STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:1713–1723
Published online 22 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2059

Generating survival times to simulate Cox proportional hazards models

Ralf Bender[1,*,†,‡], Thomas Augustin[2] and Maria Blettner[3]

[1] *Institute for Quality and Efficiency in Health Care, Cologne, Germany*
[2] *Department of Statistics, University of Munich, Germany*
[3] *Institute for Medical Biometry, Epidemiology and Informatics, Johannes-Gutenberg-University, Mainz, Germany*

SUMMARY

Simulation studies present an important statistical tool to investigate the performance, properties and adequacy of statistical models in pre-specified situations. One of the most important statistical models in medical research is the proportional hazards model of Cox. In this paper, techniques to generate survival times for simulation studies regarding Cox proportional hazards models are presented. A general formula describing the relation between the hazard and the corresponding survival time of the Cox model is derived, which is useful in simulation studies. It is shown how the exponential, the Weibull and the Gompertz distribution can be applied to generate appropriate survival times for simulation studies. Additionally, the general relation between hazard and survival time can be used to develop own distributions for special situations and to handle flexibly parameterized proportional hazards models. The use of distributions other than the exponential distribution is indispensable to investigate the characteristics of the Cox proportional hazards model, especially in non-standard situations, where the partial likelihood depends on the baseline hazard. A simulation study investigating the effect of measurement errors in the German Uranium Miners Cohort Study is considered to illustrate the proposed simulation techniques and to emphasize the importance of a careful modelling of the baseline hazard in Cox models. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Cox model; proportional hazards model; exponential distribution; Gompertz distribution; simulation; survival times; Weibull distribution

1. INTRODUCTION

One of the most important statistical models in medical research is the proportional hazards model of Cox [1]. The Cox model is intensively investigated by means of simulation studies

[*] Correspondence to: Ralf Bender, Institute for Quality and Efficiency in Health Care, Dillenburger Str. 27, D-51105 Cologne, Germany.
[†] E-mail: Ralf.Bender@iqwig.de
[‡] Statistician.

Contract/grant sponsor: German Research Foundation (DFG); contract/grant number: BE 2485/1-1
Contract/grant sponsor: German Research Foundation (DFG); contract/grant number: SFB 386

Received July 2004
Accepted August 2004

to obtain information about bias and efficiency of the estimated regression coefficients for a variety of situations, in particular when fundamental model assumptions are violated. For example, Hu *et al.* [2] compared several approaches to estimate the parameters of a Cox model when covariates are measured with error. They performed a number of simulations with exponentially distributed survival times. While the exponential distribution is widely used for the generation of survival times in simulation studies, other distributions seem to be underutilized. For instance, up to now the methods to correct for measurement errors proposed by Wang *et al.* [3], Buzas [4], Hu *et al.* [2], Kong *et al.* [5], Kong [6], Kong and Gu [7], and Huang and Wang [8] have only been considered in the case of the exponential distribution.

It is common to model survival times through the hazard function. The Cox proportional hazards model is given by

$$h(t \mid x) = h_0(t)\exp(\beta' x) \tag{1}$$

where t is the time, x the vector of covariates, β the vector of regression coefficients and $h_0(t)$ is the so-called baseline hazard function, i.e. the hazard function for $x = 0$. As model (1) is formulated through the hazard function, the simulation of appropriate survival times for this model is not straightforward. One important issue in simulation studies regarding regression models is the knowledge of the true regression coefficients. This does not present a problem in a linear regression model, where the simulated variables are directly connected with the pre-specified regression coefficients. However, in the Cox model, the effect of the covariates have to be translated from the hazards to the survival times, because the usual software packages for Cox models require the individual survival time data, not the hazard function. The translation of the regression coefficients from hazard to survival time is easy if the baseline hazard function is constant, i.e. the survival times are exponentially distributed. This may be the reason why most simulation studies regarding the Cox model consider only the exponential distribution.

Another frequently used distribution for survival times is the Weibull distribution [9]. In simulation studies, a common practice is to consider only binary covariates such as group 1 and group 2. For example, Schemper [10] simulated Weibull distributed survival times for the situation of two binary covariates in order to compare strategies for the analysis with the Cox model in the presence of non-proportional hazards. In the case of discrete covariates, the Weibull distributions can be specified with different sets of parameters for each group. The Weibull parameters can be chosen such that the hazards are proportional and the true hazard ratio (HR) for the comparison of the two groups can be calculated from the Weibull parameters. Then, the true regression coefficient for the Cox model can be obtained from log(HR).

Considering only the exponential and/or the Weibull distribution may be sufficient for some applications. However, for a realistic description of various survival time data, other distributions are required. One important field in medicine is the modelling of human mortality for which frequently the Gompertz distribution is used. Other commonly used distributions in survival time analysis are the gamma, the lognormal and the log-logistic distribution [9]. The latter distributions, however, do not have the proportional hazards property. Among the known parametric distributions, only the exponential, the Weibull and the Gompertz model share the assumption of proportional hazards with the Cox regression model [9].

A number of special distributions have been used in applications of the Cox model. In a study on problems with vaguely defined disease states Liestøl and Andersen [11] use a discrete approximation to the Cox model with a Gompertz–Makeham-type baseline hazard rate. Valenta and Weissfeld [12] explicitly simulate a piecewise-exponential model to evaluate Gray's piecewise constant time-varying coefficients model. Ng and McLachlan [13] simulate a mixture of proportional hazards models where one baseline hazard rate is bathtub-shaped. Sophisticated algorithms are also needed to generate survival times conditional on time-dependent covariates, in particular when certain additional restrictions have to be satisfied. Mackenzie and Abrahamowicz [14] present a permutational algorithm, which allows to control simultaneously the HR, the marginal distribution of failure and censoring times as well as the marginal distribution of the covariate values. This algorithm is applied in a detailed simulation study on the use of the Cox model for matched case–control studies with time-dependent covariates [15] and in investigating a method to estimate the lag duration between exposure and change of risk [16].

However, simulations based on such more complex models for the baseline hazard rates are quite rare, and even in these studies authors typically do not vary the baseline hazard. The warning that 'inference is very sensitive to the choice of $H_0(t)$' [17] seems to be widely ignored. Apparently, common opinion seems to be that the baseline hazard rate does not influence the properties of the estimators, but not much evidence is provided in favour of that conjecture. For one exception see Petersen *et al.* [18], who found not much difference between a Weibull and the exponential distribution in comparing several case–cohort estimators.

The implicitly assumed robustness of the partial likelihood estimator with respect to the underlying distribution definitely becomes questionable when non-standard situations are considered. For instance, in the case of covariate measurement errors, the induced hazard rates are usually no longer proportional to each other [19], and the induced relative risk depends on the underlying baseline hazard rate [20, 21]. This effect may be negligible under the so-called rare disease assumption [19, 20, 22]. However, since the Cox model is well-known to be quite sensitive to violations of the proportional hazards assumption, the form of the baseline hazard rate may substantially influence the properties of the estimator.

This paper provides an elegant basis for such more comprehensive simulation studies. It shows how survival times can be generated to simulate Cox models with known regression coefficients and with *any* non-zero baseline hazard rate. From this general approach the required relation for the exponential, the Weibull and the Gompertz distribution can be derived as special cases. In addition to, and far beyond this, the general relation between hazard and survival time obtained can be used to develop own distributions perfectly adapted to concrete situations. Only such techniques allow to study the behaviour of proportional hazards models in situations of complex dynamics and with flexibly parameterized baseline hazard functions. A simulation study based upon the data of the German Uranium Miners Cohort Study [23] is considered to illustrate the proposed simulation techniques and to show their relevance in simulating Cox models.

2. SIMULATING SURVIVAL TIMES

2.1. General considerations

The survival function of the Cox proportional hazards model (1) is given by

$$S(t \mid x) = \exp[-H_0(t)\exp(\beta' x)] \tag{2}$$

where

$$H_0(t) = \int_0^t h_0(u)\,\mathrm{d}u \tag{3}$$

is the cumulative baseline hazard function [24]. Thus, the distribution function of the Cox model is

$$F(t \mid x) = 1 - \exp[-H_0(t)\exp(\beta' x)] \tag{4}$$

Let Y be a random variable with distribution function F, then $U = F(Y)$ follows a uniform distribution on the interval from 0 to 1 [25], abbreviated as $U \sim \mathrm{U}[0,1]$. Moreover, if $U \sim \mathrm{U}[0,1]$, then $(1 - U) \sim \mathrm{U}[0,1]$, too [25]. Thus, let T be the survival time of the Cox model (1), then it follows from (4) that

$$U = \exp[-H_0(T)\exp(\beta' x)] \sim \mathrm{U}[0,1] \tag{5}$$

If $h_0(t) > 0$ for all t, then H_0 can be inverted and the survival time T of the Cox model (1) can be expressed as

$$T = H_0^{-1}[-\log(U)\exp(-\beta' x)] \tag{6}$$

where U is a random variable with $U \sim \mathrm{U}[0,1]$. Random numbers following a U[0,1] distribution are frequently available in statistical program packages. By applying formula (6), uniformly distributed random numbers can be transformed into survival times following a specific Cox model. It is just required to insert the inverse of an appropriate cumulative baseline hazard function into equation (6).

2.2. Application of common survival time distributions

Among the commonly used survival time distributions, only the exponential, the Weibull and the Gompertz distribution share the assumption of proportional hazards with the Cox model [9]. The main characteristics of these three distributions including the inverse cumulative hazard functions are summarized in Table I.

By using formula (6) survival time data for Cox models with exponentially (Cox–exponential model), Weibull (Cox–Weibull model), and Gompertz (Cox–Gompertz model) distributed survival times can be generated. The corresponding formulas are given in Table II. In all three models, the distribution type used for the baseline hazard is also the distribution type for the survival times, but with parameters dependent on the covariates x (see Table II).

2.3. Proportional hazards models with other distributions

While up to now techniques to run simulations based on standard parametric distributions have been reported, the result in (6) is also of great importance in the whole generality described there. Firstly, it allows to design comprehensive simulation studies for all the variants of the Cox model where the baseline hazard rate is modelled in a flexible parametric way. Then, instead of (1), one considers

$$h(t \mid x) = g(a,t)\exp(\beta' x) \tag{7}$$

Table I. Characteristics of the exponential, the Weibull and the Gompertz distribution.

	Distribution		
Characteristic	Exponential	Weibull	Gompertz
Parameter	Scale parameter $\lambda > 0$	Scale parameter $\lambda > 0$ Shape parameter $\nu > 0$	Scale parameter $\lambda > 0$ Shape parameter $\alpha \in (-\infty, \infty)$
Range	$[0, \infty)$	$[0, \infty)$	$[0, \infty)$
Hazard function	$h_0(t) = \lambda$	$h_0(t) = \lambda \nu t^{\nu - 1}$	$h_0(t) = \exp(\alpha t)$
Cumulative hazard function	$H_0(t) = \lambda t$	$H_0(t) = \lambda t^{\nu}$	$H_0(t) = \frac{\lambda}{\alpha}(\exp(\alpha t) - 1)$
Inverse cumulative hazard function	$H_0^{-1}(t) = \lambda^{-1} t$	$H_0^{-1}(t) = (\lambda^{-1} t)^{1/\nu}$	$H_0^{-1}(t) = \frac{1}{\alpha} \log(\frac{\alpha}{\lambda} t + 1)$
Density function	$f_0(t) = \lambda \exp(-\lambda t)$	$f_0(t) = \lambda \nu t^{\nu - 1} \exp(-\lambda t^{\nu})$	$f_0(t) =$ $\lambda \exp(\alpha t) \exp(\frac{\lambda}{\alpha}(1 - \exp(\alpha t)))$
Survival function	$S_0(t) = \exp(-\lambda t)$	$S_0(t) = \exp(-\lambda t^{\nu})$	$S_0(t) = \exp(\frac{\lambda}{\alpha}(1 - \exp(\alpha t)))$
Mean	$E(T) = \frac{1}{\lambda}$	$E(T) = \frac{1}{\sqrt[\nu]{\lambda}} \Gamma(\frac{1}{\nu} + 1)$ Γ denotes the gamma function	$E(T) = \frac{1}{\lambda} G\left(\frac{\lambda}{\alpha}\right)$ where $G(x) = \int_x^{\infty} \frac{1}{y} \exp(-y) \, \mathrm{d}y$
Variance	$\mathrm{Var}(T) = \frac{1}{\lambda^2}$	$\mathrm{Var}(T) =$ $\frac{1}{\sqrt[\nu]{\lambda^2}} [\Gamma(\frac{2}{\nu} + 1) - \Gamma^2(\frac{1}{\nu} + 1)]$	

Table II. Formulas for the survival time and the hazard function of Cox models using the exponential, the Weibull and the Gompertz distribution.

	Model		
Characteristic	Cox–exponential	Cox–Weibull	Cox–Gompertz
Survival time	$T = -\frac{\log(U)}{\lambda \exp(\beta' x)}$	$T = (-\frac{\log(U)}{\lambda \exp(\beta' x)})^{1/\nu}$	$T = \frac{1}{\alpha} \log[1 - \frac{\alpha \log(U)}{\lambda \exp(\beta' x)}]$
Hazard function	$h(t \mid x) = \lambda \exp(\beta' x)$	$h(t \mid x) = \lambda \exp(\beta' x) \nu t^{\nu - 1}$	$h(t) = \lambda \exp(\beta' x) \exp(\alpha t)$

U is a variable following a uniform distribution on the interval from 0 to 1.

where $g(\cdot)$ is a function known up to a multidimensional parameter a. Model (7) contains variants of the Cox model where the baseline hazard rate is determined by a finite number of parameters (see, in particular, References [26–31]). For the handling of these models in the presence of measurement errors, see Reference [32].

Secondly, result (6) enables the investigation of the behaviour and stability of estimators in the Cox model under certain additional assumptions on the baseline hazard rate. For instance, the function

$$h_0(t) = |a \sin(t)| \tag{8}$$

can be used to model regularly recurring periods of high baseline hazard of magnitude $a > 0$. Integration leads to

$$H_0(t) = a\left(2\left[\!\left[\frac{t}{\pi}\right]\!\right] + 1 + (-1)^{[\![t/\pi]\!]+1}\cos(t)\right) \tag{9}$$

where $[\![y]\!]$ is the truncation function, which returns the largest integer less or equal to y. Inverting $H_0(t)$ finally yields

$$H_0^{-1}(t) = \left[\!\left[\frac{t}{4a}\right]\!\right]\pi + \left[\!\left[0.5 + \frac{t}{4a}\right]\!\right]\pi + \arccos\left[-\left(t - 2\left[\!\left[\frac{t}{2a}\right]\!\right] - 1\right)\right] \tag{10}$$

which provides the basis for designing and performing simulations in this model.

For practical applications notice also that model (7) can be used for (almost) arbitrary functions. Neither $H_0(t)$ nor its inverse function $H_0^{-1}(t)$ are needed in analytical form. So, in principle, it is sufficient to determine both numerically, i.e. by numerical integration and inversion, respectively.

3. EXAMPLE

3.1. Simulations for the German Uranium Miners Cohort Study

The German Uranium Miners Cohort Study is one of the largest cohort studies on uranium miners with the purpose of evaluating the risks of cancer and mortality associated with low and high levels of radon exposure [23]. The cohort includes about 60 000 workers of the former Wismut uranium company in Eastern Germany, who have been exposed to different levels of radiation dependent on job, place of work, and time. Exposure to radon and its progeny was assessed by using a detailed job-exposure matrix (JEM) leading to cumulative radon exposures expressed in the so-called working level months (WLM) [23]. As in the JEM for each job a summary measure such as the annual mean is used as exposure value rather than individual exposure values, measurement errors of the Berkson type occur [33]. To investigate the effect of measurement error in the exposure values on HRs estimated by means of Cox proportional hazards models, a simulation study has been performed [34]. As the generated survival times in the simulation study should have a similar distribution like the observed survival times in the cohort study, the Gompertz distribution was applied. This was necessary, because it was impossible to generate realistic survival times by means of the exponential distribution: either the number of deaths or the attained age were too high in the simulated data. Realistic survival times reflecting the mortality of the German uranium miners can be generated by means of the Cox–Gompertz model

$$h(t \mid x) = \lambda \exp(\alpha t) \exp(\beta_{\text{age}} \times \text{AGE} + \beta_{\text{radon}} \times \text{RADON}) \tag{11}$$

which contains the vector x of the covariates age at baseline (AGE) and radon exposure (RADON). Applying the formula presented in Table II, the corresponding survival time T of model (11) is constructed from a uniformly distributed variable U by

$$T = \frac{1}{\alpha} \log \left[1 - \frac{\alpha \log(U)}{\lambda \exp(\beta_{\text{age}} \times \text{AGE} + \beta_{\text{radon}} \times \text{RADON})} \right] \tag{12}$$

One task in generating survival times with specific features is to find appropriate parameters for the model considered. Here, an obvious approach is to relate the expected value of the survival time (12) to the tables for life expectancy of German men. Unfortunately, the expected value of a Gompertz distributed random variable T is given by a formula containing an integral which has to be evaluated numerically (Table I) [35]. Thus, appropriate parameter values of the Cox–Gompertz model cannot be calculated directly. However, as the Gompertz distribution represents a left truncated extreme value distribution at time point $t=0$ [35], the extreme value distribution can be used as approximation to the Gompertz distribution. The extreme value distribution is defined for $-\infty < t < \infty$, but the hazard function for $t \geqslant 0$ is identical to that of the Gompertz distribution. The density and survival function of the extreme value distribution with parameters λ and α are given by [36]

$$f_0(t) = \lambda \exp(\alpha t) \exp\left(-\frac{\lambda}{\alpha} \exp(\alpha t) \right) \tag{13}$$

$$S_0(t) = \exp\left(-\frac{\lambda}{\alpha} \exp(\alpha t) \right) \tag{14}$$

The mean and variance of an extreme value distributed variable T are given by [36]

$$E(T) = \mu_0 = -\frac{1}{\alpha} \left(\log\left(\frac{\lambda}{\alpha} \right) + \gamma \right) \tag{15}$$

$$\text{Var}(T) = \sigma_0^2 = \frac{\pi^2}{6\alpha^2} \tag{16}$$

where $\gamma \approx 0.5772$ is Euler's constant and $\pi \approx 3.14159$. Solving (15) and (16) for λ and α leads to

$$\alpha = \frac{\pi}{\sqrt{6}\sigma_0}, \quad \lambda = \alpha \exp(-\gamma - \alpha\mu_0) \tag{17}$$

which can be used to calculate approximately the parameters of the Gompertz distribution in dependence on the mean and variance of the considered survival time. For the mean life expectancy of $\mu_0 = 66.86$ years and a standard deviation of $\sigma_0 = 6$ years, the values $\lambda = 7 \times 10^{-8}$ and $\alpha = 0.2138$ are obtained. By using the regression coefficients $\beta_{\text{age}} = 0.15$ for AGE and $\beta_{\text{radon}} = 0.001$ for RADON in the Cox–Gompertz model (11), survival times leading to realistic attained age values and numbers of deaths similar to those observed in the German Uranium Miners Cohort Study can be generated [34].

3.2. Comparison between the exponential and the Gompertz distribution

To assess the importance of using a realistic survival distribution in simulation studies, the simulation results of the Cox–Gompertz model (11) are compared with those of the corresponding Cox model with exponentially distributed survival times in the situation of the German Uranium Miners Cohort Study [23].

The goal of the simulation study was to investigate the effect of measurement errors in the radon exposure values on the estimated HRs [34]. In radon epidemiology, both classical measurement errors and Berkson-type errors may play a role [37]. Due to the skew distribution of radon data the application of usual additive measurement error models assuming normally distributed errors may be problematic. In this case multiplicative measurement error models assuming log-normally distributed errors are frequently used [38]. For demonstrating purposes, in this paper additive as well as multiplicative measurement error models for the Berkson and the classical measurement error type are considered, resulting in four different measurement error models. The data situation of the German Uranium Miners Cohort Study is used with the following characteristics: sample size $n = 58\,721$, total study time 1946–1998, mean (SD) age at study entry 24.3 (8.38) years, mean (SD) cumulative radon exposure 266.84 (507.82) WLM [23]. Cox–Gompertz models (20) as well as the corresponding model using exponentially distributed survival times were simulated. An extension would be to use radon as time-varying covariate and to consider potential confounders such as smoking. However, the focus here is to show that the results of simulation studies may depend on the choice of the distribution of the generated survival times rather than to develop the best measurement error model in a complex situation. To keep the example simple, age and radon exposure are considered here as fixed covariates (age at baseline, total cumulative radon exposure). Only a small part of the simulation study concerning the German Uranium Miners Cohort Study is presented here for illustrating purposes. The results of additive and multiplicative Berkson-type measurement errors and additive and multiplicative classical measurement errors for radon exposure are reported for one parameter situation.

For the additive models normally distributed measurement errors with mean $\mu_e = 0$ and standard deviation $\sigma_e = 359.1$ were generated. For the multiplicative models log-normally distributed errors with parameters $\mu_e = -0.2029$ and $\sigma_e = 0.637$ were generated such that the expected value of the errors amounts to $\exp(\mu_e + \sigma_e^2/2) = 1$. The values for σ_e are chosen so that in all cases the total radon exposure variance amounts to 150 per cent of the exposure variance without measurement error. For each situation 1000 simulations were performed. In Table III the relative bias of the estimated Cox regression coefficients and the difference of the relative bias between the exponential and the Gompertz distribution is shown.

In all cases, measurement errors leads to an attenuation of the true effect for both covariates, shown by the negative relative bias values. In most cases the bias values are quite similar for both distributions. Hence, in these situations, the conclusions for both distributions would be the same, although the exponentially distributed survival times did not fit the observed survival times of the German Uranium Miners Cohort Study. However, in two situations the bias difference between the exponential and the Gompertz distribution is larger than 5 per cent. Firstly, in the case of additive classical measurement errors, the relative bias for the estimated exposure effect is much higher in the simulated Cox–exponential model (−41.49 per cent) in comparison to the Cox–Gompertz model (−25.73 per cent). Secondly, in the multiplicative classical measurement error model, the relative bias for the estimated age effect is higher

Table III. Relative bias (in per cent) of estimated Cox regression coefficients due to measurement error in exposure in different measurement error situations by using the exponential and the Gompertz distribution.

Measurement error situation	Covariate	Distribution: Exponential	Distribution: Gompertz	Bias difference (Exponential − Gompertz)
Additive Berkson errors	Radon exposure	−6.81	−6.74	−0.07
	Age	−6.54	−5.78	−0.76
Multiplicative Berkson errors	Radon exposure	−26.94	−26.19	−0.75
	Age	−4.46	−8.24	3.78
Additive classical errors	Radon exposure	−41.49	−25.73	−15.76
	Age	−2.99	−5.31	2.33
Multiplicative classical errors	Radon exposure	−61.95	−66.15	4.20
	Age	−3.24	−10.28	7.04

in the Cox–Gompertz model (−10.28 per cent) than in the Cox–exponential model (−3.24 per cent). Hence, the use of the exponential distribution would lead to incorrect conclusions about the amount of attenuation due to measurement errors. In the case of additive classical measurement errors mainly the exposure itself is affected, whereas multiplicative classical measurement errors lead to different bias values mainly for the considered covariate without error. In summary, the choice of the survival time distribution in simulation studies concerning the Cox models has quite complex effects; it influences results in various directions and concerning different variables.

4. CONCLUSION

The high capacity of performing calculations by means of modern computers allows the evaluation of statistical methods via simulation studies. One of the most important statistical models in medical research is the Cox proportional hazards model, which is intensively investigated by means of simulation studies. While the exponential distribution is widely used for the generation of survival times in simulation studies, other distributions seem to be underutilized. One reason for neglecting survival distributions beyond the exponential distribution may be that the generation of survival times in dependence on pre-specified Cox regression coefficients is not obvious. In this paper, the general relation between the hazard and the survival time of the Cox model is developed, which can be used to generate survival times following any distribution compatible with proportional hazards. Examples are given by the well-known exponential, Weibull and Gompertz distribution. Additionally, own empirical distributions for special situations can be handled.

Another reason that not much attention is paid to the choice of the distribution of generated survival times in simulations studies regarding the Cox model is the independence of the partial likelihood in the classical Cox model from the baseline hazard. However, there are a lot of practical situations, where the use of more flexible distributions than the exponential distribution is required in simulation studies investigating the characteristics of the Cox proportional hazards model. When fundamental assumptions of the Cox model are violated so that the partial likelihood depends on the baseline hazard, e.g. in the presence of measurement errors, the results of the simulation study may substantially depend on the distribution of the generated survival times. In the example of the simulations concerning the German Uranium Miners Cohort Study the use of the exponential distribution would lead to incorrect conclusions about the amount of attenuation due to measurement error. Being aware of our findings that the baseline hazard rate may substantially matter, considering only the exponential distribution is much too limited to draw general conclusions from simulation studies on the properties of estimators in Cox models. The methods described in this paper provide the basis to evaluate the characteristics of the Cox model by means of simulation studies in a comprehensive way.

ACKNOWLEDGEMENTS

This work is supported in part by German Research Foundation (DFG) Grants BE 2485/1-1 and SFB 386. We thank Drs Bernd Grosche, Michaela Kreuzer and Annemarie Brachner for providing the data of the German Uranium Miners Cohort Study. We thank also Dr Jürgen Wellmann and Dirk Taeger for useful discussions on simulation techniques and an anonymous referee for helpful comments.

REFERENCES

1. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
2. Hu P, Tsiatis AA, Davidian M. Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* 1998; **54**:1407–1419.
3. Wang CY, Hsu L, Feng ZD, Prentice RL. Regression calibration in failure time regression. *Biometrics* 1997; **53**:131–145.
4. Buzas JS. Unbiased scores in proportional hazards regression with covariate measurement error. *Journal of Statistical Planning and Inference* 1998; **67**:247–257.
5. Kong FH, Huang W, Li X. Estimating survival curves under proportional hazards model with covariate measurement errors. *Scandinavian Journal of Statistics* 1998; **25**:573–587.
6. Kong FH. Adjusting regression attenuation in the Cox proportional hazards model. *Journal of Statistical Planning and Inference* 1999; **79**:31–44.
7. Kong FH, Gu M. Consistent estimation in Cox proportional hazards model with covariate measurement error. *Statistica Sinica* 1999; **9**:953–969.
8. Huang W, Wang CY. Cox regression with accurate covariates unascertainable: a nonparametric-correction approach. *Journal of the American Statistical Association* 2000; **95**:1209–1219.
9. Lee ET, Go OT. Survival analysis in public health research. *Annual Review of Public Health* 1997; **18**: 105–134.
10. Schemper M. Cox analysis of survival data with non-proportional hazard functions. *Journal of the Royal Statistical Society, Series D* 1992; **41**:455–465.
11. Liestøl K, Andersen PK. Updating of covariates and choice of time origin in survival analysis: problems with vaguely defined disease states. *Statistics in Medicine* 2002; **21**:3701–3714.
12. Valenta Z, Weissfeld L. Estimation of the survival function for Gray's piecewise time-varying coefficients model. *Statistics in Medicine* 2002; **21**:717–727.
13. Ng SK, McLachlan GJ. An EM-based semi-parametric mixture model approach to the regression analysis of competing-risks data. *Statistics in Medicine* 2003; **22**:1097–1111.
14. Mackenzie T, Abrahamowicz M. Marginal and hazard ratio specific random data generation: applications to semi-parametric bootstrapping. *Statistics and Computing* 2002; **12**:245–252.
15. Leffondré K, Abrahamowicz M, Siemiatycki J. Evaluation of Cox's model and logistic regression for matched case–control data with time-dependent covariates: a simulation study. *Statistics in Medicine* 2003; **22**: 3781–3794.

16. Rachet B, Abrahamowicz M, Sasco AJ, Siemiatycki J. Estimating the distribution of lag in the effect of short-term exposures and interventions: adaptation of a non-parametric regression spline model. *Statistics in Medicine* 2003; **22**:2335–2363.
17. Gelfand AE, Ghosh SK, Christiansen C, Soumerai SB, McLaughlin TJ. Proportional hazards models: a latent competing risk approach. *Journal of the Royal Statistical Society, Series C* 2000; **49**:385–397.
18. Petersen L, Sørensen TIA, Andersen PK. Comparison of case–cohort estimators based on data on premature death of adult adoptees. *Statistics in Medicine* 2003; **22**:3795–3803.
19. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 1982; **69**:331–342.
20. Pepe MS, Self SG, Prentice RL. Further results on covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine* 1989; **8**:1167–1178.
21. Hughes MD. Regression dilution in the proportional hazards model. *Biometrics* 1993; **49**:1056–1066.
22. Augustin T, Schwarz R. Cox's proportional hazards model under covariate measurement error: a review and comparison of methods. In *Total Least Squares and Errors-in-Variables Modeling*: *Analysis*, *Algorithms and Applications*, Van Huffel S, Lemmerling P (eds). Kluwer: Dordrecht, Boston, London, 2002; 179–188.
23. Kreuzer M, Brachner A, Lehmann F, Martignoni K, Wichmann HE, Grosche B. Characteristics of the German Uranium Miners Cohort Study. *Health Physics* 2002; **83**:26–34.
24. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980.
25. Mood AM, Graybill FA, Boes DC. *Introduction to the Theory of Statistics*. McGraw-Hill: New York, 1974.
26. Kalbfleisch JD, Prentice RL. Marginal likelihood based on Cox's regression and life model. *Biometrika* 1973; **60**:267–278.
27. Taulbee JD. A general model for the hazard rate with covariables. *Biometrics* 1979; **35**:439–450.
28. Ciampi A, Etezadi-Amoli J. A general model for testing the proportional hazards and the accelerated failure time hypothesis in the analysis of censored survival data with covariates. *Communications in Statistics Part A—Theory and Methods* 1985; **14**:651–667.
29. Whittemore AS, Keller JB. Survival estimation using splines. *Biometrics* 1986; **42**:495–506.
30. Flinn CJ, Heckmann JJ. Models for the analysis of labor force dynamics. In *Advances in Econometrics*, Basman RL, Rhodes GF (eds). JAI Press: Greenwich, CT, 1982; 35–95.
31. Gritz M. The impact of training on the frequency and duration of employment. *Journal of Econometrics* 1993; **57**:21–51.
32. Augustin T. An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics* 2004; **31**:43–50.
33. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*. Chapman & Hall: London, 1995.
34. Bender R, Blettner M. Diskussion der Messfehlerproblematik durch Verwendung einer Job-Exposure-Matrix (JEM). In *Stand der Forschung zu den* '*Deutschen Uranbergarbeiterstudien*' *1. Fachgespräch am 7./8. Mai 2001 in St. Augustin*, Geschäftsstelle der Strahlenschutzkommission beim Bundesamt für Strahlenschutz (ed.). Urban und Fischer: München, 2002; 97–105 (in German).
35. Elandt-Johnson RC, Johnson NL. *Survival Models and Data Analysis*. Wiley: New York, 1980.
36. Lawless JE. *Statistical Models and Methods for Lifetime Data*. Wiley: New York, 1982.
37. Heid IM, Küchenhoff H, Wellmann J, Gerken M, Kreienbrock L, Wichmann HE. On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. *Statistics in Medicine* 2002; **21**:3261–3278.
38. Lubin JH, Boice Jr JD, Samet JM. Errors in exposure assessment, statistical power and the interpretation of residential radon studies. *Radiation Research* 1995; **144**:329–341.