesterol)	−5.060998	21.05577	21.25574
Ln(age)*smoker	−2.996945	1.251988	0.462455
	Average 10-year survival	0.98767	0.991882

WHS, women's health study; HDL, high-density lipoprotein; TRT for HTN, treatment for hypertension; SBP, systolic blood pressure.

treatment for hypertension (systolic blood pressure >120); and two interaction terms: log-transformed total cholesterol with log-transformed age and smoking with log-transformed age. Published coefficients for the model are presented in [24] and are reproduced in Table I.

The ATP III model was developed using Framingham data. The WHS data were collected at a different time with a different population. Indeed, the average 10-year failure probability in Framingham data is 1.5 times higher than in WHS (0.8% in WHS as compared with 1.2% in Framingham data). The smoking rate is lower in the WHS with 11.6% smoking in WHS versus 32.4% in Framingham's subset of women. In a different population, one would expect to see some degree of miscalibration of the ATP III model.

Table II. Observed and expected counts in each decile for the ATP III model using women's health study data.

Decile**	Total n	Number censored	Number of events	Observed count* Kaplan–Meier	Expected count ATP III model	Observed failure rate Kaplan–Meier	Expected failure rate ATP III model
1–3	8060	3010	6	6.0	31.1	0.001	0.004
4	2686	956	7	7.0	20.2	0.003	0.008
5	2687	981	8	8.3	27.1	0.003	0.010
6	2686	935	14	14.5	36.8	0.005	0.014
7	2687	985	8	8.4	49.9	0.003	0.019
8	2686	932	25	26.1	69.9	0.010	0.026
9	2687	983	44	47.1	105.3	0.018	0.039
10	2686	1064	98	102.3	218.3	0.038	0.081

*The observed count is adjusted for censoring. It is equal to the Kaplan–Meier estimate of the number of events; had there been no censoring.

**Deciles with less than five events were collapsed.

For instance, when we estimate the survival probability in WHS using the published ATP III model, the expected counts in each decile are twofold to sixfold larger than the observed counts (Table II).

Therefore, recalibration is likely to improve the model fit. We ran the GB and GND tests for the ATP III model with four different recalibration strategies. For Run 1, we applied the published ATP III model from the website. We calculated the survival probability using the following formula; survival probability $S(t)$ for Cox model is estimated by the following formula:

$$S(t) = e^{\left(-\int_0^t \lambda(u)du\right)e^{\beta' x}} = e^{(-\mathbf{\Lambda}_0(t))e^{\beta' x}} = e^{(-\Lambda_0(t))e^{\beta' x-\beta'\bar{x}+\beta'\bar{x}}} = e^{(-\Lambda_0(t))e^{\beta'\bar{x}}e^{\beta' x-\beta'\bar{x}}}$$

We are interested in the survival probability at time 10 so we write

$$S(t=10) = \left(e^{(-\Lambda_0(10))e^{\beta'\bar{x}}}\right)^{e^{\beta' x-\beta'\bar{x}}} = A^{\left(e^{\beta' x-\beta'\bar{x}}\right)}.$$

A can be approximated by using direct substitution:

$$\bar{S}(t=10) \cong A^{\left(e^{\beta'\bar{x}-\beta'\bar{x}}\right)} = A.$$

Therefore, A can be approximated as an average survival probability at time 10. The formula for survival probability for Run 1 is thus

$$S(t=10) = 0.98767^{\left(e^{\hat{\beta}' x-\hat{\beta}'\bar{x}}\right)},$$

Where $\hat{\boldsymbol{\beta}}$ (estimates of Cox regression coefficients) and $\bar{\boldsymbol{x}}$ (averages of the predictor variables) are from the published ATP III model [2] and are presented in the second and third columns of Table I.

Run 2 also uses $\hat{\boldsymbol{\beta}}$ published for ATP III model, but we calibrate it in the large so that the average of the estimated survival probabilities is equal to the observed 10-year survival estimated by Kaplan–Meier in the WHS (=0.991751). Calibration in the large is achieved by refitting the baseline hazard for the data. We implement this model by fixing the coefficients at the published values but letting the Cox model estimate the baseline survival (the Cox PH model automatically refits the baseline hazard).

The results of Runs 1 and 2 are presented in the first two rows of Table III. It is impossible to implement the GB test for the Run 1 because the GB test requires rerunning the Cox PH model, which always calibrates in the large. This test, thus, cannot be considered a test of overall calibration. The GB test was calculated only for Run 2. The GND test is highly significant for both runs with p-values of <0.001 and 0.001, respectively, indicating miscalibration. Yet, the GB test is nonsignificant. The two tests rarely produce contradicting results in our simulations, so we use the calibration slope approach [25] as an additional test of calibration. We fit the Cox survival model with $z = \hat{\boldsymbol{\beta}}'\boldsymbol{x}$ as the only explanatory variable in the model. The regression coefficient of variable z is called the calibration slope. It is equal to 1.19 and is significantly different from 1.0 (95% confidence interval of [1.04; 1.33]), implying that Model

Table III. Women's health study data.

Run	External	ATP III model calibrated in the large	Correct functional form	GB score statistic	GB *p*-value	GND statistic	GND *p*-value	Number of deciles
1	Yes	N	N	—	—	667.463	<0.001	8
2	Yes	Y	N	7.064	0.422	23.582	0.001	8
3	Yes	Y	Y	8.291	0.308	10.301	0.172	8
4	No	Y	Y	9.628	0.211	10.055	0.185	8

Results of Grønnesby and Borgan (GB) and proposed Greenwood-Nam-D'Agostino (GND) goodness-of-fit tests of four applications of the ATP III model with varying degrees of miscalibration.

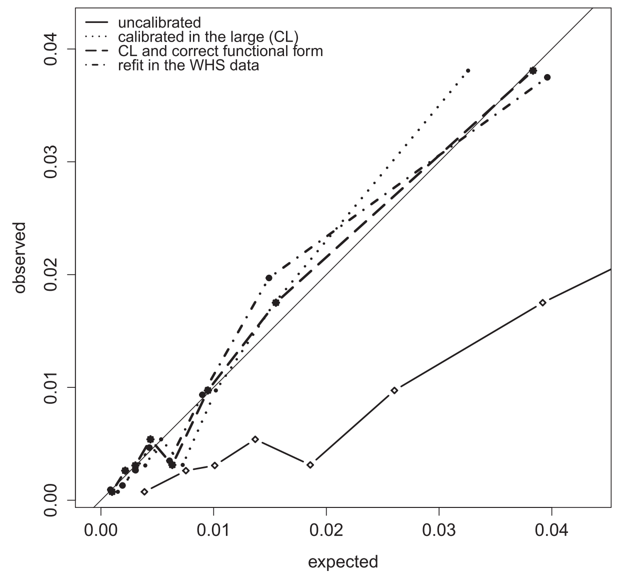

Figure 6. Observed probability of failure versus expected in each decile by four different recalibration strategies. ATP III model applied to women's health study (WHS) data.

2 is miscalibrated. To remedy this problem, we recalibrated the model using calibration slope of 1.19 as described by Janssen *et al.* [26]. For Run 3, we use Cox regression coefficients, which are equal to the product of the ATP III coefficients and the calibration slope of 1.19. The results for Run 3 are also presented in Table III. Both tests now agree (*p*-values of 0.3 and 0.2 for GND and GB, respectively), implying sufficiently good fit (calibration).

To illustrate the aforementioned findings, we plotted the observed versus expected average failure probability in each decile for all models (Figure 6). If the model is well calibrated, we would expect the plot to be close to a 45-degree line. We added Run 4 in which the ATP III model is refit in the WHS data, which ought to be well calibrated. Note that Run 1 is far off the 45-degree line. Run 2—which is only calibrated in the large—also deviates from the 45-degree line, confirming our guess that Run 2 is still miscalibrated. Of note, the lack of calibration in Run 2 in Figure 6 is driven mostly by the decile with the highest predicted event probability. Miscalibration of Run 2 was correctly picked up by GND test, but the GB test failed to detect the miscalibration. In further simulations using the WHS data, we observed that when the calibration slope is small, then the Cox model is able to readjust the baseline survival so that the survival probability remains close to the true survival probability. The GB test calculated in this setting is not significant. However, if we continue to distort $\boldsymbol{\beta}$ by increasing the value of the calibration slope, then adjustment of baseline hazard is not enough to compensate for this distortion. Distortion by a calibration slope of at least 1.8 was required to result in a significant GB test (and GND test) in this specific example. The GND test picks up the incorrect functional form for a calibration slope as low as 1.105. Looking at Figure 6, Run 3 is calibrated in the large and by the method of calibration slope and consequently closely matches the 45-degree line, illustrating that indeed with the second recalibration step [26], we achieve good calibration.

This practical example illustrates that the new GND method is robust, whereas GB method has an important limitation: it reestimates the baseline hazard, so it cannot be implemented for a situation when the baseline hazard and model coefficients are predetermined (i.e., come from an external model). We note that May and Hosmer did not suggest using the score test for externally estimated coefficients. In their two practical examples, the $\boldsymbol{\beta}$ coefficients were estimated internally. All of our simulations in which the $\boldsymbol{\beta}$ coefficients were estimated internally demonstrated consistently good performance of the GB test.

7. Discussion

Recent movement toward improvement of reproducibility of research findings [27] puts calibration measures in the spotlight. Indeed, as noted in [5], a miscalibrated model produces invalid risk estimates. In this paper, we present simulations and a practical example that illustrate whether the ND and the GB goodness-of-fit tests can be used to assess calibration of the predictive model in the survival setting and propose a new test that is a modified version of the ND test. Our simulation results are applicable to specific but commonly occurring, underlying parametric survival models but should generalize to other types of models. Note also that results from all of these goodness-of-fit tests are affected by factors such as the number of groups chosen and the sample size. Such testing should be accompanied by graphical assessments of calibration.

Nam and D'Agostino developed their test specifically for the Framingham model, which has little or no censoring (with no administrative censoring, the Framingham dataset has a censoring rate of less than 5%). Our simulations confirmed that the ND test has appropriate size in this setting. In many settings, the ND test will be adequate. But, on the other hand, existing models are routinely applied to new and sometimes heavily censored datasets. In order to apply the risk estimates to a new dataset, we need to confirm that the model is calibrated well in the new data. Our computer simulations demonstrate that for censoring of 25% and higher, the GB test performs quite well, but the ND test fails to achieve the 0.05 size even in the derivation dataset (Figure 2).

To remedy this situation, we developed a modified ND test (which we call the GND test). We investigated the performance of the three tests in a simulated and real-life cohort study. The proposed GND test has the correct size for a variety of settings: 25% and 50% censoring rates, 0.05, 0.1, and 0.4 event rates, and decreasing/increasing/constant baseline hazards. Thus, the GB and GND tests are better performing tests of fit. We focused on these two tests for the remaining analysis. Based on power plots in Figures 5A–C, we concluded that both methods can detect departures because of missing nonlinear or interaction terms, but neither can detect an omitted variable. This inability to detect an omitted covariate is common for all HL-style tests [28, 29] and was noted for binary data by Cook and Paynter [20], who demonstrated through simulations that the power plot using logistic regression is similar to the 0.05 reference line.

Our practical example illustrates a subtle but important difference between a goodness-of-fit and a calibration test. In our practical example with WHS data, Run 2 is miscalibrated (as shown by a calibration slope that is significantly different from 1.0). This miscalibration is correctly detected by GND test (p-value $= 0.001$) but missed by the GB test (p-value $= 0.4$), because in the process of running the GB test, we reestimate baseline hazard. Reestimating the baseline hazard compensates for miscalibration, even when the regression coefficients are fixed in advance. Once we adjust the ATP III coefficients by the calibration slope (Model 3), both tests are nonsignificant. This implies that the GB is a goodness-of-fit test of the selected variables but it is not a test of calibration.

Collapsing deciles worked quite well in our simulations: it helped to avoid small cells and guaranteed that the denominator in formula (5) is estimable. Several authors [11, 29] addressed this problem by reducing the number of groups. This strategy does not guarantee nonzero events cells. Collapsing to achieve five events per cell showed greater stability of the estimates versus collapsing to achieve two events per cell.

To summarize, the GND test can be used to assess calibration as well as nonlinearity in external validation sets, and therefore, it is more suitable for testing calibration. The GND test can be used for models for survival data other than the Cox model-for example, for nonparametric models in the machine learning/data mining setting-and is therefore more general. Both tests share a common limitation of failure to detect an omitted variable. SAS [30] and R [31] code for GND test is added to the Appendix and is also available at http://ncook.bwh.harvard.edu.

Appendix

Notation

Greenwood–Nam–D'Agostino test statistic.

$$\chi^2_{GND}(t) = \sum\nolimits_{g=1}^{G} \frac{\left[KM_g(t) - \overline{p(t)}_g\right]^2}{Var\left(KM_g(t)\right)} \tag{A.1}$$

where $KM_g(t) = 1 - S_g(t)_{KM}$ is the Kaplan–Meier failure probability in the gth decile at time t, $Var\left(KM_g(t)\right) = \left(1 - KM_g(t)\right)^2 \sum_{i|t_i \leqslant t} \frac{d_i}{n_i(n_i - d_i)}$, where d_i and n_i are the number of failures and number at risk at time t_i.

Hosmer–Lemeshow test statistic.

$$\chi^2_{HL=} \sum_{g=1}^{G} \frac{\left[O_g - E_g\right]^2}{n_g \pi_g (1 - \pi_g)},$$

(2) where

$$\begin{aligned} \pi_g &= \frac{E_g}{n_g} = \frac{\#\,\text{expected events in group}\ g}{\#\,\text{observations in group}\ g}, \\ O_g &= m_g = \#\text{observed events in group}\ g \\ E_g &= \text{expected number of events in group}\ g \end{aligned}$$

Lemma
In the absence of censoring, $\chi^2_{GND}(t) = \sum_{g=1}^{G} \frac{\left[O_g - E_g\right]^2}{n_g \hat{p}_g (1 - \hat{p}_g)}$

Statement 1.

In the absence of censoring, $KM_g(t) = \frac{m_g}{n_g}$ = observed proportion of events in the group g

Proof
$KM_g(t) = 1 - S_g(t)_{KM} = 1 - \prod\limits_{t_i < t} \frac{n_i - d_i}{n_i}$, with n_i is the number in group g at risk by time t_i and d_i is the number of events in group g at time t_i. In the absence of censoring, we can simplify

$$\begin{aligned} KM_g(t) = 1 - \prod\nolimits_{t_i < t} \frac{n_i - d_i}{n_i} &= 1 - \frac{n_g - 1}{n_g} \times \frac{n_g - 2}{n_g - 1} \times \frac{n_g - 3}{n_g - 4} \times \ldots \times \frac{n_g - D_g}{n_g - D_g + 1} \\ = 1 - \frac{n_g - m_g}{n_g} = \frac{m_g}{n_g} &= \text{observed proportion of events in the group}\ g. \end{aligned}$$

□

Now, let us simplify Greenwood variance formula in the absence of censoring.

$$\begin{aligned} Var\left(KM_g(t)\right) &= \left(1 - KM_g(t)\right)^2 \sum\nolimits_{i|t_i \leqslant t} \frac{d_i}{n_i(n_i - d_i)} = \text{in the absence of censoring} = \\ &= \left(\frac{n_g - m_g}{n_g}\right)^2 \left(\frac{1}{n(n-1)} + \frac{1}{(n-1)(n-2)} + \ldots + \frac{1}{(n-m+1)(n-m)}\right) \end{aligned} \tag{A.2}$$

Let us consider the second term.

Statement 2.

$$\frac{1}{n(n-1)} + \frac{1}{(n-1)(n-2)} + \ldots + \frac{1}{(n-m+1)(n-m)} = \frac{m}{n(n-m)}$$

Proof
Left-hand side contains m terms. We can write it as follows:

$$\#1 + \#2 + \#3 + \ldots + \#m$$

Using this notation, let us add terms one by one:

$$\begin{aligned}\#1 + \#2 &= \frac{1}{n(n-1)} + \frac{1}{(n-1)(n-2)} = \frac{n-2}{n(n-1)(n-2)} + \frac{n}{n(n-1)(n-2)} = \\ &= \frac{n-2+n}{n(n-1)(n-2)} = \frac{2n-2}{n(n-1)(n-2)} = \frac{2(n-1)}{n(n-1)(n-2)} = \frac{2}{n(n-2)}\end{aligned}$$

$$\begin{aligned}\#1 + \#2 + \#3 &= \frac{2}{n(n-2)} + \frac{1}{(n-2)(n-3)} = \frac{2(n-3)}{n(n-2)(n-3)} + \frac{n}{n(n-2)(n-3)} \\ &= \frac{3n-6}{n(n-2)(n-3)} = \frac{3(n-2)}{n(n-2)(n-3)} = \frac{3}{n(n-3)}\end{aligned}$$

Induction step: suppose that $\#1 + .. + \#(i-1) = \frac{i-1}{n(n-i+1)}$; let us prove that $\#1 + .. + \#i = \frac{i}{n(n-i)}$. Indeed, $\#1 + .. + \#i = \frac{i-1}{n(n-i+1)} + \frac{1}{(n-i+1)(n-i)} = \frac{(i-1)(n-i)+n}{n(n-i+1)(n-i)} = \frac{in-i(i-1)}{n(n-i+1)(n-i)} = \frac{i}{n(n-i)}$.

Therefore, by induction

$$\#1 + \#2 + \#3 + \cdots + \#m = \frac{m}{n(n-m)}$$

□

Plugging statement 2 into formula (3), we obtain

$$Var\left(KM_g(t)\right) = \left(\frac{n_g - m_g}{n_g}\right)^2 \left(\frac{m_g}{n_g(n_g - m_g)}\right)$$

Now, we can simplify GND test statistic (1):

$$\chi^2_{GND}(t) = \sum_{j=1}^{G} \frac{\left[KM_g(t) - \overline{p(t)}_g\right]^2}{Var\left(KM_g(t)\right)} = \sum_{j=1}^{G} \frac{\left[\frac{m_g}{n_g} - \overline{p(t)}_g\right]^2}{\left(\frac{n_g - m_g}{n_g}\right)^2 \left(\frac{m_g}{n_g(n_g - m_g)}\right)}$$

Statement 3.

$$\frac{\left[\frac{m_g}{n_g} - \overline{p(t)}_g\right]^2}{\left(\frac{n_g - m_g}{n_g}\right)^2 \left(\frac{m_g}{n_g(n_g - m_g)}\right)} = \frac{\left[O_g - E_g\right]^2}{n_g \hat{p}_g \left(1 - \hat{p}_g\right)}$$

Proof

$$\begin{aligned}\frac{\left[\frac{m_g}{n_g} - \overline{p(t)}_g\right]^2}{\left(\frac{n_g - m_g}{n_g}\right)^2 \left(\frac{m_g}{n_g(n_g - m_g)}\right)} &= \frac{\left[\frac{m_g}{n_g} - \overline{p(t)}_g\right]^2 \times \frac{n_g^2}{n_g^2}}{\frac{(n_g - m_g)}{n_g^2} \frac{m_g}{n_g}} = \frac{\left[\frac{n_g m_g}{n_g} - n_g \overline{p(t)}_g\right]^2 \times \frac{1}{n_g^2}}{\frac{(n_g - m_g)}{n_g^2} \frac{m_g}{n_g}} \\ &= \frac{\left[m_g - n_g \overline{p(t)}_g\right]^2}{(n_g - m_g)\frac{m_g}{n_g}} = \frac{\left[m_g - n_g \overline{p(t)}_g\right]^2}{n_g \frac{(n_g - m_g)}{n_g} \frac{m_g}{n_g}} = \frac{\left[O_g - E_g\right]^2}{n_g \hat{p}_g \left(1 - \hat{p}_g\right)}\end{aligned}$$

where $\hat{p}_g = \frac{m_g}{n_g}$.

□

Therefore, we showed that in the absence of censoring, GND can be written as follows:

$$\chi^2_{GND}(t) = \sum\nolimits_{g=1}^{G} \frac{\left[O_g - E_g\right]^2}{n_g \hat{p}_g \left(1 - \hat{p}_g\right)}.$$

Comparing it with the HL formula in (2) $\chi^2_{HL=} \sum_{g=1}^{G} \frac{\left[O_g - E_g\right]^2}{n_g \pi_g (1-\pi_g)}$, we have proved that in the absence of censoring, $\chi^2_{GND}(t)$ is very similar to χ^2_{HL}. The only difference is how proportion of events is estimated in the denominator: as observed binomial proportion in GND or as expected proportion in HL.

Code

```
##########################################################################

# R FUNCTION TO CALCULATE GREENWOOD-NAM-D'AGOSTINO CALIBRATION TEST

# FOR SURVIVAL MODEL

# TO RUN:

# GND.calib(pred,tvar,out,cens.t, groups, adm.cens)

# PARAMETERS:

# pred - PREDICTED PROBABILITIES OF AN EVENT

# out  - OUTCOME 0/1 1=EVENT

# cens.t - CENSORED/NOT CENSORED INDICATOR 1=UNCENSORED

# groups - GROUPING ASSIGNMENT FOR EACH OBSERVATION

# adm.cens - END OF STUDY TIME

# tvar,out,cens.t, groups, adm.cens

# REQUIRES AT LEAST 2 EVENTS PER GROUP, AT LEAST 5 EVENTS PER GROUP IS RECOMMENDED

# IF <2 EVENTS PER GROUP THEN QUITS

##########################################################################

kmdec=function(dec.num,dec.name, datain){

  stopped=0

  data.sub=datain[datain[,dec.name]==dec.num,]

  if (sum(data.sub$out)>1){

    avsurv=survfit(Surv(tvar,out) ~ 1, data=datain[datain[,dec.name]==dec.num,], error="g")

avsurv.est=ifelse(min(avsurv$time)<=adm.cens,avsurv$surv[avsurv$time==max(avsurv$time[avsurv$time<=adm.cens])],1)
avsurv.stderr=ifelse(min(avsurv$time)<=adm.cens,avsurv$std.err[avsurv$time==max(avsurv$ti
me[avsurv$time<=adm.cens])],0)

    avsurv.stderr=avsurv.stderr*avsurv.est
avsurv.num=ifelse(min(avsurv$time)<=adm.cens,avsurv$n.risk[avsurv$time==max(avsurv$time
[avsurv$time<=adm.cens])],0)

  } else {

    return(c(0,0,0,stopped=1))

  }

  c(avsurv.est, avsurv.stderr, avsurv.num, stopped)

}#kmdec


GND.calib = function(pred,tvar,out,cens.t, groups, adm.cens){

  datause=data.frame(pred=pred,tvar=tvar,out=out,count=1,cens.t=cens.t, dec=groups)

  datause$econtribution=ifelse(datause$cens.t==0,1,datause$tvar/time.t)

  numcat=length(unique(datause$dec))

  tzero20_categories=sort(unique(datause$dec))

  kmtab=matrix(unlist(lapply(tzero20_categories,kmdec,"dec",datain=datause)),ncol=4,
byrow=TRUE)

  if(sum(kmtab[,4])>0) stop("stopped because of less than 2 events in at least one group")

  hltab=data.frame(dec=c(1:numcat),

                  totaln=tapply(datause$count,datause$dec,sum),

                  censn=tapply(datause$cens.t,datause$dec,sum),

                  kmperc=1-kmtab[,1],

                  kmvar=kmtab[,2]^2,

                  kmnrisk=kmtab[,3],

                expected=tapply(datause$pred,datause$dec,sum),

                avgfail=tapply(datause$pred,datause$dec,mean),

                ecount=tapply(datause$econtribution,datause$dec,sum),

                numevents=tapply(datause$out,datause$dec,sum))

hltab$atrisk=hltab$totaln-hltab$censn

hltab$kmnum=hltab$kmperc*hltab$totaln

hltab$stat2=ifelse(hltab$kmvar==0, 0,(hltab$kmperc-hltab$avgfail)^2/(hltab$kmvar))

c(df=numcat-1, chi2gw=sum(hltab$stat2),pvalgw=1-pchisq(sum(hltab$stat2),numcat-1))
```

```
}#GND.calib


/**********************************************************************************************
***
# SAS MACRO TO CALCULATE GREENWOOD-NAM-D'AGOSTINO CALIBRATION TEST FOR
SURVIVAL MODEL
# FOR MORE DETAILS SEE Demler, Paynter, Cook "Tests of Calibration and Goodness of Fit
# in the Survival Setting"
# TO RUN:
# GND.calib(datain, groupvar, timevar, eventvar, predpvar)
# PARAMETERS:
# datain   - dataset with the following variables:
# groupvar - GROUPING ASSIGNMENT FOR EACH OBSERVATION
# timevar  - time of the event or censoring time
# eventvar - EVENT INDICATOR 1=EVENT 0=NONEVENT
# predpvar - PREDICTED PROBABILITIES OF AN EVENT CALCULATED FOR TIME=adm_cens
# adm_cens - TIME AT WHICH PREDICTED PROBABILITIES ARE CALCULATED
# REQUIRES AT LEAST 2 EVENTS PER GROUP, AT LEAST 5 EVENTS PER GROUP IS
RECOMMENDED
# IF <2 EVENTS PER GROUP THEN THE PROGRAM QUITS
***********************************************************************************************
**/
%macro gnd_calib(datain, groupvar, timevar, eventvar, predpvar, adm_cens);

  * KM estimates should be calculated for a fixed time=adm.cens;
  data &datain; set &datain;
    if &timevar>&adm_cens then do;
            &timevar=&adm_cens;
                  &eventvar=0;
    end;
  run;


  %let error2=0;
  %let error5=0;

  title;
  proc sort data=&datain; by &groupvar ; run;

  proc means data=&datain noprint; by &groupvar ; output out=check sum(&eventvar)=nevents; run;

  proc print data=check(drop=_TYPE_ _FREQ_); run;

  data check; set check; if nevents < 2 then call symput('error2','1'); else if nevents < 5 then call
symput('error5','1');
  run;

  %if &error2=1 %then %do; %put MACRO GND_CALIB WAS STOPPED: at least one of the groups
contains <2 events. Consider collapsing some groups.;
    %return;
  %end;
  %else %if &error5=1 %then %put WARNING: at least one of the groups contains < 5 events. GND can
become unstable.
    (see Demler, Paynter, Cook 'Tests of Calibration and Goodness of Fit in the Survival Setting')
    Consider collapsing some groups to avoid this problem;


  proc lifetest data=&datain outsurv=datain_surv stderr noprint; time &timevar*&eventvar(0); by
&groupvar; run;
```

```
 %let error2=0;
 %let error5=0;

 title;
 proc sort data=&datain; by &groupvar ; run;

 proc means data=&datain noprint; by &groupvar ; output out=check sum(&eventvar)=nevents; run;

 proc print data=check(drop=_TYPE_ _FREQ_); run;

 data check; set check; if nevents < 2 then call symput('error2','1'); else if nevents < 5 then call
symput('error5','1');
 run;

 %if &error2=1 %then %do; %put MACRO GND_CALIB WAS STOPPED: at least one of the groups
contains <2 events. Consider collapsing some groups.;
   %return;
 %end;
 %else %if &error5=1 %then %put WARNING: at least one of the groups contains < 5 events. GND can
become unstable.
   (see Demler, Paynter, Cook 'Tests of Calibration and Goodness of Fit in the Survival Setting')
   Consider collapsing some groups to avoid this problem;


 proc lifetest data=&datain outsurv=datain_surv stderr noprint; time &timevar*&eventvar(0); by
&groupvar; run;

 data datain_surv_noncens; set datain_surv; where _censor_=0 and survival ne .; pr_failure=1-survival;
run;

 data datain_surv_noncens_; set datain_surv_noncens; by &groupvar;  if last.&groupvar eq 1; run;

 proc means data=&datain noprint; output out=m mean (&predpvar)=m_&predpvar; by &groupvar;  var
&predpvar; run;

 data all; merge datain_surv_noncens_ m(keep=&groupvar m_&predpvar); by &groupvar; run;

 data all; set all; gnd_component=(((1-survival)-m_&predpvar)/sdf_stderr)**2; one=1; run;

 proc print data=all; var &groupvar. pr_failure sdf_stderr m_&predpvar gnd_component; run;

 proc means data=all noprint; output out=result sum (gnd_component one)=chi_square ngroups; run;

 %global _gnd_chisq;
 %global _gnd_df;
 %global _gnd_pvalue;
 data result; set result; df=ngroups-1; p_value=1-probchi(chi_square, df);
   call symput('_gnd_chisq',chi_square);
   call symput('_gnd_df',df);
   call symput('_gnd_pvalue',p_value);
 run;

   title "GND test: chi_square=%trim(&_gnd_chisq.) df=%trim(&_gnd_df.)
 p_value=%trim(&_gnd_pvalue.)";
   proc print data=result(drop=_TYPE_ _FREQ_); run;

 %mend gnd_calib;
```

Acknowledgements

This work was supported by grant R01HL113080 from the National Heart, Lung and Blood Institute (NHLBI). The WHS was supported by grants from the NHLBI (HL043851) and the National Cancer Institute (CA047988), all from Bethesda, Maryland. We want to thank Dr. Susanne May for the helpful discussions.

References

1. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of National Cancer Institute* 1989; **81**(24):1879–1886.
2. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham heart study. *Circulation* 2008; **117**:743–753.
3. Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. *American Heart Journal* 1991; **121**:293–298.
4. Wilson PWF, D'Agostino RB, Levy D, Belanger A, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; **97**:1837–1847.
5. Pepe MS, Janes H. "Methods for Evaluating Prediction Performance of Biomarkers and Tests" The Selected Works of Margaret S Pepe PhD, 2013. Available at: http://works.bepress.com/margaret_pepe/38 [accessed on 28 January 2015].
6. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.
7. D'Agostino RB, Byung-Ho N. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of Statistics* 2004; **23**:1–25.
8. Grønnesby JK, Borgan Ø. A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Analysis* 1996; **2**:315–320.
9. Paynter NP, Peelen LM, Cook NR. *Performance of Prediction Measures in a Survival Setting*. JSM: Vancouver, Canada, 2010.
10. Guffey D, May S, Hosmer DW. *Hosmer-Lemeshow Goodness-of-Fit Test: Translations to the Cox Proportional Hazards Model*. JSM: Montreal, Canada, 2013.
11. May S, Hosmer DW. A cautionary note on the use of the Grønnesby and Borgan goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Analysis* 2004; **10**:283–291.
12. May S, Hosmer DW. A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Analysis* 1998; **4**:109–120.
13. Hosmer DW, Lemeshow S, May S. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data* (2nd edn). John Wiley and Sons Inc.: Hoboken, New Jersey, 2011.
14. Hastie T, Friedman J, Tibshirani R. *The Elements of Statistical Learning, Vol. 2. No. 1*. Springer: New York, 2009.
15. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Annals of Internal Medicine* 2009; **150**(11):795–802.
16. Sego LH, Reynolds MR, Woodall WH. Risk-adjusted monitoring of survival times. *Statistics in Medicine* 2009; **28**(11):1386–1401.
17. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data* 2nd edn. John Wiley and Sons, Inc.: Hoboken, New Jersey, 2002:171.
18. Crowson CS, Atkinson EJ, Terneau TM. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research* 2014. [Epub ahead of print]; PMID: 23907781.
19. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods* 1980; **9**(10):1043–1069.
20. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. *Biometrical Journal* 2011; **53**(2):237–258.
21. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 1997; **16**(11):965–980.
22. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds risk score. *Journal of the American Medical Association* 2007; **297**:611–619. [PMID: 17299196].
23. Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). *Journal of the American Medical Association* 285:2486–2497.
24. Available at: http://www.framinghamheartstudy.org/risk-functions/coronary-heart-disease/hard-10-year-risk.php [21 October 2014].
25. Harrell, Jr, FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer-Verlag: New York, 2001.
26. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *Journal of Clinical Epidemiology* 2008; **61**:76–86.
27. Editorial Staff. Reducing our irreproducibility. *Nature* 2013; **496**(7):398.
28. Lin DY, Wei LJ. Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica* 1991; **1**:1–17.

29. Parzen M, Lipsitz SR. A global goodness-of-fit statistic for Cox regression models. *Biometrics* 1999; **55**:580–584.

30. SAS/STAT, Version 9.3 of the SAS System for Windows. Copyright ©2002-2010 SAS Institute Inc.

31. R Core Team (2014). R: a language and environment for statistical computing. *R Foundation for Statistical Computing*: Vienna, Austria. Available at: http://www.R-project.org/.