Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

A practical divergence measure for survival distributions that can be estimated from Kaplan-Meier curves

Trevor F. Cox[a*] and Gabriela Czanner[b]

This paper introduces a new simple divergence measure between two survival distributions. For two groups of patients, the divergence measure between their associated survival distributions is based on the integral of the absolute difference in probabilities that a patient from one group dies at time t and a patient from the other group survives beyond time t and vice versa. In the case of non-crossing hazard functions, the divergence measure is closely linked to the Harrell concordance index, C, the Mann-Whitney test statistic and the area under a Receiver Operating Characteristic curve. The measure can be used in a dynamic way where the divergence between two survival distributions from time zero up to time t is calculated enabling real-time monitoring of treatment differences. The divergence can be found for theoretical survival distributions or can be estimated non-parametrically from survival data using Kaplan-Meier estimates of the survivor functions. The estimator of the divergence is shown to be generally unbiased and approximately normally distributed. For the case of proportional hazards, the constituent parts of the divergence measure can be used to assess the proportional hazards assumption. The use of the divergence measure is illustrated on the survival of pancreatic cancer patients.

Keywords: crossing hazard functions; divergence measures; Kaplan-Meier curves; Kullback-Leibler divergence; multidimensional scaling

1. Introduction

Time-to-event data frequently occur in medicine, e.g. death-times of patients on standard care versus those for patients on a new intervention. Data collected, perhaps in a clinical trial, are generally analysed using standard techniques, such as Kaplan-Meier curves and by testing the equivalence of the survival distributions using the log-rank test, e.g. [1]. The log-rank test is optimal for the case of proportional hazards, but other tests that are more powerful have been proposed for when proportional hazards cannot be assumed; see for example the review article by [2]. However, simply carrying out a test of the null hypothesis that two survival curves are identical should not be the end of the statistical analysis, except perhaps if a test of equivalence has been carried out and it is deemed that the survival curves are the same, but even then some further description of the common survival curve would be needed. For differing survival curves, further analysis would highlight the differences in survival between the two groups and one way to measure this is by a divergence measure between the two distributions.

Survival distributions are usually compared over all time, or at least for the time period where data are available. However, some authors have considered the case where distributions are compared dynamically, see for example [3, 4] where test statistics are monitored over time. A real-time comparison measure of two survival distributions that changes over time could be used as a complement to hypothesis testing.

[a] *Cancer Research UK Liverpool Cancer Trials Unit, University of Liverpool, UK*

[b] *Department of Biostatistics, University of Liverpool, UK*

[*] *Correspondence to: Cancer Research UK Liverpool Cancer Trials Unit, University of Liverpool, Block C Waterhouse Building, 1-3 Brownlow Street, Liverpool, L69 3GL, UK. E-mail: coxt@liv.ac.uk*

In this paper, it is assumed that there are two groups, F and G, with continuous survival functions $S_F(t)$, $S_G(t)$ and corresponding density functions, $f(t)$ and $g(t)$, although for some results in this paper, continuity is not necessary. Without loss of generality, it is also assumed that f and g have support on $(0, \infty)$. Let the corresponding cumulative distribution functions (cdf) be $F(t)$ and $G(t)$.

In general, one frequently used measure of divergence between two continuous densities f and g is the Kullback-Leibler measure (KL). It is the principal information function introduced by [5] and it measures the discrepancy between two theoretical continuous density distributions, also known as relative entropy of X and Y:

$$K(f,g) = \int_0^\infty f(x) \ln\left\{\frac{f(x)}{g(x)}\right\} dx.$$

Now $K(f,g)$ is non-negative, shift and scale invariant and is asymmetric with equality of $K(f,g)$ and $K(g,f)$ holding if and only if $f(x) = g(x)\ a.e.$ Clearly, $K(f,g)$ is not a metric because it is asymmetric and it does not satisfy the triangle inequality. A symmetric extension was introduced by Kullback and Leibler as $K(f,g)+K(g,f)$. Typically, $K(f,g)$ is estimated in two steps by first estimating the parameters of the densities, f and g, and then substituting these parameter estimates into the formula of the KL divergence, see e.g. [6]. Perez-Cruz [7] proposed an estimate based on the empirical distribution functions obtained from two independent samples, and showed that this estimate converges almost surely to the actual true divergence. Apart from not being symmetric, another disadvantage of the KL divergence is that it was not designed as a function of time, i.e. it is not dynamic. However, a dynamic extension of the KL divergence principle to survival data is due to Ebrahimi and Kirmani [8] who proposed a measure of divergence between two residual-life distributions. Later, a dual KL divergence measure was proposed between two past-lives distributions by Di Crescenzo and Longobardi [9]. The divergence of Ebrahimi and Kirmani is

$$D(f,g;t) = \int_0^t \frac{f(x)}{S_F(x)} \ln\left\{\frac{f(x)/S_F(x)}{g(x)/S_G(x)}\right\} dx.$$

This measure is dynamic and for the case of proportional hazards this divergence measure can be shown to be constant. However, this measure is not symmetric, i.e. $D(f,g;t) \neq D(g,f;t)$ and in some distributions the integrand will be negative, i.e. for some x

$$\ln\{f(x)S_G(x)\} - \ln\{g(x)S_F(x)\} < 0.$$

This paper proposes a new divergence measure between two survival distributions where the main criteria for the measure were that it had to be practical, have medical motivations, be symmetric, non-negative and be dynamic enabling real-time monitoring of treatment differences. It needed to be effectively estimated from empirical distributions, i.e. without making assumptions of the parametric representation and without needing parameters to be estimated and, lastly, it needed to be easy interpreted. The new divergence measure can be used to measure the overall difference between two survival curves or used to see where curves start to separate in time, which can happen when two treatments are equally efficacious early on after start of treatment, but then one becomes more efficacious than the other as time passes. For several groups, pairwise divergences can be calculated and used to describe differences among all the groups. The new divergence measure can be used to test for equivalent survival curves, but this is not seen as the primary purpose of the measure. In fact, the measure is very closely allied to the test statistic proposed by Cox [10] for testing equivalence of survival curves and it is argued later that this test should be used in preference to the divergence measure for the purpose of hypothesis testing. In general, the estimate of the divergence measure is shown to be approximately unbiased and normally distributed. Data on survival of pancreatic cancer patients are used to illustrate the use of the new divergence measure.

2. A divergence measure for survival distributions

The functions, $S_F(t)$, $S_G(t)$, $f(t)$ and $g(t)$ have already been defined. Let $\lambda_F(t)$ and $\lambda_G(t)$ be the associated hazard functions. Consider $|S_F(t)g(t) - S_G(t)f(t)|$ which, when multiplied by δt, is essentially the difference in probabilities that a patient from one group dies in the interval $[t, t+\delta t)$ and a patient from the other group outsurvives him/her and vice versa. The proposed measure of the divergence, D_{FG}, between the survival distributions, S_F and S_G and dropping the argument t in functions for convenience, is

$$D_{FG} = \int_0^\infty |S_F g - S_G f|. \tag{1}$$

Evaluation of the integral in equation (1) is generally not straightforward because of the absolute difference between $S_F g$ and $S_G f$. However, for the case of non-crossing hazard functions with $S_F g \geq S_G f$ (or vice versa) for all t,

$$D_{FG} = \int_0^\infty (S_F g - S_G f) = 2\int_0^\infty S_F g - 1 = 2\mathrm{Pr}(T_F > T_G) - 1,$$

using integration by parts and where T_F and T_G are survival times of subjects randomly chosen from groups F and G respectively. Alternatively, $D_{FG} = 1 - 2\int_0^\infty S_G f$ and so for the case of non-crossing hazard functions, any of the following three formulae for calculating D_{FG} can be used

$$\left|\int_0^\infty (S_F g - S_G f)\right|, \left|2\int_0^\infty S_F g - 1\right|, \left|2\int_0^\infty S_G f - 1\right|. \tag{2}$$

Note that $\int_0^\infty S_F f = 0.5$ for any survival distribution and so $D_{FF} = 0$. The range of the divergence measure is $[0, 1]$.

To put the divergence measure in context, it is closely linked to $\mathrm{Pr}(T_F > T_G)$, especially for the case of non-crossing hazard functions. The quantity, $\mathrm{Pr}(T_F > T_G)$, which, together with its estimated value, have been widely used in the past. It has been suggested that $\mathrm{Pr}(T_F > T_G)$ be a measure of effect size as an alternative to the usual effect size based on means [11]. When there is no censoring, $\mathrm{Pr}(T_F > T_G)$ is equal to the Mann-Whitney test statistic [12, 13], is equal to the area under a Receiver Operating Characteristic (ROC) curve [14, 15] and equal to Harrell's concordance index C [16, 17]. These connections are discussed by Schemper *et al.* [18], together with the connection to the average hazard ratio as defined by Kalbfleisch and Prentice [19]. Here, we are using $\mathrm{Pr}(T_F > T_G)$ in the different context as a divergence measure, which possibly could be calculated from the other manifestations of $\mathrm{Pr}(T_F > T_G)$ (Harrell's C, etc.), but not when hazard functions cross.

The concept of "divergence by time t" is a useful measure to chart the progress of the divergence as time progresses. This is defined as

$$D_{FG}(t) = \int_0^t |S_F g - S_G f|,$$

or using the other formulae in (2) if appropriate.

Example Let $S_F = \exp(-t)$ and $S_G = \exp\{-(1.3t)^{3.52}\}$. Figure 1(a) shows these survival distributions, Figure 1(b) the density functions, Figure 1(c) the hazard functions, Figure 1(d) $\int_0^t S_F g$ and $\int_0^t S_G f$ and Figure 1(e) the divergence up to time t. The first distribution is an exponential and the second a Weibull chosen so that the distributions have the same median (0.69). Clearly the hazard functions cross and so the formula in (1) has to be used for calculating the divergence. The divergence up to time t rapidly increases from zero, flattens at the point where the hazard functions cross and then increases again, reaching a plateau at approximately $t = 1.1$ and giving total divergence of 0.533. The reason for choosing this somewhat contrived example is that medical practitioners (and statisticians) loosely compare survival curves by simply quoting the median survival. In this example the medians are equal, but clearly the survival curves are very different from one another as indicated by the large divergence. So in practice, it might be useful to quote both the median survival times and the divergence.

2.1. *Divergence for proportional hazards*

For the case of proportional hazards, $\lambda_G = \gamma\lambda_F$, where γ is the hazard ratio. Integrating $S_F g$ by parts gives

$$\int_0^t S_F g + \int_0^t S_G f = 1 - S_F(t)S_G(t)$$

and hence

$$\int_0^t S_F S_G \lambda_G + \int_0^t S_F S_G \lambda_F = 1 - S_F(t)S_G(t).$$

Substituting $\gamma\lambda_F$ for λ_G, gives

$$\int_0^t S_G f = [1 - S_F(t)S_G(t)]/(1+\gamma), \quad \int_0^t S_F g = [1 - S_F(t)S_G(t)]\gamma/(1+\gamma)$$

and thus $\int S_F g$ and $\int S_G f$ are in the proportions $\gamma : 1$ for all $t (> 0)$ and

$$D_{FG} = |(\gamma - 1)/(\gamma + 1)|. \tag{3}$$

Consequently, $\gamma = (1 + D_{FG})/(1 - D_{FG})$ or the reciprocal of this value, depending on which survival distribution is chosen for the baseline hazard function.

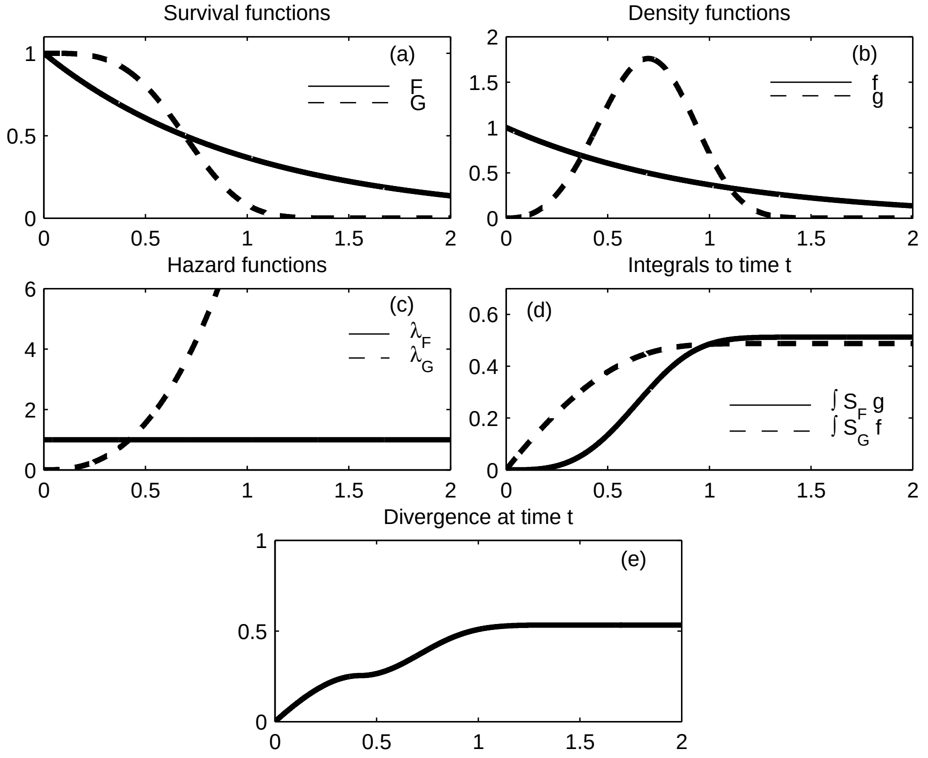

Figure 1. Plots of (a) survival curves, (b) density functions, (c) hazard functions, (d) integrals $\int_0^t S_F g$ and $\int_0^t S_G f$ to time t and (e) $D_{FG}(t)$ for the survivor distributions $S_F = \exp(-t)$ and $S_G = \exp\{-(1.3t)^{3.52}\}$.

2.2. D_{FG} is a metric for the case of proportional hazards

To show D_{FG} is a metric, the following properties have to be established: (i) $D_{FG} \geq 0$, (ii) $D_{FG} = 0 \iff D_F \equiv S_G$, (iii) $D_{FG} = D_{GF}$ and (iv) $D_{FH} \leq D_{FG} + D_{GH}$. It is possible to find counterexamples to show that D_{FG} is in general not a metric. This can be done by randomly choosing parameter values for three Weibull distributions with scale and shape parameters, (λ, γ), calculating D_{FG}, D_{FH} and D_{GH} numerically and then noting which sets of parameter values give divergencies that do not satisfy the triangle inequality. One counterexample is given by the parameter values $(1.2, 3.1)$, $(2.1, 1.7)$ and $(3.4, 2.0)$ leading to divergences 0.475, 0.647 and 0.153. It should be noted that the triangle inequality held most of the time and when it did not, the value of $D_{FG} + D_{GH} - D_{FH}$ was close to zero..

However, for the proportional hazards case, D_{FG} is a metric. Let S_F, S_G and S_H be three survival functions with proportional hazards where $S_G = S_F^{\gamma_G}$, $S_H = S_F^{\gamma_H}$ and, without loss of generality, let $\gamma_H \geq \gamma_G \geq 1$.

Conditions (i) and (iii) are clearly satisfied. For (ii), if $S_F \equiv S_G$, then $D_{FG} = 0$ trivially. Conversely, if $D_{FG} = 0$, then using the third formula in (2), $\int S_G f = \frac{1}{2}$. Hence $\int S_F^{\gamma_G} f = \frac{1}{2}$. Upon integration it is easily seen that $\gamma_G = 1$ showing S_F and S_G to be identical. To show (iv),

$$D_{FG} = \frac{\gamma_G - 1}{\gamma_G + 1} \qquad D_{FH} = \frac{\gamma_H - 1}{\gamma_H + 1} \qquad D_{GH} = \frac{\gamma_H - \gamma_G}{\gamma_H + \gamma_G}$$

and then

$$\begin{aligned} D_{FG} + D_{FH} - D_{GH} &= (\gamma_G - 1)(\gamma_G + 3\gamma_H + 3\gamma_G\gamma_H + \gamma_H^2)/A, \\ D_{FG} + D_{GH} - D_{FH} &= (\gamma_H - 1)(\gamma_G - 1)(\gamma_H - \gamma_G)/A, \\ D_{FH} + D_{GH} - D_{FG} &= (\gamma_H - \gamma_G)(3\gamma_G + 3\gamma_H + \gamma_G\gamma_H + 1)/A, \end{aligned}$$

where $A = (\gamma_G + 1)(\gamma_H + 1)(\gamma_H + \gamma_G)$ and hence all three quantities are non-negative, confirming (iv).

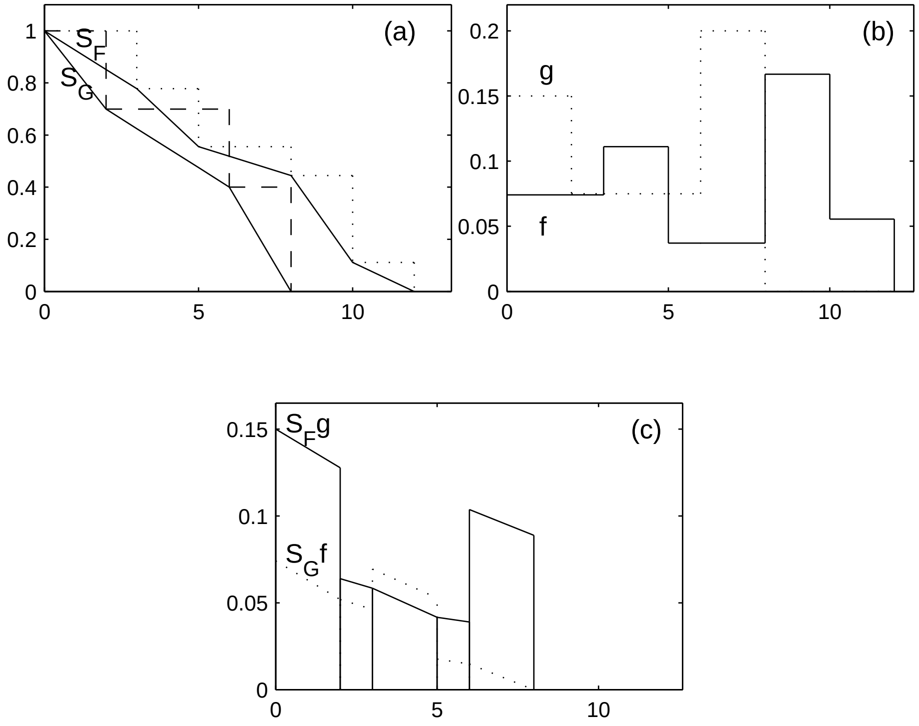

Figure 2. Two simple survival curves for illustration of the calculation of $\int \hat{S}_F \hat{g}$ and $\int \hat{S}_G \hat{f}$: (a) estimated survival functions, (b) estimated density functions and (c) estimation of $\hat{S}_F \hat{g}$ and $\hat{S}_G \hat{f}$

2.3. Kullback-Leibler divergence for proportional hazards

For the symmetric Kullback-Leibler divergence, $KL_{FG} = K(f, g) + K(g, f) = \int (f - g)\{\ln(f) - \ln(g)\}$. Using the equations, $g/S_G = \gamma f / S_F$ and $S_G = S_F^\gamma$ for the proportional hazards case, it is easily shown after some algebra that

$$KL_{FG} = \frac{(\gamma - 1)^2}{\gamma}. \tag{4}$$

3. Calculating the divergence from data

The integral $\int S_F g$ can be estimated using empirical survival distributions. Let there be N_F and N_G observations from the two groups respectively, where random censoring may have occurred. Let $\hat{S}_F$ and $\hat{S}_G$ be the Kaplan-Meier estimates of the two survival distributions, e.g. [1]. (Figure 5 shows fifteen of these for a cancer example described in Section 4.) Rather than using the Kaplan-Meier step function as an estimate of the survival distribution, where the function is constant between adjacent calculated survival probabilities, instead, survival probabilities are linearly interpolated between the adjacent points. The probability density functions, f and g will be estimated as piecewise constant functions obtained from the slopes of the estimated survival function. This is illustrated in Figure 2 where (a) shows two very simple Kaplan-Meier plots (dashed and dotted lines) and the estimated distribution functions to be used in the calculations (solid lines) and (b) the estimated density functions where the constant intervals are open to the right (vertical lines are displayed to aid visualisation). To estimate $\int S_F g$, the points of change in values of $\hat{S}_F$ and of $\hat{S}_G$ are amalgamated and then $\hat{S}_F$ and $\hat{g}$ are calculated for each of the new change point intervals and then used for the estimates $\hat{S}_F \hat{g}$ and similarly for $\hat{S}_G \hat{f}$. These estimated functions each form a sequence of trapeziums as shown in Figure 2(c). It can easily be shown that the pairwise trapeziums for $\hat{S}_F \hat{g}$ and $\hat{S}_G \hat{f}$ have their sloping top sides parallel. The integral $\int_0^t \hat{S}_F \hat{g}$ is then just a sum of areas of trapeziums for $\hat{S}_F \hat{g}$ up to time t and similarly for $\int_0^t \hat{S}_G \hat{f}$.

In practice, let there be N amalgamated intervals covering the range of the estimated survival distributions and let their values be $F_0(=1), F_1, \ldots, F_N$ and $G_0(=1), G_1, \ldots, G_N$ at the time points, $t_0(=0), t_1, \ldots, t_N$, defining the intervals. Then the trapeziums for $\int \hat{S}_F \hat{g}$ have coordinates

$$\begin{gathered}\{t_i, 0\} \quad \{t_i, F_i(G_i - G_{i+1})/(t_{i+1} - t_i)\}, \\ \{t_{i+1}, 0\} \quad \{(t_{i+1}, F_{i+1}(G_i - G_{i+1})/(t_{i+1} - t_i)\},\end{gathered}$$

with areas $(F_i + F_{i+1})(G_i - G_{i+1})/2$. Adding these areas gives the estimate of $\int S_G f$ as

$$\widehat{\int S_F g} = \{F_0G_0 + \sum_{i=0}^{N-1}(F_{i+1}G_i - F_iG_{i+1}) - F_NG_N\}/2 \tag{5}$$

and similarly for $\int S_G g$

$$\widehat{\int S_G f} = \{F_0G_0 - \sum_{i=0}^{N-1}(F_{i+1}G_i - F_iG_{i+1}) - F_NG_N\}/2. \tag{6}$$

If the hazard functions λ_F and λ_G do not cross, then D_{FG} can be estimated by $|\widehat{\int S_F g} - \widehat{\int S_G f}|$. If the hazard functions do cross, at a point t_0 say, then D_{FG} has to be estimated by the sum of $\widehat{\int S_F g} - \widehat{\int S_G f}$ calculated for $0 \leq t < t_0$ and $\widehat{\int S_G f} - \widehat{\int S_F g}$ calculated for $t \geq t_0$, or vice versa to make the quantities positive. A simple method for deciding if and where the hazard functions cross, is to plot $\widehat{\int S_F g}$ and $\widehat{\int S_G f}$ against t and look for a clear global maximum. If there is one, the crossing point of the hazard functions is estimated as the point at which this maximum occurs.

Note, the estimates (5) and (6) added together give $1 - F_NG_N$ as required since, for the population survival functions, $\int_0^t S_F g + \int_0^t S_G f = 1 - S_F(t)S_G(t)$. If the step Kaplan-Meier functions had been used, then the sum of the two estimates would not have been equal to this required value.

An alternative approach would be to calculate the estimates of $\int S_F g$ and $\int S_G f$ (or written as $\int S_F dS_G$ and $\int S_G dS_F$) using Riemann-Stieltjes integrals, $\int F dG$ and $\int G dF$, circumventing the previous arguments about areas of trapeziums. However, the argument via trapeziums is more intuitive, the same comment applies regarding the step Kaplan-Meier functions not giving correct values and a connection of $\hat{D}_{FG}$ to a test for the equivalence of survival functions, based on PP-plots, would probably have been missed. This will be described later. The Riemann-Stieltjes approach does however allow consideration of the asymptotic distribution of $\hat{D}_{FG}$.

If the last data point is censored in each group, the cut-off for the divergence will be the minimum time of the last data points, one from each group. If an assumption that the decay in the survival curves after this cut-off point, $tmax$ say, is the same for both survival curves, then the overall divergence will be the same as the divergence seen at $tmax$. For example, let $S_F(t) = S_{F1}(t)I\{t < tmax\} + S_{F1}(tmax)\exp(-\lambda t)I\{t \geq tmax\}$ and $S_G(t) = S_{G1}(t)I\{t < tmax\} + S_{G1}(tmax)\exp(-\lambda t)I\{t \geq tmax\}$ where $I\{.\}$ is the indicator function. Then it is easily seen that for $t \geq tmax$, $\int_0^t S_F g - \int_0^t S_G f = \int_0^{tmax} S_{F1}g_1 - \int_0^{tmax} S_{G1}f_1$, showing the divergence does not increase after time $tmax$.

3.1. The distribution of $\hat{D}_{FG}$

Appendix 7.1 shows how the variance of $\hat{D}_{FG}(t)$ can be estimated using the estimated variance of the Kaplan-Meier function at the jump points. First the F_i's and G_i's are related back to the original jump points, labelled $\boldsymbol{\mathcal{F}}_i$ and $\boldsymbol{\mathcal{G}}_i$. Then after some matrix algebra it is shown that

$$\mathrm{E}(\hat{D}_{FG}) \approx \boldsymbol{\mathcal{F}}^T \mathbf{A} \boldsymbol{\mathcal{G}}, \quad \mathrm{var}(\hat{D}_{FG}) \approx \mathrm{tr}(\mathbf{A}^T \mathbf{\Phi_F} \mathbf{A} \mathbf{\Phi_G}) - \mathrm{tr}(\mathbf{A}^T \boldsymbol{\mathcal{F}} \boldsymbol{\mathcal{F}}^T \mathbf{A} \boldsymbol{\mathcal{G}} \boldsymbol{\mathcal{G}}^T),$$

where $\mathbf{A}$, $\mathbf{\Phi_F}$ and $\mathbf{\Phi_G}$ are defined in the Appendix. Appendix 7.2 shows that $\hat{D}_{FG}(t)$ is asymptotically unbiased (N large rather than t large) with an asymptotic normal distribution. A counting process approach using martingales to show this is outlined.

An alternative method is to use bootstrap samples to estimate the standard deviation of $\hat{D}_{FG}$. These two approaches were compared using pairs of exponential survival distributions with various rates and amount of censoring. One thousand bootstrap draws were used. Detailed results are not given here since the agreement between the two approaches was always very good with maximum difference 0.003. For instance using $\lambda_F = 1.0$, $\lambda_G = 2.0$, $N_F = 100$, $N_G = 100$ with 20% censoring, the estimated standard deviation using the method described above was 0.081 and using the bootstrap, 0.080.

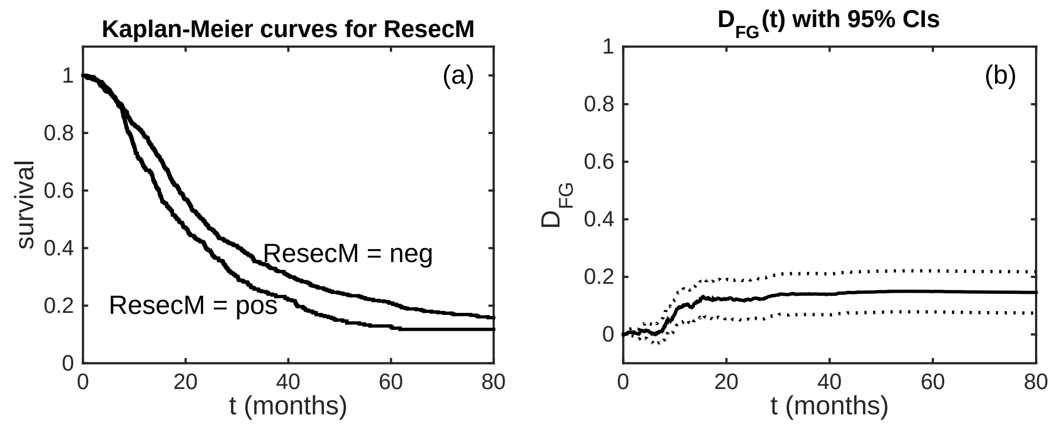

Figure 3. Kaplan-Meier plots and $\hat{D}_{FG}(t)$ for pancreatic cancer survival

3.2. *Example: Divergence of survival distributions for pancreatic cancer patients*

The ESPAC-3 trial was one in a series of pancreatic cancer trials aimed at improving treatment for patients for whom the collective five-year survival rate is less than 5%. Patients in the trial underwent surgical resection and were then randomised to chemotherapy by either Fluorouracil plus Folinic Acid (F) or Gemcitabine (G). The five-year survival rate for patients who are able to have surgery can be up to 15%. The initial results from the trial are reported in [20]. The median survival time (95% CI) for the F arm was 23.0 (21.1-25.0) months and for the G arm, 23.6 (21.4-26.4) months. The log-rank test for testing the equivalence of the survival distributions gave a test statistic value of 0.70 on 1 degree of freedom, with $p = 0.39$. Some key prognostic factors for pancreatic cancer are: *lymph node involvement* where the cancer has spread to the lymph nodes (yes/no), *resection margin* where pathology determines whether the surgical margin is clear of cancer cells (negative/positive) and *tumour grade* a measure of how closely tumour cells resemble ordinary cells, where "well differentiated" means that tumour cells appear very similar to ordinary cells, "moderately differentiated", less so and "poorly differentiated" that the tumour cells are markedly different.

The example here illustrates the divergence of the survival distributions for the two groups *Resection Margin = positive* and *Resection Margin = negative.* The data are also used in Section 4 where several groups are considered based on treatment and three prognostic factors. Figure 3(a) shows the Kaplan-Meier curves for the two groups. There is a clear survival advantage for the *negative* group. The log-rank statistic has value 26.4 on 1 df ($p < 0.0001$); the median survival times for the two groups are 23.4 and 18.2 months. Figure 3(b) shows the divergence measured up to time t. Clearly the survival curves do not diverge until approximately 7 months and then the divergence steadily increases to a value 0.13 at 15 months but then remains more or less constant but ending at a value of 0.16 overall. The pointwise 95% confidence intervals are also given. They show that the null hypothesis of zero divergence would be rejected leading to different survival curves. A point of clarity is needed here. Suppose the data had arisen from the same survival distribution or distributions that were very close, and so the divergence is zero or close to zero. The estimated divergence would have been close to zero and could be positive or negative depending on whether $\int \hat{S}_F d\hat{S}_G - \int \hat{S}_G d\hat{S}_F$ or $\int \hat{S}_G d\hat{S}_F - \int \hat{S}_F d\hat{S}_G$ was used. Upon taking the modulus, the asymptotic distribution would change from normal to half-normal with consequences for the confidence intervals. We prefer to keep things simple and allow graphs with negative divergencies, although one knows they are not possible theoretically.

3.3. *The connection with $\hat{D}_{FG}$ with a test statistic for the equality of survival functions*

Cox [10] proposed a test statistic for testing the equality of survival functions based on PP-plots, where one Kaplan-Meier curve is plotted against another. The test statistic is defined as the absolute area between the resulting curve and the diagonal from $(0, 0)$ to $(1, 1)$. For illustration, Figure 4 shows a PP-plot (solid line) for two small samples of survival times. The area between the PP-plot and the diagonal can clearly be seen. Also plotted is the piecewise linear function of $\hat{S}_G$ against $\hat{S}_F$ used in estimating the divergence (dotted line). For large sample sizes, these two functions will be very close. In the figure, the area of the trapezium highlighted by the horizontal shading is $0.5(F_i + F_{i+1})(G_i - G_{i+1})$ and the area of the trapezium highlighted by the vertical shading is $0.5(G_i + G_{i+1})(F_i - F_{i+1})$. It can be seen that the sum of the areas of the set of trapeziums parallel to the horizontal will be the estimate of $\int S_F dS_G$ and the sum of those parallel to the vertical axis, $\int S_G dS_F$. The difference in these is $\hat{D}_{FG}$ which can be seen to be twice the area between the piecewise linear function and the diagonal. Hence the connection with the test statistic. Note, the test statistic could have been defined using the piecewise linear function and then the test statistic and $\hat{D}_{FG}$ would be identical. However, if survival curves

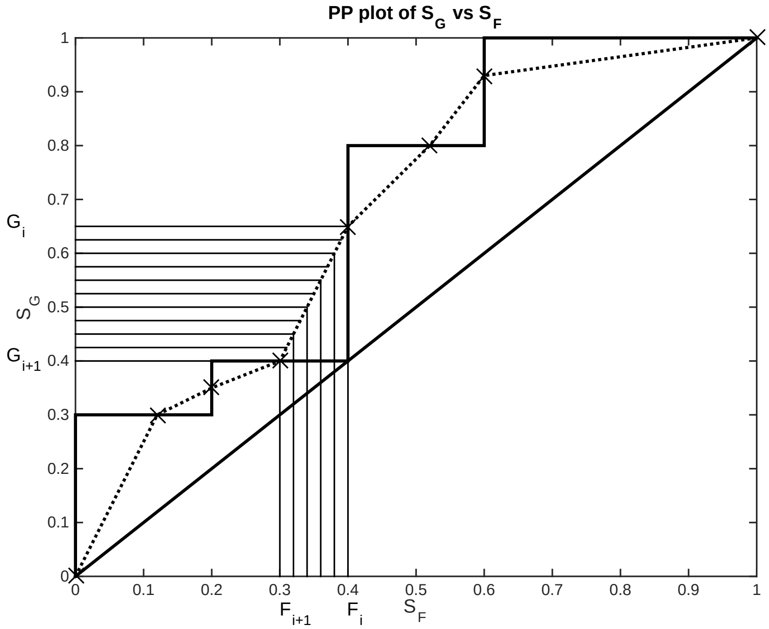

Figure 4. The connection between a test based on PP-plots and $\hat{D}_{FG}$: the solid line is the PP-plot and dotted line a plot of the $\hat{S}_G$ against $\hat{S}_F$

cross (which would be seen by the plotted functions in Figure 4 crossing the diagonal) or the hazard functions cross, then this agreement will not hold. (If hazard functions do not cross, then survival curves cannot cross, but if survival curves do not cross it does not imply that hazard functions do not cross.) In practice, testing equivalence of survival curves is best carried out using the area test statistic, leaving $\hat{D}_{FG}$ to measure divergence. Another argument against using $\hat{D}_{FG}$ to test for equivalent survival curves is that crossing hazard functions need to be detected before calculating the divergence. This is somewhat subjective but can be done. It is similar to the problem with the log-rank statistic which fails for crossing survival curves [21, 22] or where proportional hazards are far from reality. The statistic proposed by [10] is robust against all these situations.

3.4. Comparison of $\hat{D}_{FG}$ and Kullback-Leibler divergence

Pérez-Cruz [7] shows how the Kullback-Leibler divergence can be estimated from the empirical distribution functions obtained from two samples. The estimate uses a similar approach to the one used in this paper. It is shown that the estimate is better than one using estimates of density functions given by [6]. Briefly, the estimator suggested for $K(f,g)$ is as follows. Let $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$ be independent, ordered random samples from distributions F and G respectively. Let the empirical cdf for F be $\hat{F}(x)=n^{-1}\sum_{i=1}^{n}U(x-x_i)$ where $U(x)$ is the unit step function with $U(0)=0.5$ and similarly for G. The empirical distribution functions are linearly interpolated as described previously for the proposed divergence measure. Then the estimate of $K(f,g)$ is

$$\hat{K}(f,g)=\frac{1}{n}\sum_{i=1}^{n_f}\ln\left\{\frac{d\hat{F}(x_i)/dx_i}{d\hat{G}(y_i')/dy_i'}\right\}$$

where $d\hat{F}(x_i)=\hat{F}(x_i)-\hat{F}(x_i-1)$, $dx_i=x_i-x_{i-1}$, $d\hat{G}(y_{i'})=\hat{G}(y_{i'})-\hat{G}(y_{i'-1})$ and $dy_{i'}=y_{i'}-y_{i'-1}$, and where $y_{i'}$ is the sample value from G that is the smallest one of all those that are greater than x_i. Pérez-Cruz does not consider the case of censored data, but this can be easily allowed for in the formula for the estimator. The equivalent estimator, $\hat{K}(g,f)$ can be similarly constructed and then $\widehat{KL}_{FG}$ calculated by combining the two.

A comparison of $\hat{D}_{FG}$ and $\widehat{KL}_{FG}$ was carried for the case of proportional hazards. Random samples from a pair of exponential distributions with parameter values 1 and γ were simulated and then $\hat{D}_{FG}$ and $\widehat{KL}_{FG}$ calculated from (3) and (4). This was done one thousand times and means, standard deviations and mean square error (mse) found. Table 1 shows the results for sample sizes $n_1=n_2=100$, with and without censoring and for $n_1=n_2=1000$. For each value of γ the true values of D_{FG} and KL_{FG} are also shown. Although the scales of the two divergence measures are different, it can be seen that $\hat{D}_{FG}$ outperforms $\widehat{KL}_{FG}$ in terms of accuracy and precision. Note, the standard deviation for $\widehat{KL}_{FG}$ tends to be very large and for large values of γ, KL_{FG} is very much under estimated. (This reflects what was seen in [7] in that the

Table 1. Comparison of $\hat{D}_{FG}$ and $\widehat{KL}_{FG}$ for proportional hazards

	γ	D_{FG}	mean	sd	mse	KL_{FG}	mean	sd	mse
$n_1 = n_2 = 100$	1.0	0	0.064	0.049	0.006	0	0.008	0.225	0.051
	1.2	0.091	0.098	0.066	0.004	0.033	0.025	0.235	0.055
	1.5	0.2	0.199	0.077	0.006	0.167	0.079	0.246	0.068
	2.0	0.333	0.331	0.077	0.006	0.5	0.197	0.278	0.169
	3.0	0.5	0.502	0.068	0.005	1.333	0.343	0.342	1.098
$n_1 = n_2 = 100$	1.0	0	0.064	0.047	0.006	0	-0.039	0.245	0.061
with 20% censoring	1.2	0.091	0.102	0.068	0.005	0.033	-0.037	0.237	0.071
	1.5	0.2	0.202	0.077	0.006	0.167	-0.014	0.270	0.105
	2.0	0.333	0.338	0.077	0.006	0.5	0.046	0.303	0.298
	3.0	0.5	0.500	0.065	0.004	1.333	0.106	0.329	1.616
$n_1 = n_2 = 1000$	1.0	0	0.020	0.016	0.001	0	0.002	0.071	0.005
	1.2	0.091	0.091	0.027	0.001	0.033	0.036	0.076	0.006
	1.5	0.2	0.199	0.026	0.001	0.167	0.138	0.084	0.008
	2.0	0.333	0.332	0.024	0.001	0.5	0.365	0.107	0.030
	3.0	0.5	0.500	0.021	0.000	1.333	0.722	0.186	0.409

Table 2. Estimation of the hazard ratio by Cox regression and by $\hat{D}_{FG}$

			Cox PH			D_{FG}	
	γ	mean	sd	mse	mean	sd	mse
$n_1 = n_2 = 100$	1.0	1.015	0.150	0.023	1.143	0.117	0.034
	1.2	1.207	0.177	0.031	1.230	0.174	0.031
	1.5	1.518	0.221	0.049	1.521	0.249	0.063
	2.0	2.013	0.311	0.097	2.028	0.357	0.128
	3.0	3.090	0.517	0.276	3.091	0.591	0.357
$n_1 = n_2 = 1000$	1.0	1.002	0.046	0.002	1.042	0.034	0.003
	1.2	1.202	0.055	0.003	1.201	0.064	0.004
	1.5	1.500	0.068	0.005	1.499	0.081	0.007
	2.0	1.997	0.094	0.009	1.998	0.107	0.011
	3.0	3.006	0.147	0.022	3.004	0.168	0.028

estimate of $K(f, g)$ can require sample sizes of order 10^5 or 10^6 to work well.) A similar table (not shown here) to Table 1 was constructed for standardised values, i.e. dividing by the true values of D_{FG} and KL_{FG} and still $\hat{D}_{FG}$ was superior to $\widehat{KL}_{FG}$.

3.5. Estimation of the hazard rate using $\hat{D}_{FG}$

Using (3) it is possible to estimate the hazard ratio from the divergence, $\hat{D}_{FG}$, for the case of proportional hazards. Although it is not particularly proposed that the measure of divergence is used to estimate the hazard ratio generally, it is interesting to compare its performance in doing so with that using Cox proportional hazards regression. Table 2 shows the results of simulations using the exponential distributions used for Table 1. The values of γ, mean estimates of γ together with standard deviations and mse are shown for the two methods. It can be seen that the method based on $\hat{D}_{FG}$ does not perform quite so well as that for Cox PH, but the results are very close especially when γ is not close to unity.

Table 3. Divergencies/Dissimilarities $(\times 100)$

	F	G	F	F	F	G	F	G	G	G	F	G	F	F	G
	L	l	l	l	l	L	L	L	l	L	L	L	L	L	L
	r	r	r	r	r	r	r	r	R	R	R	r	r	R	R
	W	M	W	M	P	W	M	M	M	M	M	P	P	P	P
FLrW	0	11	5	39	3	20	20	21	46	36	36	43	57	55	62
GlrM		0	4	13	23	29	26	28	21	39	40	43	57	54	62
FlrW			0	8	10	23	23	24	17	37	37	41	54	53	62
FlrM				0	23	19	17	16	11	30	30	33	48	44	54
FlrP					0	13	10	10	7	20	21	24	37	34	40
GLrW						0	2	2	32	18	18	24	41	41	48
FLrM							0	1	5	13	15	22	35	43	37
GLrM								0	4	13	14	20	35	33	43
GlRM									0	16	16	18	32	30	38
GLRM										0	3	10	25	26	36
FLRM											0	8	23	36	30
GLrP												0	17	16	22
FLrP													0	1	8
FLRP														0	9
GLRP															0

4. Divergence of survival distributions for pancreatic cancer patients

The following is a further example using the pancreatic survival data described in Section 3.2. From the prognostic factors *lymph node involvement*, *resection margin* and *tumour grade*, together with treatment, twenty four sub-groups can be defined where each will be labelled by four characters: G or F for treatment, L or l for lymph node involvement as positive or negative respectively, R or r for resection margin positive or negative and W, M, or P for tumour grade of well, moderate and poorly differentiated. For example GRlM is the subgroup of patients who received Gemcitabiine, had positive resection margin, had no lymph node involvement and had moderately differentiated tumours. The survival data was split into these groups of patients, but only those fifteen groups that had more than twenty patients were retained for analysis.

Figure 5 shows the Kaplan-Meier plots for all the groups illustrating the range of survival curves. The divergence was measured between each pair of survival distributions and placed in a dissimilarity matrix as shown in Table 3. The order of the subgroups has been chosen so that the dissimilarity matrix is in anti-Robinson form, i.e. the row/column order is chosen so that the smaller dissimilarities are closer to the diagonal than the larger dissimilarities and this gives a form of clustering of the subgroups, see [23].

Nonmetric multidimensional scaling (MDS) [24], was used on the divergences considered to be dissimilarities between groups. (Multidimensional scaling covers several multivariate analysis techniques where the aim is to represent objects as points in a low dimensional space, usually Euclidean, so that distances between pairs of points match as well as possible the original dissimilarities calculated between the pairs of objects. Various dissimilarity measures are used in practice, e.g. Euclidean distance or Minkowski metric. Common methods of scaling are classical scaling and non-metric scaling. For non-metric scaling, STRESS is a measure of the fit of the configuration of points representing the objects to the dissimilarities; a value of the STRESS of 20% suggests a poor fit, 10% fair, 5% good and 2% an excellent fit.)

Figure 6 shows the resulting MDS configuration for groups where labels also contain the median survival times. The STRESS for the plot was 9%. Within the configuration of points, a curved axis of median survival time can be imagined with shortest values for the groups in the top right-hand corner, then increasing as the axis sweeps downwards and towards the left and then sweeping down to the bottom right-hand corner where groups have the longest median survival times. The groups at the top right-hand corner have poorly differentiated tumours and lymph node involvement. The groups at the bottom right-hand corner have negative resection margin, have well or moderately differentiated tumours and no lymph node involvement. The groups in the middle portion of the imaginary axis have better prognosis values than those at the beginning of the axis but worse than those at the end. Let the following scores be given to the prognosis variables: l=0, L=1; r=0, R=1; W=0, M=1, P=2, giving a minimum group score of 0 and a maximum of 4. The group scores tend to start at 4 at the beginning of the imaginary axis, reduce to 3 and then to 2 when moving along the axis, reaching 0 or 1 at the end of the axis. The subgroup that is somewhat anomalous is GLrW and to a lesser extent, FLrW, at the lower left of the plot. GLrW has a group score of 1 but is placed alongside subgroups with group score 2. Although these patients have well differentiated tumours, their lymph node involvement reduces their survival compared to those patients with moderately differentiated tumours but with no lymph node involvement.

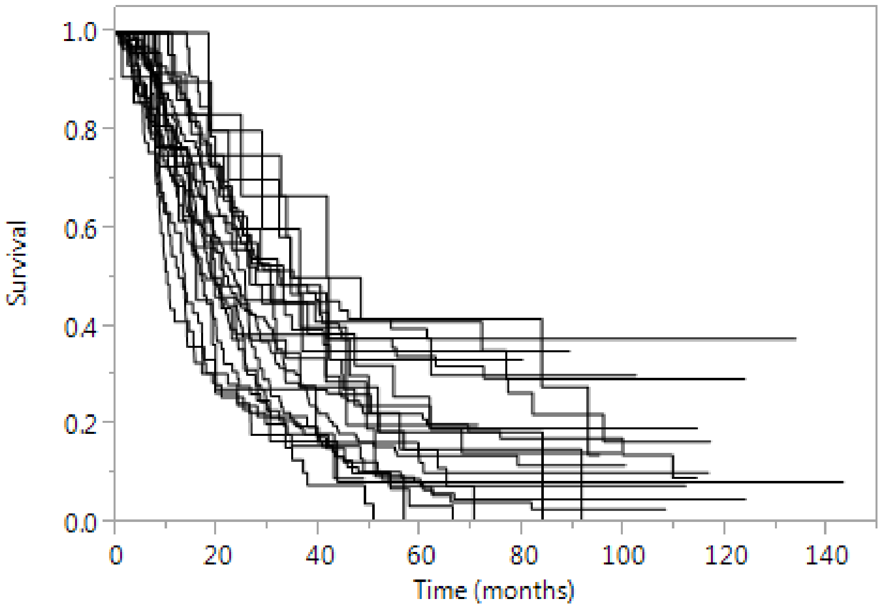

Figure 5. Survival curves for 15 subgroups of pancreatic cancer patients

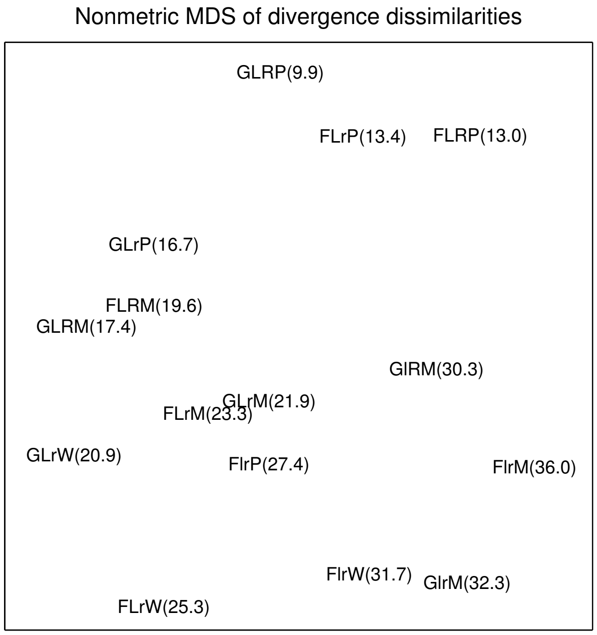

Figure 6. Nonmetric MDS of divergence dissimilarities for the 15 subgroups of pancreatic cancer patients

For the proportional hazards case, the absolute value of the log hazard ratio, $|\ln(\gamma)|$ can represent the distance between two survival distributions and is a metric. Points representing a group of distributions (all with proportional hazards) will lie on a straight line. Nonmetric MDS was carried on the the logged estimates of the hazard ratios calculated for all pairs of the pancreatic cancer subgroups. The resulting configuration of points representing the subgroups is not shown here because the STRESS was 17% suggesting a poor fit, giving evidence that proportional hazards do not apply for all of these

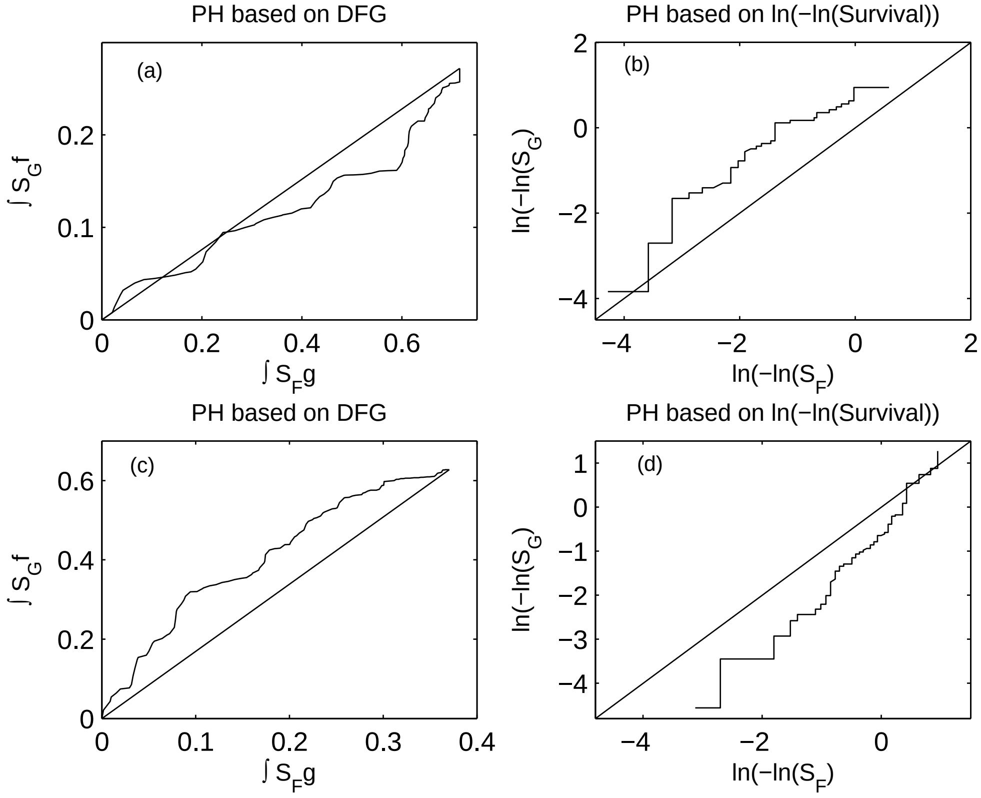

Figure 7. Assessment of the proportional hazards assumption for the groups FLrM vs FLRP (a and b) and GLRM vs FLRP (c and d) based on D_{FG} and on the log(-log(survival)) approach

fifteen survival distributions. If proportional hazards did apply, then the points in the configuration would lie close to a straight line.

5. Assessing the proportional hazards assumption

In Section 2.1 it was shown that for the proportional hazards situation, $\int_0^t S_F g$ and $\int_0^t S_G f$ are in the proportions $\gamma : 1$ where γ is the hazard ratio. So a plot of $\int_0^t \hat{S}_G \hat{f}$ against $\int_0^t \hat{S}_F \hat{g}$ should produce points more or less on a straight line of slope γ. Figure 7(a) shows this plot for two of the pancreatic cancer groups FLrM and FLRP together with a straight line joining the origin to the point defined by $(\int_0^\infty \hat{S}_F \hat{g}, \int_0^\infty \hat{S}_G \hat{f})$. Figure 7(b) shows a more traditional plot to assess proportional hazards, where $\log\{-\log(\hat{S}_G(t))\}$ is plotted against $\log\{-\log(\hat{S}_F(t))\}$ for each group [25]. Here proportional hazards would imply that the points would lie more or less on a straight line of slope unity where the vertical distance from the line of slope unity, that passed through the origin, is equal to $\ln(\gamma)$. A similar pair of plots are given for the two groups, GLRM and FLRP in Figures 7(c) and 7(d). Although interpretation of these graphs for assessing proportional hazards is subjective, the authors' view is that for the groups FLrM vs FLRP, the original method indicates proportional hazards but the new method suggests that this is not the case (Figures 7(a) and 7(b)), while for the groups GLRM vs FLRP, both methods indicate non-proportional hazards (Figures 7(c) and 7(d)).

6. Discussion

This paper has introduced a new divergence measure for survival distributions which can be calculated for theoretical distributions or non-parametrically from survival data. Although the emphasis was a measure for survival distributions, it can be easily adapted for more general distributions. The measure is simple in concept and simple to calculate. For proportional hazards, the divergence measure has been shown to be a metric. Estimation of the divergence from data was compared to the estimation of the Kullback-Leibler divergence and was shown to be superior. Estimation of the hazard ratio using the divergence was nearly as precise and accurate as using Cox proportional hazards regression. A method for estimating the standard error of the divergence measure was given together with arguments for the asymptotic distribution. The measure has been illustrated on survival data for pancreatic cancer patients where differences between survival distributions for two groups and also for multiple groups were based on the divergence measure. For the multiple group case, non-metric multidimensional scaling was used to summarise, graphically, the pairwise divergencies between the groups.

Clearly, one statistic cannot fully describe a survival distribution nor the comparison of two. The common description when comparing survival distributions is to quote the median survival times and the estimate of the hazard ratio, together with confidence intervals plus the results of a log-rank test. However, as illustrated in Figure 1, median survival times can be misleading and a hazard ratio might not be appropriate. Would it be good practice to also quote the estimated divergence, $\hat{D}_{FG}$, between the two survival distributions and its comparison with the estimated expected value under proportional hazards, as given by (3)? For the GLRM and FLRP cancer subgroups in Figures 5(c) and 5(d), the usuallly quoted statistics are: medians (95% CI) 13.0 (9.3-15.3) and (17.4 (14.8-24.4) respectively, log-rank statistic = 1.30 (p=0.254) and $\hat{\gamma} = 0.84$. The estimated divergence is 0.43 which does not match with the value of 0.09 if proportional hazards pertained.

In the context of a clinical trial, suppose an independent data monitoring committee is to analyse survival data from two or more treatment arms every six months for several years. Along with pre-specified tests of hypotheses, the divergence between the survival distributions up to the present time could be calculated, for instance, to monitor the assumption of proportional hazards. As more data becomes available, the survival curves get extended in time by those patients enrolled early in the trial and not experiencing the event of interest, as well as becoming more accurate as more patients are entered into the database. If there is a true difference in the survival distributions for patients in two different arms of the trial, then the calculated divergence will probably be small at 6-months, but then increase as more and more divergence calculations are made as the months pass by, but eventually stabilising at a constant value. The values and pattern of the dynamic divergence values would add extra insight into the relative efficacies within the two arms of the trial. The divergence might be especially useful in a multi-arm trial, or if several subgroups are to be compared.

7. Appendix

7.1. Variance of D_{FG}

First, assume the hazard functions do not cross and so $\hat{D}_{FG} = \sum_{i=0}^{N-1} F_{i+1}G_i - F_iG_{i+1}$. Some of the F_i's and G_i's will be true Kaplan-Meier jump-point values and others will be interpolated from these. Place the F_i's and G_i's in vectors, $\mathbf{F}$ and $\mathbf{G}$. Let the true jump-point values be $\mathcal{F}_i, (i = 1, \dots, n_F)$, $\mathcal{G}_i, (i = 1, \dots, n_G)$ and place these in vectors $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{G}}$. The interpolation finds weight matrices, $\mathbf{W_F}$ and $\mathbf{W_G}$ that connect $\mathbf{F}$ to $\boldsymbol{\mathcal{F}}$ and $\mathbf{G}$ to $\boldsymbol{\mathcal{G}}$ thus

$$\mathbf{F} = \mathbf{W_F}\boldsymbol{\mathcal{F}} \qquad \mathbf{W_G} = \mathbf{W_F}\boldsymbol{\mathcal{G}},$$

where most of the elements in the weight matrices will be zero.

Now $\hat{D}_{FG} = \mathbf{F}^T\mathbf{D}\mathbf{G}$, where $\mathbf{D} = [d_{ij}]$, $d_{i,i+1} = -1$, $d_{i,i-1} = 1$, $d_{i,j} = 0$ otherwise. Hence

$$\hat{D}_{FG} = (\mathbf{W_F}\boldsymbol{\mathcal{F}})^T\mathbf{D}(\mathbf{W_G}\boldsymbol{\mathcal{G}}) = \boldsymbol{\mathcal{F}}^T(\mathbf{W_F}^T\mathbf{D}\mathbf{W_G})\boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{F}}^T\mathbf{A}\boldsymbol{\mathcal{G}},$$

where $\mathbf{A} = \mathbf{W_F}^T\mathbf{D}\mathbf{W_G}$.

So far, $\mathbf{F}$, $\mathbf{G}$, $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{G}}$ have been considered as observations, but now place hats on them and consider them as random variables. The expected value of $\hat{D}_{FG}$ is given by $\mathrm{E}(\hat{\boldsymbol{\mathcal{F}}}^T)\mathbf{A}\mathrm{E}(\hat{\boldsymbol{\mathcal{G}}})$ and to calculate the variance,

$$\mathrm{E}(\hat{D}_{FG}^2) = \mathrm{E}_{\mathrm{FG}}(\hat{\boldsymbol{\mathcal{F}}}^T\mathbf{A}\hat{\boldsymbol{\mathcal{G}}}\hat{\boldsymbol{\mathcal{F}}}^T\mathbf{A}\hat{\boldsymbol{\mathcal{G}}}) = \mathrm{E}_{\mathrm{FG}}(\hat{\boldsymbol{\mathcal{F}}}^T\mathbf{A}\hat{\boldsymbol{\mathcal{G}}}\hat{\boldsymbol{\mathcal{G}}}^T\mathbf{A}\hat{\boldsymbol{\mathcal{F}}}) = \mathrm{E}_{FG}\{\mathrm{tr}(\hat{\boldsymbol{\mathcal{F}}}^T\mathbf{A}\hat{\boldsymbol{\mathcal{G}}}\hat{\boldsymbol{\mathcal{G}}}^T\mathbf{A}\hat{\boldsymbol{\mathcal{F}}})\},$$

where tr is the trace of a matrix. Hence, letting $\mathbf{\Phi_F} = \mathrm{E}(\hat{\boldsymbol{\mathcal{F}}}\hat{\boldsymbol{\mathcal{F}}}^T)$ and $\mathbf{\Phi_G} = \mathrm{E}(\hat{\boldsymbol{\mathcal{G}}}\hat{\boldsymbol{\mathcal{G}}}^T)$,

$$\mathrm{E}(\hat{D}_{FG}^2) = \mathrm{E}_F\{\mathrm{tr}(\hat{\boldsymbol{\mathcal{F}}}^T\mathbf{A}\mathbf{\Phi_G}\mathbf{A}\hat{\boldsymbol{\mathcal{F}}}^T)\} = \mathrm{E}_F\{\mathrm{tr}(\mathbf{A}^T\hat{\boldsymbol{\mathcal{F}}}\hat{\boldsymbol{\mathcal{F}}}^T\mathbf{A}\mathbf{\Phi_G})\} = \mathrm{tr}(\mathbf{A}^T\mathbf{\Phi_F}\mathbf{A}\mathbf{\Phi_G})\}.$$

Then $\mathrm{var}(\hat{D}_{FG})$ can be written as

$$\mathrm{var}(\hat{D}_{FG}) = \mathrm{tr}(\mathbf{A}^T\mathbf{\Phi_F}\mathbf{A}\mathbf{\Phi_G}) - \mathrm{tr}(\mathbf{A}^T\mathrm{E}(\hat{\boldsymbol{\mathcal{F}}})\mathrm{E}(\hat{\boldsymbol{\mathcal{F}}})^T\mathbf{A}\mathrm{E}(\hat{\boldsymbol{\mathcal{G}}})\mathrm{E}(\hat{\boldsymbol{\mathcal{G}}})^T).$$

Based on Greenwood's formula, e.g. [1],

$$\mathrm{E}(\hat{\mathcal{F}}_i) \approx \mathcal{F}_i, \quad \mathrm{cov}(\widehat{\mathcal{F}_i\mathcal{F}_k}) \approx \mathcal{F}_i\mathcal{F}_k\sum_{j=1}^{i}\frac{d_j}{n_j(n_j-d_j)} \quad (i \le k),$$

where d_j is the number of deaths in the interval $[t_{j-1}, t_j)$ and n_j is the number at risk in the interval. Hence $\mathbf{\Phi_F}$ and $\mathbf{\Phi_G}$ can be found leading to an estimate of $\mathrm{var}(D_{FG})$.

Note, if the hazard functions cross, then when the calculation is split into two parts, $\int \hat{S}_F\hat{g} - \int \hat{S}_G\hat{f}$ and $\int \hat{S}_G\hat{f} - \int \hat{S}_F\hat{g}$, the covariances where $\mathcal{F}_i$ is in one section and $\mathcal{F}_k$ is the another have to be negated.

7.2. *Asymptotic distribution of $\hat{D}_{FG}$*

This section illustrates how counting process theory using martingales can lead to the asymptotic distribution of D_{FG}. The reader is referred to texts such as [26, 27] for further details. The following relies heavily on [26] which is a good introduction to the area. The following notions from counting processes are needed. Let $\{M(s) : s \geq 0\}$ be a stochastic process and $\{\mathcal{H}_t\}$ be the filtration (an increasing sequence of σ-fields) upon which the stochastic process is defined. The filtration represents "the past history" of the process. The stochastic process $M(t)$ which is adapted to $\{\mathcal{H}_t\}$, i.e. all σ-fields for $M(s)$ $(0 \leq s \leq t)$ are contained in $\{\mathcal{H}_t\}$, is a martingale if $\mathrm{E}|M(t)| < \infty$ and $\mathrm{E}\{(M(t)|\mathcal{H}_s\} = X(s)$. In other words, the expected value of a future value is equal to the present value.

Let $\alpha(t)$ be the hazard rate and $A(t) = \int_0^t \alpha(s)ds$ the cumulative hazard rate. A counting process, $N(t)$, counts the number of events in the time interval $[0, t]$. The counting process can be considered as a succession of increments, $dN(t) = N\{(t + dt)^-\} - N(t^-)$ which has the value 1 if a point event happens at time t, and 0 otherwise. Let λ be the rate function of the counting process. Let $Y(t)$ be the number at risk at time t and assume the usual multiplicative intensity model, $\lambda(t) = \alpha(t)Y(t)$. Then $M(t) = N(t) - A(t)$ is a mean zero martingale counting process. The Nelson-Aalen estimator of $A(t)$ is $\hat{A}(t) = \int_0^t \{1/Y(s)\}^{-1}dN(s)$ and it can be shown that $\hat{A}(t) - A(t)$ is also a zero mean martingale.

A martingale representation for the Kaplan-Meier estimate of the survival function is

$$\widehat{S(t)} = S(t) - S(t)\int_0^t \frac{\hat{S}(s^-)}{S(s)Y(s)}dM(s).$$

Using this framework, asymptotically (sample size tends to infinity), the Kaplan-Meier functions will converge to $\hat{S}_F$ and $\hat{S}_G$ and $\hat{S}(s^-)/S(s) \to 1$ for both. Then

$$\begin{aligned}\int_0^t \hat{S}_F(s)d\hat{S}_G(s) \quad &\approx \quad \int_0^t [S_F - S_F(\hat{A}_F - A_F)][dS_g - d\{S_G(\hat{A}_G - A_G)\}] \\ &= \quad \int_0^t S_F dS_G - \int_0^t S_F(\hat{A}_F - A_F)dS_G - \int_0^t S_F(\hat{A}_G - A_G)dS_G - \int_0^t S_F S_G d(\hat{A}_G - A_G) \\ &+ \quad \int_0^t S_F(\hat{A}_F - A_F)(\hat{A}_G - A_G)dS_G + \int_0^t S_F S_G(\hat{A}_F - A_F)d(\hat{A}_G - A_G).\end{aligned}$$

The first term on the right in last equation is a compensator of the process (the mean of the process) and it can be shown that all the other terms are zero mean martingales, the sum of which is a martingale. Hence $\int_0^t \hat{S}_F d\hat{S}_G - \int_0^t \hat{S}_F d\hat{S}_G$ is asymptotically a zero mean martingale.

Now for a martingale $M(t)$, let the time interval $[0, t]$ be divided into n equal sub-intervals and let $dM_k = M\{k/n\} - M\{(k-1)/n\}$. Then the predictable variation process $< M >$ is

$$< M >_t = \lim_{n\to\infty}\sum_{k=1}^{n} \mathrm{var}(dM_k|\mathcal{H}_{(k-1)/n}).$$

If (i) the sizes of jumps go to zero and (ii) the predictable variation converges to a deterministic function, then M tends to a normal distribution. This is the martingale central limit theorem.

The predictable variation for the Aalen estimator is $\int_0^t \alpha(s)ds/Y(s)$. To intuitively show that $\hat{A}(t)$ tends to a normal distribution, consider $\sqrt{n}\hat{A}(t)$. The jump sizes are $\sqrt{n}/Y(s)$. For the sample size n tending to infinity, assume $Y(t)/n$ tends to a predictable process. Then the jump sizes for $\sqrt{n}\hat{A}(t)$ are of order $1/\sqrt{n}$ and the predictable variation is $\int_0^t \{Y(s)/n\}^{-1}\alpha(s)ds$ which tends to a deterministic function. Thus conditions (i) and (ii) for the martingale central limit theorem are satisfied.

The martingales within the equation for $\int_0^t \hat{S}_F(s)d\hat{S}_G(s)$ above depend upon $(\hat{A}_F - A_F)$ and $(\hat{A}_G - A_G)$. A similar intuitive argument suggests that jump sizes for $\int_0^t S_F\hat{A}_F dS_G$ in the first martingale are of order $1/\sqrt{n}$, assuming the jump dS_G is order $1/\sqrt{n}$ and as $\hat{A}_F$ tends to a deterministic process, so does the whole integral. These arguments apply to all the martingales in the equation and hence this leads to asymptotic normality. In turn, this implies that the distribution of $\hat{D}_{FG}$ tends to normality when the hazard functions do not cross. It is a more complicated situation when the hazard functions do cross. It the true crossing point were known, then $\hat{D}_{FG}$ would tend towards normality. Estimating the crossing point introduces another random variable upon which $\hat{D}_{FG}$ has to be conditioned and then the unconditioned distribution of $\hat{D}_{FG}$ found.

Two small simulation exercises were carried out to check the normality assumptions. Firstly, one thousand simulated values of $\hat{D}_{FG}$ were generated using pairs of exponential distributions and the normality assumptions checked using QQ-plots and the Jarque-Bera and Lilliefors tests of normality. For sample sizes greater than fifty, normality was confirmed for $\hat{D}_{FG}$ not too close to the extremities, 0 and 1. Also confirmed was that $\mathrm{var}(\hat{D}_{FG})$ is asymptotically of order n^{-1}. Secondly, data were generated for crossing Weibull survival distributions, with the crossing point estimated at the point where the global maximum of $\widehat{\int S_F g}$ (or $\widehat{\int S_G f}$ as appropriate) was seen. For various sets of parameters of the distributions, including those used in Figure 1, normality of $\hat{D}_{FG}$ was tested. Generally, normality was very good for sample sizes greater than fifty. Note, normality fails when the two underlying survival distributions are equal or very close.

Acknowledgements

The authors would like to thank Prof. JP Neoptolemos for permission to use the pancreatic cancer data, to Prof. PJ Diggle for helpful comments and two very helpful referees and an associate editor who have been responsible for us producing a much better version of the paper than that originally submitted for publication.

References

1. Collett D. *Modelling Survival Data in Medical Research 3rd Edition*, Chapman and Hall/CRC: Boca Raton, 2015.
2. Suciu GP, Lewmeshow S, Moeschberger M. Statistical tests of the equality of survival curves: reconsidering the options. In *Handbook of Statistics 23*, Balakrishnan N, Rao CR (eds). Elsevier B.V: Amsterdam, 2004; 251–262.
3. Brittain E, Follmann D, Yang S. Dynamic comparison of Kaplan-Meier proportions: monitoring a randomized clinical trial with a long-term binary endpoint. *Biometrics* 2008; **64**:189–197. DOI: 10.1111/j.1541-0420.2007.00874.x
4. Gu MG, Follman D, Geller NL. Monitoring a general class of two-sample survival statistics with applications. *Biometrika* 1999; **86**:45–57. DOI: 10.1093/biomet/86.1.45
5. Kullback S, Leibler RA. On information and sufficiency. *Annals of Mathematical Statistics* 1951; **22**:79-86.
6. Wang Q, Kulkarni S, Verdu S. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory* 2005; **51**:3064–3074. DOI: 10.1109/TIT.2005.853314
7. Perez-Cruz F. Kullback-Leibler divergence estimation of continuous distributions. *Proceedings of IEEE International Symposium on Information Theory* 2008; 1666-1670.
8. Ebrahimi N, Kirmani SNUA. A characterisation of the proportional hazards model through a measure of discrimination between two residual life distributions. *Biometrika* 1996; **83**:233–235. DOI: 10.1093/biomet/83.1.233
9. Di Crescenzo A, Longobardi M. A measure of discrimination between past lifetime distributions. *Statistics & Probability Letters* 2004; **67**:173–182. DOI: 10.1016/j.slp.2003.11.019
10. Cox TF. Testing the equivalence of survival distributions using PP- and PPP-plots. *International Journal of Statistics in Medical Research* 2014; **3**:161–173. http://dx.doi.org/10.6000/1929-6029.2014.03.02.10
11. Acion L, Peterson JJ, Temple S and Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine* 2006; **25**:591–602. DOI: 10.1002/sim.2902
12. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 1947; **18**:50–60.
13. Newcombe RG. 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* 2006; **25**:543–557. DOI: 10.1002/sim.2323
14. Hanley JA, McNeil BJ. The meaning and use of the area under an ROC curve. *Radiology* 1982; **143**:29-36.
15. Brumback LC, Pepe MS, Alonzo TA. Using the ROC curve for guaging treatment effect in clinical trials. *Statistics in Medicine* 2006; **25**:575–590. DOI: 10.1002/sim.2345
16. Harrell Jr FE. *Regression Modeling Strategies – with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer: New York, 2001.
17. Koziol JA, Jia Z. The concordance index C and the Mann-Whitney parameter Pr(X¿Y) with randomly censored data. *Biometrical Journal* 2009; **51**:467–474. DOI: 10.1002/bimj.200800228
18. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine* 2009; **28**:2473-2489. DOI: 10.1002/sim.3623
19. Kalbfleisch JD, Prentice RL. Estimation of the average hazard ratio. *Biometrika* 1981; **68**:105-112. DOI: 10.1093/biomet/68.1.105
20. Neoptolemos JP, Stocken DD, Bassi C *et al.* Adjuvant chemotherapy with Fluorouracil plus Folinic Acid vs Gemcitabine following pancreatic cancer resection. *Journal American Medical Association* 2010; **304**(10):1073–1081. DOI: 10.1001/jama.2010.1275
21. Mantel N, Stablein DM. The crossing hazard function problem. *Statistician* 1988; **37**:59–64. DOI: 10.2307/2348379

22. Cheng MY, Qiu P, Tan X, Tu D. Confidence intervals for the first crossing point in two hazard functions. *Lifetime Data Analysis* 2014; **15**:441–454. DOI: 10.1007/s10985-009-9132-6
23. Hahsler M, Hornik K and Buchta C. Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software* 2008; **25**:3952.
24. Cox TF, Cox MAA. *Multidimensional Scaling, 2nd edn*. Chapman and Hall/CRC: Boca Raton, 2001.
25. Cantor AB. *Survival Analysis Techniques, 2nd edn*. Cary, NC: SAS Institute Inc., 2003.
26. Aalen OO, Andersen PK, Borgan Ø, Gill RD, Keiding N. History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics* 2009; **5**(1):www.jehps.net.
27. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: Hoboken, 1991.