

Comparing Cox and parametric models in clinical studies

Alessandra Nardi^{1,2,‡} and Michael Schemper^{2,*,†}

¹*Department of Systems Theory, University of Teramo, Viale Crucoli 122, I-64100 Teramo, Italy*

²*Section of Clinical Biometrics, Department of Medical Computer Sciences, Vienna University, Spitalgasse 23, A-1090 Vienna, Austria*

SUMMARY

Parametric models are only occasionally used in the analysis of clinical studies of survival although they may offer advantages over Cox's model. In this paper, we report experiences that we have made fitting parametric models to data sets from different clinical trials mainly performed at the Vienna University Medical School. We emphasize the role of residuals for discriminating among candidate models and judging their goodness of fit. The effect of misspecification of the baseline distribution on parameter estimates and testing has been explored. The results from parametric analyses have always been contrasted with those from Cox's model. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: goodness of fit; model choice; normal-deviate residuals; survival analysis

1. INTRODUCTION

Although Cox's semi-parametric model [1] is the most frequently employed regression tool for survival data, fully parametric models [2, 3] may offer some advantages. Based on asymptotic results, Efron [4] and Oakes [5] showed that, under certain circumstances, parametric models lead to more efficient parameter estimates than Cox's model. With decreasing sample sizes, relative efficiencies may further change in favour of parametric models. When empirical information is sufficient, parametric models can provide some insight into the shape of the baseline hazard. Furthermore, extrapolations of survival functions become possible which—though speculative—may be of interest in some applications (see e.g., Reference [6]).

However, fully parametric models involve stronger assumptions than semi-parametric models. Compared to Cox regression there is the added requirement of checking the

*Correspondence to: M. Schemper, Section of Clinical Biometrics, Department of Medical Computer Sciences, Vienna University, Spitalgasse 23, A-1090 Vienna, Austria.

†E-mail: Michael.Schemper@akh-wien.ac.at

‡E-mail: Alessandra.Nardi@akh-wien.ac.at

Contract/grant sponsor: Marie-Curie Research Training Grant of the European Commission; contract/grant number: ERBFMBICT983189

Received December 2001

Accepted April 2003

appropriateness of the chosen family of distributions. To avoid increased efforts with model checking and because dangers of misspecification are assumed to increase with parametric survival models, statisticians tend to prefer Cox's model.

We show that the task of checking goodness of fit for parametric models is not formidable and that it can be standardized. For this purpose, we use and recommend residuals recently developed for Cox regression [7].

By means of the statistical analysis of four studies of survival, we re-examine the role of parametric models versus Cox regression and investigate in which aspects there is a net benefit of parametric survival models over Cox regression. We have selected three studies from the many clinical trials of the Vienna University Medical School that we have analysed in recent years and also deal with a trial performed in the Killingbeck hospital in Leeds, U.K. [8].

In the following section, we briefly review the models and diagnostic tools for model checking used in the four case studies of Section 3. Finally, we summarize our experiences of using parametric models versus Cox regression in Section 4.

2. METHODS

2.1. Cox and parametric models

We briefly recall the definitions of Cox and parametric models. Assume that for a sample of n individuals ($i=1, \dots, n$) the survival time t , the status indicator δ (1 for dead and 0 for alive) and a vector of explanatory variables \mathbf{z} are recorded. We also assume that survival times are continuously distributed and that the explanatory variables are linked to survival through $\exp(\beta'\mathbf{z})$ or $\exp(\gamma'\mathbf{z})$, where β and γ are vectors of unknown regression parameters.

The basic Cox model assumes that

$$\lambda(t; \mathbf{z}) = \lambda_0(t)e^{\beta'\mathbf{z}}$$

where $\lambda_0(t)$ denotes the baseline hazard function obtained for an individual with $\mathbf{z}=\mathbf{0}$. The explanatory variables act multiplicatively on the baseline hazard that is completely unspecified. This basic model can be extended by including time-dependent covariates and/or time-dependent coefficients [9].

The parametric accelerated failure time models can be represented in the form

$$T_Z = T_0 e^{-\gamma'\mathbf{z}}$$

where T_Z denotes the survival (or failure) time with explanatory variables \mathbf{z} and $\mathbf{z}=\mathbf{0}$ again denotes the baseline level. In this case, the explanatory variables act multiplicatively and directly on the survival time so that their effect is to accelerate or decelerate time to death relative to the time associated with the baseline level. The baseline distribution is completely specified up to an unknown shape parameter. We shall consider the exponential, Weibull, log-normal and log-logistic regression models. The exponential and Weibull models correspond to the assumptions of a constant and of a monotone baseline hazard, respectively. In the log-normal model and the log-logistic model the baseline hazard has value 0 at $t=0$, increases to a maximum and then decreases, approaching 0 as t becomes large.

Results from analyses by Cox and parametric models can be summarized in terms of ‘relative risk’ ($= \exp(\beta)$) and ‘acceleration factor’ ($= \exp(-\gamma)$), respectively.

2.2. Normal-deviate residuals

For the purpose of discriminating among alternative parametric models, we used graphical methods based on residuals. It is well established for the general linear model that, in order to verify the assumption of normally distributed outcomes, the distribution of residuals is checked for approximate normality. Similarly, in survival analysis a reference distribution is needed to verify any hypothesis for the distribution of survival times, the knowledge of the expected value of a residual being not enough for this purpose. Among the different types of residuals available in survival analysis, we have chosen normal-deviate residuals [7] since they most closely resemble the residuals in the general linear model. They are defined as

$$N_i = \Phi^{-1}[\hat{S}(t_i; \mathbf{z}_i)]$$

where $\hat{S}(t_i; \mathbf{z}_i)$ denotes the estimated survival function of the fitted model for individual i evaluated at his observed failure time and Φ is the standard normal cumulative distribution function. Defined in terms of probability integral transform, the suggested residuals can be regarded as *generalized residuals* in the sense of Cox–Snell [10]. Compared to the classical Cox–Snell residuals they offer two advantages: they can be intuitively interpreted as a ‘distance’ between predicted median time and observed survival time. Assuming a correctly specified model, their distribution converges in probability to a standard normal distribution, the symmetry of which is of help in any graphical procedure [11].

For censored observations normal-deviate residuals can be replaced by their conditional mean or median values [7]. However, this option results in an anomalous concentration of residuals so that their empirical distribution departs from normality even for perfectly fitting models. Therefore, we prefer to randomly generate m normal-deviate residuals from their conditional reference distribution. The idea is that, given $T_i > c_i$, $S(T_i)$ is uniformly distributed in $[0, S(c_i)]$, c_i being the observed censored time. Then each of the imputed residuals for individual i is weighted by $1/m$. In our experience, three or four imputations are enough to reconstruct the left tail of the residuals’ distribution, removing the confounding effect of censoring. As imputation is done under the null hypothesis of a correctly specified model, it may lead to a conservative behaviour in assessing departures from model assumptions when the percentage of censoring is high. The conservative effect of multiple imputation can be recognized because it mainly affects the left tail of the residuals’ distribution that corresponds to long survival times. There a possibly spurious agreement between reference and empirical distributions simply indicates that the empirical information is not enough to reject the fitted model in that area. In our experience good discrimination among parametric models requires the censoring percentage not to exceed 40–50 per cent. For higher censoring the goodness of fit of a parametric model can only be judged for relatively short failure times.

We prefer to compare empirical and reference distributions of residuals by means of histograms and kernel estimates rather than by cumulative (P–P or Q–Q) plots, despite some instability for small samples. Histograms give a more detailed description of the shape of the residuals’ empirical distribution and, in particular, of its tails.

3. COX AND PARAMETRIC MODELS IN FOUR SURVIVAL STUDIES

For the comparative analysis by Cox and parametric models, we start by defining a suitable predictor for Cox regression using the relevant main effects terms plus those terms that describe significant departures (interactions, non-linear effects, time-dependent effects) from the main effects model. Plots of martingale [11] and weighted Schoenfeld residuals [12, 13] were also used to judge the appropriateness of the linear predictor. Then exponential, Weibull, log-normal and log-logistic models were similarly fitted to each data set. We fitted Cox and parametric models by PHREG and LIFEREG procedures of SAS [14], respectively, and developed an S-plus macro for the analysis of residuals from parametric models.

As the scales of the parameters in Cox's model and in the parametric models differ, neither parameter estimates nor their estimated variances are suitable for comparisons. The efficiency of parameter estimates can be better compared by Wald-type χ^2 -values and by standardized measures of variability, sv ($sv = \hat{\sigma}_\beta / |\hat{\beta}|$), the latter analogous to the coefficient of variation. Although the two measures are related to each other, they offer a different, complementary perspective in the evaluation of relative efficiency. Neither measure is superior: sv tends to overemphasize cases where parameter estimates are close to zero while χ^2 -values intuitively exaggerate the gains in precision for highly significant effects.

For each study prognostic factors are given in descending order of corresponding Wald-type χ^2 -values from Cox regression.

3.1. A study of non-small cell lung cancer

The Yorkshire lung cancer study [8] consists of 272 patients who did not receive curative treatment for their condition. Survival is the endpoint of interest. The estimated median follow-up is 22.5 months and results in 17 per cent censoring. Table I gives results from Cox's model and from alternative parametric models. Neither interactions nor time dependencies nor non-linear effects of continuous factors were observed.

Fitting a Weibull model, a scale parameter estimate of 0.98 suggests the more parsimonious exponential model in agreement with the residual plot (Figure 1(a)), which indicates satisfactory fitting. On the other hand, the empirical distribution of residuals from the log-normal model departs from the reference normal density (Figure 1(b)), showing that the goodness of fit cannot be improved by assuming a peak in the baseline hazard function. A similar message comes from fitting a log-logistic model (not shown). Thus, the assumption of a constant baseline hazard appears to be justified both from the above analysis and the limited follow-up time.

In Table I the χ^2 - and sv -values of the prognostic factors can be compared across different models. The well-fitting exponential model and Cox regression produce similar values, with a tendency towards higher significances for the stronger factors in the exponential model. The results by the log-normal model are different, although the departure observed in Figure 1(b) is not substantial. Note, that in this case also the apparent relative importance of factors is not fully preserved. In conclusion, a simple exponential model here shows a satisfactory fit, provides more efficient estimates for parameters and the additional message of constant hazards.

The estimated acceleration factors ($= \exp(-\gamma)$) from the exponential and the log-normal regression models differ moderately as shown in Table I.

Table I. Analysis of the non-small cell lung cancer study.

Factors	Range and interpretation of values	Wald χ^2 (standardized variability)				Estimated accelerations	
		Cox	Exponential	Log-normal	Log-normal	Exponential	Log-normal
Activity score	0-4 (low is 'good')	17.55 (0.25)	18.08 (0.24)	24.19 (0.20)		1.40	1.51
Sex	0 = male, 1 = female (20 per cent)	14.47 (0.27)	15.96 (0.26)	16.86 (0.24)		0.49	0.46
Hoarseness	0 = absent, 1 = present (11 per cent)	10.43 (0.31)	11.17 (0.30)	6.83 (0.38)		1.99	1.88
Anorexia	0 = absent, 1 = present (45 per cent)	4.68 (0.45)	4.38 (0.47)	8.86 (0.33)		1.35	1.62
Metastases	0 = absent, 1 = present (10 per cent)	3.54 (0.54)	3.58 (0.52)	1.92 (0.71)		1.52	1.40
Age	38-84 years; median 67	1.51 (0.80)	1.10 (0.89)	0.91 (1.13)		1.00	1.00

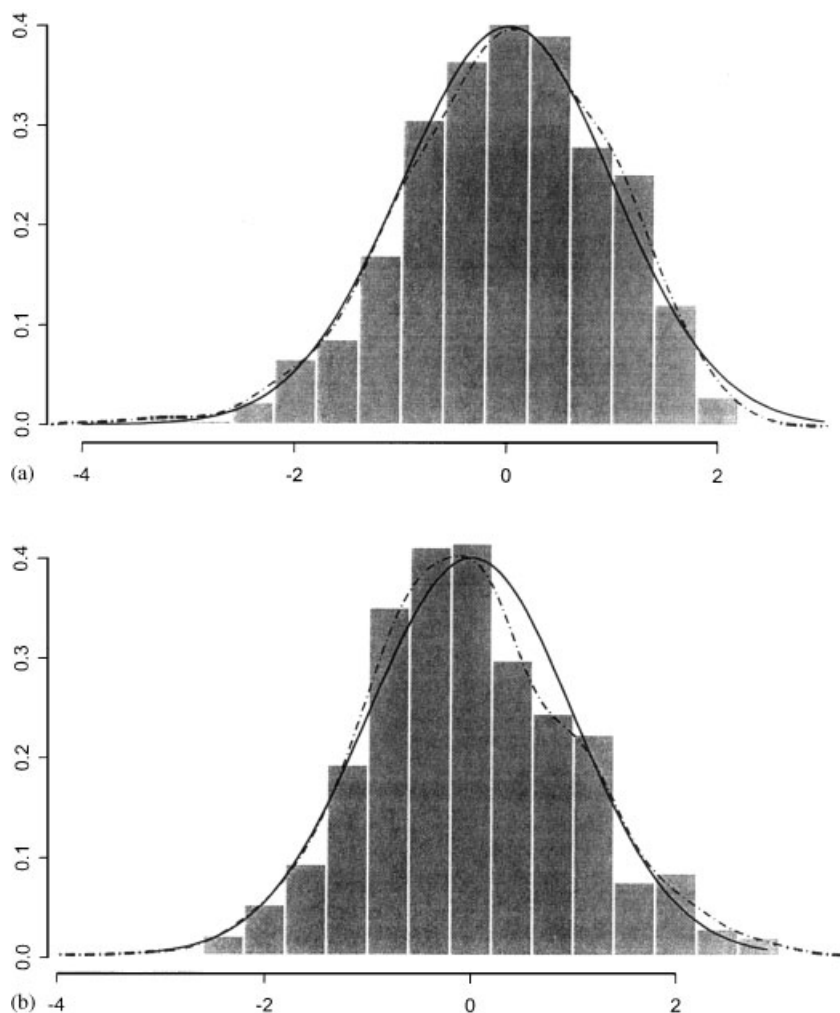


Figure 1. Empirical distributions of normal-deviate residuals from the exponential (a) and the log-normal (b) model for the Yorkshire lung cancer study. The reference standard normal density (solid line) and a kernel estimate (dashed line) are superimposed.

The current data set can be suitably fitted by both a Cox and an exponential model. Furthermore, follow-up does not depend on any of the covariates, which was checked by means of a Cox model with the meaning of the status indicator reversed [15]. Therefore, in order to study the differential effect of follow-up dependent on covariates, we artificially censored the survival times of all male patients at 6 months. As a consequence, overall censoring increased to 44 per cent (49 and 28 per cent censoring of males and females, respectively) and the χ^2 -values of both models tended to moderately decrease. For the factor 'sex', however, the χ^2 -value dropped from 14.47 to 6.50 for Cox's model but from 15.97 to only 13.84 for the exponential regression, which impressively demonstrates that substantial covariate-associated

Table II. Analysis of the heart transplant study.

Factors	Range and interpretation of values	Wald χ^2 (standardized variability)		
		Cox	Weibull	Log-normal
PAT	0 = no, 1 = yes (19 per cent)	8.89 (0.34)	10.28 (0.31)	13.29 (0.27)
CMV	0 = no, 1 = yes (5 per cent)	7.85 (0.36)	8.87 (0.34)	6.46 (0.39)
Adiposity	0 = no, 1 = yes (21 per cent)	5.00 (0.45)	6.37 (0.40)	5.80 (0.42)
log(WT)	0–7.4; median 4.6	4.59 (0.47)	3.81 (0.51)	8.27 (0.35)

follow-up reduces the efficiency of parameter estimates by Cox's model much more than that by parametric models.

3.2. Infections after heart transplantation

In an observational study of 526 patients who underwent heart transplantation at the Section of Heart and Thoracic Surgery of the Vienna University Surgery Department between 1990 and June 1999 and had full documentation of important prognostic factors, we analyzed time till infection (22 per cent censoring). The results of this analysis are shown in Table II. Four prognostic factors were identified: time on a waiting list before transplantation (WT), adiposity, presence of cytomegalovirus antibodies in the heart of the donor but recipient uninfected previously (CMV) and prophylactic antibody therapy (PAT). None of the parametric distributions fits well and we observe peaks in the empirical distributions of the residuals (Figure 2), which reflect an excess of early infections the models cannot describe. However, there is a clear improvement moving from the Weibull (max. $\log L = -1127$) to the log-normal (max. $\log L = -1072$) model, which is consistent with the medical experience of a non-monotone baseline hazard. A more satisfactory fit could only be obtained by a mixture model or by separate analyses of early and late infections. It is interesting to note that results from the Cox and log-normal models mainly differ for PAT and log(WT). Indeed, inspection of Schoenfeld residuals suggests a decaying effect for both factors. When modelling this effect by a log(time) interaction, both terms are significant (0.016 and 0.013, respectively). On the other hand, when introducing the corresponding time-by-covariate interactions using restricted cubic splines [16, 17] with three nodes, the tests for non-constant effects do not reach significance. This may be a consequence of over-parameterizing time dependencies. For Cox regression the most efficient and also flexible way to model time-dependency is still an open issue.

3.3. A randomised trial for node-positive breast cancer

Between December 1990 and January 1996, the Department of Surgery of the Vienna University Medical School recruited 2019 primary operable, non-metastatic, post-menopausal and receptor-positive breast cancer patients to participate in a randomized trial of Nolvadex versus Nolvadex+Orimeten. Patients are still under follow-up. For the current analysis a subset of 239 patients with at least four positive lymph nodes was selected and the endpoint of interest is time to distant metastases (65% censoring). For this rather homogeneous subgroup only the diameter of the original tumour and the estrogen receptor values (ER) are of prognostic

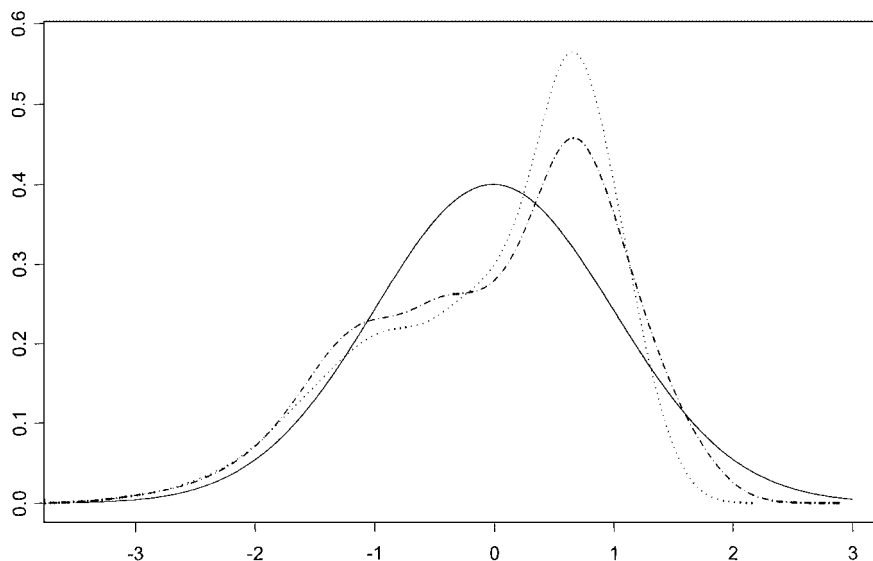


Figure 2. Heart transplant study: reference standard normal density (solid line) and kernel estimates for the empirical density of residuals from the Weibull (dotted line) and the log-normal models (dashed line).

Table III. Analysis of the node-positive breast cancer study.

Factors	Range and interpretation of values	Wald χ^2 (standardized variability)			
		Cox 1	Cox 2	Weibull	Log-normal
ER	≤ 70 versus > 70 fmol (41 per cent)	2.58 (0.62)	7.01 (0.38)	2.71 (0.61)	4.87 (0.45)
ER*log(time)	—	—	5.77 (0.42)	—	—
Diameter	≤ 2 (25 per cent) versus 2–5 (63 per cent)	0.05 (4.41)	0.16 (1.92)	0.06 (4.15)	0.23 (2.07)
	≤ 2 versus > 5 (12 per cent)	0.11 (0.51)	3.18 (0.56)	3.86 (0.51)	1.65 (0.78)

interest, the latter also showing a time-dependent effect. Table III gives the results from Cox and alternative parametric models.

If the time-dependent effect of ER is ignored in Cox's model (see Cox 1 in Table III), the constant effect of ER loses its significance. The suitability of the $\log(t)$ shape of time dependency was checked by both a smoothed plot of weighted Schoenfeld residuals versus time and by introducing a time-by-covariate interaction in the predictor. A residual plot similar to Figure 1(a) shows a good fit to the observed data by the log-normal model, which outperformed the Weibull model. Time dependence is automatically catered for in the log-normal model in which the (constant) effect of ER is significant. It is interesting to note that the p -value of ER from the log-normal model ($p = 0.0272$) is very close to the joint test of significance of

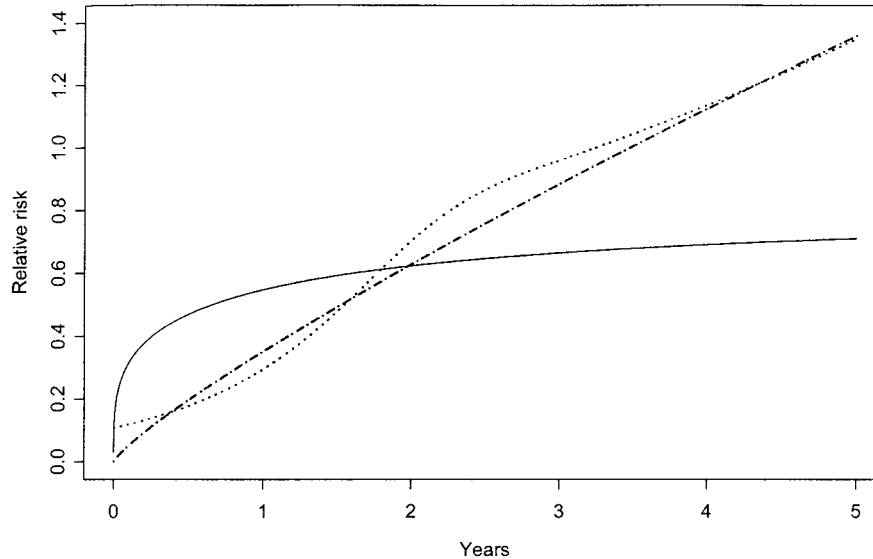


Figure 3. Node-positive breast cancer study: estimated relative risk of the oestrogen receptor over time from the log-normal model (solid line) and from the Cox model with time dependence modelled by $\log(t)$ (dashed line) and by a restricted cubic spline (dotted line).

ER and the time-dependent term from Cox's model ($p=0.0265$). This relative efficiency of parameter estimates from Cox's model is quite good in this example.

Do the appropriate log-normal and Cox models lead to different conclusions with respect to the effect of ER? Both models suggest that the favourable effect of higher ER is decreasing in time (Figure 3). The shapes of the estimated relative risks are different: the relative risk for Cox's model (dashed line) is almost linearly increasing, leading to an unfavourable effect for long survival time. According to the log-normal model (continuous line) the relative risk becomes constant after about 2 years, being still lower than one.

Finally, we remark that the estimated relative risk from Cox's model with a restricted cubic spline for the time-dependent ER effect (dotted line) is in agreement with the results based on the corresponding specification by $\log(t)$.

3.4. An observational study of liver transplants

Between 1982 and 1999, 217 patients with a primary diagnosis of liver cancer were transplanted at the Department of Surgery of the Vienna University Medical School. Of these patients, 178 with complete information on prognostic factors were selected. We analyzed survival as an endpoint and observed 30 per cent censored times. Results by Cox's model, the Weibull and the log-normal models are given in Table IV.

As shown in Figure 4, we observe an excess of residuals close to zero. This may be an indication of local overfitting to a subset of the data. Note that at the same time the correlations among the parameters for tumour stage, tumour diameter and their interaction become larger than 0.94, indicating overlapping effects, which are not easy to accommodate. If the second interaction term is omitted, the excess of residuals close to zero is reduced.

Table IV. Analysis of the liver transplant study.

Factors	Range and interpretation of values	Wald χ^2 (standardized variability)			Parameter estimates	
		Cox	Weibull	Log-normal	Weibull	Log-normal
Lymph nodes (LN)	0 = no, 1 = yes (14 per cent)	12.96 (0.28)	17.89 (0.24)	8.06 (0.35)	-3.46	-3.11
Tumour diameter (TD)	85-197; median 174	8.18 (0.35)	9.71 (0.32)	6.98 (0.38)	-0.11	-0.12
Tumour stage (TS)	0 = T1, T2; 1 = T3, T4 (82 per cent)	6.46 (0.39)	7.57 (0.36)	5.26 (0.44)	-18.49	-20
Interaction LN*TS	—	7.64 (0.36)	10.38 (0.31)	3.05 (0.57)	2.85	2.08
Interaction TS*TD	—	5.53 (0.43)	6.65 (0.39)	4.47 (0.47)	0.10	0.11

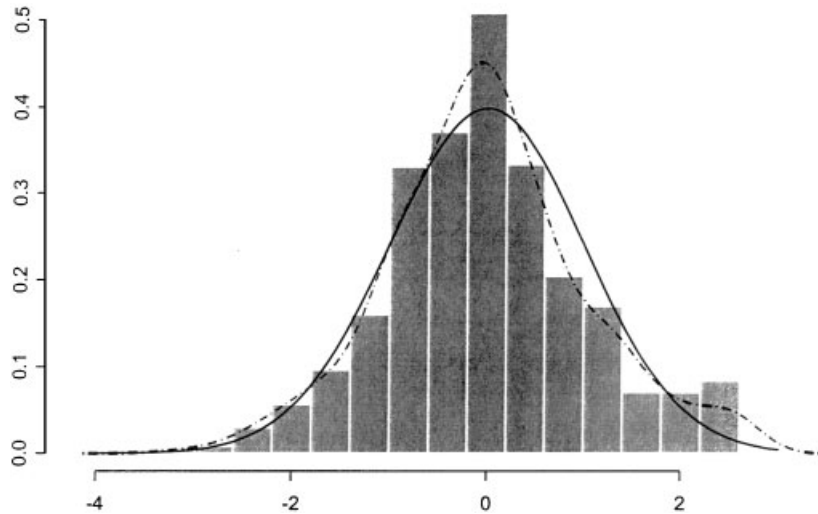


Figure 4. Empirical distribution of normal-deviate residuals from the Weibull model for the liver transplant study. The reference standard normal density (solid line) and a kernel estimate (dashed line) are superimposed.

Because of approximate proportionality of hazards, the fit for the log-normal model is even less satisfactory. From Table IV we learn that there is a clear gain in using the Weibull model over Cox regression in terms of χ^2 - and sv-values for all parameters. The log-normal model, on the other hand, performs more poorly than Cox regression. We notice again that a non-satisfactory fit leads to lower efficiency in parameter testing. As shown in Table IV, log-normal and Weibull models lead to moderately differing parameter estimates which affect the estimation of survival functions. These, however, differ substantially between the Weibull and the less-appropriate log-normal models as shown in Figure 5.

4. CONCLUDING REMARKS

Currently, parametric models are not often used in the analysis of medical studies of survival. Applied statisticians seem to be concerned about the required assumptions on the baseline distribution and the assumed effort in arriving at an appropriate model. In this respect normal deviate residuals might play an important role. Although originally developed for Cox regression [7], they have proved very useful to judge and compare the goodness of fit of alternative parametric models. These residuals can be intuitively interpreted as a ‘distance’ between predicted median time and observed survival time. Departures of the distributions of the residuals from their normal reference distribution indicate that a particular model does not fit well. While we have seen a nearly perfect fit for study 1 (Figure 1(a)), none of the parametric models considered could fit the excess of early infections in study 2 resulting in peaks of positive residuals in Figure 2. We have observed a similar shape of the residuals’ distribution in other studies also, where both early events and long survival times occurred

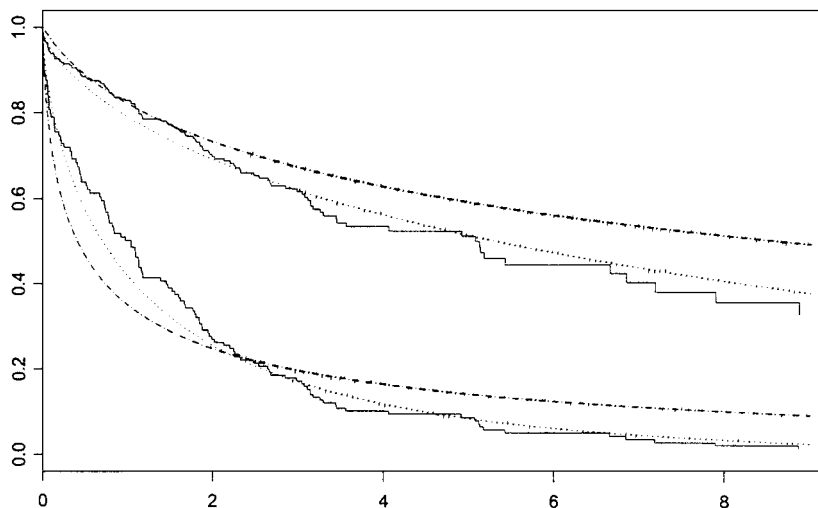


Figure 5. Estimated survival functions from Cox's model (solid line), from the Weibull (dotted line) and the log-normal model (dashed line) for low-risk patients (no lymph nodes, tumour diameter 90 and tumour stage $T1/T2$) and for high-risk patients (lymph nodes, tumour diameter 190 and tumour stage $T1/T2$).

which the model cannot describe simultaneously. In Figure 4, we have demonstrated the likely result of local overfitting, an anomalous concentration of residuals around zero.

Sometimes we observed a clear superiority of one parametric model that showed a very good fitting to observed data. It also happened that we did not get a satisfactory fit by any of the parametric models considered, in which case one could move to a richer model for the baseline hazard or simply use Cox's model. Still, the different perspective from parametric models reveals features of a data set, which remain undiscovered if only Cox models are fitted.

A major goal of this paper is to investigate the comparative performance of Cox and parametric survival models under the typical conditions of clinical studies. This can be studied by small sample simulations or by means of real data sets. In this paper, we have taken the latter approach, which may be more subjective by the choice of data sets but also more realistic with respect to the typical conditions of such studies. The sample sizes of the studies ranged between 217 and 526, the numbers of estimated parameters between three and six and censoring proportions between 17 and 65 per cent.

Cox and Oakes [18] and Oakes [5] concluded that, asymptotically, well-fitting parametric models should yield more efficient parameter estimates than Cox regression if either:

- (i) parameter values are far from zero,
- (ii) follow-up depends on covariates,
- (iii) there is a strong time trend in covariates.

They further concluded that for finite samples the loss in precision from using Cox regression is rather greater.

In study 1 none of these conditions is fulfilled and, indeed, the well-fitting exponential model does not lead to a real gain in precision as shown by the standardized variability in Table I. However, if follow-up is made dependent on a covariate, the gain becomes substantial in agreement with item (ii). In study 3, where some parameter estimates are far from 0, the Weibull model produces lower standardized variability (Table III) than Cox's model (item (i)). If time-dependent effects are present in a data set and are not appropriately taken into account in Cox regression, study 3 confirms item (iii). If such a time dependence is appropriately modelled in Cox regression, it does not appear to reduce the efficiency.

It is interesting to compare parameter estimates across different parametric models. For large samples and small effects, it has been proved [19] that the ratios of regression coefficients from a misfitted model approximate the ratios from the proper model. This result does not continue to hold for our studies where even the ratios of regression parameters were sensitive to the chosen baseline hazard.

An essential property of a model is the interpretability of its parameters. Clinicians are familiar with the relative risk interpretation of parameters in Cox's model. In the considered parametric models, parameters can be interpreted as factors accelerating or decelerating a patient's biological clock. This interpretation too is intuitive although it may require a clinician to get used to it.

ACKNOWLEDGEMENTS

The authors are grateful to R. Henderson, Lancaster, for helpful comments on an earlier version of this paper and to G. Laufer, Section of Heart and Thoracic Surgery, R. Jakesz, Section of General Surgery, and F. Muehlbacher, Section of Transplant Surgery, all of Vienna University's Surgery Department, for permission to use the heart transplant, the breast cancer and the liver transplant data sets, respectively.

REFERENCES

1. Cox DR. Regression models and life-tables. (with discussion). *Journal of the Royal Statistical Society B* 1972; **34**:187–220.
2. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980.
3. Lawless JF. Parametric models in survival analysis. In *Encyclopaedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: New York, 1998; 3254–3264.
4. Efron B. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* 1977; **72**:557–565.
5. Oakes D. The asymptotic information in censored survival data. *Biometrika* 1977; **64**:441–448.
6. Gelber RD, Goldhirsch A, Cole BF. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. *Controlled Clinical Trials* 1993; **14**:485–499.
7. Nardi A, Schemper M. New residuals for Cox regression and their application to outlier screening. *Biometrics* 1999; **55**:523–529.
8. Muers MF, Shevlin P, Brown J. Prognosis in lung cancer: physicians' opinions compared with outcome and a predictive model. *Thorax* 1996; **51**:894–902.
9. Therneau TM, Grambsch PM. *Modeling Survival Data. Extending the Cox Model*. Springer: New York, 2000.
10. Cox DR, Snell EJ. A general definition of residuals (with discussion). *Journal of the Royal Statistical Society B* 1968; **30**:248–275.
11. Grambsch PM, Therneau TM, Fleming TR. Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* 1995; **51**:1469–1482.
12. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982; **69**:239–241.
13. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**:515–526.
14. SAS. *SAS/STAT User's Guide*, Version 8. SAS Institute Inc.: Cary, NC, USA.
15. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. *Controlled Clinical Trials* 1996; **17**:343–346.

16. Hess KR. Assessing time-by-covariate interactions in proportional hazard regression models using cubic spline functions. *Statistics in Medicine* 1994; **13**:1045–1062.
17. Heinzl H, Kaider A. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine* 1997; **54**:201–208.
18. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman & Hall: London, 1984.
19. Cox DR. Tests of separate families of hypotheses. *Proceedings of the 4th Berkeley Symposium*, Vol. 1. University of California Press: Berkeley, U.S.A, 1961; 105–123.