The Concordance Index C and the Mann–Whitney Parameter $\Pr(X > Y)$ with Randomly Censored Data

James A. Koziol[1], and **Zhenyu Jia***[,2]

[1] Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA

[2] Department of Pathology and Laboratory Medicine, University of California, Irvine, CA 92697, USA

Received 22 October 2008, revised 22 January 2009, accepted 28 January 2009

Harrell's c-index or concordance C has been widely used as a measure of separation of two survival distributions. In the absence of censored data, the c-index estimates the Mann–Whitney parameter $\Pr(X > Y)$, which has been repeatedly utilized in various statistical contexts. In the presence of randomly censored data, the c-index no longer estimates $\Pr(X > Y)$; rather, a parameter that involves the underlying censoring distributions. This is in contrast to Efron's maximum likelihood estimator of the Mann–Whitney parameter, which is recommended in the setting of random censorship.

Key words: c-index; Random censorship model; Survival analysis.

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1002/bimj.200800228.

1 Introduction

Harrell *et al*. (1982) and Harrell, Lee, and Mark (1996) introduced to the biomedical community the c-index or concordance C, a measure of the separation of two survival distributions. Their measure has been widely adopted and extensively used for assessing prediction performance in survival analysis settings (D'Agostino and Nam, 2004; Pencina and D'Agostino, 2004). With full information on survival, that is, no censorship, the c-index is equivalent to the Mann–Whitney parameter $\Pr(X > Y)$, which is well known and widely used in settings far removed from the comparison of survival distributions. In the survival setting with randomly censored data, the Mann–Whitney parameter remains a valid measure of separability of the underlying survival distributions independent of the censoring distributions, whereas the c-index has a less pellucid interpretation, dependent on the underlying censoring distributions, as we show subsequently. Efron (1967) has masterfully addressed technical details relating to distributional properties of the Mann–Whitney parameter in the random censorship setting, and we here supplement his theoretical results with a brief simulation study contrasting operating characteristics of the Mann–Whitney parameter to the c-index. We illustrate use of the Mann–Whitney parameter in the censored data setting with data from a recent study of ensemble tree classifiers for prostate cancer prognosis (Koziol *et al*., 2009).

* Correspondence author: e-mail: zjia@uci.edu, Phone: +019498249837, Fax: +019498242160

2 The *C*-Index and the Random Censorship Model

To describe C, we introduce some standard notation relative to randomly censored data. Let $X_1^0, X_2^0, \ldots, X_m^0$ be independent, identically distributed (iid) random variables having a common survival function (or right-sided cumulative distribution function) $F^0(s) = \Pr\{X_i^0 \geq s\}$, and let $Y_1^0, Y_2^0, \ldots, Y_n^0$ be iid random variables having common survival function G^0. The X_i^0 are censored on the right by independent random variables U_i, $1 \leq i \leq m$, and the Y_j^0 by independent random variables V_j, $1 \leq j \leq n$, where the U_i and V_j are drawn from survival functions H and I, respectively. Hence the available data consist of the pairs (X_i, δ_i), $1 \leq i \leq m$, where $X_i = \min(X_i^0, U_i)$, and

$$\delta_i = \begin{cases} 1 \text{ if } X_i = X_i^0 \\ 0 \text{ if } X_i = U_i, \end{cases}$$

and the pairs (Y_j, ε_j), $1 \leq j \leq n$, where $Y_j = \min(Y_j^0, V_j)$, and

$$\varepsilon_j = \begin{cases} 1 \text{ if } Y_j = Y_j^0 \\ 0 \text{ if } Y_j = V_j. \end{cases}$$

The observed X_i and Y_j constitute random samples from the survival functions F and G, respectively, where $F(s) = F^0(s)H(s) = \Pr\{X_i \geq s\}$, $1 \leq i \leq m$, and $G(s) = G^0(s)\,I(s) = \Pr\{Y_j \geq s\}$, $1 \leq j \leq n$. With these preliminaries, Harrell's C is easily described. First, form all mn possible pairs of observations from the two samples $(X_i, Y_j, \delta_i, \varepsilon_j)$. $1 \leq i \leq m$, $1 \leq j \leq n$. Omit all pairs for which the shorter event time is censored (*i.e.* $X_i < Y_j$, $\delta_i = 0$, or, $Y_j < X_i$, $\varepsilon_j = 0$). For each of the remaining "evaluable" pairs, score 1 if $X_i > Y_j$ and $\varepsilon_j = 1$, 0 if $X_i < Y_j$ and $\delta_i = 1$. C is then the sum of these scores over all evaluable pairs, divided by the number of evaluable pairs.

In the absence of censoring, all pairs are evaluable, and it should be clear that C is the U-statistic estimate of the value $\Pr(X^0 > Y^0)$, where $X^0 \sim F^0$ and $Y^0 \sim G^0$. This parameter has a long and distinguished history in the statistical literature related to the Mann–Whitney statistic (Acion *et al.*, 2006; Newcombe, 2006a, b).

Under the random censorship model, what does C estimate? To answer this, let us first invoke Efron's shorthand notation $\Pr(F \geq G)$ to denote the probability that a random variable X with survival function F is greater than or equal to an independent random variable Y with survival function G. Next, let us invoke an equivalence relation in the random censorship model, namely,

$$\{X_i \geq Y_j, \varepsilon_i = 1\} \approx \{Y_j^0 \leq \min(X_i^0, U_i, V_j)\}. \tag{1}$$

This relation is ascribed to John Gilbert in his unpublished University of Chicago Ph.D. thesis, and leads to the two-sample Gehan–Gilbert Wilcoxon test with randomly censored data.

The equivalence relation Eq. (1) leads to a compact representation of the expected values of the components of C. In this regard, note that in the formulation of C, the number of evaluable pairs is actually a random variable, with expected value

$$mn[\Pr\{F^0 HI \geq G^0\} + \Pr\{G^0 HI \geq F^0\}] \leq mn. \tag{2}$$

In Eq. (2), we are invoking Efron's shorthand notation introduced earlier in an obvious fashion, since $F^0 HI$ and $G^0 HI$ represent survival functions. An immediate implication is that there may well be an information loss incurred by discarding "non-evaluable" pairs. To first order, then, C estimates

$$\frac{\Pr\{F^0 HI \geq G^0\}}{\Pr\{F^0 HI \geq G^0\} + \Pr\{G^0 HI \geq F^0\}}. \tag{3}$$

If $F^0 = G^0$, the parameter (3) is 1/2, independent of H and I; however, if $F^0 \neq G^0$, C does not estimate the usual Mann–Whitney parameter $\Pr(F^0 \geq G^0)$. Here, interpretation of Eq. (3) is rather opaque, and dependent on the censoring distributions H and I. This is unfortunate, since under the random censorship formulation the censoring mechanisms are generally taken to be extraneous, with primary

focus solely on the underlying survival distributions F^0 and G^0 (notably, as with Kaplan–Meier estimators).

3 Efron's Estimator

We have shown that Harrell's C typically does not estimate the Mann–Whitney parameter $\Pr(X^0 > Y^0)$ in the random censorship setting. Nevertheless, efficient estimation of this parameter is available through the elegant derivation of Efron (1967) (and has subsequently been periodically invoked in this setting, *e.g.* Brown, Hollander, and Korwar, 1974; Begg *et al.*, 2000). Efron's solution follows from the representation

$$\Pr(X^0 > Y^0) = -\int F^0(s)\, dG^0(s), \tag{4}$$

and substitution of the Kaplan–Meier estimates $\hat{F}^0$ and $\hat{G}^0$ into the integral representation Eq. (4). Efron gives convenient computing forms for $\hat{W} = -\int \hat{F}^0(s)\, d\hat{G}^0(s)$, following from his self-consistent formulation of the Kaplan–Meier estimates. He also gives asymptotic distribution theory relating to $\hat{W}$, though with small samples sizes a bootstrap approach might be preferable. For completeness, we include these results in the Appendix. As a reviewer has presciently pointed out, "any consistent estimators of F^0 and G^0, in particular parametric ones if available, would also inherit the desired properties and lead to alternative possible estimators for C."

4 Numerical Comparisons

To elucidate some of the differences between the Efron and Harrell concordance estimators with randomly censored survival data, we undertook a limited simulation study. We took the underlying survival distributions F^0 and G^0, as well as the underlying censoring distributions H and I, to be standard two-parameter Weibull. Recall that the survival function of a two-parameter Weibull has representation $S(t) = \exp[-(\lambda t)^{\beta}]$, where λ and β are referred to as the scale and shape parameter, respectively. We parameterized F^0 with $(\lambda, \beta) = (4, 2)$, and G^0 with $(\lambda, \beta) = (2, 4)$. With these choices for (λ, β), Efron's C should estimate $\Pr(F^0 \geq G^0) = 0.807$, irrespective of H and I. (Numerical

Table 1 Results based on 1000 simulated data sets.

	H	I	Efron's C		Harrell's C	
	(λ, β)	(λ, β)	Mean	SD	Mean	SD
$m = n = 20$	(4, 2)	(2, 4)	0.812	0.084	0.740	0.123
	(2, 4)	(4, 2)	0.769	0.109	0.741	0.124
	(2, 2)	(3, 3)	0.802	0.103	0.744	0.131
	(3, 3)	(2, 2)	0.816	0.084	0.751	0.122
$m = n = 50$	(4, 2)	(2, 4)	0.810	0.051	0.740	0.074
	(2, 4)	(4, 2)	0.791	0.065	0.743	0.080
	(2, 2)	(3, 3)	0.808	0.060	0.745	0.083
	(3, 3)	(2, 2)	0.809	0.050	0.742	0.074
$m = n = 100$	(4, 2)	(2, 4)	0.810	0.037	0.742	0.053
	(2, 4)	(4, 2)	0.797	0.042	0.739	0.052
	(2, 2)	(3, 3)	0.804	0.040	0.738	0.056
	(3, 3)	(2, 2)	0.806	0.035	0.739	0.052

integrations were carried out in Mathematica 6.0, Wolfram Research Inc., 2007.) We then varied (λ, β) values for H and I, and undertook a simulation study to compare the Efron and Harrell estimates, with underlying sample sizes $m = n = 20$, 50, or 100. Findings based on 1000 simulations in each instance are given in Table 1.

As noted above, Efron's index should be estimating 0.807, and numerical agreement seems quite satisfactory. In comparison, from Eq. (3), Harrell's index should be consistent for $0.3703/(0.3703+0.1297) = 0.741$ for the first two selections of (λ, β) for H and I, and, similarly, Harrell's index should be consistent for $0.3237/(0.3237+0.1134) = 0.741$ in the next two selections of (λ, β) for H and I. (Note the interchangeability in parameters, as can be appreciated from inspection of Eq. (4).) Again, agreement between observed and expected is quite nice.

Censorship impacts the MSE of Harrell's index adversely compared with Efron's index, as can be seen in Fig. 1. Here, we have smoothed the simulated values that we had summarized in Table 1. The discrepancies in both estimated values and variances are readily apparent. From Eq. (2), the expected proportion of evaluable pairs for Harrell's C is 0.5 for the (λ, β) = (4, 2) or (2, 4) simulations, and 0.44 for the (λ, β) = (2, 2) or (3, 3) simulations. Therefore, the expected censoring rates for these two simulation scenarios are 0.5 (1–0.5) and 0.56 (1–0.44), respectively. In fact, Eq. (2) also

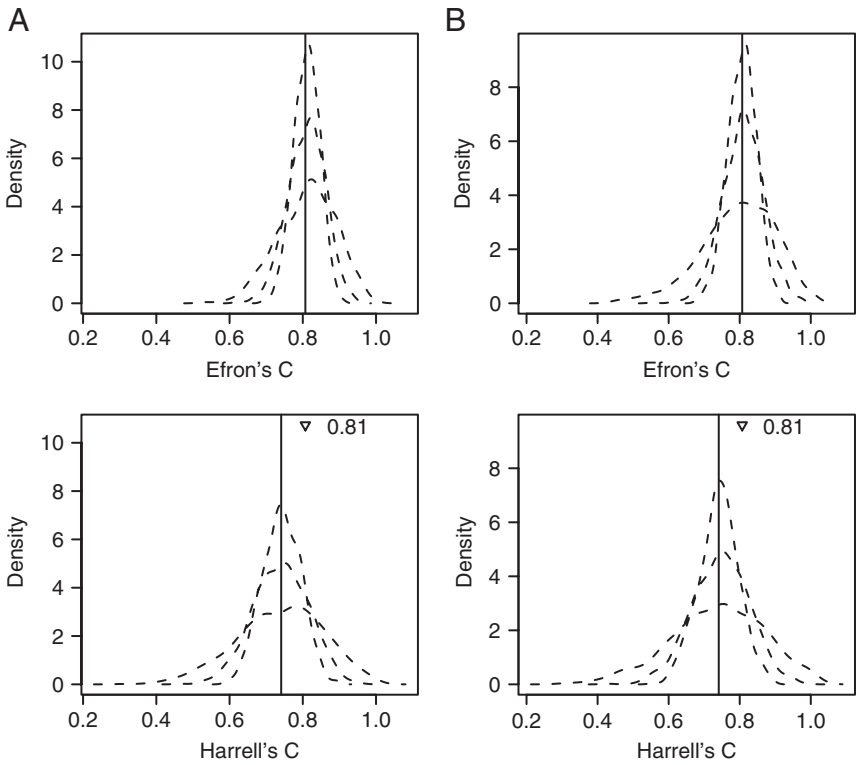

Figure 1 Empirical distributions of Efron's and Harrell's C from 1000 simulations. Plots in column (A) with (λ, β) = (4, 2) for H and (2, 4) for I. Plots in column (B) with (λ, β) = (2, 2) for H and (3, 3) for I. The solid vertical lines represent the expected values of C in respective cases, namely, 0.81 for Efron and 0.74 for Harrell. The dashed curves with low, median, and high peaks correspond to the empirical distributions of C with sample sizes 20, 50, and 100, respectively. For the plots of Harrell's C, we mark Efron's "true" values (0.81) with triangles.

leads to a crude approximation to the loss of information in Harrell's C relative to Efron's C. Recall that Efron's C utilizes all mn pairs of observations regardless of censoring status. Heuristically, then, the expression $\Pr\{F^0 HI \geq G^0\} + \Pr\{G^0 HI \geq F^0\}$ is an approximation to the relative information of Harrell compared to Efron. Informally, this quantity also represents an approximation to the ratio of the variance of Efron's C to the variance of Harrell's C.

We remark that the interplay between survival and censoring distribution affecting Harrell's index can lead to even more pronounced differences relative to Efron's than the cases we have reported here. For example, we were to select $(\lambda, \beta) = (2, 3)$ for H, and $(1, 3)$ for I, Harrell's index would estimate 0.531 from Eq. (3), whereas the expected value of Efron's index would remain unchanged at 0.807.

5 An Example

The impetus for this research was a project that we recently undertook, and reported in Koziol *et al*. (2009). In this project, we began with five distinct lists of genes putatively associated with prostate cancer diagnosis or prognosis. We had two independent clinical datasets of prostate cancer patients (referred to below as Set 1 and Set 2) who had undergone radical prostatectomy. Post-surgery follow-up data were available for all patients, with the primary outcome of interest being time to biochemical progression. In addition, for each patient we had complete expression data relating to the gene lists, derived from Affymetrix microarrays. We utilized recursive partitioning methodology to construct prognostic classification trees from expression data corresponding to each of the gene lists, and found that each of the trees produced tripartite classification schemes. We then utilized an ensemble classifier to arrive at a consensus classification of patients into one of three ordered subgroups that manifest low, intermediate, or

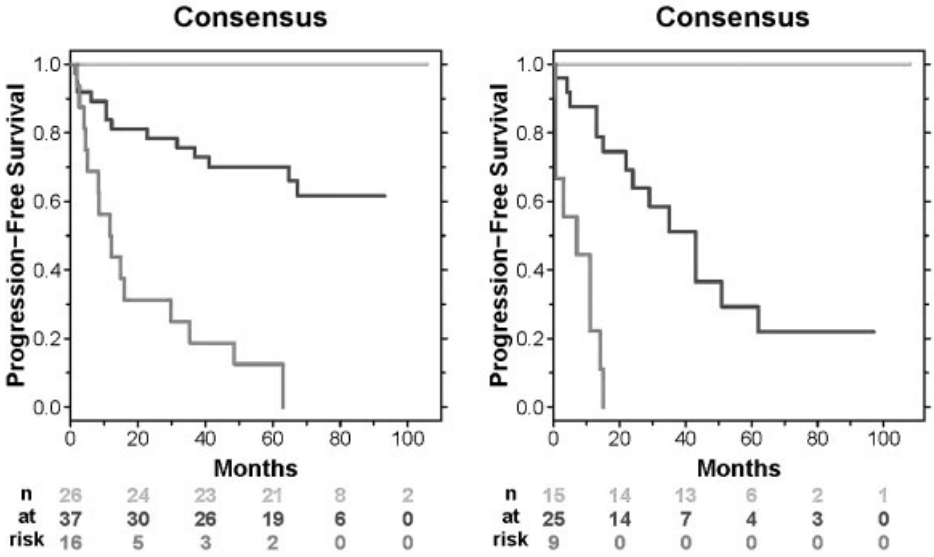

Figure 2 Progression-free survival curves from the consensus classifiers derived from the ensemble for Set 1 and Set 2. Numbers of subjects at risk in the three subgroups at the various time points are indicated beneath the x-axis of each graph.

Table 2 Concordance indices for the consensus classifiers.[a)]

	L *versus* H	L *versus* M	M *versus* H
Set 1	0.94 (0.013)	0.90 (0.014)	0.81 (0.016)
Set 2	1 (0)	1 (0.049)	0.90 (0.016)

a) For each classifier, L denotes the low-risk subgroup, M the intermediate-risk subgroup, and H, the high-risk subgroup. Tabulated values are Efron's concordance indices calculated from the survival curves presented in Fig. 2; values in parentheses are bootstrap standard errors of these estimates.

high risk of progression relative to one another. We refer the interested reader to Koziol *et al.* (2009) for details.

During the course of this investigation, we wanted to summarize the degree of separation between the time and progression curves with a suitable index. A Medline search uncovered the popularity of Harrell's *c*-index for this purpose. But upon closer examination of the underlying statistic, we found disconcerting the possibility that the *c*-index could ignore some data, a black flag to statisticians cautioned to utilize all available information. Even more disturbing was the misconception that the *c*-index ought to be utilized in this setting: for example, Graefen *et al.* (2002) state, ''Because the data were censored, the traditional area under the receiving operating characteristic curve (AUC) is problematic, and Harrell's version is calculated.'' The results reported herein are meant to counter this unintentionally ironic statement.

In Fig. 2, we present the progression-free survival curves from the consensus classifier derived from the ensemble for Set 1 and Set 2. The clear separation among the subgroups is readily apparent. In Table 2 we give the concordance indices for these two consensus classifiers. The separation is reflected in the concordance indices. (We remark that the stark separation is maintained with validation sets; see Koziol *et al.* (2009) for details.) Nevertheless, we remain somewhat cautious in our interpretation, preferring to place greater emphasis on the identification of a subgroup of patients that remain progression-free, for whom adjuvant chemotherapy subsequent to prostatectomy might well be unnecessary.

6 Conclusion

The concordance C is a useful measure of the discrepancy between two survival curves. The underlying parameter, $\Pr(X^0 > Y^0)$, enjoys an eminent and storied position in the statistical literature, dating from Kendall, Mann, and Whitney, with more recent applications to ROC curves and effect sizes. Indeed, Harrell and colleagues have utilized concordance C in the context of logistic regression, wherein the Mann–Whitney parameter is readily interpretable, and random censorship is not present. Interrelations with other statistical measures of association abound; see, for example, Schemper and Stare (1996), O'Quigley and Flandre (1994), O'Quigley, Xu, and Stare (2005), Schumacher, Graf, and Gerds (2003), Royston and Sauerbrei (2004), Newson (2006).

In the survival setting with randomly censored data, Efron's construction yields the maximum likelihood estimator of the Mann–Whitney parameter, and hence is fully efficient. In contrast, the expected value of Harrell's concordance index is dependent upon the underlying censoring distributions, hence may be far removed from the Mann–Whitney parameter; and, censorship adversely affects the precision of Harrell's C. To obviate potential concerns relating to computational complexities, we have prepared programs in `R` and `SAS` that calculate Efron's estimator with randomly censored data, as well as the bootstrap standard error estimate.

Table A1 Values of $Q(X_i, Y_j, \delta_i, \varepsilon_j)$.

$(\delta_i, \varepsilon_j)$	$(X_i \geq Y_j)$	$(X_i < Y_j)$
(1, 1)	1	0
(0, 1)	1	$\hat{F}^0(Y_j)/\hat{F}^0(X_i)$
(1, 0)	$1 - \hat{G}^0(X_i)/\hat{G}^0(Y_j)$	0
(0, 0)	$1 - \hat{G}^0(X_i)/\hat{G}^0(Y_j) - \int_{X_i}^{\infty} \hat{F}^0(s)\, d\hat{G}^0(s)/\hat{F}^0(X_i)\hat{G}^0(Y_j)$	$-\int_{Y_j}^{\infty} \hat{F}^0(s)\, d\hat{G}^0(s)/\hat{F}^0(X_i)\hat{G}^0(Y_j)$

The programs are freely obtainable from Z. J., and are also available through this journal's web site. Insofar as computing costs attending calculation of Efron's estimator are negligible, and especially as the estimator enjoys favorable statistical properties, we urge that his estimate receive more widespread attention.

Appendix

In Efron's paper (1967), the scoring function was defined as

$$\hat{Q}(X_i, Y_j, \delta_i, \varepsilon_j) = \Pr\{X_i^0 \geq Y_j^0 | X_i, Y_j, \delta_i, \varepsilon_j, \hat{F}^0, \hat{G}^0\},$$

where $\hat{F}^0$ and $\hat{G}^0$ represent the self-consistent estimates of F^0 and G^0, respectively. The values of $Q(X_i, Y_j, \delta_i, \varepsilon_j)$ in all eight possible different cases are listed in Table A1. Efron's C statistic (denoted by $\hat{W}$) is then defined as:

$$\hat{W} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \hat{Q}(X_i, Y_j, \delta_i, \varepsilon_j).$$

Two accompanying theorems that had been proved by Efron are restated as follows:

Theorem 1. $\hat{W} = -\int_{-\infty}^{\infty} \hat{F}^0(s)\, d\hat{G}^0(s) = \Pr\{\hat{F}^0 \geq \hat{G}^0\}$, and is the maximum likelihood estimate of $\Pr\{\hat{F}^0 \geq \hat{G}^0\}$.

Theorem 2. Let m and n go to infinity in such a way that $\lim m/(m+n) = \lambda$, with $0 < \lambda < 1$. Then $(m+n)^{1/2}[\hat{W} - \Pr\{\hat{F}^0 \geq \hat{G}^0\}] \longrightarrow N(0,\, 1/\lambda\,\sigma_1^2 + 1/1 - \lambda\,\sigma_2^2)$, where under the null hypothesis, $\sigma_1^2 = \frac{1}{4}\int_0^1 z^2\, dz/H[F^{0^{-1}}(z)] = \frac{1}{4}\int_0^1 z^3\, dz/F[F^{0^{-1}}(z)]$ and $\sigma_2^2 = \frac{1}{4}\int_0^1 z^2\, dz/I[G^{0^{-1}}(z)] = \frac{1}{4}\int_0^1 z^3\, dz/G[G^{0^{-1}}(z)]$. Here $F^{0^{-1}}$ and $G^{0^{-1}}$ are the inverse functions of F^0 and G^0, respectively, and are identical under the null hypothesis.

Acknowledgements The authors thank the editor and reviewers for insightful comments, and Dr. Dan Mercola (UCI) and the NIH for support (U01CA114810; P01CA104898).

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Acion, L., Peterson, J. J., Temple, S. and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine* **25**, 591–602.

Begg, C. B., Cramer, L. D., Venkatraman, E. S. and Rosai, J. (2000). Comparing tumour staging and grading systems: a case study and a review of the issues, using thymoma as a model. *Statistics in Medicine* **19**, 1997–2014.

Brown, B. W., Hollander, M. and Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies. in: F. Proschan and R. J. Serfling (Eds.), *Reliability and Biometry*, SIAM, Philadelphia, 327–354.

D'Agostino, R. B. and Nam, B.-H. (2004). Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of Statistics* **23**, 1–24.

Efron, B. (1967). The two-sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **4,** 831–853.

Graefen, M., Karakiewicz, P. I., Cagiannos, I., Quinn, D. I., Henshall, S. M., Grygeil, J. J., Sutherland, R. L., Stricker, P. D., Klein, E., Kupelian, P. , Skinner, D. G., Lieskovsky, G., Bochner, B., Huland, H., Hammerer, P. G., Haese, A., Erbersdobler, A., Eastham, J. A., de Kernion, J., Cangiano, T., Schröder, F. H., Wildhagen, M. F., van der Kwast, T. H., Scardino, P. T. and Kattan, M. W. (2002). International validation of a preoperative nomogram for prostate cancer recurrence after radical prostatectomy. *Journal of Clinical Oncology* **20**, 3206–3212.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982). Evaluating the yield of medical tests. *The Journal of the American Medical Association* **247**, 2543–2546.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

Koziol, J. A., Feng, A. C., Jia, Z., Wang, Y., Goodison, S., McClelland, M. and Mercola, D. (2009). The wisdom of the commons: ensemble tree classifiers for prostate cancer prognosis. *Bioinformatics* **25**, 54–60.

Newcombe, R. G. (2006a). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 1: general issues and tail-area-based methods. *Statistics in Medicine* **25**, 543–557.

Newcombe, R. G. (2006b). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: asymptotic methods and evaluation. *Statistics in Medicine* **25**, 559–573.

Newson, R. (2006). Confidence intervals for rank statistics: Somers' D and extensions. *Stata Journal* **6**, 309–334.

O'Quigley, J. and Flandre, P. (1994). Predictive capability of proportional hazards regression. *Proceedings of the National Academy of Sciences* **91,** 2310–2314.

O'Quigley, J., Xu, R. and Stare, J. (2005). Explained randomness in proportional hazards models. *Statistics in Medicine*. **24**, 479–489.

Pencina, M. J. and D'Agostino, R. B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine* **23**, 2109–2123.

Royston, P. and Sauerbrei, W. (2004). A new measure of prognostic separation in survival data. *Statistics in Medicine* **23**, 723–748.

Schemper, M. and Stare, J. (1996). Explained variation in survival analysis. *Statistics in Medicine* **15**, 1999–2012.

Schumacher, M., Graf, E. and Gerds, T. (2003). How to assess prognostic models for survival data: a case study in oncology. *Methods of Information in Medicine* **42**, 564–571.

