Testing the Equivalence of Survival Distributions using PP- and PPP-Plots

Trevor F. Cox*

Cancer Research UK Liverpool Cancer Trials Unit, University of Liverpool, Liverpool, UK

Abstract: This paper discusses the use of PP-plots for survival distributions where for a pair of survival distributions, one is plotted against the other. This is another way of visualizing the nature of the relationship between the two survival distributions along with typical Kaplan-Meier plots. For three survival distributions, the PPP-plot is introduced where the survival distributions are plotted against each other in three-dimensions. At the population level, measures of divergence between distributions are introduced based on areas and lengths associated with the PP- and PPP- plots. At the sample level, two test statistics are defined, based on these areas and lengths, to test the null hypothesis of equivalent survival curves. A simulation exercise showed that, overall, the new tests are worthy competitors to the log-rank and Wilcoxon tests and also to a Levine-type test and a Kolmogorov-Smirnov type test for the case of crossing survival curves. The paper also shows how the PP-plot can be used to estimate the hazard ratio and to assess the ratio of hazard functions if proportional hazards are not appropriate. Finally, the methods introduced are illustrated on two cancer data sets.

Keywords: Crossing survival curves, Hazard ratio, Kaplan-Meier, Log-rank test, PP-plot, Wilcoxon test.

INTRODUCTION

The typical way of drawing two or more survival curves for comparison purposes, is to plot the survival functions against time, as illustrated in Figure **1i** where two survival curves from Weibull distributions $(\text{pdf}: \lambda\beta(\lambda t)^{\beta-1}\exp\{(-\lambda t)^{\beta}\})$ with parameters $\lambda = 0.3, \beta = 1.0$ (i.e. an exponential distribution) and $\lambda = 0.2, \beta = 1.5$ are plotted. In addition to the survival curves, a QQ-plot can also be constructed where corresponding quantiles of the distributions are plotted against each other. This plot is shown in Figure **1ii** for the two curves of Figure **1i**. However, the additional plot proposed here is the PP-plot where the corresponding survival probabilities at time *t* for the two curves are plotted against each other as shown in Figure **1iii**. For identical survival curves, the plot would simply be the diagonal from (0, 0) to (1, 1), or following the course of time, from (1, 1) to (0, 0). When corresponding times are added along the curve, this PP-plot contains the same information as in the standard survival plot of Figure **1i** although the shapes of the individual curves are harder to assimilate. For instance, the median survival times for the two curves can be read from the curve in Figure **1i** and from the curve in Figure **1iii** (if plotted on a larger scale) as 2.3 and 3.9. The PP-plot has a long history; two early papers describing their use are Wilk and Gnanadesikan [1] and Michael [2], but note that usual PP-plots, plot cumulative distribution functions (CDF) against each other, whereas here, one minus the CDFs are used. The only consequence is that the graph starts at (1, 1) and moves towards (0, 0), instead of from (0, 0) to (1, 1). The PP-plot extends to three or more curves. Figure **1iv** shows a PPP-plot in three dimensions for the two Weibull distributions of Figure **1i** and another with parameters $\lambda = 0.3, \beta = 0.5$.

The PP-plot lends itself to assessing differences in the survival probabilities and distributions, both at the population and sample levels. At the population level, the PP-plot can be used to define a new divergence measure for distributions. At the sample level, the PP-plot gives rise to test statistics for the hypothesis of identical survival curves. The first statistic to be considered will be the estimated absolute area between the PP-curve and the diagonal of identical curves, the minimum theoretically being zero for identical survival curves, and with a maximum of 0.5. The second statistic is the estimated length of the PP-plot, with a theoretical minimum of $\sqrt{2}$ and maximum of 2.

Clearly, the PP-area is linked to the area under a Receiver Operator Characteristic curve (ROC curve), widely used in analyses of medical and other data [3]. There exists a research monograph on the subject [4]. The concept of a ROC curve has been extended to a ROC surface for the case of three populations [5] but this will not be analogous to the PPP-curve discussed here. The PP-area will only be equivalent to the much used area under the ROC curve (subtracting 0.5) when the PP-curve does not cross the diagonal line.

The outline of the rest of the paper is as follows: the divergence measures are explored in the population PP- and PPP-plots section which is followed by a section on the use of sample PP- and PPP-plots where it is shown how lengths of curves and areas are calculated. Testing the equivalence of survival curves

*Address correspondence to this author at the Cancer Research UK Liverpool Cancer Trials Unit, University of Liverpool, Block C Waterhouse Building, 1-3 Brownlow Street, Liverpool, L69 3GL, UK; Tel: +44 (0)151 7948891;
E-mail: coxt@liv.ac.uk

E-ISSN: 1929-6029/14

is covered in the section on hypothesis testing which is followed by a section on estimation of the hazard ratio. There is a section describing the use of the PP- and PPP-plots on two cancer examples. Lastly, there is a discussion section. To calculate the length of a PP- or PPP- curve, or the area of the surface defined by the curve and the diagonal line requires some basic differential geometry; an outline of what is required for this paper is given in the Appendix. Note, the terms PP-plot and PP-curve will be used interchangeably.

POPULATION PP- AND PPP-PLOTS

A standard method for measuring the "difference" between two distributions is to use the Kullback-Leibler (KL) divergence [6,7]. The KL divergence of the distribution F, that has probability density function $f(t)$, from the distribution G, that has probability density function $g(t)$, is given by

$$D_{KL}(F\|G) = \int f(t)\left\{\frac{f(t)}{g(t)}\right\}dt.$$

Now $D_{KL}(G\|F)$ is different from $D_{KL}(F\|G)$ and so the symmetric measure of divergence, $D_{KL}(F\|G) + D_{KL}(G\|F)$ is often used.

Here it is suggested that the absolute area defined by the PP curve and the diagonal can be used as a measure of divergence. This area will be termed the PP-area. This could be useful when fitting parametric distributions to survival data for various groups and wishing to have a measure of the differences between the fitted distributions.

Example

Consider a PP-curve based on two exponential distributions, $[\exp(-\lambda_1 t), \exp(-\lambda_2 t)]$. The symmetric KL divergence is

$$\frac{(\lambda_1-\lambda_2)^2}{\lambda_1\lambda_2} = \frac{(1-\gamma)^2}{\gamma},$$

where $\gamma = \lambda_2/\lambda_1$..

The divergence as given by the PP-area is easily calculated to be

$$\frac{|\lambda_1-\lambda_2|}{2(\lambda_1+\lambda_2)} = \frac{|1-\gamma|}{2(1+\gamma)}.$$

The length of the PP-curve (termed the PP-length) and subtracting $\sqrt{2}$ could be an alternative measure of divergence. For the two exponential distributions, the length of the curve is given by (see Appendix)

$$\int_0^\infty \{\lambda_1^2\exp(-2\lambda_1 t) + \lambda_2^2\exp(-2\lambda_2 t)^{\frac{1}{2}}\,dt,$$

or if the curve is reparameterised as (u,u^γ), the length is

$$\int_0^1 \{1+\gamma^2 u^{2(\gamma-1)}\}^{\frac{1}{2}}\,du. \quad (1)$$

Now, this particular PP-curve is of prime interest because it is the one obtained for any proportion hazards situation, since, in this case, for survival curves $S_1(t), S_2(t)$ and hazard ratio, γ, $S_2(t) = S_1(t)^\gamma$. The PP-curve is given by $[S_1(t), S_1(t)^\gamma]$, or reparameterized as (u,u^γ). Reversing the roles of $S_1(t)$ and $S_2(t)$, gives the parameterization $(u,u^{\frac{1}{\gamma}})$. The integral (1) can be found. For $\gamma = 1$ the length is clearly $\sqrt{2}$. When $\gamma = 2$, the indefinite integral (excluding the constant of integration) of the integrand is

$$\{2\sqrt{1+4u^2} + \sinh^{-1}(2u)\}/4$$

and hence the length of the curve is 1.4789. For general γ, the indefinite integral of the integrand is

$$\frac{\sqrt{\gamma^2u^{2\gamma-2}-1}}{\gamma} - \frac{(\gamma-1)u^{2-\gamma}}{\gamma^2(\gamma-2)}F_{21}(a,b,c,z) \quad (2)$$

where $F_{21}(a,b,c,z)$ is the hypergeometric function with $a = 1/2, b = (\gamma-2)/\{2(\gamma-1)\}, c = (3\gamma-4)/\{2(\gamma-1)\}$ and $z = -u^{2(1-\gamma)}/\gamma^2$.

For two arbitrary survival distributions, the PP-curve can be parameterized as $[u, S_2(S_1^{-1}(u))]$, or alternatively by $[u, S_1(S_2^{-1}(u))]$. Generally, the length of the curve will be impossible to find explicitly and so numerical integration will have to be used. This was done for the PP-curve in Figure **1iii**, the length being 1.470.

The difference or divergence between three (or more) distributions based on length is easily generalised from that for two distributions. The length of the curve in Figure **1iv** is 1.903.

In three dimensions, the area between the PPP-curve and the diagonal has to be defined. Imagine a

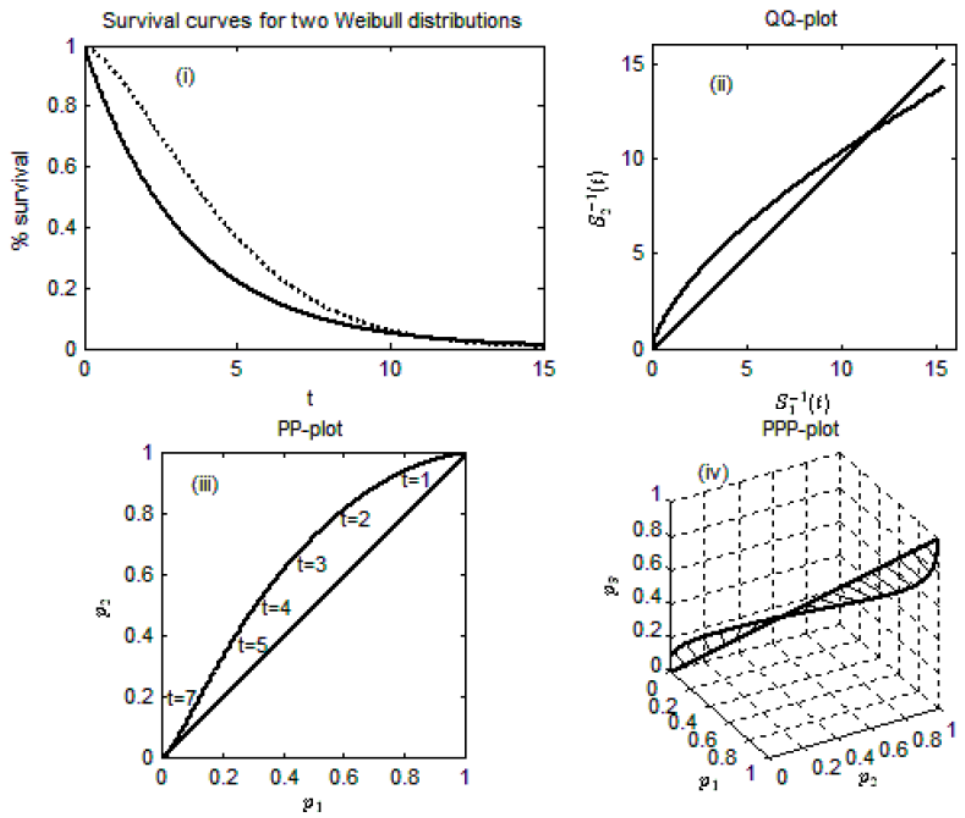

Figure1: Plots of (**i**) survival curves, (**ii**) QQ-plot, (**iii**) PP-plot and (**iv**) PPP-plot based on Weibull distributions

closed wire frame made to follow the PPP-curve and the diagonal. A soap film held by the frame would have minimal surface area [8] and this could be used to define the PPP-area. However, this minimal surface area is difficult to calculate and so a pragmatic alternative is to use the ruled surface (see Appendix) where a straight line is drawn from a point on the diagonal, perpendicular to the diagonal, and meeting the PPP-curve. As this line moves along the diagonal, always perpendicular, and meeting the PPP-curve, it sweeps out the "ruled surface". Using the formula (Appendix) for calculating the area of a ruled surface, the area between the PPP-curve and the diagonal in Figure **1iv** is 0.0729.

Finally, it would be possible to calculate various properties of these PP- and PPP-curves, e.g. curvature for PP-curves and curvature and torsion for the PPP-curves using the Frenet formulae [8, 9] but this is not entered into here.

SAMPLE PP- AND PPP-PLOTS

In practice, survival curves are estimated using the Kaplan-Meier method (KM), the Nelson-Aalen or some other method (see for example [10]). Figure **4i** shows the KM curves for two groups of cancer patients, details of which will be given later. Figure **4ii** shows the corresponding PP-plot using the KM curves. Note, theoretically, the sample PP-plot will be a sequence of unconnected points and so an area or length will not exist, even in the limit of an infinite sample size. But, to be practical, the sample PP-plot is taken as the PP-plot that has the disjoint sequence of points connected together.

Calculation of the Area between PP- and PPP-Curves and the Diagonal

The absolute area between the PP-plot and the diagonal can be easily found. One method is to view the PP-plot as a series of trapeziums with bases parallel to the x-axis, sides parallel to the y-axis and the fourth side formed by part of the diagonal. The width of the bases of the trapesiums are equal to the "jumps" parallel to the x-axis and the length of the sides are related to the "jumps" in the curve parallel to the y-axis (see Figure **4ii**). Care has to be taken during programming of the procedure when the PP-plot crosses the diagonal. However, the preferred method here for finding the area is to orthogonally transform the PP-plot so that the diagonal becomes the x-axis and the y-axis is the difference in survival distributions. The resulting "saw-tooth" function is split into trapeziums in order to find the total area. Making this

transformation allows the procedure to be easily extended to three or more survival curves.

Let $(p_{1i}, p_{2i}), i = 1,\ldots,N$ be the points in the original PP-plot. The coordinates of these points in the transformed plot are $[\frac{p_{1i}+p_{2i}}{\sqrt{2}}, \frac{p_{2i}-p_{1i}}{\sqrt{2}}] = (x_i, y_i)$ say. The area of the trapezium with base from (x_i, y_i) to (x_{i+1}, y_{i+1}), if the x-axis is not crossed, is $|(y_{i+1}+y_i)(x_{i+1}-x_i)|/2$, which can then be expressed in terms of the p_{ij}'s. If the x-axis is crossed then the area has to be calculated taking this into account. Note, the minimum of the calculated area is of the order $1/\max(n_1, n_2)$ where n_1 and n_2 are the number of steps in the x-axis and y-axis directions of the original PP-plot.

The sample PPP-plot, like the sample PP-plot, consists of a sequence of points, and again these will be joined. The area between the curve and the diagonal from (0,0,0) to (1,1,1) can be calculated in a similar manner to that for the PP-plot, although not in such a straightforward manner. First, the coordinates of the plot are transformed to a new set of orthogonal coordinates, (x_i, y_i, z_i) given by

$$[(p_{1i}+p_{2i}+p_{3i})/\sqrt{3}, (p_{1i}-p_{2i})/\sqrt{2}, (p_{1i}+p_{2i}-2p_{3i})/\sqrt{6}].$$

The area now comprises a series of "twisted" trapeziums, each with its base on the x-axis and sides orthogonal to the x-axis. Consider the *i*th twisted trapezium, and let its base be the line from $(x_{i-1}, 0, 0)$ to $(x_i, 0, 0)$. The sides are the lines from $(x_{i-1}, 0, 0)$ to $(x_{i-1}, y_{i-1}, z_{i-1})$ and from $(x_i, 0, 0)$ to (x_i, y_i, z_i) and the fourth side is the line from $(x_{i-1}, y_{i-1}, z_{i-1})$ to (x_i, y_i, z_i). The area of this twisted trapezium will be defined as the area of the ruled surface obtained as the first side of the twisted trapezium is moved along the x-axis, following the top edge and ending at the second side.

The ruled surface is given by

$$D(u,v) = [(x_{i-1}+u(x_i-x_{i-1}), 0, 0] + [0, v\{y_{i-1}+u(y_i-y_{i-1})\}, v\{z_{i-1}+u(z_i-z_{i-1})\}],$$

from which the partial derivatives with respect to u and v, D_u and D_v, can be found and then the surface area found numerically using formula (1) in the Appendix. The area required is the sum of the individual areas of the twisted trapeziums, noting that in the special case where the top edge of the trapezium crosses the diagonal, the area is then formed as the area of two triangles.

Calculation of the Length of PP- and PPP-Curves

It would be pointless calculating the length of the sample PP-curve as the sum of the actual lengths of the horizontal and vertical steps within the plot as this would nearly always give the value 2. Instead, the PP-curve is smoothed by fitting an appropriate function. There are several possible choices of function, including splines, but here the following polynomial is used and is fitted to the transformed points of the curve, (x_i, y_i):

$$y = x(\sqrt{2}-x)(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3).$$

Note, the polynomials are forced to pass through (0, 0) and $(\sqrt{2}, 0)$. The function can easily be fitted as a multiple linear regression.

Similarly, for PPP-curves, multivariate multiple regression can be used to fit the polynomial

$$(y,z) = [x(\sqrt{3}-x)(\beta_{10}+\beta_{11}x+\beta_{12}x^2+\beta_{13}x^3), x(\sqrt{3}-x)(\beta_{20}+\beta_{21}x+\beta_{22}x^2+\beta_{23}x^3)].$$

Higher order polynomials could of course be used.

HYPOTHESIS TESTING FOR SURVIVAL CURVES

Testing the equality of two or more survival curves has a long history, with the log-rank test (Mantel-Cox test, Peto-Mantel-Haenszel test) probably being the most widely used test for this purpose [11, 12]. The construction of this test and its relationship to the Cox proportional hazards model is described in many texts. Briefly, the test for the two group case can be constructed from a series of 2×2 tables, one for each of the distinct event times within the data, the entries in the table being the number of events for each group at the specific event time and the number surviving beyond the specific event time. The statistic is

$$Q = \frac{\left\{\sum_{j=1}^{r} w_j (d_{1j} - e_{1j})\right\}^2}{\sum_{j=1}^{r} w_j^2 v_{1j}} \quad (4)$$

where d_{ij} is the number of events in group *i* at the *j*th event time, e_{ij} is the expected number of events within group *i* at the *j*th event time, v_{ij} is the variance of the number of events within group *i* at the *j*th event time and $w_j = 1$. Let n_{ij} be the number of patients at risk in

the *i*th group just before the *j*th event time. Under the null hypothesis of identical survival functions and if censoring is independent of group, Q has, asymptotically, a chi-squared distribution with one degree of freedom.

This test is optimal under the proportional hazards assumption and so the new tests proposed here, like any other new test, are not expected to be more powerful in this particular case. However, in many situations, the log-rank test is not optimal and other tests are superior. Different weights in (4) give rise to various other tests: $w_j = n_j$ gives rise to the Wilcoxon test [13]; $w_j = \sqrt{n_j}$ gives rise to the Tarone and Ware test [14] and these can outperform the log-rank test in various situations. For instance, the Wilcoxon test is more appropriate when the hazard ratio is more extreme at early survival than at later survival since early survival is given more weight; the Tarone and Ware test lies in between. A class of tests were introduced that were optimum for a range of alternatives to proportional hazards including the log-rank and Wilcoxon tests [15]. Similarly, [16] introduced a class of tests that incorporated many of the existing tests.

Several authors have considered the case of crossing survival curves or hazard functions [17, 18, 19, 20]. [21] shows how the curves can be compared after pre-specifying a time point, t_0, where the curves are expected to cross. [22] shows how Cox proportional hazards models can be adjusted by allow for crossing survival curves by introducing an interaction of treatment group with a covariate involving time (e.g. log of the time since surgery). However, this covariate has to be specified and modelled. [23] compared several methods for crossing survival curves and recommended a Levene-type test for this situation. [24] introduced a model that parameterises short-term and long-term hazard ratios and used this to develop a log-rank type test with adaptive weights [25]. On a different track from rank-type tests, Kolmogorov-Smirnov type tests have been developed [26]. [27] proposed a test based on the absolute difference of the area under the survival curves and show that the test outperforms the log-rank, Wilcoxon and Kolmogorov-Smirov tests for situations where survival curves cross and, where survival curves are close to start with but then diverge. The test is not very inferior under proportional hazards. The test statistic estimates the absolute difference under the survival curves using Kaplan-Meier estimates of the curves, with the variance of the statistic calculated using Greenwood's formula. There is a large drawback in that, unfortunately, the calculation needs the value of the correlations between absolute differences in the two estimated survival curves at all pairs of observed survival times. The authors take this correlation to be 0.5 for all pairs - but is this realistic?

The two test statistics proposed here are the PP-area and PP-length tests for the two group case and the PPP-area and PPP-length tests for the three group case which could be extended to four or more groups. These new tests are explored in the next two subsections. Under the proportional hazards assumption, there is a strong linear relationship between the squared PP-area and the PP-length and so the two tests based on these would be expected to perform similarly. The test using PP-length will probably perform better when the PP-curve keeps crossing or continually being attracted towards the diagonal, i.e. when the area is small but the length is large. One attractive feature of these tests is that the values of the test statistics have an immediate geometric interpretation, unlike that for the log-rank test say, which appears as just a number.

Tests Based on PP-Area and PP-Length

The distribution of the PP-area under the null hypothesis of identical survival curves is best determined by simulation since it would be very difficult to find its true distribution although some progress on this has been made and some simulation studies have given a hint as to the nature of the distribution (further details are available from the author). However, simulation to find critical values for the distribution of the PP-area under the null hypothesis is very easy since the only distribution needed for the simulation is the uniform distribution on the interval [0, 1]. This is because under the null hypothesis, the PP-curve, $[S_1(t), S_2(t)]$, can be re-parameterised as (u,u). Table **1** shows the 90%, 95% and 99% critical values for $n_1, n_2 = 50, 100, 200, 500$, based on one hundred thousand simulations for each combination of sample sizes. Clearly, the critical values decrease as sample size increases. A similar simulation exercise was carried out for PP-length with critical values also shown in Table **1**.

When random censoring occurs, then the critical values do not significantly change for the PP-area test, for instance with 50% censoring the critical values for PP-area in Table **1** become 0.074, 0.083 and 0.101. For the PP-length test the changes are slight for a small amount of censoring, but become more pronounced with increased censoring, for example, with 50% censoring the three critical values in Table **1** for $n_1 = n_2 = 100$ would be 1.449, 1.470 and 1.515. When censoring is heavy towards the end of the survival distributions, the PP-plot will not reach the origin and a choice has to be made as to whether (i) the final point is connected to the origin, essentially modelling the tail of the distributions and then the PP-

Table 1: Critical Values for PP-Area and PP-Length Tests

	n2=50			n2=100			n2=200			n2=500		
PP-area	**90%**	**95%**	**99%**	**90%**	**95%**	**99%**	**90%**	**95%**	**99%**	**90%**	**95%**	**99%**
n1=50	0.100	0.117	0.151	0.087	0.101	0.130	0.079	0.092	0.119	0.074	0.086	0.111
n1=100				0.071	0.083	0.107	0.061	0.071	0.092	0.055	0.064	0.082
n1=200							0.050	0.058	0.076	0.042	0.049	0.063
n1=500										0.031	0.037	0.047
PP-length	**90%**	**95%**	**99%**	**90%**	**95%**	**99%**	**90%**	**95%**	**99%**	**90%**	**95%**	**99%**
n1=50	1.470	1.481	1.504	1.456	1.465	1.484	1.450	1.457	1.473	1.446	1.452	1.466
n1=100				1.443	1.449	1.461	1.436	1.440	1.450	1.431	1.435	1.443
n1=200							1.429	1.432	1.439	1.424	1.427	1.431
n1=500										1.420	1.421	1.424

area calculated from the origin to (1,1), or (ii) simply calculate the PP-area from the point where the PP-plot ends to the point (1,1) and similarly for PP-length.

Also, in practice, it is recommended that the critical values of the distribution for particular values of n_1 and n_2 are found by simulation from the uniform distribution, but applying the pattern of censored values that has been seen in the data. So, if the *i*th ordered observation is censored in a particular group, then the *i*th ordered simulated observation is also censored for that group.

Many simulations were carried out to investigate the dependence of the critical points on n_1 and n_2. To illustrate the findings, The 90%, 95% and 99% points for PP-area and PP-length - $\sqrt{2}$ were plotted (not shown here) against $\left[n_1 + n_2(1-\frac{n_1}{n_2})\right]^{-\frac{1}{2}}, (n_1 \le n_2)$ for various values of n_1 and n_2. The values for (n_1, n_2) were randomly generated with each n_1 and n_2 taken from the discrete uniform distribution on (30, 500) with n_1 taken as the minimum of the two sample sizes. Some equal sample size cases were also added. A very clear linear relationship was found. Also, it was noted from QQ-plots that $k\left[n_1 + n_2(1-\frac{n_1}{n_2})\right]^{\frac{1}{2}} \times \text{PP} - \text{area}, k$ a constant, approximately follows a chi-squared distribution on six degrees of freedom. However the approximation is not good enough in the tail of the distribution to allow the distribution to give required critical values.

The power of the PP-area and PP-length tests was investigated using simulation and compared to that for the log-rank and Wilcoxon tests, the Levine-type test proposed by Le [23] and the Kolmogorov-Smirnov type test of Fleming [26]. The following three scenarios were considered: (i) proportional hazards using exponential distributions for the two distributions, the first with parameter $\lambda_1 = 1$ and the second with various values of λ_2; (ii) an exponential distribution with parameter $\lambda_1 = 1$ for the first distribution and for the second, a Weibull distribution with $\lambda_2 = 1$ and various values of β_2 giving crossing survival curves; (iii) an increasing hazards scenario without the survival curves necessarily crossing, using an exponential distribution for the first distribution with parameter $\lambda_1 = 1$ and an equal mixture of two exponential distributions with parameters $\lambda_1 - a$ and $\lambda_1 + a$ respectively for the second distribution.

The results are shown in Figure **2i**, **ii** and **iii** for the three scenarios where estimated power is plotted against the various distribution parameters. (The key is shown in Figure **2iii**; A - PP-area test, L - PP-length test, LR - log-rank test, W - Wilcoxon test, LV - Levine, KS - Kolmogorov-Smirnov). The significance level has been chosen as 0.05 and the estimated power is given by the proportion of times, in ten thousand simulations, a test rejects the null-hypothesis of equivalent survival curves. The sample sizes were $n_1 = n_2 = 100.$ Figure **2iv** shows the power for scenario (ii) but with different levels of censoring. Figure **2i** shows that for the proportional hazards situation (i), the log-rank test is more powerful than the others, but the PP-area and the Wilcoxon tests have only slightly less power, the KS and PP-length test slightly less power again and, as expected, the Levine-type test has poor power. For the crossing survival curves, (ii), the Levine-type test and the PP-length tests have very similar and best power, the PP-area and KS tests have less power and the log-

rank test and the Wilcoxon test do badly. For the increasing hazard ratio situation (iii), the PP-area, the PP-length and the KS tests all perform similarly with the highest power, the Levine-type and Wilcoxon tests have less power and the log-rank test performs badly. Figure **2iv** shows that for crossing survival curves (where $\beta = 1.6$), power decreases as the censoring increases for all the tests except the PP-area test where power tends to increase. Clearly, the PP-length test has higher power than the Levine-type test when censoring occurs. At 50% censoring the power of the PP-area test matches that of the PP-length test.

ESTIMATION OF THE HAZARD RATIO

The PP-plot is ideal for estimating the hazard ratio in the case of proportional hazards, where the plot is characterized as (u, u^{γ}). This function can easily be fitted in a variety of ways, for instance, using simple linear regression through the origin of $\log(p_2)$ on $\log(p_1)$ to estimate γ, or non-linear regression minimising $\sum (p_{2i} - p_{1i}^{\gamma})^2$. The latter was the method used here but it does imply that the two distributions take different roles, one the "response" variable and one the "regressor" variable. It is straightforward to put the two distributions on an equal footing by taking the "error" associated with a point on the PP-plot as the distance from the point to the closest point on the diagonal (i.e. the line perpendicular to the diagonal that passes through the point on the PP-plot). The regression can be easily carried out on the transformed data.

Many data sets were simulated for a variety of hazard ratios and sample sizes and the estimated hazard ratio compared to the true value. The results were also compared to the results from Cox proportional hazards models for estimating the hazard ratio. The bias and mean square error (MSE) were compared. The estimates based on the PP-plot were not quite as good as those based on Cox proportional hazards, Table **2** showing a small set of the comparisons. It is noted that the bias in estimating the hazard ratio is always positive. This means that the bias and MSE for the PP-curve estimate can be reduced by regressing p_2 on p_1 to obtain $\hat{\gamma}$, the estimate of γ and then regressing p_1 on p_2 to obtain $\hat{\delta}$, the estimate of $1/\gamma$. The two estimates are then combined as $\sqrt{\hat{\gamma}/\hat{\delta}}$ which does reduce the bias but is still slightly greater than that from the Cox model.

If proportion hazards are not appropriate in a particular situation, then the PP-plot allows a

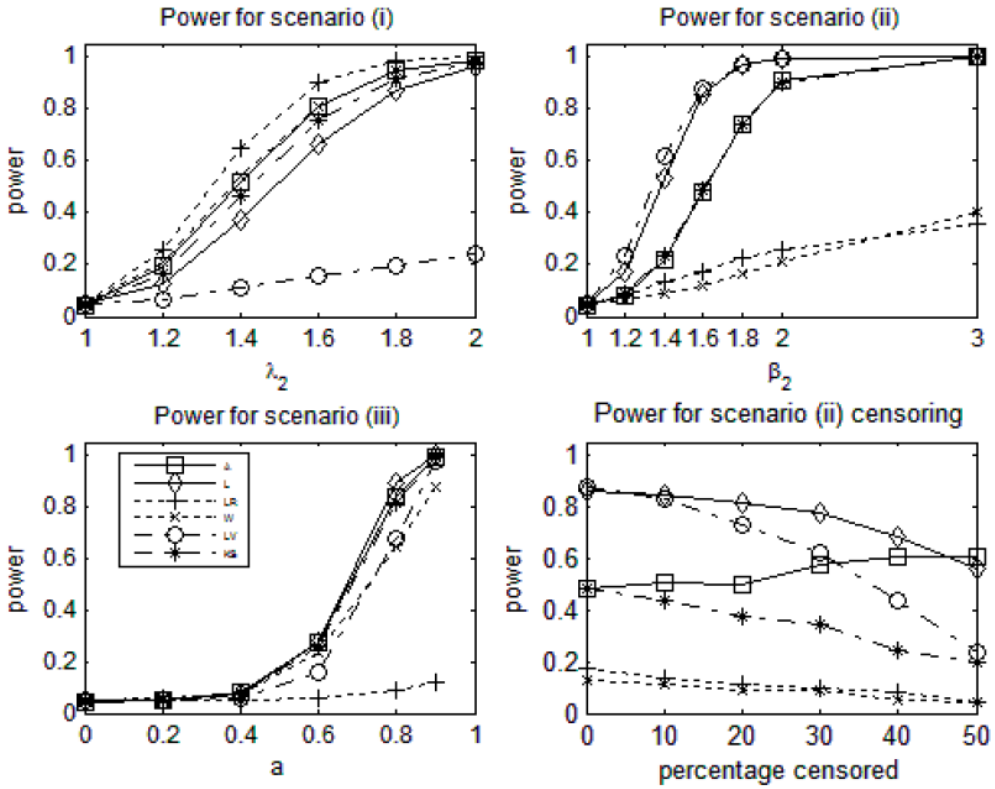

Figure 2: Plots of power, for significance level 0.05 for the following scenarios: (**i**) proportional hazards, (**ii**) crossing survival curves, (**iii**) an exponential distribution against a mixture of two exponential distributions, (**iv**) crossing survival curves with censoring.

Table 2: Bias and MSE when Estimating the Hazard Ratio

$\beta = 1.2$		$\gamma = 1.2$				$\gamma = 1.5$				$\gamma = 2.0$			
		Cox PH		**PP-curve**		**Cox PH**		**PP-curve**		**Cox PH**		**PP-curve**	
n1	n2	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
50	50	0.043	0.068	0.049	0.094	0.062	0.114	0.072	0.160	0.077	0.215	0.096	0.335
100	100	0.018	0.033	0.018	0.046	0.015	0.064	0.028	0.070	0.030	0.095	0.043	0.138
200	200	0.009	0.017	0.012	0.021	0.017	0.025	0.023	0.035	0.019	0.045	0.030	0.067
$\beta = 1.5$													
50	50	0.017	0.015	0.032	0.095	0.041	0.115	0.064	0.172	0.044	0.227	0.082	0.357
100	100	0.019	0.031	0.016	0.042	0.027	0.054	0.038	0.080	0.036	0.098	0.038	0.137
200	200	0.001	0.015	0.004	0.020	0.014	0.025	0.015	0.036	0.014	0.045	0.019	0.070

description of how the ratio of the hazard functions deviates from a constant (proportional hazards). Let the two survival distributions be S_1 and $S_2 = f(S_1)$ where f is an unknown monotonically non-decreasing function. Now a standard result is $S'(t) = -S(t)h(t)$ and so from $S_2(t) = f\{S_1(t)\}$,

$$S_2'(t) = f'\{S_1(t)\}S_1'(t)$$
$$S_2(t)h_2(t) = f'\{S_1(t)\}S_1(t)h_1(t)$$

$$h_2(t)/h_1(t) = \frac{f'\{S_1(t)\}S_1(t)}{f\{S_1(t)\}} \quad (5)$$

To model the ratio of the hazard functions, the function f has to be chosen and fitted and ideally it passes through (0, 0) and (1, 1). Here, the chosen function is $f(S_1) = S_1^{\exp\{\gamma(S_1)\}}$, where γ is a polynomial in S_1. Using equation (5), $h_2/h_1 = \{1 + S_1\log(S_1)\gamma'(S_1)\}\exp\{\gamma(S_1)\}$. This function is easily fitted to (p_{1i}, p_{2i}) using standard least squares on the complementary log-log transformation for p_1 and for p_2. Once fitted, the plot of h_2/h_1 against p_1 or against the corresponding times visually describes how the ratio of the hazard functions varies. Care has to be taken so that overfitting does not occur giving rise to an unrealistic plot. For illustartion, data were simulated from various exponential and Weibull distributions and the ratio of the hazard functions modelled. Figure **3** shows some of these. Figure **3i** shows the ratio of hazard functions obtained for data simulated from two exponential distributions with parameter values of 1.0 and 1.25 using a polynomial of order 1 for $\gamma(S_1)$. As expected, the ratio is more or less constant. Figure **3ii** show a PP-plot for data simulated from an exponential distributions with parameter value of 1.0 and a Weibull distribution with parameter values of $\lambda = 1.25$ and $\beta = 1.5$, together with a curve fitted for an order 3 polynomial for γ. Figure **3iv** shows the ratio of hazard functions based on this fitted curve, whilst Figure **3iii** shows the equivalent when a polynomial of order only 1 is chosen for γ.

TWO CANCER EXAMPLES

[21] uses data from a study comparing disease-free survival for autologous and allogenetic bone marrow transplants for follicular lymphoma [28]. Details are given in their paper where the emphasis is on survival after a pre-specified time point, but here the data are only used to illustrate the PP-plot, PP-area and PP-length tests, comparing them to standard tests. There were 596 observations in the autologous group and 175 in the allogenic group.

Figure **4i** shows the Kaplan-Meier survival curves and Figure **4ii** the corresponding PP-plot. The various standard tests for equality of survival distributions gave the following results: log-rank, p = 0.443; Wilcoxon, p = 0.169; Tarone, p = 0.666; Peto, p = 0.406; Modified Peto, p = 0.403; Fleming, p = 0.361. The area between the curve in the PP-plot and the diagonal (ending where the PP-plot ends and with a bounding line perpendicular to the diagonal) is 0.066 with estimated p-value <0.0001 based on the simulation of 100,000 values under the null-hypothesis. The length of the PP-curve was calculated as 1.028 with associated p-value of 0.005. The Levine-type test and the Kolmogorov-Smirnov test both rejected the null hypothesis with p < 0.0001.

Figure **4iii** shows how the ratio of hazard functions varies with p_1 and equivalently, Figure **4iv** shows how the ratio of hazard functions varies with time. The ratio starts with a value of approximately 2.3, drops rapidly to 0.1 and increases steadily to approximately 0.5.

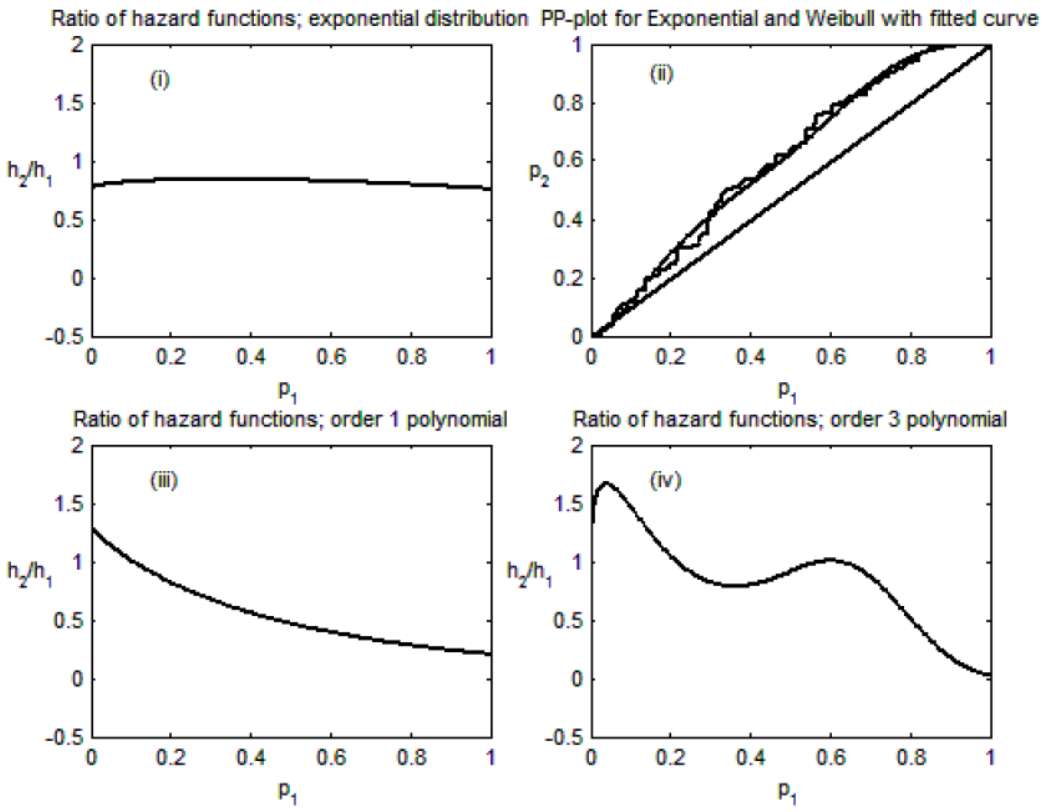

Figure 3: Ratio of hazard functions: (**i**) plot for two exponential distributions $(\lambda_1 = 1, \lambda_2 = 1.2)$, (**ii**) PP-plot and fitted curve for an exponential distribution $(\lambda_1 = 1)$ versus a Weibull distribution $(\lambda_1 = 1, \beta = 1.3)$ and fitted curve for a polynomial of order 3, (**iii**) ratio of hazard functions if the fitted curve is chosen to be order 1 and (**iv**) ratio of hazard functions if the fitted curve is chosen to be order 3.

The second example relates to survival times for a subgroup of patients within a pancreatic cancer trial. The patients of this subgroup have lymph nodes involved with the cancer, have had a resection where there has been a positive resection margin and then received chemotherapy. Interest here is in the difference in survival according to tumour differentiation, which is essentially how the cells of the tumour compare to cells of normal tissue. The categories are: *well-differentiated* which means the tumour cells are not too different from normal cells, then in decreasing order, *moderately-differentiated* and *poorly-differentiated*. Figure **5** shows the K-M curves for the three groups and the corresponding PPP-plot. Clearly, survival is worse when differentiation is poor. The log-rank statistic for differences in survival curves is 7.41 (p=0.025), the Wilcoxon test statistic is 15.91 (p <0.001), the PPP-area statistic is 0.259 (p=0.011) and the PPP-length statistic is 1.815 (p=0.014). All four tests reject the null hypothesis of equivalent survival distributions. Further analyses are not shown here where pairs of survival curves are compared using PP-plots and comparisons of hazard functions made.

DISCUSSION

The main thrust of this paper has been to introduce two new tests for testing the null hypothesis of equal survival distributions. Alongside this, has been the proposal that PP-plots might be a useful graphical method for displaying differences in survival curves alongside standard Kaplan-Meier plots. Also a new way of estimating the hazard ratio has been suggested, or, if proportional hazards are not appropriate, a way of displaying the behaviour of the changing ratio of the hazard functions.

The search for the ultimate test statistic that is the most powerful in all situations for testing for equal survival distributions is doomed to failure. Such a test does not exist. The new tests are in the same family as the log-rank test and its associated tests because they are non-parametric and only depend on the order of events and not on the times of the events. An appealing property of the new tests is that they are not designed for any particular situation for the survival distributions, unlike the log-rank test that is designed to perform well under proportional hazards. Another

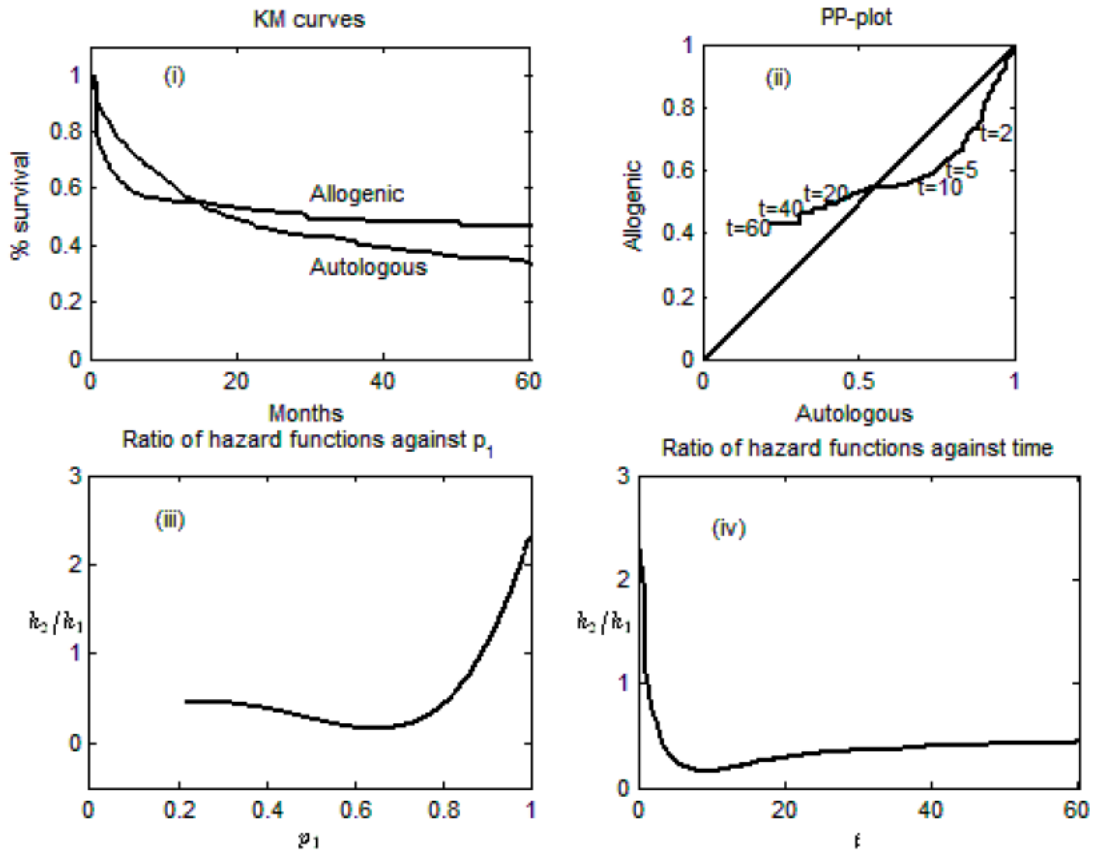

Figure 4: Follicular lymphoma example: (**i**) KM curves, (**ii**) PP-plot, (**iii**) ratio of hazard functions plotted against p_1 and (**iv**) ratio of hazard functions plotted against time.

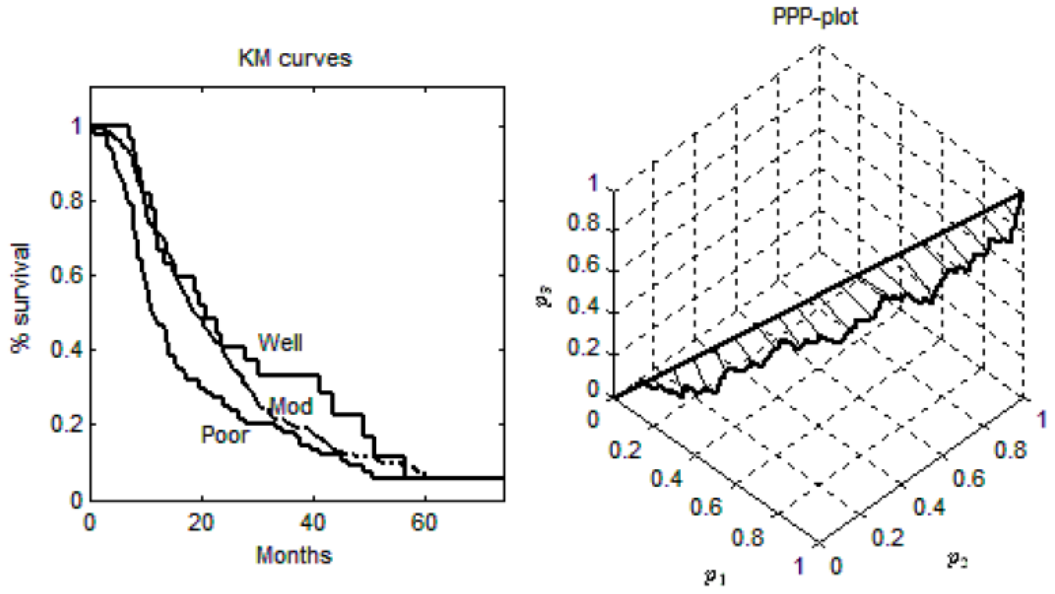

Figure 5: Pancreatic cancer example: KM curves and PPP-plot.

property is that the tests can be conditioned on the observed censoring pattern within the survival data and do not rely on the assumption of random censoring, although non-random censoring can lead to bias in the estimation of the Kaplan-Meier curve, a problem for all log-rank type tests. From power considerations, it was shown that the tests perform well in various situations, outperforming other tests.

The PP- and PPP-area, the PP- and PPP-length tests were used on two data sets for illustrative purposes, one relating to follicular lymphoma and the other to pancreatic cancer. Whether there will be widespread use of the new tests and the PP- and PPP-plots in the future will depend on availability of software. MATLAB (v 7.11.0584) programs were written for the analyses and graphs in this paper. However, an R-package will be built and deposited in the Comprehensive R Archive Network (CRAN) for others to use. Also further research is being undertaken on the distributional properties of the area and length tests.

ACKNOWLEDGEMENTS

The author would like to thank RJ Jackson for a useful discussions, Prof. BR Logan and the Center for International Blood and Marrow Transplant Research for use of the lymphoma data and Prof. JP Neoptolemos for use the pancreatic cancer data.

APPENDIX

Some basic differential geometry

This section outlines some basic differential geometry of curves and surfaces needed for the PP- and PPP-plots.There are many introductory books (for example [8, 9]).

Let I be an open interval in the real line R. Then a *curve* in three-dimensional Euclidean space, E^3, is defined by the mapping $x: I \to E^3, t \to [x_1(t), x_2(t), x_3(t)] = \mathbf{x}(t)$. One can imagine the curve as the path mapped out as a particle moves in time t in the Euclidean space. For example, a particle starting at *t* = 0 and finishing at *t* = 2, when moving along the curve, (2+*t*, 3+4*t*, -5-*t*), will have described a straight line from (2,3, -5) to (4,11, -7). In a similar manner, curves can be defined in E^2 and in E^n in general.

As the particle describes the curve, there is a notion of velocity and speed. The *velocity vector* of x at *t* is

$$\mathbf{x}'(t) = [x_1'(t), x_2'(t), x_3'(t)].$$

This is the velocity at the point $\mathbf{x}(t)$. The *speed* of the curve at $\mathbf{x}(t)$ is

$$\{x'_1(t)^2 + x'_2(t)^2 + x'_3(t)^2\}^{\frac{1}{2}}.$$

Note, we require that curves are *regular*, which means at no point is the velocity vector zero.

Example

Consider the curve given by

$$[\exp(-\lambda_1 t), \exp(-\lambda_2 t), \exp(-\lambda_3 t)] \text{ for } 0 < t < \infty.$$

This is the PPP-plot for three exponential survival functions. It has velocity vector

$$[-\lambda_1 \exp(-\lambda_1 t), -\lambda_2 \exp(-\lambda_2 t), -\lambda_3 \exp(-\lambda_3 t)]$$

and speed

$$\{\lambda_1^2 \exp(-2\lambda_1 t) + \lambda_2^2 \exp(-2\lambda_2 t) + \lambda_3^2 \exp(-2\lambda_3 t)\}^{\frac{1}{2}}$$ at the point $[\exp(-\lambda_1 t), \exp(-\lambda_2 t), \exp(-\lambda_3 t)]$.

The *arc length*, or distance along the path of the curve, from $t = t_0$ to $t = t_1$ is given by

$$\int_{t_0}^{t_1} \{x_1'(t)^2 + x_2'(t)^2 + x_3'(t)^2\}^{1/2} dt$$

Example

The arc length of the above curve defined by $0 < t < 1$ is

$$\int_0^1 \{\lambda_1^2 \exp(-2\lambda_1 t) + \lambda_2^2 \exp(-2\lambda_2 t) + \lambda_3^2 \exp(-2\lambda_3 t)\}^{\frac{1}{2}} dt$$

and this, like for most curves, has no explicit solution and has to be found numerically for particular values of the parameters.

Curves can be re-parameterised, changing the parameter t to another. In the example, if $u = \exp(-t)$, then the new parameterisation is

$$(u^{\lambda_1}, u^{\lambda_2}, u^{\lambda_3})$$

and the arc length is now given by

$$\int_{1/e}^1 -\{\lambda_1^2 u^{2\lambda_1} + \lambda_2^2 u^{2\lambda_2} + \lambda_3^2 u^{2\lambda_3}\}^{\frac{1}{2}} du.$$

The *Frenet formulas* allow the geometry of a curve to be studied, assessing properties such as curvature and torsion, but these are not needed for this paper.

Let *D* be an open set in E^2. A *surface* (or part of a surface) in E^3 is defined by a one-to-one mapping

$$x: D \to E^3, (u,v) \to [x_1(u,v), x_2(u,v), x_3(u,v)] = D(u,v).$$

Example

The mapping $[u,v,\exp(-\lambda_1 u-\lambda_2 u-vuv)]$, for $0<u,v<\infty$ and $0\le v\le\lambda_1,\lambda_2$, gives the surface of a bivariate joint survival function.

The surface area of a surface over a particular region is given by

$$\int_u\int_v\{(D_u\cdot D_u)(D_v\cdot D_v)-(D_u\cdot D_v)^2\}^{\frac{1}{2}} \quad (1)$$

where D_u and D_v are partial derivatives and $\cdot$ is the dot product.

Example

For the bivariate exponential survival distribution above, with $\lambda_1=\lambda_2=2$ and $v=1$, the surface is given by

$$D=[u,v,\exp\{-(2u+2v+uv)\}]$$

and hence the partial derivatives are given by

$$D_u=[1,0,-(2+v)\exp\{-(2u+2v+uv)\}]$$
$$D_v=[1,0,-(2+u)\exp\{-(2u+2v+uv)\}].$$

Then after some algebra the surface area between $0<u,v<1$ is given by

$$\int_0^1\int_0^1[1+(4+u+v)\exp\{-(4u+4v+2uv)\}]^{\frac{1}{2}}$$

which when calculated numerically has the value 1.1017.

Here, only one type of surface will be needed and that is a *ruled surface* which is a surface generated by the motion of a straight line called the *generator* or *ruling*. Let $\mathbf{x}:I\to E^3$ be a curve in E^3 and let $y:I\to E^3$ be a smooth function so that $y(u)\neq 0$ for all $u\in I$. The ruled surface is given by

$$D(u,v)=x(u)+vy(u),\quad u\in I, v\in R.$$

The curve $x(u)$ is called the *directrix* and $y(u)$ is the *ruling*. The surface area of the ruled surface is found as in the example above using formula (1).

REFERENCES

[1] Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis of data. Biometrika 1968; 55; 1-17.

[2] Michael JR. The stabilized probability plot. Biometrika 1983; 70; 11-17. http://dx.doi.org/10.1093/biomet/70.1.11

[3] Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. 2nd ed. Hoboken: Wiley 2011.

[4] Krzanowski WJ, Hand DJ. ROC curves for continuous data. Boca Raton: Chapman and Hall/CRC 2009.

[5] Yang H, Carlin D. ROC surface: a generalisation of ROC curve analysis. J Biopharm Stat 2000; 10; 183-96. http://dx.doi.org/10.1081/BIP-100101021

[6] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951; 22; 79-86. http://dx.doi.org/10.1214/aoms/1177729694

[7] Kullback S. The Kullback-Leibler distance. Am Stat 1987; 41: 341-42.

[8] Struik DJ. Lectures on Classical Differential Geometry. 2nd ed. London: Dover 1961.

[9] O'Neill B. Elementary Differential Geometry. Revised 3rd ed. Burlington: Academic Press 2006.

[10] Collett D. Modelling Survival Data in Medical Research. 2nd ed. Boca Raton: Chapman and Hall/CRC 2003.

[11] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Report 1966; 50; 163-70.

[12] Peto R. Contribution to the discussion of a paper by D.R. Cox. JRSS A 1972; 34; 205-207.

[13] Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. Biometrika 1965; 52; 203-23. http://dx.doi.org/10.1093/biomet/52.1-2.203

[14] Tarone RE, Ware J. On distribution free tests for equality of survival distributions. Biometrika 1977; 64; 156-60. http://dx.doi.org/10.1093/biomet/64.1.156

[15] Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika 1982; 69; 553-66. http://dx.doi.org/10.1093/biomet/69.3.553

[16] Jones MP, Crowley J. A general class of nonparametric tests for survival analysis. Biometrics 1989; 45; 157-70. http://dx.doi.org/10.2307/2532042

[17] Stablein DM, Carter WH Jr, Novak JW. Analysis of survival data with nonproportional hazard functions. Control Clin Trials 1981; 2: 149-59. http://dx.doi.org/10.1016/0197-2456(81)90005-2

[18] Mantel N, Stablein DM. The crossing hazard function problem. Statistician 1988; 37: 59-64. http://dx.doi.org/10.2307/2348379

[19] Liu K, Qiu P, Sheng J. Comparing two crossing hazard rates by Cox proportional hazards modelling. Stat Med 2007; 26: 375-91. http://dx.doi.org/10.1002/sim.2544

[20] Bouliotis G, Billingham L. Crossing survival curves: alternatives to the log-rank test. Trials 2011; 12(Suppl. 1): A137. http://dx.doi.org/10.1186/1745-6215-12-S1-A137

[21] Logan BR, Klein JP, Zhang MJ. Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. Biometrics 2008; 64: 733-40. http://dx.doi.org/10.1111/j.1541-0420.2007.00975.x

[22] Putter H, Sasako M, Hartgrink HH, van de Velde CJ H, van Houwelingen JC. Long-term survival with non-proportional hazards: results from the Dutch gastric cancer trial. Stat Med 2005; 24: 2807-21. http://dx.doi.org/10.1002/sim.2143

[23] Le CT. Statistical methods for the comparison of crossing survival curves. In: Balakrishnan N, Rao CR, Eds. Handbook of Statistics, Amsterdam: Elsevier 2004; vol. 23: pp. 277-289.

[24] Yang S, Prentice R. Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. Biometrika 2005; 92: 1-17. http://dx.doi.org/10.1093/biomet/92.1.1

[25] Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. Biometrics 2010; 66: 30-38. http://dx.doi.org/10.1111/j.1541-0420.2009.01243.x

[26] Fleming T, O'Fallon JR, O'Brien PC. Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. Biometrics 1980; 36: 607-625. http://dx.doi.org/10.2307/2556114

[27] Lin X, Xu Q. A new method for the comparison of survival distributions. Pharmaceut Statist 2010; 9: 67-76. http://dx.doi.org/10.1002/pst.376

[28] Van Besien K, Loberiza F, Bajorunaite, R, *et al*. Comparison of autologous and allogenetic hematopoietic stem cell transplantation for follicular lymphoma. Blood 2003; 102: 3521-29. http://dx.doi.org/10.1182/blood-2003-04-1205

Received on 28-03-2014 Accepted on 30-04-2014 Published on 14-05-2014

http://dx.doi.org/10.6000/1929-6029.2014.03.02.10