

Accuracy of point predictions in survival analysis

Robin Henderson^{1,*}, Margaret Jones² and Janez Stare³

¹*Medical Statistics Unit, Mathematics and Statistics, Lancaster University, LA4 4YF, U.K.*

²*Biometrics Department, Pfizer, Kent, CT13 9NJ, U.K.*

³*Faculty of Medicine, University of Ljubljana, Vrazov trg2, 1104 Ljubljana, Slovenia*

SUMMARY

Survival time prediction is important in many applications, particularly for patients diagnosed with terminal diseases. A measure of prediction error taken from the medical literature is advocated as a practicable method of quantifying reliability of point predictions. Optimum predictions are derived for familiar survival models and the accuracy of these predictions is investigated. We argue that poor predictive capability is inherent to standard survival models with realistic parameter values. A lung cancer example is used to illustrate difficulties in prediction in practice. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Following a recent prospective cohort study, Christakis and Lamont [1] concluded that doctors are inaccurate in their prognoses for terminally ill patients, with only 20 per cent of predictions accurate to within 33 per cent of survival time. This is disappointing given that the ability to predict future survival times is often of considerable importance. For cancer patients, for example, MacKillop and Quirt [2] state there are three reasons why accurate prediction is required. First, prognostic judgement can influence choice of treatment, especially in the terminal stages of a disease in deciding what palliative treatment is offered. Many such treatments have significant side-effects which may only be considered acceptable if the patient is likely to live long enough to experience any subsequent benefit. Second, accurate prediction can be important in the effective use of limited health care resources, since treating patients with expensive regimens without benefit not only subjects patients to unnecessary toxicity, but also wastes valuable resources. Finally, accurate prediction may help patients and their families come to terms with impending death and make suitable plans for their remaining lifespan.

There are also specific pressures on clinicians to make accurate predictions of survival in order that a patient may qualify for certain financial benefits. In the U.K. for instance, a patient

*Correspondence to: Robin Henderson, Medical Statistics Unit, Mathematics and Statistics, Lancaster University, LA4 4YF, U.K.

† E-mail: robin.henderson@lancaster.ac.uk

can claim additional financial support without the usual waiting time if a doctor certifies that the patient has 'progressive disease' and is not expected to live longer than six months'. Another occasion is when a patient with a terminal disease is discharged from hospital where, again in the U.K., responsibility for paying for continuing care falls on either the local Health Authority or the Department of Social Security. In 1995 the Department of Health issued guidelines for continuing care responsibilities [3]: each local Health Authority is required to clarify when it will pay for continuing care, generally requiring a doctor to certify that the patient has less than a certain time to live. In the Leeds area, for example, the time span is six weeks. Thus the accuracy of survival predictions has financial implications for the Health Authority and for the Department of Social Security.

In the U.S.A. similar pressure was brought to bear on doctors when in 1982 the Medicare insurance programme introduced hospice care for its beneficiaries. Kinzbrunner [4] states that Medicare benefit will apply if the patient is 'terminally ill, with a life expectancy of six months or less, as certified by two physicians, the patient's attending physician and the hospice medical director'. He cites the ability to predict patient survival with sufficient accuracy as a major issue adversely affecting the referrals of patients to hospice. Christakis and Escarce [5] found the actual survival of Medicare patients enrolled in a hospice to have median 36 days, with 16 per cent dying within 7 days and 15 per cent surviving to at least 6 months. These authors suggest that the short survival of patients in hospice may mean they have made an inadequate use of a desirable type of care, and may have undergone unnecessary and costly aggressive treatment for an unduly long period before enrolment into a hospice.

Given the importance of accurate prediction, an interesting question arises as to whether the use of objective methods based on statistical models and with the attachment of probability statements can replace or inform subjective clinical judgement. In discussion of reference [1], Parkes suggests clinicians should 'stop guessing' and make more use of model-based methods. Several attempts have been made to compare subjective and objective methods (for example, references [6–9]) but no clear and consistent conclusion has emerged. Thus, almost always in practice, survival prediction remains as a subjective assessment made by the patient's doctor. The validity of such assessments has been investigated in a number of studies covering a variety of situations (for example, references [1, 2] and [10–13]) with a common finding being that predictive accuracy is poor.

Why then are formal statistical models not used more often in practice in survival time prediction? Survival analysis is of course in widespread use to determine covariate effects, to compare different groups, and to form prognostic indices. None the less, in our experience it is rare for these models and results to be translated into individual point or interval lifetime predictions, even given a prognostic index. Presumably the answer is that statistical methods have not convincingly been demonstrated to lead to accurate and thus helpful prediction. In this paper we discuss why this might be the case and attempt to demonstrate that poor predictive accuracy is *inherent* for commonly used survival models in realistic situations. This is evinced in two ways:

1. Poor point estimates. Accuracy of point predictions can be measured by an appropriate loss function. In this work we advocate a particularly simple and easily interpreted loss, due to Parkes [10], which essentially classifies a prediction as seriously in error or otherwise. We show that the probability of serious error in practice is high.

2. Low explained variation unless hazard ratios are extremely large. Covariate knowledge, even if statistically highly significant, rarely improves predictive capability by a practically useful amount.

In Section 2 we introduce the example which originally motivated our interest in this topic. In Section 3 we describe Parkes' definition of serious error and show that its probability is high for a variety of statistical models. The common two-group treatment/control situation is considered in some detail in Section 4, with point predictions, predictive intervals and explained variation all investigated under accelerated failure and proportional hazards models. We return to the motivating example in Section 5 with some final remarks in Section 6 completing the paper.

Throughout, the focus is on model properties, not issues of sampling, estimation or model validity.

2. EXAMPLE: LUNG CANCER PREDICTION

This work was motivated by a study into the accuracy of survival time prediction for patients diagnosed with non-small-cell lung cancer, described by Muers *et al.* [14] and also discussed by Henderson [15] and Henderson and Jones [16]. Non-small-cell lung cancer is usually terminal and provides a good illustration of the type of situation where there is a genuine need for survival time predictions. Hence one purpose of the original study was to obtain subjective point predictions made at diagnosis by experienced clinicians, and compare these with objective predictions obtained from a proportional hazards model. The overall finding was that both forms of prediction were in general poor, worse than anticipated, leading to our investigation into why this might be the case, and whether improved prediction might be obtained under alternative statistical models.

Survival data are available for 272 patients together with the values of six covariates: age; sex (0 = F, 1 = M); activity score (0–4), and presence/absence of anorexia, hoarseness and metastases. Median survival time was 6 months and 17 per cent of patients had censored responses within the approximate three year follow-up. We concentrate on prediction given fitted models and omit other detail of data analysis other than to report high statistical significance for most covariate effects (z -statistics 1.24, -3.81 , 4.21, 2.16, 3.24 and 1.89 for age, sex, activity score, anorexia, hoarseness and metastases, respectively), with an overall likelihood ratio test 64.4 on 6d.f. under a Cox model.

Figure 1 shows Kaplan–Meier plots for patients categorized into roughly equally sized high, medium and low risk groups according to prognostic index as determined by fit of the following four models:

- (i) Cox proportional hazards

$$S(t|x) = S_0(t)^{\exp\{\beta x\}}$$

- (ii) Weibull

$$S(t|x) = e^{-\exp\{\beta x^*\}t^{\gamma}}$$

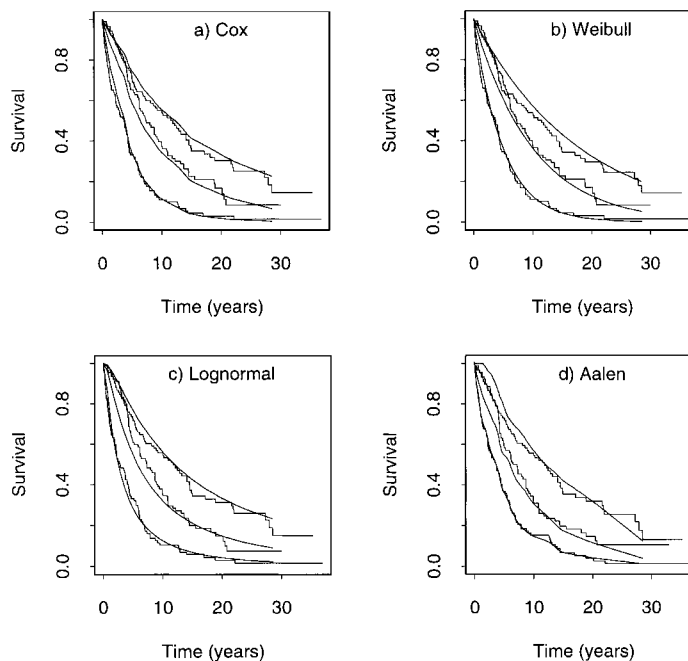


Figure 1. Observed (steps) and fitted (smooth) lung cancer survival by prognostic index group.

(iii) Log-normal

$$S(t|x) = 1 - \Phi\left(\frac{\log t - \beta x^*}{\sigma}\right)$$

(iv) Aalen semi-parametric linear [17]

$$S(t|x) = \exp\left\{-A_0(t) - \sum_{j=1}^6 A_j(t)x_j\right\}$$

In the above x^* is the covariate vector x augmented by 1 for an intercept term, and x_j represents the j th of the six elements of x . There is no unique prognostic index for the Aalen model since covariate effects change over time: Figure 1(d) is based on prognostic index at median survival time six months.

Superimposed on the observed survival curves in Figure 1 are fitted curves at the mean covariate values for each group, after smoothing for the Cox and Aalen semi-parametric models. The fit seems good on the basis of this plot for all four models and there seems to be little advantage to any one over the others.

Predictive accuracy under each of these models is discussed in Section 5, after consideration of some more general issues in the following two sections.

3. ACCURACY OF POINT PREDICTIONS: PARKES' DEFINITION OF SERIOUS ERROR

Accuracy of point survival time predictions can be assessed through expected loss if some suitable function $L(t, p)$ is available to compare outcome t with prediction p . For patient survival, however, it is very difficult to quantify loss associated with inaccurate prediction, given that costs cannot and should not be exclusively expressed financially. We have experimented with various types of loss function which might be of quite general use, consulted with clinicians, and eventually concluded that a very simple assessment of 'serious error' due to Parkes [10] provides a realistic method of measuring prediction accuracy which can be acceptable in a wide variety of circumstances. Parkes' definition is that a prediction is in serious error if it differs from outcome by a multiplicative factor of two, that is, if $t > 2p$ or $t < 0.5p$. This is equivalent to additive error of $\pm \log(2)$ on a log scale, and to a binary loss with $L(t, p) = 0$ if $0.5p < t < 2p$ and $L(t, p) = 1$ otherwise. Such a definition might be considered naive; the factor two is arbitrary, the function is not continuous so a small change in t near the boundaries of $(0.5p, 2p)$ can have a dramatic effect, and there is no continued increase in loss as t moves away from p by large amounts. None the less, the definition is simple, very easy to use, straightforward to explain to collaborators, allows for increasing uncertainty as the prediction horizon increases, and most of all is eminently sensible. Most people for example would accept that a lifetime prediction of, say, 2 months, was reasonably accurate if death occurs between about 1 and 4 months.

Taken together, the advantages of Parkes' definition outweigh the disadvantages and we recommend its use in practice in assessing predictive accuracy. Actually we might go a little further and replace the factor two by some other value κ , referring to a 'factor κ multiplicative error' E_κ if $t < p/\kappa$ or $t > \kappa p$. Usually $\kappa = 2$ seems a good choice however.

Having decided upon a measure of prediction accuracy, a natural next step is to consider the optimum point prediction, say p_κ , for any given κ . This is the value of p which minimizes the probability of factor κ prediction error

$$\begin{aligned} P(E_\kappa | p) &= P(T < p/\kappa) + P(T > \kappa p) \\ &= 1 - S(p/\kappa) + S(\kappa p) \end{aligned}$$

and hence solves

$$f(p/\kappa) = \kappa^2 f(\kappa p)$$

Explicit expressions for p_κ can be derived along with the corresponding values of $P(E_\kappa | p_\kappa)$ for a variety of commonly used statistical models, as the following examples illustrate.

3.1. Example 1: accelerated failure, symmetric on log scale

Suppose survival time T can be represented as

$$\log T = \mu + \varepsilon \quad (1)$$

where ε is a random variable, independent of the scale parameter μ , with distribution which is symmetric about zero. Then for all κ the optimum prediction is the median survival time,

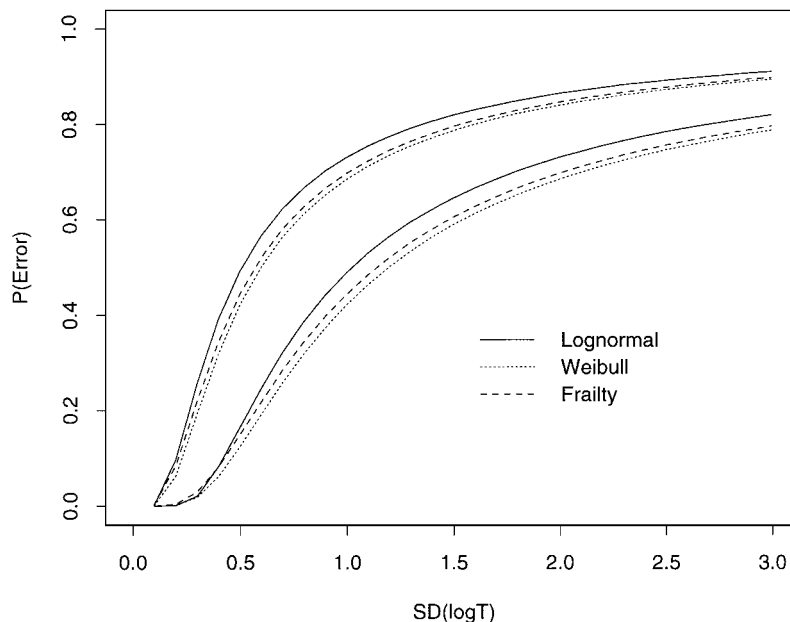


Figure 2. Probability of prediction error for log-normal, Weibull and Weibull/gamma frailty models, standardized to common standard deviation for log(survival time). The lower group of lines shows the probability of Parkes' error, and the upper group shows the probability of prediction not being within 33 per cent of outcome.

$m = e^\mu$. The probability of serious error does of course depend on the chosen value of κ but is independent of μ . For instance, if T has log-normal distribution with $\log T \sim N(\mu, \sigma^2)$ then

$$P(E_\kappa | p_\kappa) = 1 + \Phi\left(\frac{-\log \kappa}{\sigma}\right) - \Phi\left(\frac{\log \kappa}{\sigma}\right)$$

The probability of Parkes' error, with $\kappa = 2$, is shown in Figure 2 for a range of values of σ . If the standard deviation is small then clearly there is little variability in response and good prediction is possible, whereas there is a high probability of serious error when σ is large. In practice values of σ around one are not unusual, in which case E_2 is expected to occur for almost 50 per cent of patients.

The probability of predictions not being within 33 per cent of actual survival, the accuracy measure selected by Christakis and Lamont [1], is also shown in Figure 2. With this stricter measure of accuracy the probability of error is necessarily greater, now over 70 per cent at $\sigma = 1$.

3.2. Example 2: Weibull

Whilst the Weibull distribution can be expressed in accelerated failure form, the random term ε has extreme value distribution, which is not symmetric about zero, and so the previous

results do not apply. Instead, with the survivor function parameterized as

$$S(t) = \exp\{-\lambda t^\gamma\}$$

we can show that the optimum prediction is

$$p_\kappa = \left(\frac{2\gamma \log(\kappa)}{\lambda(\kappa^\gamma - \kappa^{-\gamma})} \right)^{1/\gamma}$$

no longer the median, and depending now upon the choice of κ , though difference from the median is slight.

The probabilities of Parkes' error and of predictions not being within 33 per cent of outcome are shown in Figure 2. This plot is based on the optimum predictions above. If the median is used for prediction the Weibull error probabilities are only slightly higher and there is very little loss in using this simpler value in practice. In both cases the error probabilities do not depend on the scale parameter λ . For comparability with log-normal we selected the shape parameter γ to give specified standard deviations for $\log(T)$, and with this scaling we see close similarity between the models. For reference, shape parameters of 0.75, 1 and 1.25 correspond to standard deviations of 2.92, 1.65 and 1.05, respectively, for $\log(T)$. At shape 1 the probability of Parkes' error is 0.528 if the optimum prediction is used, 0.543 if median survival time is used. Thus even without the additional complication of sampling and estimation error, if one accepts that survival times in practice can be approximately exponentially distributed then Parkes' error will occur for at least 50 per cent of cases, even with optimal point prediction.

3.3. Example 3: gamma frailty mixture of Weibulls

Now suppose that subject-specific unobserved frailty terms act multiplicatively on the hazard functions, and take the common assumption that frailties have gamma distribution of unit mean but variance ξ [18, 19]. Assume that, conditional upon the frailty, survival times have Weibull distribution parameterized as in the previous example. The unconditional, marginal survival distribution is then

$$S(t) = \left(\frac{1}{1 + \xi \lambda t^\gamma} \right)^{\frac{1}{\xi}}$$

and optimal point predictions can again be obtained. The exact expressions are not given here because, just as for Weibull, there is little difference from the median survival time, which is simple to use and is independent of the choice of κ . Once more the probability of error is independent of the scale parameter λ , whether median or optimal prediction is used.

Error probabilities are shown in Figure 2, with ξ chosen so that half of the variance in $\log(T)$ is due to frailty, that is, the conditional variance given frailty is half the overall variance. Use of medians rather than optimal predictions makes very little difference and in both cases the values are very close to those given for Weibull, with the same comments applying. Clearly the error probability depends on the standard deviation of $\log(T)$, but conditional on that it seems to be fairly insensitive to choice of model.

4. TREATMENT/CONTROL STUDIES: COMPARISON OF PROPORTIONAL HAZARDS AND ACCELERATED FAILURE MODELS

In the previous section we considered prediction for individuals given a correctly specified statistical model, incorporating covariate information if required into subject-specific parameter choices. We turn now to the important situation of a two-group clinical trial and compare predictive accuracy under two alternative models for the treatment effect: proportional hazards (PH) and accelerated failure (AF). Given any functional form for survival in the control group we assume an additive effect on the log scale for an AF treatment model, or a multiplicative effect on the hazard for a PH treatment model. More specifically we consider the following parametric models:

- (i) control group – $\log(T) \sim N(0, \sigma_0^2)$, that is

$$S_0(t) = 1 - \Phi\left(\frac{\log t}{\sigma_0}\right)$$

- (ii) treatment model 1 – proportional hazards

$$S_1(t) = S_0(t)^{1/r}$$

- (iii) treatment model 2 – accelerated failure, $\log(T) \sim N(\mu, \sigma_0^2)$ with

$$S_2(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma_0}\right)$$

For given relative risk (hazard ratio) r under treatment model 1, PH, for comparability we select (log) location parameter μ to give the same median survival time in treatment model 2, AF.

First we consider accuracy of point predictions. Table I illustrates $P(E_2 | p_2)$ as relative risk changes from one to five for three choices of standard deviation σ_0 for $\log(T)$ in the control group. Since the AF treatment model is log-normal, example 1 above shows $p_2 = m$ and $P_{AF}(E_2 | m)$ depends only on σ_0 , not location change on the log scale, and hence not relative risk r in this illustration. Optimal prediction p_2 is tabulated for the PH treatment model 1, together with $P_{PH}(E_2 | p_2)$ and for reference $P_{PH}(E_2 | m)$. Note that for $r > 1$ the optimum prediction p_2 is smaller than m but $P_{PH}(E_2 | p_2)$ and $P_{PH}(E_2 | m)$ are close.

The central part of Table I has $\sigma_0 = 1$, which is reasonably realistic for practical applications. Error probabilities are near 50 per cent for both the control group and the treatment group under the AF model. If the PH treatment model is assumed, the probability of E_2 increases with r to almost 70 per cent at relative risk five. Clearly the chance of serious error will be reduced if there is less variability in response, as illustrated by the first part of the table with $\sigma_0 = 0.5$. Control and AF error probability is now reasonable at under 20 per cent, but note that under the PH treatment model $P_{PH}(E_2 | p_2)$ can still be over 40 per cent at $r = 5$. In practice there is usually (perhaps always) relatively large variability between individuals and small σ_0 is unlikely. The final section of Table I with $\sigma_0 = 1.5$ is included for reference, to illustrate how high the probability of Parkes' error can quickly become if the survival distribution in the control group itself is long tailed.

Second, we consider predictive intervals. In our opinion $P(E_k | p)$ gives a practically useful measure of confidence one might have in point prediction p , but obviously another method

Table I. Prediction error probabilities for accelerated failure (AF) and proportional hazards (PH) treatment group models: σ_0 = standard deviation of $\log(T)$ in log-normal control group; r = relative risk; m = median survival = optimal AF prediction; p_2 = optimal PH prediction.

σ_0	r	m	$P_{AF}(E_2 m)$	p_2	$P_{PH}(E_2 p_2)$	$P_{PH}(E_2 m)$
0.5	1	1.00	0.166	1.00	0.166	0.166
	2	1.40	0.166	1.38	0.267	0.267
	3	1.78	0.166	1.72	0.337	0.337
	4	2.15	0.166	2.05	0.389	0.389
	5	2.54	0.166	2.37	0.429	0.431
1	1	1.00	0.488	1.00	0.488	0.488
	2	1.96	0.488	1.86	0.580	0.580
	3	3.16	0.488	2.86	0.632	0.633
	4	4.64	0.488	4.01	0.667	0.669
	5	6.44	0.488	5.35	0.693	0.695
1.5	1	1.00	0.644	1.00	0.644	0.644
	2	2.75	0.644	2.52	0.712	0.713
	3	5.62	0.644	4.77	0.750	0.750
	4	9.99	0.644	7.94	0.775	0.776
	5	16.35	0.644	12.22	0.793	0.794

Table II. Eighty per cent prediction intervals and ratio of limits for two-group scenario (as Table I).

σ_0	r	Accelerated failure			Proportional hazards		
		$q_{0.1}$	$q_{0.9}$	$q_{0.9}/q_{0.1}$	$q_{0.1}$	$q_{0.9}$	$q_{0.9}/q_{0.1}$
0.5	1	0.53	1.90	3.6	0.53	1.90	3.6
	2	0.74	2.66	3.6	0.64	3.20	5.0
	3	0.94	3.37	3.6	0.74	4.69	6.4
	4	1.13	4.09	3.6	0.82	6.42	7.9
	5	1.34	4.82	3.6	0.89	8.44	9.5
1	1	0.28	3.60	13.0	0.28	3.60	13.0
	2	0.54	7.07	13.0	0.42	10.24	24.6
	3	0.88	11.38	13.0	0.54	21.98	40.5
	4	1.29	16.70	13.0	0.67	41.22	61.6
	5	1.79	23.20	13.0	0.80	71.16	89.5
1.5	1	0.15	6.84	46.7	0.15	6.84	46.7
	2	0.40	18.80	46.7	0.27	32.77	122.3
	3	0.82	38.39	46.7	0.40	103.06	257.3
	4	1.46	68.27	46.7	0.55	264.68	483.6
	5	2.39	111.77	46.7	0.71	600.24	846.0

is to consider the width of predictive intervals (PI). For completeness Table II shows 80 per cent PIs, providing 10 per cent and 90 per cent quantiles and their ratio for the PH and AF treatment models, the ratio being independent of r for the AF model. The combination of high

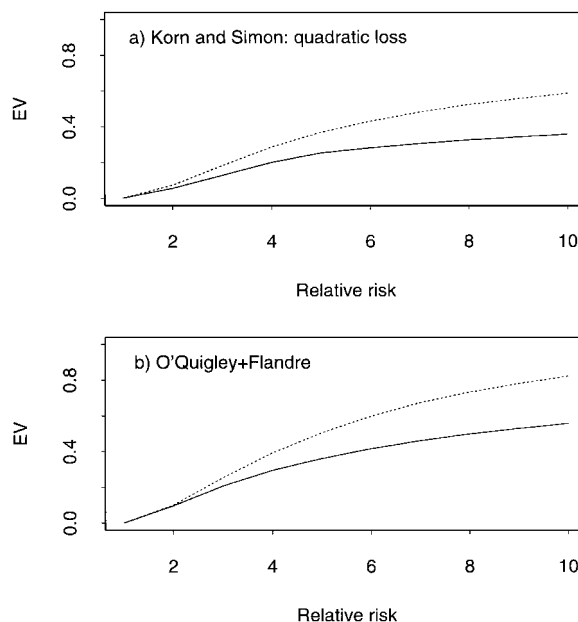


Figure 3. Explained variation under accelerated failure (dotted) and proportional hazards (solid) treatment/control models.

σ_0 and high r and hence extremely wide PI is unlikely to occur in practice, but even at more moderate values of σ_0 the PI width quickly becomes large as r and hence variance increases. Note that given common median survival time the PH model for treatment effect leads to much wider predictive intervals than AF, assuming the same log-normal control (baseline) distribution.

Finally we consider the relative increase in predictive capability obtained through knowledge of treatment group. This can be assessed by a measure of explained variation (EV) akin to R^2 for linear regression, used to assess the relative increase in predictive accuracy (in some respect) due to covariate information. Unfortunately, although a variety of EV measures have been suggested for survival analysis [20], no single proposal has yet been widely accepted and therefore we shall consider two alternatives:

1. The Korn and Simon measure based on expected loss [21]. Suppose $L(t, p)$ is the loss associated with point prediction p and actual outcome t . The Korn and Simon explained variation measure compares minimum expected loss with and without covariate information. For this illustration we will select quadratic loss $L(t, p) = (t - p)^2$.
2. O'Quigley and Flandre's R^2 [22], based on the variability of Schoenfeld residuals at each observed failure time with and without covariate information.

Our purpose is not to compare performance of these measures, but rather to investigate how quickly explained variation increases with relative risk for the different survival models. Figure 3 illustrates how EV increases with relative risk for the two selected measures. Val-

ues shown are expected (population) quantities as sample size increases indefinitely, and two features are common to both plots:

- (i) EV is low unless relative risk is very large (for example, $r = 2$ or 3 would be considered effective in practice);
- (ii) given the same log-normal baseline group and equal medians in the control group, EV is lower for PH than AF.

Qualitatively similar findings are obtained with other suggested measures of explained variation, including the Korn and Simon measure with alternative loss functions, Schemper's V measures [23] and an information gain statistic proposed by Kent and O'Quigley [24]. The reason for low EV is readily explained. Essentially, EV can only be expected to be high if the *between density* variation is high in comparison with the *within density* variation [16]. Even when the separation between control and treatment group survival curves is high, in practice there is often relatively little difference between the *locations* of the treatment and control group densities, for proportional hazards models in particular. The main difference is an increase in *variance* in the treatment group in comparison with control, and thus low EV should be expected. Survival analysts usually work with hazards and survival functions and there is rarely any need to consider density functions explicitly. Hence the effect of AF and PH model assumptions on densities is not always considered.

5. LUNG CANCER EXAMPLE CONTINUED

We now return to the lung cancer example introduced in Section 2. Covariate effects are highly statistically significant, as stated earlier, but explained variation is low in agreement with the findings above: the Korn and Simon measure with quadratic loss is 0.21, O'Quigley and Flandre's statistic is 0.23, Schemper's V1 is 0.13, and Kent and O'Quigley's information gain is 0.22, all under a Cox PH model.

Figure 4 shows the median predicted lifetime m under each model for 17 individuals, at the 10, 15, 20, ..., 90 per cent points of prognostic index as obtained under the Cox model, with points joined for clarity. The same individuals are considered in all four plots for comparability. Median survival time is selected as it is optimum prediction under log-normal, close to optimum under Weibull, and because no explicit expression for optimum prediction is available under Cox and Aalen models. The outcome interval $(0.5m, 2m)$ within which m would be considered an acceptable prediction under Parkes' definition is indicated, and where possible an 80 per cent predictive interval is shown. These intervals are not available under the semi-parametric Cox and Aalen models when estimated survival probability at the maximum observed death time does not fall below 0.1, without further untested assumptions at least.

Median predictions are generally quite similar under all four models, as are PIs for high risk patients. There are differences between the two parametric models for low risk individuals however, with the log-normal model yielding wider intervals than Weibull.

Taking prediction m , estimated Parkes' error probabilities $P(E_2 | m)$ depend on covariates for the Cox and Aalen models but not for Weibull and log-normal. For the 17 individuals

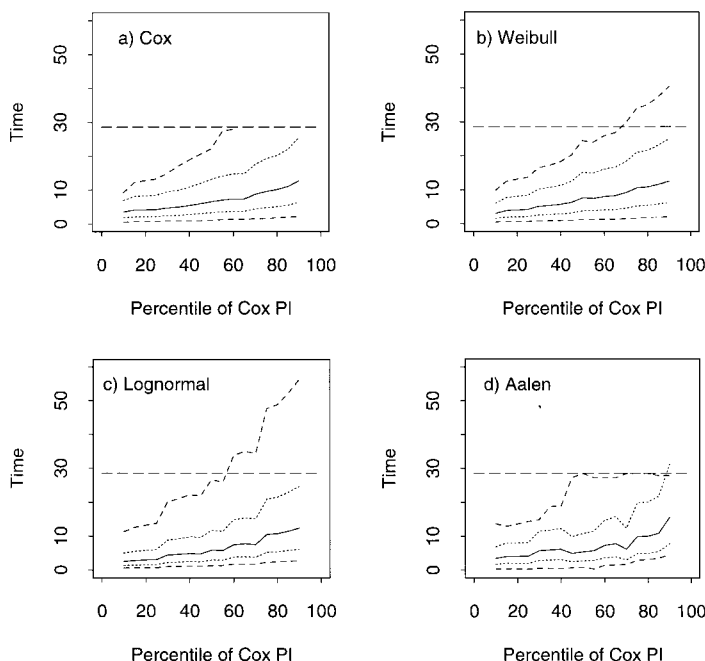


Figure 4. Point prediction (solid), 80 per cent intervals (short dashed) and E_2 bounds (dotted) for lung cancer patients at 10, 15, ..., 90 per cent points of prognostic index under Cox model. Upper limit to 80 per cent interval not available for semi-parametric models ((a), (d)) if in excess of maximum observed death time (long dashed line).

considered in Figure 4, estimated values are:

Model	$P(E_2 m)$
Cox	0.48–0.59
Weibull	0.53
Log-normal	0.56
Aalen	0.48–0.59

For reference, if optimal prediction p_2 is used for the Weibull model in place of m , error probability is only slightly lower at $P(E_2 | p_2) = 0.52$. Hence for all models and all individuals there is a substantial probability of predictions being in ‘serious error’ according to Parkes’ definition. This is also true for subjective clinicians’ predictions, which are available for these data. After omission of 36 cases which could not be classified because of censoring, 52 per cent of the remaining predictions were in Parkes’ error, with 37 per cent of predictions seriously high and 15 per cent seriously low.

6. DISCUSSION

Christakis and Lamont [1] found that doctors’ predictions of survival times were accurate to within 33 per cent of the outcome for only 20 per cent of terminally ill patients, and were

subject to Parkes' factor two error for 67 per cent of patients. These values are broadly similar to those seen in this paper for statistical models with realistic parameter values. It seems that Parkes' suggestion [1] that clinicians should stop guessing and make more use of statistical models is unlikely to lead to much improvement in accuracy in point predictions.

One can argue, and given the results above perhaps *should* argue, that prediction on the time axis is unnecessary, and it is preferable to consider the probability axis and the full predicted survival distribution for any individual. In practice, however, the time axis remains the most natural measure for many people. Thus it is usually easier to elicit expected patient survival time from experienced clinicians rather than a subjective assessment of probability of survival to a certain time point; and to many patients the natural question remains 'How long have I got Doctor?' [25].

Clearly point predictions alone are inadequate and some type of reliability measure should be attached. A prediction interval is of course one option, but another is to quote the associated probability of error by a practically important amount. We suspect that such a method of qualifying predictions may be easily understood by patients. Consider for instance the PH results in the central rows of Tables I and II ($\sigma_0 = 1, r = 3$), and assume the time unit is months. The information from clinician to patient might be paraphrased as 'my best guess is that you will live 3 months but there is a 60 per cent chance I will be seriously wrong' or 'my best guess is that you will live 3 months but there is an 80 per cent chance of you dying sometime between two weeks and two years'. In either case in this example the point of qualifying the point prediction is to indicate its unreliability, and it may be that the more direct first statement is clearer. In the absence of other information as to what constitutes a practically important prediction error, we recommend use of the Parkes definition as serious error if prediction and outcome differ by a multiplicative factor of two or more. As discussed above, this measure can be criticized but in our view its advantages outweigh its disadvantages.

Whatever measure or method is selected, it is clear from the results in the previous sections that poor predictive capability should be expected for the well known and much used survival models considered, however much effort is spent on data collection and analysis. Models are of course useful for the identification of risk factors and for group comparisons, but at the individual level even if complete covariate information is available and all parameters are fully known we should anticipate poor point predictions and wide predictive intervals. Perhaps this is an unavoidable consequence of real survival data.

ACKNOWLEDGEMENTS

We are grateful to the Thoracic Group of the Yorkshire Regional Cancer Organisation for providing the data used in Section 5. We thank the editor and two referees for detailed and valuable comments on previous versions of this paper. In particular, we are grateful to one referee for a suggestion that models can be compared by fixing the standard deviation of $\log(T)$, leading to Figure 2.

REFERENCES

1. Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *British Medical Journal* 2000; **320**:469–473.
2. MacKillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. *Journal of Clinical Oncology* 1997; **50**:21–29.
3. Department of Health. Health Service Guidelines 95(8). Department of Health Guidelines. HMSO, 1995.
4. Kinzbrunner BM. Ethical dilemmas in hospice and palliative care. *Supportive Care in Cancer* 1995; **3**:28–36.

5. Christakis NA, Escarce JJ. Survival of Medicare patients after enrollment in hospice programs. *New England Journal of Medicine* 1996; **335**:172–178.
6. Knaus WA, Harrell FE, Lynn J, Goldman L, Phillips RS, Connors AF, Dawson NV, Fulkerson WJ, Califf RM, Desbiens N, Layde P, Oye RK, Bellamy PE, Hakim RB, Wagner DP. The SUPPORT prognostic model—objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine* 1995; **122**: 191–203.
7. Barrett BJ, Parfrey PS, Morgan J, Barre P, Fine A, Goldstein MB, Handa SP, Jindal KK, Kjellstrand CM, Levin A, Mandin H, Muirhead N, Richardson RMA. Prediction of early death in end-stage renal disease patients starting dialysis. *American Journal of Kidney Diseases* 1997; **29**:214–222.
8. Hallen S, Hasberg A, Edna TH. Estimating the probability of acute appendicitis using clinical criteria of a structured record sheet: the physician against the computer. *European Journal of Surgery* 1997; **163**:427–432.
9. Poses RM, Smith WR, McClish DK, Huber EC, Clemo FLW, Schmitt BP, AlexanderForti D, Racht EM, Colenda CC, Centor RM. Physicians' survival predictions for patients with acute congestive heart failure. *Archives of Internal Medicine* 1997; **157**:1001–1007.
10. Parkes CM. Accuracy of predictions of survival in later stages of cancer. *British Medical Journal* 1972; **2**: 29–31.
11. Heyes-Moore LH, Johnson-Bell VE. Can doctors accurately predict the life expectancy of patients with terminal cancer? *Palliative Medicine* 1987; **1**:165–166.
12. Forster I, Lynn J. Predicting lifespan for applicants to inpatient hospice. *Archives of Internal Medicine* 1988; **148**:2540–2543.
13. Poses RM, McClish DK, Bekes C, Scott WE, Morley JN. Ego bias, reverse ego bias, and physicians prognostic judgements for critically ill patients. *Critical Care Medicine* 1991; **19**:1533–1539.
14. Muers MF, Shevlin P, Brown J. Prognosis in lung cancer: physicians' opinions compared with outcome and a predictive model. *Thorax* 1996; **51**:894–902.
15. Henderson R. Modelling conditional distributions in bivariate survival. *Lifetime Data Analysis* 1996; **2**:241–259.
16. Henderson R, Jones M. Prediction in survival analysis: model or medic? In *Lifetime Data: Models in Reliability and Survival Analysis*. Kluwer: Dordrecht, 1996; 125–129.
17. Aalen OO. Further results on the nonparametric linear regression model in survival analysis. *Statistics in Medicine* 1993; **12**:125–129.
18. Aalen OO. Effects of frailty in survival analysis. *Statistical Methods in Medical Research* 1994; **3**:227–243.
19. Hougaard P. Frailty models for survival data. *Lifetime Data Analysis* 1995; **1**:255–273.
20. Schemper M, Stare J. Explained variation in survival analysis. *Statistics in Medicine* 1996; **15**:1999–2012.
21. Korn LK, Simon R. Measures of explained variation for survival data. *Statistics in Medicine* 1990; **9**:487–503.
22. O'Quigley J, Flandre P. Predictive capability in proportional hazards regression. *Proceedings of the National Academy of Science* 1994; **91**:2310–2314.
23. Schemper M. The explained variation in proportional hazards regression. *Biometrika* 1990; **77**:216–218.
24. Kent JT, O'Quigley J. Measures of dependence for censored survival data. *Biometrika* 1988; **75**:525–534.
25. Maher EJ. How long have I got Doctor? *European Journal of Cancer* 1994; **30A**:283–284.