An Introduction to Survival Analysis

Mark Stevenson, Simon Firestone, Anke Wiethoelter and Caitlin Pfeiffer
Faculty of Veterinary and Agricultural Sciences
The University of Melbourne
Victoria 3010, Australia

THE UNIVERSITY OF
MELBOURNE

Contact information:

Mark Stevenson (mark.stevenson1@unimelb.edu.au)
Faculty of Veterinary and Agricultural Sciences
University of Melbourne
Victoria 3010, Australia

Tel.: +61 (03) 9035 4114
Fax: +61 (03) 8344 7374

URL: fvas.unimelb.edu.au/veam

1 August 2019

Contents

1 General principles

Survival analysis is the name for a collection of statistical techniques used to describe and quantify time to event data. In survival analysis we use the term 'failure' to define the occurrence of the event of interest (even though the event may actually be a 'success' such as recovery from therapy). The term 'survival time' specifies the length of time taken for failure to occur. Situations where survival analyses have been used in epidemiology include:

- Survival of patients after surgery.
- The length of time taken for cows to conceive after calving.
- The time taken for a farm to experience its first case of an exotic disease.

1.1 Describing time to event

Instantaneous failure rate

When the variable under consideration is the length of time taken for an event to occur (e.g. death) a frequency histogram can be constructed to show the count of events as a function of time. A curve fitted to this histogram produces a plot of the instantaneous failure rate *f(t)*, as shown in Figure 1. If we set the area under the curve of the death density function to equal 1 then for any given time *t* the area under the curve to the left of *t* represents the proportion of individuals in the population who have experienced the event of interest. The proportion of individuals who have died as a function of *t* is called the failure function *F(t)*.

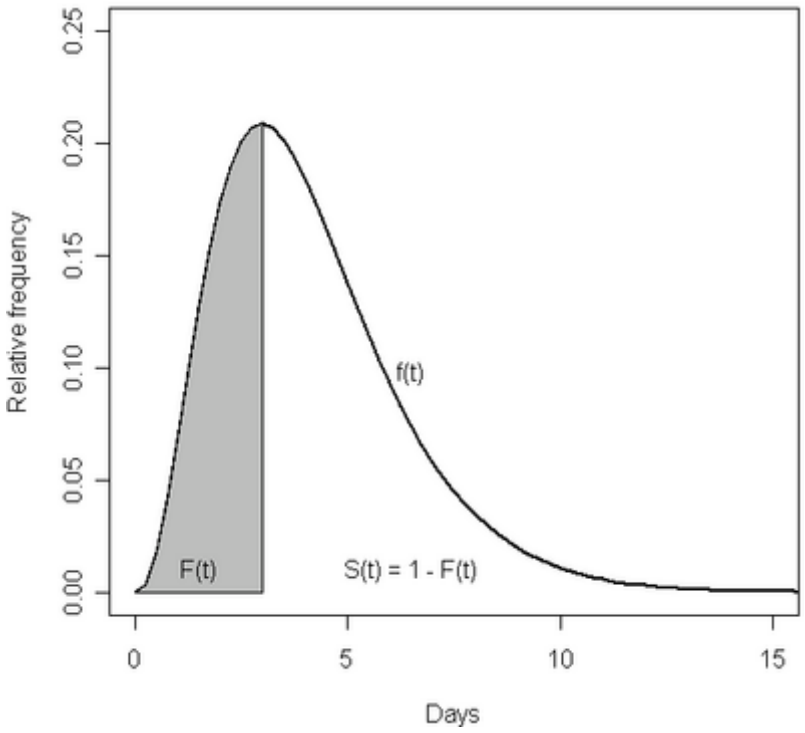


Figure 1: Line plot of $f(t)$ (instantaneous failure rate) as a function of time. The cumulative proportion of the population that has died up to time t equals $F(t)$. The proportion of the population that has survived to time t is $S(t) = F(t) - 1$.

Survival

Consider again the plot of instantaneous failure rate shown in Figure 1. The area under the curve to the right of time *t* is the proportion of individuals in the population who have survived to time *t*, *S(t)*. *S(t)* can be plotted as a function of time to produce a survival curve, as shown in Figure 2. At $t = 0$ there have been no failures so $S(t) = 1$. By day 15 all members of the population have failed and $S(t) = 0$. Because we use counts of individuals present at discrete time points, survival curves are usually presented in step format.

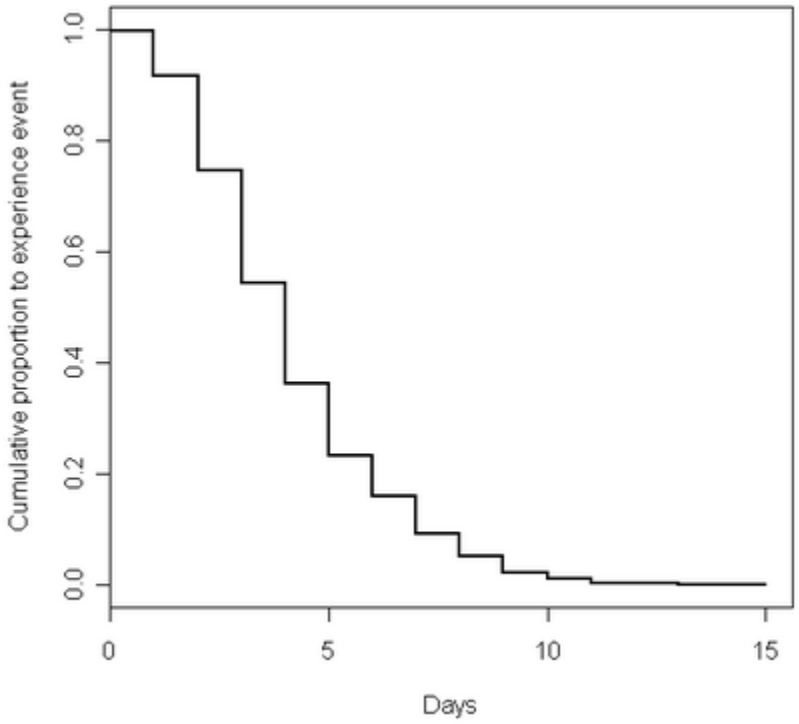


Figure 2: Survival curve showing the cumulative proportion of the population who have 'survived' (not experienced the event of interest) as a function of time.

Hazard

The instantaneous rate at which a randomly-selected individual known to be alive at time $(t - 1)$ will die at time *t* is called the conditional failure rate or instantaneous hazard, *h(t)*. Instantaneous hazard equals the number that fail between time *t* and time $t + \Delta(t)$ divided by the size of the population at risk at time *t*, divided by $\Delta(t)$. This gives the proportion of the population present at time *t* that fail per unit time.

An example of an instantaneous hazard curve is shown in Figure 3. Figure 3 shows the weekly probability of foot-and-mouth disease occurring in two farm types in Cumbria (Great Britain) in 2001. You should interpret this curve in exactly the same way you would an epidemic curve. The advantage of plotting instantaneous hazard as a function of time is that it shows how disease risk changes, correcting for changes in the size of the population at risk (an important issue when dealing with foot-and-mouth disease data, particularly when stamping out is carried out as a means for disease control).

Cumulative hazard (also known as the integrated hazard) at time *t*, *H(t)* equals the area under the instantaneous hazard curve up until time *t*. The cumulative hazard curve shows the (cumulative) probability that the event of interest has occurred up to any point in time.

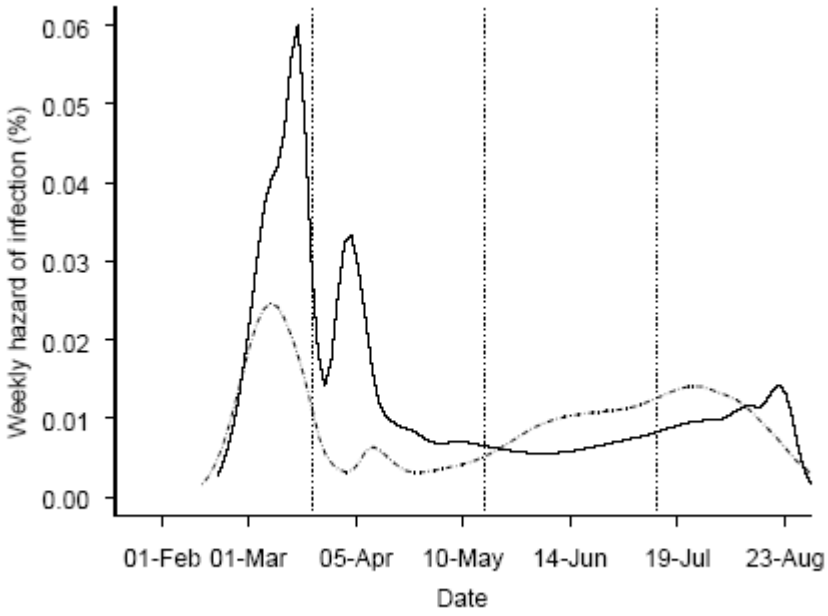


Figure 3: Weekly hazard of foot-and-mouth disease infection for cattle holdings (solid lines) and 'other' holdings (dashed lines) in Cumbria (Great Britain) in 2001. Reproduced from Wilesmith et al. (2003).

1.2 Censoring

In longitudinal studies exact survival time is only known for those individuals who show the event of interest during the follow-up period. For others (those who are disease free at the end of the observation period or those that were lost) all we can say is that they did not show the event of interest during the follow-up period. These individuals are called censored observations. An attractive feature of survival analysis is that we are able to include the data contributed by censored observations right up until they are removed from the risk set. The following terms are used in relation to censoring:

- Right censoring: a subject is right censored if it is known that the event of interest occurs some time *after* the recorded follow-up period.
- Left censoring: a subject is left censored if it is known that the the event of interest occurs some time *before* the recorded follow-up period. For example, you conduct a study investigating factors influencing days to first oestrus in dairy cattle. You start observing your population (for argument's sake) at 40 days after calving but find that several cows in the group have already had an oestrus event. These cows are said to be left censored at day 40.
- Interval censoring: a subject is interval censored if it is known that the event of interest occurs between two times, but the exact time of failure is not known. In effect we say 'I know that the event occurred between date A and date B: I know that the event occurred, but I don't know exactly when.' In an observational study of EBL seroconversion you sample a population of cows every six months. Cows that are negative on the first test and positive at the next are said to have seroconverted. These individuals are said to be interval censored with the first sampling date being the lower interval and the second sampling date the upper interval.

We should distinguish between the terms censoring and truncation (even though the two events are handled the same way analytically). Censoring is when an observation is incomplete due to some random cause. The

cause of the censoring must be independent of the event of interest if we are to use standard methods of analysis. Truncation is a variant of censoring which occurs when the incomplete nature of the observation is due to a systematic selection process inherent to the study design.

- Left truncation: a subject is left truncated if it enters the population at risk some stage after the start of the follow-up period. For example, in a study investigating the date of first BSE diagnosis on a group of farms, those farms that are established after the start of the study are said to be left truncated (the implication here is that there is no way the farm can experience the event of interest before the truncation date).
- Right truncation: a subject is right truncated if it leaves the population at risk some stage after the study start (and we know that there is no way the event of interest could have occurred after this date). For example, in a study investigating the date of first foot-and-mouth disease diagnosis on a group of farms, those farms that are pre-emptively culled as a result of control measures are right truncated on the date of culling.

Consider a study illustrated in Figure 4. Subjects enter at various stages throughout the study period. An 'X' indicates that the subject has experienced the outcome of interest; a 'O' indicates censoring. Subject A experiences the event of interest on day 7. Subject B does not experience the event during the study period and is right censored on day 12 (this implies that subject B experienced the event sometime after day 12). Subject C does not experience the event of interest during its period of observation and is censored on day 10. Subject D is interval censored: this subject is observed intermittantly and experiences the event of interest sometime between days 5 – 6 and 7 – 8. Subject E is left censored — it has been found to have already experienced the event of interest when it enters the study on day 1. Subject F is interval truncated: there is no way possible that the event of interest could occur to this individual between days 4 – 6. Subject G is left truncated: there is no way possible that the event of interest could have occurred before the subject enters the study on day 3.

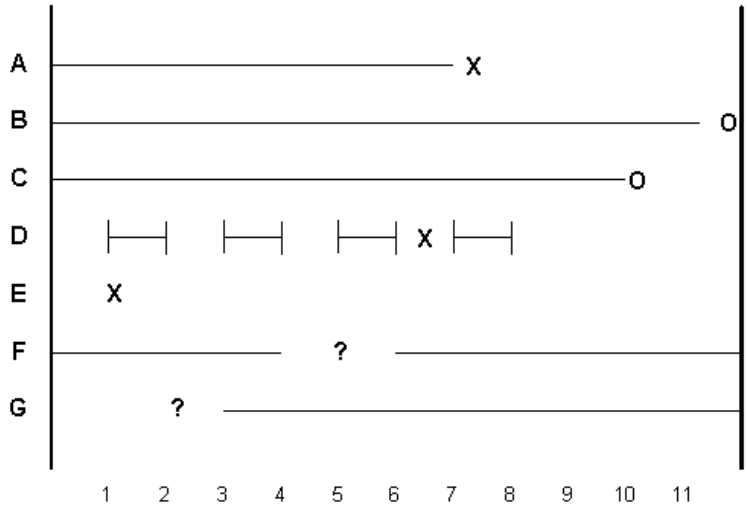


Figure 4: Left-, right-censoring, and truncation (Dohoo, Martin and Stryhn 2003).

2 Non-parametric survival

Once we have collected time to event data, our first task is to describe it — usually this is done graphically using a survival curve. Visualisation allows us to appreciate temporal pattern in the data. It also helps us to identify an appropriate distributional form for the data. If the data are consistent with a parametric distribution, then parameters can be derived to efficiently describe the survival pattern and statistical inference can be based on the chosen distribution. Non-parametric methods are used when no theoretical distribution adequately fits the data. In epidemiology non-parametric (or semi-parametric) methods are used more frequently than parametric methods.

There are three non-parametric methods for describing time to event data: (1) the Kaplan-Meier method, (2) the life table method, and (3) the Nelson-Aalen method.

2.1 Kaplan-Meier method

The Kaplan-Meier method is based on individual survival times and assumes that censoring is independent of survival time (that is, the reason an observation is censored is unrelated to the cause of failure). The Kaplan-Meier estimator of survival at time *t* is shown in Equation 14. Here t_j, *j = 1, 2, ..., n* is the total set of failure times recorded (with t^+ the maximum failure time), d_j is the number of failures at time t_j, and r_j is the number of individuals at risk at time t_j. A worked example is provided in Table 1. Note that: (1) for each time period the number of individuals present at the start of the period is adjusted according to the number of individuals censored and the number of individuals who experienced the event of interest in the previous time period, and (2) for ties between failures and censored observations, the failures are assumed to occur first.

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{(r_j - d_j)}{r_j}, \text{ for } 0 \leq t \leq t^+ \tag{1}$$

Table 1: Details for calculating Kaplan-Meier survival estimates as a function of time.

Time	Start	Fail	Censored	At risk	Surv prob	Cumulative survival
	n_j	d_j	w_j	r_i	$P_j = (r_j - d_j)/r_j$	$S_j = P_j \times P_{j-1}$
0	31	2	3	31 - 3 = 28	(28 - 2) / 28 = 0.93	0.93 $\times$ 1.00 = 0.93
1	26	1	2	26 - 2 = 24	(24 - 1) / 24 = 0.96	0.96 $\times$ 0.93 = 0.89
2	23	1	2	23 - 2 = 21	(21 - 1) / 21 = 0.95	0.95 $\times$ 0.89 = 0.85
3	20	1	2	20 - 2 = 18	(18 - 1) / 18 = 0.94	0.94 $\times$ 0.85 = 0.80
etc						

2.2 Life table method

The life table method (also known as the actuarial or Cutler Ederer method) is an approximation of the Kaplan-Meier method. It is based on grouped survival times and is suitable for large data sets. Calculation details are shown in Table 2.

The life table method assumes that subjects are withdrawn randomly throughout each interval — therefore, on average they are withdrawn half way through the interval. This is not an important issue when the time intervals are short, but bias may introduced when time intervals are long. This method also assumes that the

Table 2: Details for calculating life table survival estimates as a function of time.

Time	Start	Fail	Censored
	n_i	d_i	w_i
0 to 1	31	3	4
2 to 3	24	2	4
etc			

Time	Failure prob	Survival prob	Cumulative survival
	$q_i = d_i/[n_i - (w_i/2)]$	$p_i = 1 - q_i$	$S_i = p_i \times S_{i-1}$
0 to 1	3 / [31 - (4/2)] = 0.10	1 - 0.10 = 0.90	0.90 $\times$ 1 = 0.90
2 to 3	2 / [24 - (4/2)] = 0.09	1 - 0.09 = 0.91	0.90 $\times$ 0.91 = 0.82
etc			

rate of failure within an interval is the same for all subjects and is independent of the probability of survival at other time periods. Life tables are produced from large scale population surveys (e.g. death registers) and are less-frequently used these days (the Kaplan-Meier method being preferred because it is less prone to bias).

2.3 Nelson-Aalen estimator

Instantaneous hazard is defined as the proportion of the population present at time *t* that fail per unit time. The cumulative hazard at time *t*, *H(t)* is the summed hazard for all time up to time *t*. The relationship between cumulative hazard and survival is as follows:

$$H(t) = -ln[S(t)], \text{ or } S(t) = e^{-H(t)} \tag{2}$$

The Nelson-Aalen estimator of cumulative hazard at time *t* is defined as:

$$\hat{H}(t) = \sum_{j:t_j \leq t} \frac{d_j}{r_j}, \text{ for } 0 \leq t \leq t^+ \tag{3}$$

The Flemington-Harrington estimate of survival can be calculated using the Nelson-Aalen estimate of cumulative hazard using the relationship between survival and cumulative hazard described in Equation 2.

2.4 Worked examples

An Australian study by Caplehorn and Bell (1991) compared retention in two methadone treatment clinics for heroin addicts. A patient's survival time was determined as the time in days until the patient dropped out of the clinic or was censored at the end of the study. The two clinics differed according to their overall treatment policies. Interest lies in identifying factors that influence retention time: clinic, maximum daily methadone dose, and presence of a prison record.

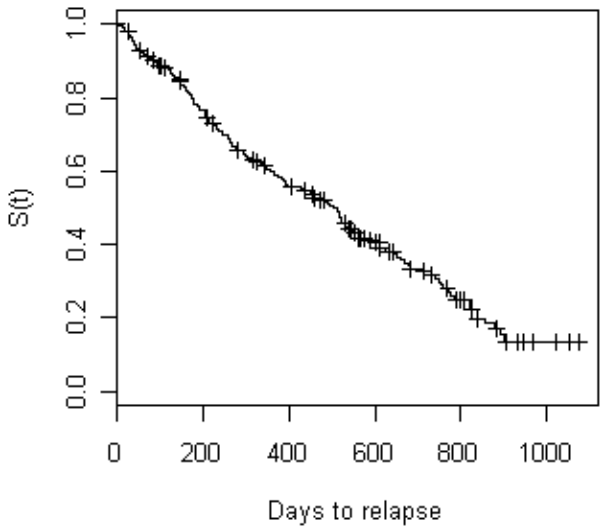


Figure 5: Kaplan-Meier survival curve showing the cumulative proportion of heroin addicts retained in two methadone treatment clinics (Caplehorn and Bell 1991).

Kaplan-Meier method

Figure 5 is a Kaplan-Meier survival curve showing the cumulative proportion of addicts retained in the clinics over time. Figure 5 shows that the rate of loss of patients over time is relatively constant and that approximately 15% remain in treatment by 1000 days post admission.

Figure 5 was produced using the following code:

```
library(survival)
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Inspect the first six rows of data:

```
head(dat)
```

id	start	stop	status	clinic	prison	dose
1	0	428	1	1	0	50
2	0	275	1	1	1	55
3	0	262	1	1	0	55
4	0	183	1	1	0	30
5	0	259	1	1	1	65
6	0	714	1	1	0	55

id: patient identifier.
start: day of entry to clinic.
stop: day of relapse.
status: 0 = censored, 1 = died.
clinic: 1 = clinic 1, 2 = clinic 2.
prison: 0 = prison record absent, 1 = prison record present.
dose: maximum daily methadone dose (mg/day).

The function `Surv` creates a survival object in R linking survival time and censoring:

```
Surv(dat$stop, dat$status)
```

The output above shows survival times for each subject. A plus sign after time indicates that the subject was censored. Plot the Kaplan-Meier survival function of days from discharge from clinic to relapse:

```
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)")
```

Kaplan-Meier survival function with confidence intervals:

```
addict.km <- survfit(Surv(stop, status) ~ 1, type = "kaplan-meier", data =
dat)
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)", conf.int = TRUE)
```

Summary estimates of survival at different time points throughout the follow up period:

```
summary(addict.km, times = seq(from = 0, to = 1000, by = 250))
```

Kaplan-Meier survival function of days to relapse, stratifying by clinic:

```
addict.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)", lty = c(1,2))
legend(x = "topright", legend = c("Clinic 1","Clinic 2"), lty = c(1,2), bty =
"n")
```

Kaplan-Meier survival function of days to relapse, stratifying by methadone dose:

```
dat$cdose[dat$dose < 60] <- 0
dat$cdose[dat$dose >= 60] <- 1
addict.km <- survfit(Surv(stop, status) ~ cdose, type = "kaplan-meier", data =
dat)
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)", lty = c(1,2))
legend(x = "topright", legend = c("Low dose methadone","High dose methadone"),
lty = c(1,2), bty = "n")
```

The `survminer` package can be used to make more attractive survival plots (Figure 6).

Kaplan-Meier survival function of days to relapse with confidence intervals:

```
library(survminer)
ggsurvplot(addict.km)
```

Add a dashed line to indicate median survival time. The setting `surv.median.line = hv` draws both horizontal and vertical lines on the plot:

```
ggsurvplot(addict.km, surv.median.line = "hv")
```

Add a table to show the number of individuals at risk at different follow-up times:

```
ggsurvplot(addict.km, surv.median.line = "hv",
risk.table = TRUE, tables.height = 0.1, tables.theme = theme_cleantable(),
risk.table.pos = "in")
```

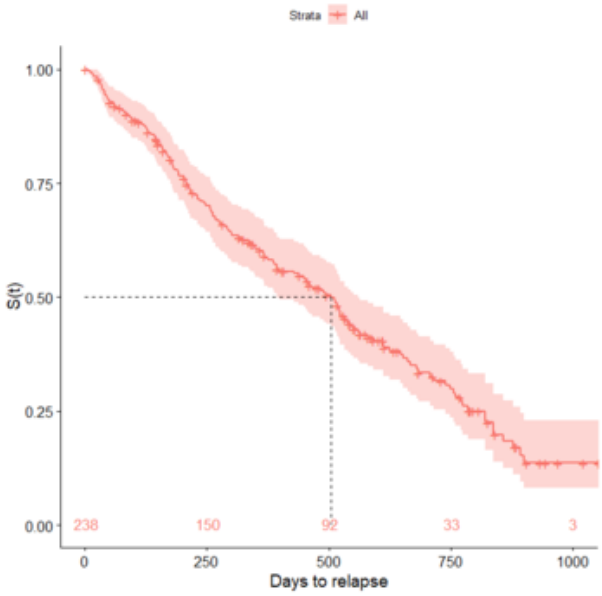


Figure 6: Kaplan-Meier survival curve for the addict data plotted using the survminer package. The dashed line indicates the median time of relapse. The numbers of individuals at risk at different follow-up times are shown in red just above the horizontal axis.

Flemington-Harrington estimator

```
addict.km <- survfit(Surv(stop, status) ~ 1, type = "kaplan-meier", data =
dat)
addict.fh <- survfit(Surv(stop, status) ~ 1, type = "fleming-harrington", data
= dat)

par(pty = "s", mfrow = c(1,2))
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)", conf.int = FALSE)
plot(addict.fh, xlab = "Days to relapse", ylab = "S(t)", conf.int = FALSE)
```

With this data set the difference between the Kaplan-Meier and the Flemington-Harrington estimate of survival is not obvious. A closer comparison of the two functions:

```
tmp <- as.data.frame(cbind(km = addict.km$surv, fh = addict.fh$surv))
head(tmp)
tail(tmp)
```

Instantaneous hazard

```
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
```

Work out the proportion that fail at each evaluated time period:

```
names(addict.km)
prop.fail <- addict.km$n.event/addict.km$n.risk
```

Work out the length of time over which these failure occur:

```
time <- addict.km$time
time
[1] 2, 7, 13, 17, ..., 1076

time0 <- c(0, time[-length(time)])
time0
[1] 0, 2, 7, 13, ..., 1052
```

Divide `prop.fail` by the time interval over which those failures occur (that is, `time - time0`) to get the probability of failing per unit time, i.e. the instantaneous hazard:

```
haz <- prop.fail/(time - time0)
```

Plot the result:

```
plot(time, haz, ylim = c(0,0.03), type = "s", xlab = "Days to relapse", ylab =
"h(t)")
lines(lowess(time[-1], haz[-1], f = 0.10))
```

Tidier plot:

```
plot(time, haz, type = "n", xlab = "Days to relapse", ylab = "h(t)", ylim =
c(0,0.03))
lines(lowess(time[-1], haz[-1], f = 0.10))
```

A simpler way to plot instantaneous hazard using the `epiR` package:

```
library(epiR)
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
addict.haz <- epi.insthaz(addict.km)

library(epiR)
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
addict.haz <- epi.insthaz(addict.km)
plot(addict.haz$time, addict.haz$est, xlab = "Days to relapse", ylab = "h(t)",
ylim = c(0,0.03), type = "n")
lines(lowess(addict.haz$time, addict.haz$est, f = 0.10), lty = 1)
lines(lowess(addict.haz$time, addict.haz$lower, f = 0.10), lty = 2, col =
"gray")
lines(lowess(addict.haz$time, addict.haz$upper, f = 0.10), lty = 2, col =
"gray")
```

Cumulative hazard

A cumulative hazard plot shows the (average) cumulative number of events likely to be experienced by a subject as a function of time. A cumulative hazard of 2.0 at day 1000 means that, on average, a subject will have experienced 2.0 events by the time it reaches day 1000.

Plot the cumulative hazard of relapse:

```
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
addict.haz <- epi.insthaz(addict.km)

plot(addict.km, fun = "cumhaz", xlab = "Days to relapse", ylab = "H(t)", lty =
c(1,2))
```

Compare cumulative hazard as a function of time with instantaneous hazard as a function of time:

```
par(pty = "s", mfrow = c(1,2))
plot(addict.km, fun = "cumhaz", xlab = "Days to relapse", ylab = "H(t)")
plot(addict.haz$time, addict.haz$est, xlab = "Days to relapse", ylab = "h(t)",
ylim = c(0,0.06), type = "s")
```

3 Parametric survival

On some occasions the pattern of survivorship for our study subjects follows a predictable pattern. In this situation parametric distributions can be used to describe time to event. An advantage of using a parametric distribution is that we can reliably predict time to event well after the period during which events occurred for our observed data. Several parametric distributions are used to describe time to event data. Each parametric distribution is defined by a different hazard function, as shown in Table 3.

Table 3: Parametric survival distributions used in epidemiology.

Distribution	$f(t)^a$	$h(t)^b$	$H(t)^c$	$S(t)^d$
Exponential	λ exp[-λt]	λ	λt	exp[-λt]
Weibull	λpt^{p-1} exp[-λt^p]	λpt^{p-1}	λt^p	exp[-(λt)^p]
Gompertz	a exp[bt] exp[-a/b (exp[bt] - 1)]	a exp[bt]	a/b (exp[bt] - 1)	exp[-a/b (exp[bt] - 1)]
Log-logistic	[ab(at)$^{b-1}$] / [1 + (at)^b]^2	[ab(at)$^{b-1}$] / [1 + (at)^b]	log[1 + (at)^b]	[1 + (at)^b]$^{-1}$

[a] $f(t)$ instantaneous failure rate.
[b] $h(t)$ instantaneous hazard.
[c] $H(t)$ cumulative hazard.
[d] $S(t)$ survival.

The Gompertz distribution provides a convenient way of describing survival in human subjects and is frequently used in demography. The Gompertz distribution can be generalised to the Gompertz-Makeham distribution by adding a constant to the instantaneous hazard: $h(t) = c + a$ exp(bt).

As a general approach to the analysis of time to event data you should plot the hazard function for the observed data and determine whether or not it is consistent with a parametric distribution. If the data follows a parametric distribution, parametric methods are preferred to non-parametric methods for describing and quantifying factors that influence time to event. In veterinary epidemiology, the most important parametric forms are the exponential and Weibull distributions.

3.1 The exponential distribution

The exponential distribution is described by the mean, λ. A feature of the exponential distribution is that the instantaneous hazard does not vary over time (Figure 7). Observed survival distributions can be checked for consistency with the exponential distribution by plotting instantaneous hazard as a function of time: exponential distributions in this case will yield a straight line. Alternatively, the log of cumulative hazard can be plotted as a function of the log of time: exponential distributions will yield a 45° line.

3.2 The Weibull distribution

The Weibull distribution is described by a scale parameter λ and shape parameter *p*. If *p* is less than 1 instantaneous hazard monotonically decreases with time, if *p* equals 1 instantaneous hazard is constant over time (equivalent to the exponential distribution) and if *p* is greater than 1 instantaneous hazard increases with time. Figure 8 is an example of a Weibull distributed survival pattern with $p < 1$. Time to event data can be checked for consistency with the Weibull distribution by plotting the log cumulative hazard as a function of log time: Weibull distributions in this case will yield a straight line.

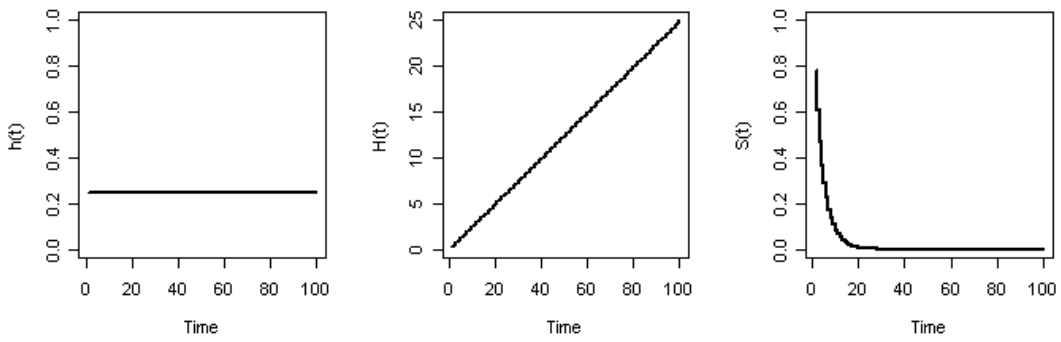


Figure 7: Instantaneous hazard, cumulative hazard and survival as a function of time for the exponential distribution. In this example λ = 0.25.

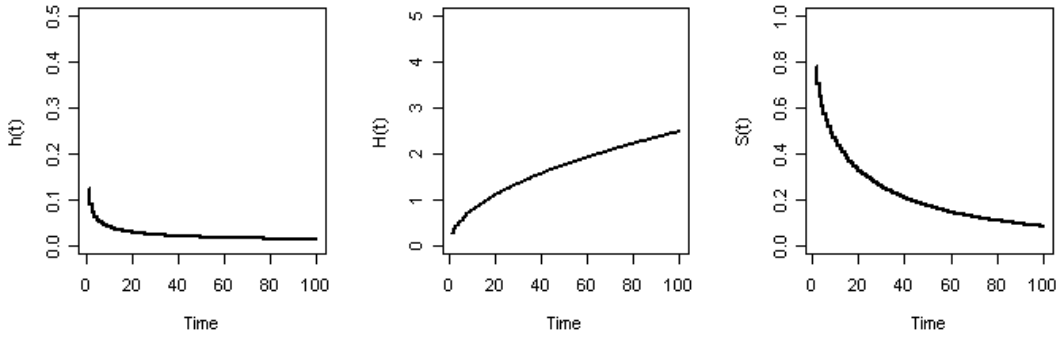


Figure 8: Instantaneous hazard, cumulative hazard and survival as a function of time for the Weibull distribution. In this example λ = 0.25 and p = 0.5.

3.3 Worked examples

The exponential distribution

Figure 7 was produced using the following code:

```
t <- seq(from = 1, to = 100, by = 1)
lambda = 0.25
ht <- lambda
Ht <- lambda * t
St <- exp(-lambda * t)

par(mfrow = c(1,3), pty = "s")
plot(t, rep(ht, times = length(t)), ylim = c(0, 1), lwd = 2, type = "s", xlab
= "Time", ylab = "h(t)")
plot(t, Ht, ylim = c(0, 25), lwd = 2, type = "s", xlab = "Time", ylab =
"H(t)")
plot(t, St, ylim = c(0, 1), lwd = 2, type = "s", xlab = "Time", ylab = "S(t)")
```

The Weibull distribution

Figure 8 was produced using the following code:

```
t <- seq(from = 1, to = 100, by = 1)
lambda = 0.25; p = 0.5
ht <- lambda * p * t^(p - 1)
Ht <- lambda * t^p
St <- exp(-lambda * t^p)

par(mfrow = c(1,3), pty = "s")
plot(t, ht, ylim = c(0, 0.5), lwd = 2, type = "s", xlab = "Time", ylab =
"h(t)")
plot(t, Ht, ylim = c(0, 5), lwd = 2, type = "s", xlab = "Time", ylab = "H(t)")
plot(t, St, ylim = c(0, 1), lwd = 2, type = "s", xlab = "Time", ylab = "S(t)")
```

Plots of hazard using different values of `lambda` and `p`:

```
t <- seq(from = 0, to = 10, by = 0.1)
lambda <- 1; p05 <- 0.5; p10 <- 1.0; p15 <- 1.5; p30 <- 3.0
h05 <- lambda * p05 * (lambda * t)^(p05 - 1)
h10 <- lambda * p10 * (lambda * t)^(p10 - 1)
h15 <- lambda * p15 * (lambda * t)^(p15 - 1)
h30 <- lambda * p30 * (lambda * t)^(p30 - 1)

plot(t, h05, type = "l", ylim = c(0, 6), xlab = "Time", ylab = "h(t)", lty =
1, lwd = 1)
lines(t, h10, lty = 2, lwd = 1)
lines(t, h15, lty = 3, lwd = 1)
lines(t, h30, lty = 4, lwd = 1)
legend(x = "topright", legend = c("lambda = 1, p = 0.5", "lambda = 1, p =
1.0", "lambda = 1, p = 1.5", "lambda = 1, p = 3.0"), lty = c(1,2,3,4), lwd =
c(1,1,1,1), bty = "n", cex = 0.75)
```

Comparison of Kaplan-Meier and Weibull estimates of survival:

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Fit parametric (Weibull) and non-parametric (Kaplan-Meier) survival functions to the observed data:

```
addict.we <- survreg(Surv(stop, status) ~ 1, dist = "weib", data = dat)
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
```

Using the Weibull distribution μ (the intercept) = -log(λ) and σ (scale) = 1 / p. Thus the scale parameter λ = exp(-μ) and p = 1 / σ. See Venables and Ripley p 360 and Tableman and Kim p 78 for details.

```
p <- 1 / addict.we$scale
lambda <- exp(-addict.we$coeff[1])
t <- 1:1000
St <- exp(-(lambda * t)^p)
addict.we <- as.data.frame(cbind(t = t, St = St))
```

Compare the two estimates of survival:

```
plot(addict.km, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event")
lines(addict.we$t, addict.we$St, lty = 2)
legend(x = "topright", legend = c("Kaplan-Meier", "Weibull"), lty = c(1,2),
bty = "n")
```

The Weibull distribution provides an adequate fit to the observed data up to day 500, then appears to underestimate survivorship.

Cumulative hazard plots can provide an alternative method for assessing the appropriateness of a parametric approach to describe survivorship. Here we plot cumulative hazard as a function of time to check for consistency with the exponential distribution, log cumulative hazard as a function of log time to check for consistency with the Weibull distribution and the inverse normal transformation of the Kaplan-Meier estimates as a function of log time to check for consistency with the log-normal distribution.

```
par(pty = "s", mfrow = c(2,2))
plot(addict.km, conf.int = FALSE, fun = "cumhaz", xlab = "Days to relapse",
main = "Exponential")
plot(addict.km, conf.int = FALSE, fun = "cumhaz", log = "xy", xlab = "Days to
relapse (log)", main = "Weibull")
plot(addict.km, conf.int = FALSE, fun = qnorm, log = "x", xlab = "Days to
relapse (log)", main = "Log-normal")
```

4 Comparing survival distributions

It is frequently of interest to compare the survival of one group of study subjects with another.

- Did animals survive longer in one herd compared with another?
- Did disease take longer to develop in one region of a country compared with another?
- Did patients survive longer after one therapy compared with another?

In addition to providing useful information about how time to event distributions differ among groups, separate survival curves for different levels of covariates provide an effective screening process that helps one to identify factors that are influential in determining survival. Once influential factors are screened using these methods their influence can then be tested using multivariate analyses.

When there are no censored observations, standard non-parametric tests can be used to compare two survival distributions. If the groups are independent, a Wilcoxon or Mann-Whitney U test may be used. If the groups are dependent the Sign Test may be used.

4.1 The log-rank test

The log-rank test (also known as the Mantel log-rank test, the Cox Mantel log-rank test, and the Mantel-Haenszel test) is the most commonly used test for comparing survival distributions. It is applicable to data where there is progressive censoring and gives equal weight to early and late failures. It assumes that hazard functions for the two groups are parallel. The test takes each time point when a failure event occurs and a 2×2 table showing the number of deaths and the total number of subjects under follow up is created. For each table the observed deaths in each group, the expected deaths and the variance of the expected number are calculated. These quantities are summed over all tables to yield a χ^2 statistic with 1 degree of freedom (known as the Mantel-Haenszel or log-rank test statistic). The log-rank test calculations also produce for each group the observed to expected ratio which relates the number of deaths observed during the follow up with the expected number under the null hypothesis that the survival curve for that group would be the same as that for the combined data.

4.2 Other tests

Breslow's test (also known as Gehan's generalised Wilcoxon test) is applicable to data where there is progressive censoring. It is more powerful than the log-rank test when the hazard functions are not parallel and where there is little censoring. It has low power when censoring is high. It gives more weight to early failures.

The Cox Mantel test is similar to the log-rank test. It is applicable to data where there is progressive censoring. More powerful than Gehan's generalised Wilcoxon test. The Peto and Peto modification of the Gehan-Wilcoxon test is similar to Breslow's test and is used where the hazard ratio between groups is not constant. Cox's F test is more powerful than Breslow's test if sample sizes are small.

4.3 Worked examples

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Kaplan-Meier survival function of days to relapse, stratifying by clinic:

```
addict.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)
plot(addict.km, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", lty = c(1,2))
legend(x = "topright", legend = c("Clinic 1", "Clinic 2"), lty = c(1,2), bty =
"n")
```

In the `survdiff` function the argument `rho = 0` returns the log-rank or Mantel-Haenszel test, `rho = 1` returns the Peto and Peto modification of the Gehan-Wilcoxon test. Mantel-Haenszel test:

```
survdiff(Surv(stop, status) ~ clinic, data = dat, na.action = na.omit, rho =
0)
```

Peto and Peto modification of the Gehan-Wilcoxon test:

```
survdiff(Surv(stop, status) ~ clinic, data = dat, na.action = na.omit, rho =
1)
```

5 Non-parametric and semi-parametric regression

Survival models are used to quantify the effect of one or more explanatory variables on failure time. This involves specification of a linear-like model for the log hazard. A parametric model based on the exponential distribution may be parameterised as follows:

$$\log h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \tag{4}$$

or, equivalently:

$$h_i(t) = \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}) \tag{5}$$

In this case the constant α represents the log-baseline hazard since $\log h_i(t) = \alpha$ when all the *x*'s are zero. The Cox proportional hazards model is a semi-parametric model where the baseline hazard $\alpha(t)$ is allowed to vary with time:

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \tag{6}$$

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}) \tag{7}$$

If all of the *x*'s are zero the second part of the above equation equals 1 so $h_i(t) = h_0(t)$. For this reason the term $h_0(t)$ is called the baseline hazard function. With the Cox proportional hazards model the outcome is described in terms of the hazard ratio. We talk about the hazard of the event of interest at one level of an explanatory variable being a number of times more (or less) than the hazard of the specified reference level of the explanatory variable.

Assumptions of the Cox proportional hazards model are as follows:

- The ratio of the hazard function for two individuals with different sets of covariates does not depend on time.
- Time is measured on a continuous scale.
- Censoring occurs randomly.

Table 4 presents the results of a Cox proportional hazards regression model for the Caplehorn `addict` data set (Caplehorn and Bell 1991). Here the authors have quantified the effect of clinic, methadone dose, and prison status on the daily hazard of relapse (re-using heroin). Clinic is a categorical variable with Clinic 1 as the reference category. The results of the model show that, compared with patients from Clinic 1 and after adjusting for the effect of methadone dose and prison status, Clinic 2 patients had 0.36 (95% CI 0.24 – 0.55) times the daily hazard of relapse. Similarly, for unit increases in the daily dose of methadone, after adjusting for the effect of clinic and the presence of a prison record the daily hazard of relapse was reduced by of factor of 0.96 (95% CI 0.95 – 0.98).

5.1 Model building

Selection of covariates

We now discuss how a set of variables are selected for inclusion in a regression model of survival. Begin with a thorough univariate analysis of the association between survival time and all important covariates. For categorical variables this should include Kaplan-Meier estimates of the group-specific survivorship functions. Tabulate point and interval estimates of the median and quartiles of survival time. Use one or more of the significance tests to compare survivorship among the groups defined by the variable under investigation. Continuous covariates should be broken into quartiles (or other biologically meaningful groups) and the same methods applied to these groups.

Table 4: Cox proportional hazards regression model showing the effect of clinic, methadone dose and prison status on the daily hazard of relapse (adapted from Caplehorn and Bell 1991).

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Clinic				< 0.01 [a]	
Clinic 1	163	122			1.0
Clinic 2	75	28	-1.0092 (0.2147)		0.36 (0.24 – 0.55) [b]
Prison record				0.06	
Absent	127	81			1.0
Present	111	69	0.3146 (0.1672)		1.37 (0.98 – 1.90)
Dose	238	150	-0.0352 (0.0064)	< 0.01	0.96 (0.95 – 0.98)

[a] The significance of inclusion of the two clinic variables in the model.
[b] Intepretation: compared with the reference category (patients from Clinic 1) after adjusting for the effect of methadone dose and prison status patients from Clinic 2 had 0.36 (95% CI 0.24 – 0.55) times the daily hazard of relapse.
SE: standard error.
CI: confidence interval.

Tied events

A central assumption in survival analysis is that time is continuous. Sometimes (particularly in veterinary epidemiological research) the outcome of interest is not measured on a continuous scale and outcome events may occur simultaneously (e.g. service number when conception occurred). When the number of tied events is large, approximate methods yield regression coefficients that are biased towards zero. There are three common methods for dealing with ties:

1. Breslow approximation. There is a contribution to the partial likelihood from each of the tied failure times. For each failure time, the risk set comprises all subjects failing at or after the failure time. This includes all subjects whose failure times are tied with that of the subject contributing to the numerator.
2. Efron approximation. In the Breslow approximation, if *m* subjects share the same survival time, they all contribute to the risk set for each of the *m* failure times as if each one of the *m* subjects failed, all others were still alive. In the Efron approximation, the contribution to the denominator from the *m* subjects with tied survival times is weighted down by a factor of *(m - k)/m* for the *k*th term.
3. Exact partial likelihood. Assuming that no two subjects ever failed simultaneously (this would be the case if we measured the time of failure down to milliseconds), there is a true (unknown) unique ordering of the tied survival times. The exact partial likelihood can be obtained by taking the sum (or average) of the partial likelihoods for all possible orderings of the tied survival times. Computationally intensive.

Fitting a multivariable model

A multivariable model should contain at the outset all covariates significant in the univariate analyses at the P = 0.20 to 0.25 level and any others that are thought to be of clinical importance. You should also include any covariate that has the potential to be an important confounder.

Following the fit of the multivariable model, use the P values from the Wald tests of the individual coefficients to identify covariates that might be deleted from the model. The partial likelihood ratio test should confirm that the deleted covariate is not significant. Also check if removal of a covariate produces a 'significant' change (say 20%) in the coefficients of the covariates remaining in the model. Continue until no covariates can be deleted from the model. At this point, work backwards and add each of the deleted covariates back into the model one at a time — checking that none of them are significant or show evidence of being a confounder.

Check the scale of continuous covariates

The next thing is to examine the scale of the continuous covariates in the preliminary model. Here we need to check that the covariate is linear in its log hazard. Replace the continuous covariate with three design variables using Q1, Q2, and Q3 as cutpoints. Plot the estimated coefficients for the design variables versus the midpoint of the group. A fourth point is included at zero using the midpoint of the first group. If the correct scale is linear, then the line connecting the four points should approximate a straight line. Consider transforming the continuous variable if this is not the case. Another method to check this property of continuous covariates uses fractional polynomials.

Another method is to use two residual-based plots: (1) a plot of the covariate values versus the Martingale residuals (and their smooth) from a model that excludes the covariate of interest, and (2) a plot of the covariate values versus the log of the ratio of smoothed censor to smoothed cumulative hazard. To construct the second plot: (1) fit the preliminary main effects model, including the covariate of interest (e.g. 'age'), (2) save the Martingale residuals (M_i) from this model, (3) calculate $H_i = c_i - M_i$, where c_i is the censoring variable, (4) plot the values of c_i versus the covariate of interest and calculate a lowess smooth (called c_{LSM}), (5) plot the values of H_i versus the covariate of interest and calculate a lowess smooth (called H_{LSM}), (6) the smoothed values from these plots are used to calculate:

$$y_i = ln\left(\frac{c_{LSM}}{H_{LSM}}\right) + \beta_{age} \times age_i \tag{8}$$

and the pairs (y_i, age_i) are plotted and connected by straight lines. There should be a linear relationship between the covariate values and each of the described parameters.

Interactions

The final step is to determine whether interaction terms are required. An interaction term is a new variable that is the product of two other variables in the model. Note that there can be subject matter considerations that dictate that a particular interaction term (or terms) should be included in a given model, regardless of their statistical significance. In most settings there is no biological or clinical theory to justify automatic inclusion of interactions.

The effect of adding an interaction term should be assessed using the partial likelihood ratio test. All significant interactions should be included in the main-effects model. Wald statistic P-values can be used as a guide to selecting interactions that may be eliminated from the model, with significance checked by the partial likelihood ratio test.

At this point we have a 'preliminary model' and the next step is to assess its fit and adherence to key assumptions.

5.2 Testing the proportional hazards assumption

Once a suitable set of covariates has been identified, it is wise to check each covariate to ensure that the proportional hazards assumption is valid. To assess the proportional hazards assumption we examine the extent to which the estimated hazard curves for each level of strata of a covariate are equidistant over time.

A plot of the scaled Schoenfeld residuals (and a loess smoother) as a function of time may be used to test proportionality of hazards. In a 'well-behaved' model the Schoenfeld residuals are scattered around 0 and a regression line fitted to the residuals has a slope of approximately 0. The idea behind this test is that if

the proportional hazards assumption holds for a particular covariate then the Schoenfeld residuals for that covariate will not be related to survival time. The implementation of the test can be thought of as a three-step process: (1) run a Cox proportional hazards model and obtain the Schoenfeld residuals for each predictor, (2) create a variable that ranks the order of failures (the subject who has the first (earliest) event gets a value of 1, the next gets a value of 2, and so on, (3) test the correlation between the variables created in the first and second steps. The null hypothesis is that the correlation between the Schoenfeld residuals and ranked failure time is zero. An important point about this approach is that the null hypothesis is never proven with a statistical test (the most that can be said is that there is not enough evidence to reject the null) and that p-values are driven by sample size. A gross violation of the null assumption may not be statistically significant if the sample is very small. Conversely, a slight violation of the null assumption may be highly significant if the sample is very large.

For categorical covariates the proportional hazards assumption can be visually tested by plotting -log[-log *S(t)*] vs time for strata of each covariate. If the proportionality assumption holds the two (or more) curves should be approximately parallel and should not cross. Alternatively, run a model with each covariate (individually). Introduce a time-dependent interaction term for that covariate. If the proportional hazards assumption is valid for the covariate, the introduction of the time-dependent interaction term won't be significant. This approach is regarded as the most sensitive (and objective) method for testing the proportional hazards assumption.

What do you do if a covariate violates the proportional hazards assumption? The first option is to stratify the model by the offending covariate. This means that a separate baseline hazard function is produced for each level of the covariate. Note you can't obtain a hazard ratio for the covariate you've stratified on because its influence on survival is 'absorbed' into the (two or more) baseline hazard functions in the stratified model. If you are interested in quantifying the effect of the covariate on survival then you should introduce a time-dependent interaction term for the covariate, as described above.

5.3 Residuals

Residuals analysis provide information for evaluating a fitted proportional hazards model. They identify leverage and influence measures and can be used to assess the proportional hazards assumption. By definition, residuals for censored observations are negative and residual plots are useful to get a feeling for the amount of censoring in the data set — large amounts of censoring will result in 'banding' of the residual points. There are three types of residuals:

1. Martingale residuals. Martingale residuals are the difference between the observed number of events for an individual and the conditionally expected number given the fitted model, follow up time, and the observed course of any time-varying covariates. Martingale residuals may be plotted against covariates to detect non-linearity (that is, an incorrectly specified functional form in the parametric part of the model). Martingale residuals are sometimes referred to as Cox-Snell or modified Cox-Snell residuals.
2. Score residuals. Score residuals should be thought of as a three-way array with dimensions of subject, covariate and time. Score residuals are useful for assessing individual influence and for robust variance estimation.
3. Schoenfeld residuals. Schoenfeld residuals are useful for assessing proportional hazards. Schoenfeld residuals provide greater diagnostic power than unscaled residuals. Sometimes referred to as score residuals.

5.4 Overall goodness-of-fit

To assess the overall goodness-of-fit of a Cox proportional hazards regression model Arjas (1988) suggests plotting the cumulative observed versus the cumulative expected number of events for subjects with observed

(not censored) survival times. If the model fit is adequate, then the points should follow a 45° line beginning at the origin. The methodology is as follows: (1) create groups based on covariate values (e.g. treated yes, treated no) and sort on survival time within each group, (2) compute the cumulative sum of the zero-one censoring variable and the cumulative sum of the cumulative hazard function within each group, (3) plot the pairs of cumulative sums within each group only for subjects with an observed survival time.

As in all regression analyses some sort of measure analogous to R^2 may be of interest. Schemper and Stare (1996) show that there is not a single simple, easy to calculate, easy-to-interpret measure to assess the goodness-of-fit of a proportional hazards regression model. Often, a perfectly adequate model may have what, at face value, seems like a very low R^2 due to a large amount of censoring. Hosmer and Lemeshow (1999) recommend the following as a summary statistic for goodness of fit:

$$R^2_M = 1 - exp\left[\frac{2}{n}(L_0 - L_M)\right] \tag{9}$$

Where:

L_0: the log partial likelihood for the intercept-only model,
L_M: the log partial likelihood for the fitted model,
n: the number of cases included.

5.5 Worked examples

Selection of covariates

Load the `survival` library. Read the `addict` data file into R:

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Set contrasts for clinic and prison. Set the reference category for `clinic`, making Clinic 1 (`base = 1`) the reference category:

```
dat$clinic <- factor(dat$clinic, levels = c(1, 2), labels = c("1", "2"))
contrasts(dat$clinic) <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(dat$clinic)
```

Same for `prison`, making absence of a prison record the reference category:

```
dat$prison <- factor(dat$prison, levels = c(0, 1), labels = c("0", "1"))
contrasts(dat$prison) <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(dat$prison)
```

Assess the influence of `clinic`, `prison` and `dose` on days to relapse. First of all categorise `dose` into four classes based on quartiles:

```
quantile(dat$dose, probs = c(0.25, 0.50, 0.75))
hist(dat$dose)
```

Quartiles for `dose` are 50, 60 and 70. Create a categorical variable based on `dose`:

```
cdose <- rep(0, time = nrow(dat))
cdose[dat$dose < 50] <- 1
cdose[dat$dose >= 50 & dat$dose < 60] <- 2
cdose[dat$dose >= 60 & dat$dose < 70] <- 3
cdose[dat$dose >= 70] <- 4
dat <- cbind(dat, cdose)
```

Assess the effect of `clinic`, `prison` and `cdose` on days to relapse:

```
addict.km01 <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier",
data = dat)
addict.km02 <- survfit(Surv(stop, status) ~ prison, type = "kaplan-meier",
data = dat)
addict.km03 <- survfit(Surv(stop, status) ~ cdose, type = "kaplan-meier", data
= dat)
```

Plot all Kaplan-Meier curves on one page. The `mark.time = FALSE` argument disables the censor marks:

```
par(pty = "s", mfrow = c(2,2))
plot(addict.km01, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", main = "Clinic", lty = c(1,2), mark.time = FALSE)
legend(x = "topright", legend = c("Clinic 1","Clinic 2"), lty = c(1,2), bty =
"n", cex = 0.80)
plot(addict.km02, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", main = "Prison", lty = c(1,2), mark.time = FALSE)
legend(x = "topright", legend = c("Prison absent","Prison present"), lty =
c(1,2), bty = "n", cex = 0.80)
plot(addict.km03, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", main = "Dose categories", lty = c(1,2,3,4), mark.time =
FALSE)
legend(x = "topright", legend = c("Dose 1","Dose 2", "Dose 3", "Dose 4"), lty
= c(1,2,3,4), bty = "n", cex = 0.80)
```

Log-rank tests:

```
survdiff(Surv(stop, status) ~ clinic, data = dat, na.action = na.omit, rho =
0)
survdiff(Surv(stop, status) ~ prison, data = dat, na.action = na.omit, rho =
0)
survdiff(Surv(stop, status) ~ cdose, data = dat, na.action = na.omit, rho = 0)
```

The variables `clinic` and `dose` (as a categorical variable) influence days to relapse. The variable `prison` is not significant when tested with a log-rank test (P = 0.28), but since it is considered to be biologically important it is retained in our model.

Fit multivariable model

Days to relapse depends on `clinic`, `prison` and `dose`:

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
summary(addict.cph01)
```

Variables `clinic` and `dose` significantly influence time to relapse (P = 2.6E-06 and 3.1E-08, respectively). Variable `prison` approaching significance (P = 0.06). Drop variable `prison` (using the `update` function):

```
addict.cph02 <- update(addict.cph01, ~. - prison)
summary(addict.cph02)
```

Now include an interaction term for `clinic` and `prison`:

```
addict.cph03 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose + (clinic * prison), method = "breslow", data = dat)
summary(addict.cph03)
```

Does `addict.cph03` provide a better fit to the data than `addict.cph01`?

```
x2 <- -2 * (addict.cph01$loglik[2] - addict.cph03$loglik[2])
1 - pchisq(x2,2)
```

Inclusion of the interaction terms has no significant effect on model fit (P = 0.17). Now compare observed survival (as estimated by the Kaplan-Meier technique) and predictions from Cox model `addict.cph01`. Start by creating two data frames. The first for a subject from clinic 1 without a prison record and on a maximum daily methadone dose of 50 mg. The second for a subject from clinic 2 without a prison record and on a maximum daily methadone dose of 50 mg:

```
addict.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)
dat.clin01 <- data.frame(clinic = factor(c(1)), prison = factor(c(0)), dose =
50)
dat.clin02 <- data.frame(clinic = factor(c(2)), prison = factor(c(0)), dose =
50)

plot(addict.km01, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", lty = c(1,2), mark.time = FALSE, main = "")
lines(survfit(addict.cph01, newdata = dat.clin01), conf.int = FALSE, mark.time
= FALSE, col = "red", lty = 1)
lines(survfit(addict.cph01, newdata = dat.clin02), conf.int = FALSE, mark.time
= FALSE, col = "red", lty = 2)
legend(x = "topright", legend = c("Clinic 1: Kaplan Meier", "Clinic 2: Kaplan
Meier", "Clinic 1: Cox model", "Clinic 2: Cox model"), lty = c(1,2,1,2), col =
c("black", "black", "red", "red"), bty = "n", cex = 0.80)
```

The Cox model underestimates the pattern of relapse, particularly for patients from Clinic 2.

Check scale of continuous covariates (method 1)

Replace the continuous covariate `dose` with design (dummy) variables. Plot the estimated coefficients versus the midpoint of each group:

```
dat$clinic <- factor(dat$clinic, levels = c(1, 2), labels = c("1", "2"))
contrasts(dat$clinic) <- contr.treatment(2, base = 1, contrasts = TRUE)
dat$prison <- factor(dat$prison, levels = c(0, 1), labels = c("0", "1"))
contrasts(dat$prison) <- contr.treatment(2, base = 1, contrasts = TRUE)

cdose <- rep(0, length(dat[,1]))
cdose[dat$dose < 50] <- 1
cdose[dat$dose >= 50 & dat$dose < 60] <- 2
cdose[dat$dose >= 60 & dat$dose < 70] <- 3
cdose[dat$dose >= 70] <- 4
dat <- cbind(dat, cdose)

dat$cdose <- factor(dat$cdose, labels = c("1", "2", "3", "4"))
contrasts(dat$cdose) <- contr.treatment(4, base = 1, contrasts = TRUE)

addict.cph03 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
cdose, method = "breslow", data = dat)
summary(addict.cph03)

addict.cph03$coefficients
addict.cph03$coefficients[3:5]

x <- c(((50 + min(dat$dose))/2), 55, 65,((max(dat$dose) + 70)/2))
y <- c(0, addict.cph03$coefficients[3:5])
plot(x, y, xlim = c(0, 100), type = "l", xlab = "Dose", ylab = "Regression
coefficient")
```

Scale of continuous covariates linear — no transformations required for variable `dose`.

Check scale of continuous covariates (method 2)

Plot covariate values versus the Martingale residuals (and their smooth) from a model that excludes the covariate of interest, `dose`:

```
addict.cph04 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison,
method = "breslow", data = dat)
addict.mg04 <- residuals(addict.cph04, type = "martingale")
plot(dat$dose, addict.mg04, ylim = c(-3,3))
lines(lowess(dat$dose, addict.mg04, iter = 0))
abline(h = 0, lty = 2, col = "gray")
```

Plot of the covariate values versus the log of the ratio of smoothed censor to smoothed cumulative hazard:

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
addict.mg01 <- residuals(addict.cph01, type = "martingale")
addict.hi01 <- dat$status - addict.mg01
addict.clsm01 <- lowess(dat$dose, dat$status, iter = 0)
addict.hlsm01 <- lowess(dat$dose, addict.hi01, iter = 0)
addict.yi01 <- log(addict.clsm01$y / addict.hlsm01$y) +
(addict.cph01$coefficients[3] * dat$dose)
plot(addict.yi01, dat$dose)
```

Now the two plots together:

```
par(pty = "s", mfrow = c(1,2))
plot(dat$dose, addict.mg04, ylim = c(-3,3))
lines(lowess(dat$dose, addict.mg04, iter = 0))
abline(h = 0, lty = 2, col = "gray")
plot(addict.yi01, dat$dose)
```

A linear relationship between the covariate values and each of the calculated parameters is evident, indicating that the continuous variable `dose` is linear in its log hazard.

Testing the proportional hazards assumption

Proportional hazards assumption test based on Schoenfeld residuals:

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
addict.zph <- cox.zph(addict.cph01)

par(pty = "s", mfrow = c(2,2))
plot(addict.zph[1], main = "Clinic"); abline(h = 0, lty = 2)
plot(addict.zph[2], main = "Prison"); abline(h = 0, lty = 2)
plot(addict.zph[3], main = "Dose"); abline(h = 0, lty = 2)
```

The variability band for `clinic` displays a negative slope over time, suggesting non-proportionality of hazards. Formally test the proportional hazards assumption for all variables in `addict.cph01`:

```
cox.zph(addict.cph01, global = TRUE)
```

Using the `cox.zph` function, `rho` is the Pearson product-moment correlation between the scaled Schoenfeld residuals and time. The hypothesis of no correlation is tested using test statistic `chisq`. In the above example, the significant `cox.zph` test for `clinic` ($P < 0.01$) implies that the proportional hazards assumption as been violated for the `clinic` variable. This notion is supported by the Schoenfeld residual plots. An alternative (and less sensitive) means of testing the proportional hazards assumption is to plot log[-log *S(t)*] versus time (or the log of time). If the proportional hazards assumption holds the curves should be reasonably parallel. First we plot log[-log *S(t)*] versus the log of time:

```
clinic.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)

plot(clinic.km, fun = "cloglog", lty = c(1,2), xlab = "Days to relapse (log
scale)", ylab = "Log cumulative hazard")
legend(x = "topleft", legend = c("Clinic 1", "Clinic 2"), lty = c(1,2), bty =
"n")
```

Now plot log[-log *S(t)*] versus time:

```
clinic.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)
clinic <- c(rep(1, times = clinic.km$strata[1]), rep(2, times =
clinic.km$strata[2]))
clinic.haz <- data.frame(clinic, time = clinic.km$time, surv = clinic.km$surv)
clinic1 <- log(-log(clinic.haz$surv[clinic.haz$clinic == 1]))
clinic2 <- log(-log(clinic.haz$surv[clinic.haz$clinic == 2]))

plot(c(clinic.haz$time[clinic.haz$clinic == 1],
clinic.haz$time[clinic.haz$clinic == 2]), c(clinic1, clinic2), type = "n",
ylim = c(-5, 2), xlab = "Days to relapse", ylab = "Log cumulative hazard",
main = "Clinic")
lines(clinic.haz$time[clinic.haz$clinic == 1], clinic1, type = "s", lty = 1)
lines(clinic.haz$time[clinic.haz$clinic == 2], clinic2, type = "s", lty = 2)
legend(x = "topleft", legend = c("Clinic 1", "Clinic 2"), lty = c(1, 2), bty =
"n")
```

We could be talked into concluding that the -log[-log *S(t)*] vs time plots for `clinic` are parallel — conflicting with the findings of the Schoenfield residual test above.

Residuals

Deviance residuals:

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
addict.res <- residuals(addict.cph01, type = "deviance")

par(pty = "s", mfrow = c(2, 2))
boxplot(addict.res ~ dat$clinic, main = "Clinic"); abline(h = 0, lty = 2)
boxplot(addict.res ~ dat$prison, main = "Prison"); abline(h = 0, lty = 2)
plot(dat$dose, addict.res, xlab = "Dose", ylab = "Deviance residual", main =
"Dose"); abline(h = 0, lty = 2)
```

The following plots show the change in each regression coefficient when each observation is removed from the data (influence statistics). The changes plotted are scaled in units of standard errors and changes of less than 0.1 are of little concern. Plot influence statistics (using a common scale for the vertical axis: -0.1 to +0.1):

```
addict.res <- resid(addict.cph01, type = "dfbeta")
par(mfrow = c(2, 2))
main <- c("Clinic", "Prison", "Dose")
for (i in 1:3){
plot(1:238, addict.res[,i], type = "h", ylim = c(-0.1,0.1), xlab =
"Observation", ylab = "Change in coefficient")
    title(main[i])
  }
```

The above plots give an idea of the influence individual observations have on the estimated regression coefficients for each covariate. Data sets where the influence plot is tightly clustered around zero indicate an absence of influential observations. Now plot the Martingale residuals:

```
res <- residuals(addict.cph01, type = "martingale")
X <- dat[,c("clinic", "prison", "dose")]

par(mfrow = c(2,2))
for(j in 1:3){
plot(X[,j], res, xlab = c("Clinic", "Prison", "Dose")[j], ylab = "Martingale
residuals")
  abline(h = 0, lty = 2)
  lines(lowess(X[,j], res))
  }

X$clinic <- as.numeric(levels(X$clinic))[X$clinic]
X$prison <- as.numeric(levels(X$prison))[X$prison]

par(mfrow = c(2,2))
b <- coef(addict.cph01[1:3])
for(j in 1:3){
  plot(X[,j], b[j] * X[,j] + res, xlab = c("Clinic", "Prison", "Dose")[j],
    ylab = "Component + residual")
  abline(lm(b[j] * X[,j] + res ~ X[,j]), lty = 2)
  lines(lowess(X[,j], b[j] * X[,j] + res, iter = 0))
  }
```

Overall goodness of fit

Cox model:

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
summary(addict.cph01)
```

Log partial likelihood for the [intercept-only] model and for the fitted model:

```
addict.cph01$loglik[1]; addict.cph01$loglik[2]
```

Schemper and Stare (1996) R^2:

```
r.square <- 1 - exp( (2/length(dat[,1])) * (addict.cph01$loglik[1] -
addict.cph01$loglik[2]))
r.square
```

Dealing with violation of the proportional hazards assumption

From the analyses conducted so far, we conclude that the proportional hazards assumption has been violated for the variable `clinic`. One method of dealing with this is to stratify the model by `clinic`. This means that we produce a separate baseline hazard function for each level of `clinic`. Note that by stratifying we cannot obtain a hazard ratio for `clinic` since the 'clinic effect' is absorbed into the baseline hazard.

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
addict.cph04 <- coxph(Surv(stop, status, type = "right") ~ strata(clinic) +
prison + dose, method = "breslow", data = dat)
summary(addict.cph04)
```

Compare the `clinic + dose + prison` model with the stratified model:

```
x2 <- 2 * (addict.cph04$loglik[2] - addict.cph01$loglik[2])
1 - pchisq(x2, 1)
```

The stratified model provides a significantly better fit. Parameterising `clinic` as a time dependent covariate would be one option for dealing with non-proportionality of hazards and retaining the ability to quantify the effect of clinic. Plot Kaplan-Meier survival curves for each clinic, adjusting for the effect of `prison` and `methadone` dose:

```
plot(survfit(addict.cph04), lty = c(1,2), xlab = "Days to relapse", ylab =
"Cumulative proportion to experience event")
legend(x = "topright", legend = c("Clinic 1","Clinic 2"), lty = c(1,2), bty =
"n")
```

6 Parametric regression

Semi-parametric models make no assumption about the distribution of failure times, but do make assumptions about how covariates change survival experience. Parametric models, on the other hand, make assumptions about the distribution of failure times and the relationship between covariates and survival experience. Parametric models fully specify the distribution of the baseline hazard/survival function according to some (defined) probability distribution. Parametric models are useful when we want to predict survival rather than identify factors that influence survival. Parametric models can be expressed in: (1) proportional hazard form, where a one unit change in an explanatory variable causes a proportional change in hazard; and (2) accelerated failure time (AFT) form, where a one unit change in an explanatory variable causes a proportional change in survival time. The advantage of the accelerated failure time approach is that the effect of covariates on survival can be described in absolute terms (e.g. numbers of years) rather than relative terms (a hazard ratio).

6.1 Exponential model

The exponential model is the simplest type of parametric model in that it assumes that the baseline hazard is constant over time:

$$h(t) = h_0 \text{ exp}^{\beta X} \text{ where } h_0 = \lambda \tag{10}$$

The assumption that the baseline hazard is constant over time can be evaluated in several ways. The first method is to generate an estimate of the baseline hazard from a Cox proportional hazards model and plot it to check if it follows a straight, horizontal line. A second approach is to fit a model with a piecewise-constant baseline hazard. Here, the baseline hazard is allowed to vary across time intervals (by including indicator variables for each of the time intervals). The baseline hazard is assumed to be constant within each time period, but can vary between time periods.

6.2 Weibull model

In a Weibull model it is assumed that the baseline hazard has a shape which gives rise to a Weibull distribution of survival times:

$$h(t) = h_0 \text{ exp}^{\beta X} \text{ where } h_0 = \lambda p t^{p-1} \tag{11}$$

Where βX includes an intercept term β_0. The suitability of the assumption that survival times follow a Weibull distribution can be assessed by generating a log-cumulative hazard plot. If the distribution is Weibull, this function will follow a straight line. The estimated shape parameter from the Weibull model gives an indication of whether hazard is falling ($p < 1$), constant ($p = 1$), or increasing ($p > 1$) over time.

6.3 Accelerated failure time models

The general form of an accelerated failure time model is:

$$log(t) = \beta X + log(\tau) \text{ or } t = \text{exp}^{\beta X} \tau \tag{12}$$

Table 5: Accelerated failure time model showing the effect of clinic, methadone dose and prison status on expected retention time on the program (adapted from Caplehorn and Bell 1991). Note that the term 'hazard' in the last column of the table is replaced with 'survival.'

Variable	Subjects	Failed	Coefficient (SE)	P	Survival (95% CI)
Intercept	238	250	4.7915 (0.2782)	< 0.01	
Clinic				< 0.01	
Clinic 1	163	122			1.0
Clinic 2	75	28	0.7198 (0.1595)		2.05 (1.50 – 2.81) [b]
Prison record				0.07	
Absent	127	81			1.0
Present	111	69	-0.2232 (0.1224)		0.80 (0.63 – 1.02)
Dose	238	150	0.0247 (0.0046)	< 0.01	1.02 (1.01 – 1.03)

[a] The significance of inclusion of the two clinic variables in the model.
[b] Intepretation: after adjusting for the effect of methadone dose and prison status retention time for patients from Clinic 2 was twice that of patients from Clinic 1 (95% CI 1.50 – 2.81).
SE: standard error.
CI: confidence interval.

where log(*t*) is the natural log of the time to failure event, βX is a linear combination of explanatory variables and log (τ) is an error term. Using this approach τ is the distribution of survival times when βX = 0. If we assume that τ follows a log-normal distribution, then the log of survival times will have a normal distribution, which is equivalent to fitting a linear model to the natural log of survival time (assuming that you can ignore the problem of dealing with censored observations). Equation 12 can be re-expressed as follows:

$$\tau = \text{exp}^{-\beta X} t \text{ or } ln(\tau) = -\beta X + log(t) \tag{13}$$

The linear combination of predictors in the model (βX) can act additively or multiplicatively on the log of time: they speed up or slow down time to event by a multiplicative factor. In this case $\text{exp}^{-\beta X}$ is called the acceleration parameter such that if $\text{exp}^{-\beta X} > 1$ time passes more quickly, if $\text{exp}^{-\beta X} = 1$ time passes at a normal rate, and if $\text{exp}^{-\beta X} < 1$ time passes more slowly.

Exponential and Weibull models can be parameterised as either proportional hazards models or as accelerated failure time models. Other parametric models (e.g. the log-normal, the log-logistic, and gamma) can only be expressed as accelerated failure time models (the predictors in these models do not necessarily multiply the baseline hazard by a constant amount).

Accelerated failure time coefficients represent the expected change in ln(*t*) for a one unit change in the predictor. Consider an accelerated failure time model fitted to the `addict` data, as shown in Table 5. What was the effect of being treated at Clinic 2 in terms of additional retention time?

log(t) = 4.7915 + (0.7198 $\times$ 1)
log(t) = 5.5113
t = exp(5.5113)
t = 247 days

Being treated at Clinic 2 extended retention time by 247 days.

6.4 Worked examples

Exponential and Weibull models

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Set contrasts for clinic and prison. Set the reference category for `clinic`, making Clinic 1 (`base = 1`) the reference category:

```
dat$clinic <- factor(dat$clinic, levels = c(1, 2), labels = c("1", "2"))
contrasts(dat$clinic) <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(dat$clinic)
```

Same for `prison`, making absence of a prison record the reference category:

```
dat$prison <- factor(dat$prison, levels = c(0, 1), labels = c("0", "1"))
contrasts(dat$prison) <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(dat$prison)
```

Cox proportional hazards model (for comparison):

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
summary(addict.cph01)
```

Exponential model:

```
addict.exp01 <- survreg(Surv(stop, status, type = "right") ~ clinic + prison +
dose, dist = "exponential", data = dat)
summary(addict.exp01)

shape.exp = 1 / addict.exp01$scale
shape.exp
```

R outputs the parameter estimates for the accelerated failure time form of the exponential model. We multiply the estimated regression coefficients by -1 to get estimates consistent with the Cox proportional hazards parameterisation. For example, the estimated acclceration factor comparing a patient with a prison record to one without a prison record is exp(-0.2526) = 0.78. Having a prison record accelerates time to event by a factor of 0.78. The estimated hazard ratio comparing a patient with a prison record to one without a prison record is exp(-1 $\times$ 0.2526) = 1.29. Having a prison record increases the daily hazard of relapse by a factor of 1.29. We now run a Weibull model:

```
addict.wei01 = survreg(Surv(stop, status, type = "right") ~ clinic + prison +
dose, dist = "weibull", data = dat)
summary(addict.wei01)

shape.wei = 1 / addict.wei01$scale
shape.wei
```

The acceleration factor comparing clinic 2 with clinic 1 patients is exp(0.8806) = 2.41. The estimated median survival time (time off heroin) is double for patients from clinic 2 compared with those from clinic 1. We now plot the estimated survival for patients attending clinic 1 and clinic 2. Both have a prison record and received a maximum dose of 50 mg methadone per day:

```
addict.km01 <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier",
data = dat)
dat.clin01 <- data.frame(clinic = as.factor(c(1)), prison = as.factor(c(0)),
dose = 50)
dat.clin02 <- data.frame(clinic = as.factor(c(2)), prison = as.factor(c(0)),
dose = 50)

days.clin01 <- predict(addict.wei01, newdata = dat.clin01, type = "quantile",
p = 0:100/100)
days.clin02 <- predict(addict.wei01, newdata = dat.clin02, type = "quantile",
p = 0:100/100)
surv.wei01 <- 1 - 0:100/100

plot(addict.km01, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", lty = c(1,2), mark.time = FALSE, main = "")
lines(days.clin01, surv.wei01, type = "s", col = "red", lty = 1)
lines(days.clin02, surv.wei01, type = "s", col = "red", lty = 2)
legend(x = "topright", legend = c("Clinic 1: Kaplan Meier","Clinic 2: Kaplan
Meier", "Clinic 1: Weibull model", "Clinic 2: Weibull model"), lty =
c(1,2,1,2), col = c("black", "black", "red", "red"), bty = "n", cex = 0.80)
```

Now compare the three models using AIC:

```
extractAIC(addict.cph01)
extractAIC(addict.exp01)
extractAIC(addict.wei01)
```

The AIC for the Cox model is the smallest, indicating that this model provides the best fit with the data (this is consistent with the diagnostics we ran earlier to assess how consistent the data was with the exponential and Weibull distributions).

Accelerated failure time models

Here we use the `psm` function in the `rms` library to develop an AFT model. The `psm` function is a modification of `survreg` and is used for fitting the accelerated failure time family of parametric survival models.

```
library(rms)
addict.aft01 <- psm(Surv(stop, status) ~ clinic + prison + dose, dist =
"weibull", data = dat)
addict.aft01
```

Compare `addict.aft01` with `addict.wei01`:

```
addict.aft01 <- psm(Surv(stop, status) ~ clinic + prison + dose, dist =
"weibull", data = dat)
addict.aft01
```

What is the effect of Clinic 2 on retention time (after adjusting for the effect of presence of a prison record and maximum daily methadone dose)?

```
exp(addict.aft01$coefficients[2])
```

Treatment at Clinic 2 doubles patient retention time. What does this mean in terms of calendar time?

```
log.t <- as.numeric(addict.aft01$coefficients[1]) +
(as.numeric(addict.aft01$coefficients[2]) * 1)
exp(log.t)
```

Treatment at Clinic 2 results in patients remaining on the program for an additional 250 days (compared with those treated at Clinic 1).

7 Time dependent covariates

As discussed, stratification as a method for dealing with a covariate that violates the proportional hazards assumption is not an option when you want to include it in a model, in order to describe and/or test its effect on survival. In this case it can be useful to look at a plot of the hazard function versus time for different strata of the variable. This may indicate the type of deviation from proportional hazards that is occurring. Two common types of departure from proportional hazards in clinical situations are: (1) the time to peak hazard varies between prognostic groups, i.e. strata, or (2) the influence of a covariate diminishes with time.

If the covariate is fixed (i.e. the covariate itself does not change over time, but its *effect* varies over time) we can explore this time-dependent effect by dividing the time period into distinct intervals. We then fit proportional hazards models to the survival in each interval and compare the coefficients for each covariate across the different time intervals. If the coefficient changes with time, we have evidence of non-proportional hazards. This approach has been called the step function proportional hazards or piecewise Cox model. In each interval, patients who die or who are censored before the interval are treated as usual — i.e. coded as censored or died and survival times for patients who live through the interval to the next one are censored at the end of the interval. The number and length of the intervals is arbitrary, but each interval should contain enough deaths to enable regression coefficients to be estimated reliably.

Covariates themselves may change over time. In this case the survival period for each individual is divided up into a sequence of shorter 'survival spells', each characterised by an entry and an exit time, and within which covariate values remain fixed. Thus the data for each individual are represented by a number of shorted censored intervals and possibly one interval ending with the event of interest (death, for example).

It may be thought that the observations, when organised in this way, are 'correlated' and so not suitable for Cox regression. Fortunately, this is not an issue, since the partial likelihood on which estimation is based has a term for each unique death or event time, and involves sums over those observations that are available or at risk at the actual event date. Since the intervals for a particular individual do not overlap, the likelihood will involve at most only one of the observations for the individual, and so will be based on independent observations. The values of the covariates between event times do not enter the partial likelihood.

7.1 Worked examples

Piecewise Cox models

Load the `survival` library. Read the `addict` data file (in counting format) into R:

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Recode the `addict` data set into counting process (start, stop) format. Here we define a cutpoint at day 365 creating a new data frame called `dat.cp`:

```
dat.cp <- survSplit(data = dat, cut = 365, end = "stop", event = "status",
start = "start", id = "id")
```

Recode the `clinic` variable to make Clinic 2 the reference category:

```
dat.cp$clinic <- as.vector(ifelse(dat.cp$clinic == 2, 0, 1))
```

First of all we'll consider the period before 365 days. Create a new variable called `t1` such that `t1 = 1` if the time to event is less than or equal to 365 days and zero otherwise:

```
t1 <- rep(0, nrow(dat.cp))
t1[dat.cp$stop <= 365] <- 1
dat.cp$t1 <- t1
```

Interpretation of the `clinic` $\times$ `t1` interaction is as follows:

Clinic (code)	Time (code)	Interaction
Clinic 1 (1)	Less than 365 (1)	$1 \times 1 = 1$
Clinic 1 (1)	Greater than 365 (0)	$1 \times 0 = 0$
Clinic 2 (0)	Less than 365 (1)	$0 \times 1 = 0$
Clinic 2 (0)	Greater than 365 (0)	$0 \times 0 = 0$

Using this coding, the reported hazard for the `clinic` $\times$ `t1` interaction will be for Clinic 1 when time is less than or equal to 365 days. Next consider the period after 365 days. Create a new variable called `t2` such that `t2 = 1` if the time to event is greater than 365 days and one otherwise:

```
t2 <- rep(0, nrow(dat.cp))
t2[dat.cp$stop > 365] <- 1
dat.cp$t2 <- t2
```

Interpretation of the `clinic` $\times$ `t2` interaction is as follows:

Clinic (code)	Time (code)	Interaction
Clinic 1 (1)	Less than 365 (0)	$1 \times 0 = 0$
Clinic 1 (1)	Greater than 365 (1)	$1 \times 1 = 1$
Clinic 2 (0)	Less than 365 (0)	$0 \times 0 = 0$
Clinic 2 (0)	Greater than 365 (1)	$0 \times 1 = 0$

Using this coding, the reported hazard for the `clinic` $\times$ `t2` interaction will be for Clinic 1 when time is greater than 365 days. Now fit the model:

```
addict.cph05 <- coxph(Surv(start, stop, event = status, type = "counting") ~
prison + dose + I(clinic * t1) + I(clinic * t2), method = "breslow", data =
dat)
summary(addict.cph05)
```

The effect of `clinic` is borderline when days on treatment is less than 365 days (P = 0.06). When days on treatment is greater than 365 days the effect of `clinic` is highly significant (P 0.01). Note that the confidence interval for days on treatment 365 days is considerably larger than that for days on treatment 365 days. These results suggest a large difference in clinic survival times after one year on treatment with Clinic 2 always doing better than Clinic 1 at any one time.

Table 6: Cox proportional hazards regression model showing the effect of prison, methadone dose and the variable effect of time on clinic on the daily hazard of relapse.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio (95% CI)
Clinic					
Clinic $\times$ t1	118	87	0.4802 (0.2548)	0.06	1.62 (0.98 – 2.66) [a]
Clinic $\times$ t2	120	63	1.8103 (0.3861)	$<$ 0.01	6.11 (2.87 – 13.03)
Prison	238	150	0.3650 (0.1684)	0.03	1.44 (1.04 – 2.00)
Dose	238	150	-0.0353 (0.0064)	$<$ 0.01	0.96 (0.95 – 0.98)

[a] Interpretation: compared with the reference category (patients from Clinic 2) when days on treatment is less than 365, after adjusting for the effect of methadone dose and prison record, patients from Clinic 1 had 1.62 (95% CI 0.98 – 2.66) times the daily hazard withdrawing from the treatment program.
SE: standard error.
CI: confidence interval.

Now compare observed survival (as estimated by the Kaplan-Meier technique) and predictions from the Cox model:

```
addict.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)
dat.clin01 <- data.frame(clinic = 1, prison = factor(c(0)), dose = 50, t1 = 1,
t2 = 0)
dat.clin02 <- data.frame(clinic = 0, prison = factor(c(0)), dose = 50, t1 = 1,
t2 = 0)

plot(addict.km01, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", lty = c(1,2), mark.time = FALSE, main = "")
lines(survfit(addict.cph05, newdata = dat.clin01), conf.int = FALSE, mark.time
= FALSE, col = "red", lty = 1)
lines(survfit(addict.cph05, newdata = dat.clin02), conf.int = FALSE, mark.time
= FALSE, col = "red", lty = 2)
legend(x = "topright", legend = c("Clinic 1: Kaplan Meier","Clinic 2: Kaplan
Meier", "Clinic 1: Cox model", "Clinic 2: Cox model"), lty = c(1,2,1,2), col =
c("black", "black", "red", "red"), bty = "n", cex = 0.80)
```

The Cox model with a time dependent covariate provides a reasonable fit to the survival experience of patients from Clinic 2. Survival is underestimated for Clinic 1 patients up to day 550, then survival is substantially overestimated. Introduction of a third time period at around day 550 might improve this problem.

Counting process formulation

The most common type of time-dependent covariates are repeated measurements on a subject or a change in the subject's treatment. Both of these situations are easily handled by the counting process formulation. As an example consider the Stanford heart transplant study where the objective is to determine if patients receiving transplants live longer than those that don't (Crowley and Hu 1977). This analysis is a little tricky, because patients who are enrolled into the study spend a variable amount of time on a waiting list (that is, waiting for a suitable donor).

Consider the first three patients (details shown below). Patient 1 entered the program on day 0 and died, waiting for a transplant on day 50. Patient 2 entered the program on day 0 and died on 6 (again, while waiting for a transplant). Patient 3 entered on day 0, spent 1 day on the waiting list and was then transplanted. This patient died 16 days later.

id	start	stop	event	age	year	surgery	transplant
1	0	50	1	-17.155	0.123	0	0
2	0	6	1	3.836	0.255	0	0
3	0	1	0	6.297	0.266	0	0
3	1	16	1	6.297	0.266	0	1

id: patient identifier.
start: days from entry into program.
stop: days from entry into program to event.
event: 0 = censored, 1 = died.
age: age in years minus 48.
year: date of acceptance into the program (years from 1 October 1967).
surgery: 0 = no previous surgery, 1 = previous surgery.
transplant: 0 = no transplant, 1 = transplant.

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("heart.csv", header = TRUE, sep = ",")
```

The first thing we could do is to stratify the analysis by transplant status, that is to consider `transplant` as a time-dependent strata:

```
heart.cph01 <- coxph(Surv(start, stop, event, type = "counting") ~ age +
surgery + strata(transplant), data = dat, method = "breslow")
summary(heart.cph01)
```

Table 7: Stratified Cox proportional hazards regression model showing the effect of age and prior surgery on the daily hazard of death.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio(95% CI)
Age	172	75	0.0321 (0.0141)	0.023	1.03 (1.00 - 1.06) $^{\alpha}$
Surgery	172	75	-0.7678 (0.3619)	0.03	0.46 (0.23 – 0.94)

R^2 = 0.062.

$^{\alpha}$ Interpretation: for yearly increases in the age at enrolment the daily hazard of death was 1.03 (95% CI 1.00 – 1.06).

An alternative is to consider `transplant` as a time-dependent covariate (rather than as a time-dependent strata). This provides a direct way of making comparisons between failure rates of the transplanted and untransplanted groups. In this model the effect of `transplant` status on each covariate (that is `age` and `surgery`) is assessed:

```
heart.cph02 <- coxph(Surv(start, stop, event, type = "counting") ~ (age +
surgery) * transplant, data = dat, method = "breslow")
summary(heart.cph02)
```

Now assess the effect of year of entry into the program:

```
heart.cph03 <- coxph(Surv(start, stop, event, type = "counting") ~ (age +
year) * transplant, data = dat, method = "breslow")
summary(heart.cph03)
```

A feature of these data is the dependence of survival time on the date of acceptance into the study. Unfortunately, the date of acceptance interaction with transplantation is also approaching significance but in the opposite direction. Together, these suggest that the overall quality of patient being admitted to the study was improving over calendar time, but the survival time of transplanted patients was not improving at the same rate. In fact, the sum of the two coefficients for year of acceptance would suggest a nearly constant survival pattern for the transplanted patients.

Table 8: Cox proportional hazards regression model showing the effect of age, prior surgery, transplantation status, and the age-transplantation, prior surgery-transplantation interaction on the daily hazard of death.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio (95% CI)
Age	172	75	0.0138 (0.0181)	0.45	1.01 (0.98 – 1.05)
Surgery	172	75	-0.5457 (0.6109)	0.37	0.58 (0.17 – 1.92)
Transplant	172	75	0.1181 (0.3277)	0.72	1.12 (0.59 – 2.14)
Age $\times$ transplant	172	75	0.0348 (0.0273)	0.20	1.03 (0.98 – 1.09) [a]
Surgery $\times$ transplant	172	75	-0.2916 (0.7582)	0.70	0.75 (0.17 - 3.30) [b]

R^2 = 0.070.

[a] Interpretation: for yearly increases in the age at enrolment the daily hazard of death after transplantation was 1.03 (95% CI 0.98 – 1.09).

[b] Interpretation: where a patient had prior surgery, the daily hazard of death after transplantation was 0.75 (95% CI 0.17 – 3.30).

Table 9: Cox proportional hazards regression model showing the effect of age, prior surgery, transplantation status, and the age-transplantation, prior surgery-transplantation interaction on the daily hazard of death.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio(95% CI)
Age	172	75	0.0155 (0.0173)	0.37	1.02 (0.98 – 1.05)
Year	172	75	-0.2735 (0.1058)	$<$ 0.01	0.76 (0.62 – 0.94)
Transplant	172	75	-0.5884 (0.5427)	0.28	0.55 (0.19 – 1.61)
Age $\times$ transplant	172	75	0.0339 (0.0279)	0.23	1.03 (0.98 – 1.09) [a]
Surgery $\times$ transplant	172	75	0.2013 (0.1425)	0.16	1.22 (0.92 – 1.62) [b]

R^2 = 0.083.

[a] Interpretation: for yearly increases in the age at enrolment the daily hazard of death after transplantation was 1.03 (95% CI 0.98 – 1.09).

[b] Interpretation: for yearly increases in the date of acceptance into the program, the daily hazard of death after transplantation was 1.22 (95% CI 0.92 – 1.62).

Compare the survival of two patients. Both have prior surgery and are enrolled into the program on 1 October 1967. The first patient is 30 years of age at enrolment, the second patient is 50 years of age. Both have not received a transplant.

```
dat.age30 <- data.frame(start = 0, stop = 183, age = 30 - 48, year = 0,
surgery = 1, transplant = 0)
dat.age50 <- data.frame(start = 0, stop = 183, age = 50 - 48, year = 0,
surgery = 1, transplant = 0)

plot(survfit(heart.cph03, newdata = dat.age30), xlim = c(0,200), xlab =
"Survival (days)", ylab = "Cumulative proportion to experience event", lty =
1, conf.int = FALSE, mark.time = FALSE, main = "")
lines(survfit(heart.cph03, newdata = dat.age50), conf.int = FALSE, mark.time =
FALSE, col = "red", lty = 1)
legend(x = "topright", legend = c("30 years of age, prior surgery", "50 years
of age, prior surgery"), col = c("black", "red"), lwd = 1, lty = 1, bty = "n")
```

8 Correlated data

The ordinary Cox model treats each outcome event as being independent. When the survival likelihood for subjects is not independent (say subjects are grouped into herds, regions, households) the assumption of independence may not hold. Failure to account for this lack of independence means that the variance of computed regression coefficients will be underestimated, making us more likely to make a Type I error (rejecting the null hypothesis when, in fact, it is really true). Two techniques may be applied to account for non-independent (i.e. correlated) data: (1) robust sandwhich estimators, and (2) frailty terms.

8.1 Robust variance

With robust variance estimates the observed data is resampled (with replacement) to achieve a sample of the same size each time, and to use the variation in the estimated parameters across the set of samples to obtain a value for the sampling variability of the estimates. With correlated data the sample needs to be drawn with replacement from the set of independent subjects (not observations) so that intra-subject correlation is preserved in the samples that are taken.

In R, the `cluster` argument within the `cph` function is used to compute a robust variance for a Cox proportional hazard regression model.

This example is taken from a paper by Wei, Lin and Weissfeld (1989). The study is of time to recurrence of bladder cancer. The data frame `bladder2` has multiple rows for each subject. Many subjects had recurrences of bladder cancer, sometimes as many as four, and were followed beyond the fourth recurrence. Reference: Wei LJ, Lin DY, Weissfeld L (1989) Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. Journal of the American Statistical Association. 84:1065-1073.

id	rx	size	number	start	stop	event	enum
1	1	3	1	0	1	0	1
2	1	1	2	0	4	0	1
3	1	1	1	0	7	0	1
4	1	1	5	0	10	0	1
5	1	1	4	0	6	1	1
5	1	1	4	6	10	0	2

id: patient identifier.
rx: 1 = placebo, 2 = thiopet.
size: size of the largest initial tumour.
number: number of initial tumours.
start: entry into the study or the time of last recurrence.
stop: time to event (months).
event: 0 = censored, 1 = event.
enum: event number.

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("bladder2.csv", header = TRUE, sep = ",")

bladder.cph01 <- coxph(Surv(start, stop, event) ~ rx + size + number, data =
dat)
summary(bladder.cph01)

bladder.cph02 <- coxph(Surv(start, stop, event) ~ rx + size + number +
cluster(id), data = dat)
summary(bladder.cph02)
```

Table 10: Cox proportional hazards regression model showing the effect of treatment, tumour size, and number of initial tumours on the monthly hazard of tumour recurrence.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Treatment	190	112	-0.4116 (0.1999)	0.04	0.66 (0.45 – 0.98)
Size	190	112	-0.0411 (0.0703)	0.56	0.96 (0.84 – 1.10)
Number	190	112	0.1637 (0.0478)	< 0.01	1.18 (1.07 – 1.29)

R^2 = 0.074.

Likelihood ratio test = 14.7 on 3 df, P < 0.01.

Table 11: Cox proportional hazards regression model showing the effect of treatment, tumour size, and number of initial tumours on the monthly hazard of tumour recurrence (robust variance).

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Treatment	190	112	-0.4116 (0.2488)	0.04	0.66 (0.41 – 1.08)
Size	190	112	-0.0411 (0.0742)	0.58	0.96 (0.83 – 1.11)
Number	190	112	0.1637 (0.0584)	< 0.01	1.18 (1.05 – 1.32)

R^2 = 0.074.

Likelihood ratio test = 14.7 on 3 df, P < 0.01.

8.2 Frailty

In recent times there has been active research concerning the addition of random effects to survival models. In this setting, a random effect is a continuous variable that describes excess risk or frailty for distinct categories such as individuals, families or herds. The idea is that individuals have different frailties, and those who are most 'frail' will experience failure earlier than others. The inclusion of the frailty term in a Cox model allows for the possible correlation between the recurrence times of an individual. The model that could be fitted is as follows:

$$h(t, X) = h_0(t)\text{exp}\left(\beta_1 \text{age}_{ij} + \beta_2 \text{sex}_{ij} + u_j\right) \tag{14}$$

Above, the term u_j represents the frailty term for the j^{th} cluster. Frailty terms can be specified as following a normal or gamma distribution. Under the normal assumption, frailty estimates are symmetric around zero. Under the gamma assumption, frailty estimates are asymmetric (allowing for groups displaying exceptionally low or high risk — as in the case of genetic disease studies where the presence of a high-risk allele markedly increases the hazard of failure).

Table 12: Fixed-effects Cox proportional hazards regression model showing the effect of sex, physician estimate of Karnofsky score and patient estimate of Karnofsky score on the daily hazard of death.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio (95% CI)
Sex	228	165	-0.5118 (0.1693)	< 0.01	0.60 (0.43 – 0.83)
Physician karno	228	165	-0.0062 (0.0068)	0.37	0.99 (0.98 – 1.01)
Patient karno	228	165	-0.0170 (0.0065)	< 0.01	0.98 (0.97 – 1.00)

R^2 = 0.097.

Likelihood ratio test = 22.9 on 2 df, P < 0.01.

As an illustration, consider the data in the `lung.csv` file. These data relate to mortality among advanced lung cancer patients, conducted by the North Central Cancer Treatment Group. The subset used here contains details for 228 patients. Reference: Loprinzi CL, Laurie JA et al. (1994) Prospective evaluation of prognostic variables from patient-completed questionnaires. J. Clinical Oncol. 12:601-607.

inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
3	306	2	74	1	1	90	100	1175	.
3	455	2	68	1	0	90	90	1225	15 3
1010	1	56	1	0	90	90	.	15	

inst: enrolling institution.
time: day of event.
status: 1 = alive, 2 = dead.
age: patient age at enrolment.
sex: 1 = male, 2 = female.
ph.ecog: physician's estimate of ECOG performance score (0 to 4).
ph.karno: physician's estimate of Karnofsky score (an alternative to ECOG performance score). Values are 20, 30, ... , 100.
pat.karno: patient's estimate of Karnofsky score.
meal.cal: calories consumed at meals (excluding beverages and snacks).
wt.loss: weight loss in the last six months (a negative number implies weight gain).

There are 18 separate institutions that enrolled at least one subject in this trial. Because the enrolling institutions range from community practices to a large tertiary care centre, differences in the baseline risk of enrollees might be a concern. A fixed-effects Cox proportional hazards model (i.e. one that ignores intra-institutional correlation) would be called as follows:

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("lung.csv", header = TRUE, sep = ",")
\end{Schunk} \begin{Schunk}
\begin{Sinput}
lung.cph01 <- coxph(Surv(time, status) ~ sex + ph.karno + pat.karno, data =
dat)
```

Include enrolling institution as fixed effect:

```
dat$inst <- factor(dat$inst)
contrasts(dat$inst) <- contr.treatment(18, base = 1, contrasts = TRUE)
lung.cph02 <- coxph(Surv(time, status) ~ sex + ph.karno + pat.karno + inst,
data = dat)
summary(lung.cph02)
```

Treating enrolling institution as a categorical variable (that is, a factor) is a satisfactory approach when there is only a small number of them. Presentation of the model becomes clumsy when there are large numbers (as in the example above). An alternative is to treat the variable `inst` as a frailty term:

```
lung.cph03 <- coxph(Surv(time, status) ~ sex + ph.karno + pat.karno +
frailty(inst), data = dat)
summary(lung.cph03)
```

The variance of the frailty term is 0.00149. The p-value for the frailty term is 0.36. We conclude that the frailty term is not significant, that is the variance of the frailty term does not differ significantly from zero.

Table 13: Fixed-effects Cox proportional hazards regression model showing the effect of sex, physician estimate of Karnofsky score, patient estimate of Karnofsky score and enrolling institution on the daily hazard of death.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio (95% CI)
Sex	190	112	-0.5197 (0.1763)	< 0.01	0.59 (0.42 – 0.84)
Physician karno	190	112	-0.0108 (0.0073)	0.14	0.99 (0.97 – 1.00)
Patient karno	190	112	-0.0190 (0.0069)	< 0.01	0.98 (0.97 – 0.99)
Institution:					
Inst 2	5	4	0.5159 (0.5427)	0.34	1.67 (0.58 – 4.85)
Inst 3	19	15	-0.4168 (0.3315)	0.21	0.66 (0.34 – 1.26)
...					
Inst 33	2	1	-2.9509 (9.2902)	0.75	0.05 (0.00 - 4.23E+06)

R^2 = 0.173.

Likelihood ratio test = 42.4 on 20 df, P < 0.01.

Table 14: Mixed-effects Cox proportional hazards regression model showing the effect of sex, physician estimate of Karnofsky score and patient estimate of Karnofsky score on the daily hazard of death. Identity of the patient's enrolling institution is trated as a frailty term.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard ratio (95% CI)
Sex	228	165	-0.5098 (0.1695)	< 0.01	0.60 (0.43 - 0.84)
Physician karno	228	165	-0.0062 (0.0068)	0.37	0.99 (0.98 – 1.00)
Patient karno	228	165	-0.0170 (0.0065)	< 0.01	0.98 (0.97 – 1.00)

R^2 = 0.099.

Variance of random effect: 0.00149.

Likelihood ratio test = 23.2 on 3.21 df, P < 0.01.

A ranked error bar plot of the frailty terms for each institution allow us to better-appreciate those institutiions that are prone to ‘earlier failures’ than others, as shown in Figure 9.

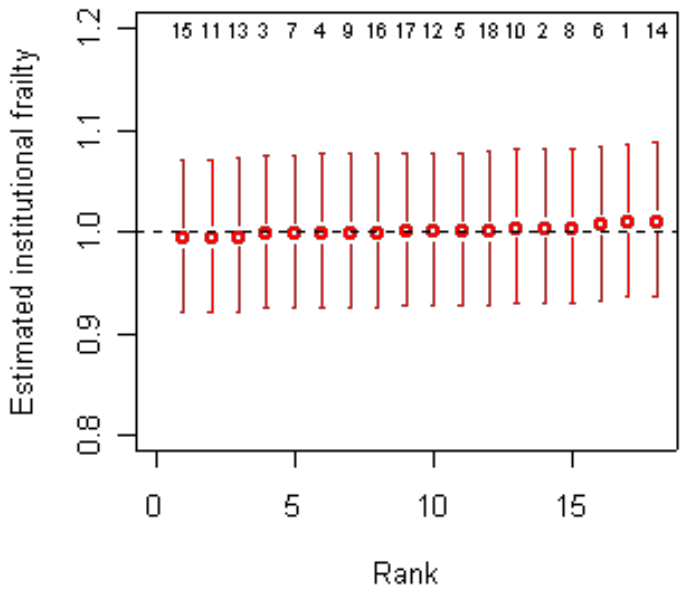


Figure 9: Influence of institution on the daily hazard of death for patients enrolled in a study of advanced lung cancer patients, conducted by the North Central Cancer Treatment Group. Institution identifiers are shown at the top of the plot. The effect of institution on hazard of death is subtle, with institutions 15, 11, and 3 demonstrating the lowest daily hazard of failure. Confidence intervals are wide, demonstrating no clear evidence that one institution is associated with lower or higher hazards of failure over another.

9 Penalised Cox models

An alternative to parameterising regression models using terms applied in an additive fashion is to use indicator variables or polynomials. A particularly useful class of functions for this purpose is smoothing splines.

Polynomials are one of the easiest smooth functions to fit: simply add $x, x^2, x^3, ...$ to the right-hand side of the model equation. Polynomials have major flaws however. Firstly, the data fits are not local and secondly the fitting process for polynomials can be numerically ill-conditioned. Penalised Cox models offer a means for avoiding this problem by fitting non-parametric functions (for example, spline smoothers) to account for relationships between explanatory and outcome variables. A nice feature of this technique is that results can be displayed graphically to illustrate the multivariable functional form of these relationships (e.g. linear, quadratic or cubic).

Spline curves have a natural approximate analogue: imagine one were to outline a shape by placing a small number of nails onto a wooden board, and then interpolating them with a thin flexible metal strip (or spline). Spline curves have several important properties. The first is locality of influence. The second is that useful property is that the curve can be constrained to be linear beyond the last control point. The final important property is that splines can be fit using existing programs simply by creating an appropriate set of dummy variables, based on the prespecified locations of the control points (the nails). These control points are called knots and the dummy variables are known as basis functions. One can then regress the data on the basis of the dummy variables. The degrees of freedom for the fit is given by the number of basis functions, equal to the number of fitted regression coefficients. For regression splines, the degrees of freedom equals the number of knots plus 1; for natural splines, it is one less than the number of knots.

The data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants. Missing data items are denoted by a period. Reference: Fleming TR, Harrington DP (1991). Counting Processes and Survival Analysis. Comm. Stat. Theory, Methods, 13:2469 - 2486.

id	futime	status	age	bili
1	400	2	21464	14.5
2	4500	0	20617	1.1
3	1012	2	25594	1.4
4	1925	2	19994	1.8
5	1504	1	13918	3.4

id: patient identifier.

futime: follow-up time (number of days between registration and the earlier of death, transplantion, or study termination date in July 1986).

status: 0 = alive, 1 = liver transplant, 2 = dead.

age: age of patient at enrolment (in days).

bili: serum bilirubin (mg/dL) at enrolment.

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("pbc.csv", header = TRUE, sep = ",")

pbc.cph01 <- coxph(Surv(futime, status == 2) ~ age + bili, data = dat)
summary(pbc.cph01)
```

Table 15: Fixed-effects Cox proportional hazards regression model showing the effect of age and bilirubin concentration on the daily hazard of death in primary biliary cirrhosis patients.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Age	418	161	0.0001 (0.00002)	< 0.01	1.00 (1.00 – 1.00)
Bilirubin	418	161	0.1436 (0.0114)	< 0.01	1.15 (1.13 – 1.18)

R^2 = 0.260.

Likelihood ratio test = 126 on 2 df, P < 0.01.

Check the scale of `bili`:

```
quantiles <- signif(quantile(dat$bili, probs = c(0.25, 0.50, 0.75)), digits =
3)
hist(dat$bili)
```

Quartiles for `bili` are 0.8, 1.4, and 3.4. Create a categorical variable based on `bili`:

```
bili.cat <- rep(0, length(dat[,1]))
bili.cat[dat$bili < quantiles[1]] <- 1
bili.cat[dat$bili >= quantiles[1] & dat$bili < quantiles[2]] <- 2
bili.cat[dat$bili >= quantiles[2] & dat$bili < quantiles[3]] <- 3
bili.cat[dat$bili >= quantiles[3]] <- 4
dat <- cbind(dat, bili.cat)
dat$bili.cat <- factor(dat$bili.cat, labels=c("1", "2", "3", "4"))
contrasts(dat$bili.cat) <- contr.treatment(4, base = 1, contrasts = TRUE)

pbc.cph02 <- coxph(Surv(futime, status == 2) ~ age + bili.cat, method =
"breslow", data = dat)
summary(pbc.cph02)

pbc.cph02$coefficients
pbc.cph02$coefficients[2:4]
x <- c((quantiles[1] + min(dat$bili))/2, (quantiles[1] + quantiles[2])/2,
(quantiles[2] + quantiles[3])/2, (max(dat$bili) + quantiles[3])/2)
y <- c(0, pbc.cph02$coefficients[2:4])
plot(x, y, type = "l", xlim = c(0,25), ylim = c(0,4), xlab = "Serum bilirubin
(mg/dL)", ylab = "Regression coefficient (log hazard)")
```

On the basis of the above plot, doesn't appear valid to fit `bili` directly into the model as a continuous variable. Two options: log transform `bili` or apply a smoothing spline.

Option 1. Log transform `bili`:

```
logbili <- log(dat$bili)
dat <- cbind(dat, logbili)
```

Work out quantiles for `bili`, create dummy variables and add to `dat`:

```
quantiles <- signif(quantile(dat$logbili, probs = c(0.25, 0.50, 0.75)), digits
= 3)
hist(dat$logbili)
```

Quartiles for `bili` are -0.223, 0.336, and 1.220. Create a categorical variable based on `logbili`:

```
logbili.cat <- rep(0, length(dat[,1]))
logbili.cat[dat$logbili < quantiles[1]] <- 1
logbili.cat[dat$logbili >= quantiles[1] & dat$logbili < quantiles[2]] <- 2
logbili.cat[dat$logbili >= quantiles[2] & dat$logbili < quantiles[3]] <- 3
logbili.cat[dat$logbili >= quantiles[3]] <- 4
dat <- cbind(dat, logbili.cat)
dat$logbili.cat <- factor(dat$logbili.cat, labels=c("1", "2", "3", "4"))
contrasts(dat$logbili.cat) <- contr.treatment(4, base = 1, contrasts = TRUE)

pbc.cph03 <- coxph(Surv(futime, status == 2) ~ age + logbili.cat, method =
"breslow", data = dat)
summary(pbc.cph03)

pbc.cph03$coefficients
pbc.cph03$coefficients[2:4]
x <- c((quantiles[1] + min(dat$bili))/2, (quantiles[1] + quantiles[2])/2,
(quantiles[2] + quantiles[3])/2, (max(dat$logbili) + quantiles[3])/2)
y <- c(0, pbc.cph03$coefficients[2:4])
plot(x, y, type = "l", xlim = c(0,3), ylim = c(0,3), xlab = "Log transformed
serum bilirubin (mg/dL)", ylab = "Regression coefficient (log hazard)")
```

Option 2. Smoothing spline for `bili`:

```
pbc.cph02 <- coxph(Surv(futime, status == 2) ~ age + pspline(bili, df = 4),
data = dat)
summary(pbc.cph02)
```

Plot the log hazard as a function of `bili`. To do this we first list the fitted values (and the lower and upper limits of the fitted values) into a data frame called `temp`:

```
temp <- predict(pbc.cph02, type="terms", se.fit = TRUE)
```

For each subject, extract the fitted values for the second explanatory variable in the model (in this case `bili`):

```
tmat <- cbind(temp$fit[,2], temp$fit[,2] - 1.96 * temp$se.fit[,2],
temp$fit[,2] + 1.96 * temp$se.fit[,2])
```

List the unique values of bili, sorting them from lowest to highest:

```
jj <- match(sort(unique(dat$bili)), dat$bili)
```

And plot the log hazard as a function of `bili`:

```
matplot(dat$bili[jj], tmat[jj,], type = l, lty = c(1,2,2), col = c(1,1,1),
xlim = c(0,25), ylim = c(0,4), xlab = "Serum bilirubin (mg/dL)", ylab =
"Regression coefficient (log hazard)")
```

A graphical comparison of the two methods is shown in Figure 10.

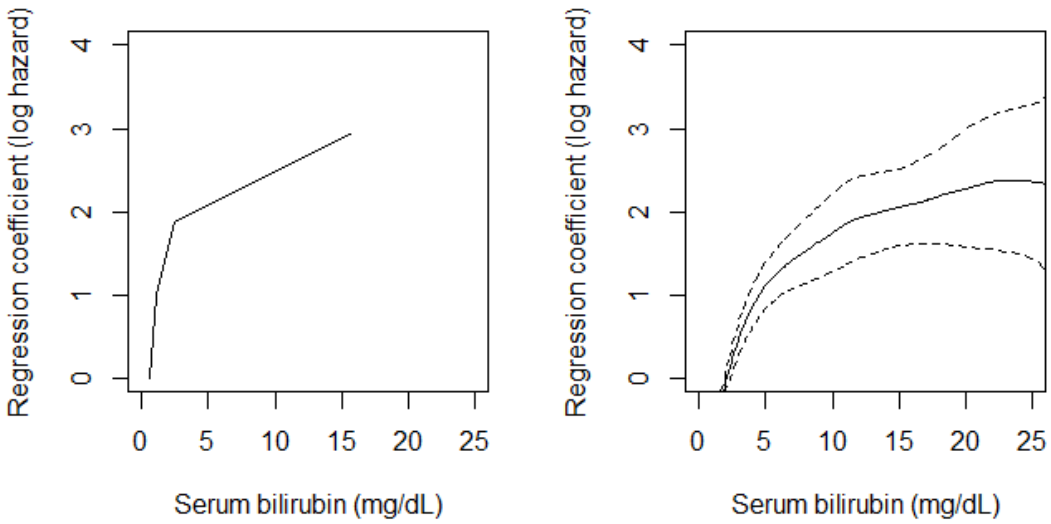


Figure 10: Log hazard as a function of serum bilirubin for: (a) Cox model where bilirubin has been parameterised using a categorical variable of four levels, and (b) Cox model where bilirubin has been parameterised using a smoothing spline.

10 Competing risks

So far, we have described analytical approaches where the event of interest has been the same for all study subjects (e.g. where calving to conception interval has been the outcome of interest, we monitor the time to the single event, conception). Sometimes, it may be desirable to distinguish between several different kinds of events.

Consider studies of wastage in dairy cows. We monitor a population of dairy cows for a defined study period and record both the date and timing of removal from the herd. Since dairy cows may be removed from the herd for a variety of reasons, we might also record the reason for removal. In analysing a study of this type, we might be interested in the different hazard for cows removed for reproductive failure, compared with cows that were removed for udder disorders. In this section we consider the method of competing risks for handling this situation. What is most characteristic is that the occurrence of one type of event removes the individual from being at risk of all other event types (people who die from cancer are no longer at risk of death from heart disease; employees who resign are no longer at risk of being sacked and so on).

The classification of events into different types will vary according to the specific goals of an analysis. Lets assume that the events we are interested in are deaths, and we have classified them into five types according to cause: heart disease, cancer, stroke, accident and a residual category called 'other'. We assign the numbers 1 through to 5 respectively, to these death types. For each type of death we define a separate hazard function which will be called a type-specific or cause-specific hazard. Let T_i be a random variable denoting the time of death for person *i*. Let J_i be a variable denoting the type of death person *i* experienced. Thus $J_5 = 2$ means that person 5 died from death reason 2, cancer. We can determine a hazard function for cancer, and in doing so we treat deaths for all other reasons are censored observations. This gives us a type-specific survival function, giving the probability that an event of type *J* occurs later than time *t*. Now that type-specific hazards can be determined, we can proceed to formulate models for their dependence on covariates. Both proportional hazards and accelerated failure time models may be fitted. For example, we can specify a general proportional hazards model for all five different death types:

$$h(t, X) = h_{J0} \exp(\beta_{J1} X_{J1} + \beta_{J2} X_{J2} + ... + \beta_{Jk} X_{Jk}) \tag{15}$$

Note that the coefficients β are subscripted according to death type to indicate that the effects of each of the covariates may be different for the different death types. In addition, some coefficients may be set to zero, thereby excluding the influence of a covariate for a specific death type. The baseline survival function $h_{J0}(t)$ is also subscripted to allow the dependence of the hazard on time to vary across death types.

Although it would be unusual, there would be nothing to prevent one from specifying a Weibull model for heart disease, a gamma model and a proportional hazards model for strokes. What makes this possible is that the models may be estimated separately for each event type, with no loss of statistical precision. This is perhaps the most important principle of competing risks analysis. A further implication is that you don't need to estimate models for all event types unless you really need to. For example, if you are really interested in the effects of covariates on deaths from heart disease, then just estimate a single model for heart disease, treating all other death types as censored observations.

Kyle (1993) studied 241 cases of monoclonal gammopathy identified at the Mayo Clinic before 1 January 1971 with between 20 and 35 years of total followup on each patient. The response variable is the time to the first of various adverse events death ($n = 130$), multiple myeloma ($n = 39$), and 'other' ($n = 20$). Reference: Kyle RA (1993) 'Benign' monoclonal gammopathy — after 20 to 35 years of follow-up. Mayo Clinic Proceedings, 68:26 - 36. Most subjects in the study were discovered incidentally in the process of being examined for other indications. The laboratory values (albumin, creatinine, etc.) may be related to the severity of those other indications, but have shown less relationship to MGUS per se.

In a competing risks analysis, we assign one stratum for each outcome type and all subjects appear in each stratum. In the following example data set the first subject experiences death at day 760 and the second subject experiences lymphoproliferative disease at day 2160.

id	time	status	endpoint	sex	age	hgb	creat	mspike
1	760	1	death	2	79	1.5	1.2	2.0
1	760	0	myeloma	2	79	1.5	1.2	2.0
1	760	0	other	2	79	1.5	1.2	2.0
2	2160	0	death	2	76	13.3	1.0	1.8
2	2160	0	myeloma	2	76	13.3	1.0	1.8
2	2160	1	other	2	76	13.3	1.0	1.8

id: patient identifier.
time: day of event.
status: 0 = censored, 1 = died.
endpoint: reason for failure — death, myeloma or other lymphoproliferative disorder.
sex: 1 = male, 2 = female.
age: patient age at enrolment.
hgb: plasma haemoglobin concentration at enrolment.
creat: plasma creatinine concentration at enrolment.
mspike: size of monclonal spike.

The code for a competing risks analysis is as follows:

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("mgus2.csv", header = TRUE, sep = ",")

mgus2.km <- survfit(Surv(time, status) ~ endpoint, type = "kaplan-meier", data
= dat)
plot(mgus2.km, xlab = "Days to endpoint", ylab = "Cumulative proportion to
experience event", lty = c(1,2,3), mark.time = FALSE)
legend(x = "topright", legend = c("Death","Myeloma","Other"), lty = c(1,2,3),
bty = "n")
```

Even though the data set contains three observations for each subject, because the endpoint times are independent it is valid to apply a log rank test to compare survivorship for the three outcomes. This is similar to the situation when one has paired data where the two measurements are independent (or uncorrelated) and a two-sample test would be used (rather than a paired sample t-test).

```
survdiff(Surv(time, status) ~ endpoint, data = dat, na.action = na.omit, rho =
0)
```

Time to endpoint differs according to outcome (Chi-squared test statistic = 111; df 2; P 0.001).

```
mgus2.cph01 <- coxph(Surv(time, status, type = "right") ~ sex + age + hgb +
mspike + cluster(id) + strata(endpoint), data = dat)
summary(mgus2.cph01)
```

Note the use of `cluster(id)`, to apply a robust (sandwhich) estimate to correct for multiple events that an individual experiences. This correction is not required in the competing risks case when each subject can have at most one outcome event.

Table 16: Competing risks regression model showing the effect of sex, age, haemoglobin concentration and size of monoclonal spike on the daily hazard of 'endpoint'.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Sex	720	188	-0.3399 (0.1563)	0.03	0.71 (0.53 – 0.96)
Age	720	188	0.0514 (0.0073)	$<$0.01	1.05 (1.04 – 1.07)
Haemoglobin	720	188	-0.1664 (0.0443)	$<$0.01	0.85 (0.79 – 0.91)
Monoclonal spike	720	188	-0.0878 (0.1869)	0.63	0.92 (0.64 – 1.31)

R^2 = 0.100.

Likelihood ratio test = 74.4 on 4 df, P $<$0.01.

Table 17: Competing risks regression model showing the effect of sex, age (for deaths or for all other reasons), haemoglobin concentration and size of monoclonal spike on the daily hazard of 'endpoint'.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Sex	720	188	-0.3274 (0.1554)	0.03	0.72 (0.53 – 0.98)
Age 1	720	188	0.0760 (0.0093)	$<$0.01	1.08 (1.06 – 1.10)
Age 2	720	188	0.0026 (0.0123)	0.83	1.00 (0.98 – 1.03)
Haemoglobin	720	188	-0.1613 (0.0446)	$<$0.01	0.85 (0.78 – 0.93)
Monoclonal spike	720	188	-0.0887 (0.1870)	0.64	0.91 (0.63 – 1.32)

R^2 = 0.128.

Likelihood ratio test = 96.7 on 5 df, P $<$0.01.

The advantage of a large data set is that it allows for easy estimation of within-event-type coefficients. For instance, one might ask if the effect of age is identical for both outcomes, while controlling for the common effect of haemoglobin. This can be investigated by coding two dummy variables:

```
age1 <- dat$age * (dat$endpoint == "death")
age2 <- dat$age * (dat$endpoint != "death")
mgus2.cph02 <- coxph(Surv(time, status, type = "right") ~ sex + age1 + age2 +
hgb + mspike + strata(endpoint), data = dat)
summary(mgus2.cph02)
```

Age is a significant predictor of the overall death rate. Age is, however, of far less importance in predicting the likelihood of plasma cell malignancy.

11 Recurrent events

Examples of recurrent events:

- Recurrence of cancer in treated patients.
- 'Episodes' of illness: hypoglycaemic episodes in insulin-dependent diabetics, seizure events in epilepsy.
- Rehospitalisation.

A major issue in extending proportional hazards regression models to deal with recurrent events relates to the area of intrasubject correlation. Several approaches to such data have appeared in the literature:

- Marginal model approaches. Regression coefficients are estimated by Cox regression and the correlation is handled by robust variance estimators.
- Random effects or frailty models. The model includes a random per-subject effect. Multiple outcomes are assumed to be independent, conditional on the per-subject coefficient.
- Proportional hazards models where the subject's correlation is modelled directly (not often used).

These notes concentrate on marginal model approaches. In each case the analysis is based on three steps:

- Decide on a model framework (strata, time dependent covariates) and structure the data accordingly.
- Fit the data as an ordinary Cox model, ignoring the possible correlation.
- Replace the standard variance estimate with one which is corrected for the possible correlation.

11.1 Selecting a model

One aspect of multiple event data sets is that there are a range of choices available when setting a model up. These include the choice of strata and membership within strata, time scales within strata, constructed time-dependent covariates, and stratum-by-covariate interactions. Stratification, if used, tends to be based on external variables such as enrolling institution or disease subtype. These generally correspond to predictors for which we need a flexible adjustment, but not an estimate of the covariate effect. Time-dependent covariates usually reflect directly measured data such as repeated lab tests. Strata by covariate interactions (that is, separate coefficients per stratum for some covariate) are infrequently encountered. The first issue is to distinguish between data sets where the multiple events have a distinct ordering and those where they do not. An example of correlated *unordered* events are paired survival data, such as a subject's two eyes in a diabetic retinopathy study. An example of *ordered* events is cancer recurrence: one can't have a recurrence of cancer until a primary diagnosis has been made.

11.2 Multiple event models

Anderson-Gill (AG) model: Using the counting process style of data input, each subject is represented as a series of observations (rows of data) with time intervals of (entry time, first event], (first event, second event], (second event, third event] ... (m^{th} event, last followup]. A subject with zero events would have a single observation. A subject with one event would have one or two observations (depending on whether there was

Table 18: Representation of a hypothetical subject for AG, PWP and WLW models.

Model	Start, stop	Status	Stratum
AG	0, 2	1	1
	2, 20	1	1
	20, 30	0	1
PWP	0, 2	1	1
	2, 20	1	2
	20, 30	0	3
WLW	0, 2	1	1
	0, 20	1	2
	0, 30	0	3
	0, 30	0	4

additional followup after the first event), and so on. This model is ideally situated to the situation of mutual independence of the observations within a subject.

Conditional (PWP) model: The conditional model was proposed in Prentice, Williams, and Peterson (1981). It assumes that a subject cannot be at risk for a second event until a first event has occurred (that is, a subject is not at risk of a k^{th} event until he/she has experienced event (k - 1). To accomplish this, the counting process style of input is used, as in the AG model, but each event is assigned to a separate stratum. The use of time dependent strata means that the underlying hazard function is allowed to vary from event to event (unlike the AG model, which assumes that all events are identical).

Wei, Lin, and Weissfeld (WLW) model: Here, one treats the ordered outcome data set as though it were an unordered competing risks problem. If there is a maximum of four events (for example) in the data set, then there will be four strata in the analysis. Every subject has four observations - one for each stratum (barring deletion for missing covariates).

The key to fitting these models is to set the data up appropriately. Assume we have a subject with events at times 2, 20 and 23 with the study is terminated on day 30. Over the entire data set, the maximum number of experienced per subject is four. This subject will be represented in the data set by three observations, but the time intervals and strata differ according to the model used, as shown in Table 18.

11.3 Worked examples

This example is taken from a paper by Wei, Lin and Weissfeld (1989). The study is of time to recurrence of bladder cancer. The data frame `bladder` has either 4 or 5 rows for each subject. Many subjects had recurrences of bladder cancer, sometimes as many as four, and were followed beyond the fourth recurrence.

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("bladder.csv", header = TRUE, sep = ",")
head(dat)
```

id	rx	size	number	start	stop	event	obsnum
1	1	3	1	0	1	0	1
1	1	3	1	1	1	0	2
1	1	3	1	1	1	0	3
1	1	3	1	1	1	0	4
2	1	1	2	0	4	0	1
2	1	1	2	4	4	0	2
2	1	1	2	4	4	0	3
2	1	1	2	4	4	0	4

id: patient identifier.
rx: 1 = placebo, 2 = thiopet.
size: size of the largest initial tumour.
number: number of initial tumours.
start: entry into the study or the time of last recurrence.
stop: time to event (months).
event: 0 = censored, 1 = event.
obsnum: observation number.

We create two data frames for analysis. The first one has only the first four rows for each subject and has `start` removed:

```
dat1 <- dat[dat$obsnum < 5,]
dat1$start <- NULL
dat1[1:10,]
```

The second data frame has removed all rows for which `start` and `stop` are equal:

```
dat2 <- dat[dat$start < dat$stop,]
dat2[1:10,]
```

Anderson-Gill model:

```
bladder.ag <- coxph(Surv(start, stop, event) ~ rx + size + number +
cluster(id), data = dat2)
summary(bladder.ag)
```

Conditional risk sets or PWP model:

```
bladder.pwp <- coxph(Surv(start, stop, event) ~ rx + size + number +
cluster(id) + strata(obsnum), data = dat2)
summary(bladder.pwp)
```

Marginal risk sets or WLW model:

```
bladder.wlw <- coxph(Surv(stop, event) ~ rx + size + number + cluster(id) +
strata(obsnum), data = dat1, method = "breslow")
summary(bladder.wlw)
```

Estimated regression coefficients from the Anderson Gill, conditional risk set and marginal risk set models are shown in Tables 19, 20 and 21.

Table 19: Cox proportional hazards regression model showing the effect of treatment, tumour size, and number of initial tumours on the monthly hazard of tumour recurrence, Anderson-Gill model.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Treatment	228	63	-0.4647 (0.2656)	0.08	0.66 (0.41 – 1.08)
Size	228	63	-0.0437 (0.0776)	0.57	0.96 (0.83 – 1.11)
Number	228	63	0.1750 (0.0630)	< 0.01	1.18 (1.05 – 1.32)

R^2 = 0.074.
Likelihood ratio test = 14.7 on 3 df, $P < 0.01$.

Table 20: Cox proportional hazards regression model showing the effect of treatment, tumour size, and number of initial tumours on the monthly hazard of tumour recurrence, conditional risk sets model.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Treatment	228	63	-0.3335 (0.2048)	0.10	0.72 (0.48 – 1.07)
Size	228	63	-0.0085 (0.0616)	0.89	0.99 (0.88 – 1.12)
Number	228	63	0.1196 (0.0514)	0.02	1.13 (1.02 – 1.25)

R^2 = 0.034.
Likelihood ratio test = 6.51 on 3 df, P = 0.09.

Table 21: Cox proportional hazards regression model showing the effect of treatment, tumour size, and number of initial tumours on the monthly hazard of tumour recurrence, marginal risk sets model.

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Treatment	228	63	-0.5798 (0.3034)	0.06	0.56 (0.31 – 1.01)
Size	228	63	-0.0509 (0.0930)	0.58	0.95 (0.79 – 1.14)
Number	228	63	0.2085 (0.0657)	< 0.01	1.23 (1.08 – 1.40)

R^2 = 0.070.
Likelihood ratio test = 24.7 on 3 df, $P < 0.01$.

Plot the hazard ratios for each of the three models using `ggplot2`:

```
library(ggplot2); library(scales)
rval <- rbind(summary(bladder.ag)$coefficients[1,],
summary(bladder.pwp)$coefficients[1,], summary(bladder.wlw)$coefficients[1,])



ggplot(data = rval, aes(x = est, y = model)) +
   geom_point() +
   geom_errorbarh(aes(xmax = low, xmin = upp, height = 0.2)) +
   labs(x = "Hazard ratio", y = "Model") +
scale_x_continuous(trans = log2_trans(), breaks = c(0.25,0.5,1.0,2), labels
= c(0.25,0.50,1.0,2.0), limits = c(0.25,3), name = "Hazard ratio") +
   geom_vline(xintercept = 1, lwd = 1) +
   coord_fixed(ratio = 0.5 / 1) +
   theme(axis.title.y = element_text(vjust = 0))
```

The resulting plot is shown in Figure 11.

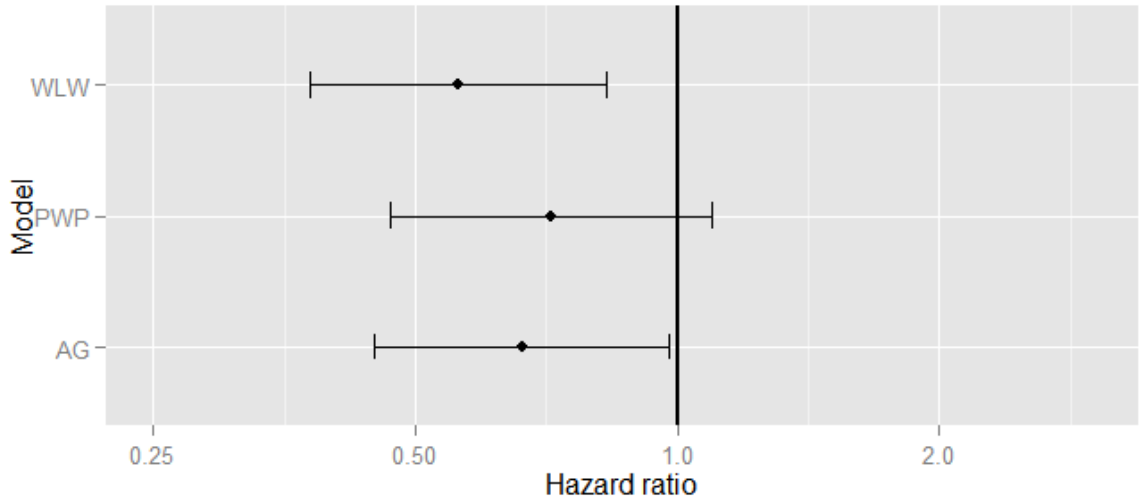


Figure 11: Estimated hazard ratios for effect of treatment computed from the Anderson-Gill (AG), conditional risk sets (PWP) and marginal risk sets (WLW) models.

12 Sample size and power estimation for survival analysis

Consider an experiment comparing two groups, for example a clinical trial comparing a new treatment with a standard (control). We wish to detect a log-hazard ratio of β. That is, $h_{treat}(t) = h_{control}(t).\exp^{\beta}$, where $h_{treat}(t)$ and $h_{control}(t)$ denote the hazards in the treatment and control groups, respectively. The total number of events required (not the total number of study subjects) is given by:

$$d = \frac{(c_\alpha + z_{power})^2}{pq\beta^2} \tag{16}$$

Where:

p: the proportion of study subjects in the experimental group.
q: the proportion of study subjects in the control group (1 - *p*).
c_α: the critical value for the test (where α is the probability of making a type I error).
z_{power}: the upper quartile of the standard normal distribution.

Suppose we wish to compare an experimental treatment to a standard. Five-year survival under standard treatment is approximately 30% and we anticipate that the new treatment will increase five-year survival to 45%. Assume that the study will use a two-sided test at a = 0.05 and we want a sample size to produce results with 90% power. The study will allocate patients equally to the treatment groups, so p = 0.5 and q = 0.5.

```
library(epiR)
epi.studysize(treat = 0.45, control = 0.30, n = NA, sigma = NA, power = 0.90,
r = 1, design = 1, sided.test = 2, conf.level = 0.95, method = "survival")
```

The study requires around 125 events in the treatment group and 125 subjects in the control group (250 events in total) to be 95 percent certain of detecting a difference of 0.15 units on 90 percent of occasions.

A company approaches you to analyse the results of a trial to evaluate the effect of a trace element supplement on dairy cow fertility. There are 259 animals enrolled in the trial: 134 are assigned to the treatment group, 125 are assigned to the control group. At 40 days after the start of mating, 60% of the treatment group and 50% of the control group are confirmed in-calf. What is the power of this study at the alpha = 0.05 level of significance?

```
epi.studysize(treat = 0.60, control = 0.50, n = 259, sigma = NA, power = NA, r
= 125/134, design = 1, sided.test = 1, conf.level = 0.95, method = "survival")
```

The study contains enough subjects ('events') to be 95 percent certain of detecting a survival difference of 10 percent on 43 percent of occasions.

For the dairy cow fertility study described above, imagine only 150 animals ended up enrolling in the trial (100 in the treatment group and 50 in the control group). What is the power of this study to detect 60% of the treatment group and 50% of the control group in-calf at 40 days after the start of mating?

```
epi.studysize(treat = 0.60, control = 0.50, n = 150, sigma = NA, power = NA, r
= 100/50, design = 1, sided.test = 1, conf.level = 0.95, method = "survival")
```

The study contains enough subjects ('events') to be 95 percent certain of detecting a survival difference of 10 percent on 28 percent of occasions.

British Journal of Cancer (2003) 89, 232–238

www.bjcancer.com

Tutorial Paper

Survival Analysis Part I: Basic concepts and first analyses

TG Clark*[,1], MJ Bradburn[1], SB Love[1] and DG Altman[1]

[1]*Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, University of Oxford, Old Road, Oxford OX3 7LF, UK*

British Journal of Cancer (2003) **89,** 232–238. doi:10.1038/sj.bjc.6601118 www.bjcancer.com

Keywords: survival analysis; statistical methods; Kaplan-Meier

INTRODUCTION

In many cancer studies, the main outcome under assessment is the time to an event of interest. The generic name for the time is *survival time*, although it may be applied to the time 'survived' from complete remission to relapse or progression as equally as to the time from diagnosis to death. If the event occurred in all individuals, many methods of analysis would be applicable. However, it is usual that at the end of follow-up some of the individuals have not had the event of interest, and thus their true time to event is unknown. Further, survival data are rarely Normally distributed, but are skewed and comprise typically of many early events and relatively few late ones. It is these features of the data that make the special methods called *survival analysis* necessary.

This paper is the first of a series of four articles that aim to introduce and explain the basic concepts of survival analysis. Most survival analyses in cancer journals use some or all of Kaplan–Meier (KM) plots, logrank tests, and Cox (proportional hazards) regression. We will discuss the background to, and interpretation of, each of these methods but also other approaches to analysis that deserve to be used more often. In this first article, we will present the basic concepts of survival analysis, including how to produce and interpret survival curves, and how to quantify and test survival differences between two or more groups of patients. Future papers in the series cover multivariate analysis and the last paper introduces some more advanced concepts in a brief question and answer format. More detailed accounts of these methods can be found in books written specifically about survival analysis, for example, Collett (1994), Parmar and Machin (1995) and Kleinbaum (1996). In addition, individual references for the methods are presented throughout the series. Several introductory texts also describe the basis of survival analysis, for example, Altman (2003) and Piantadosi (1997).

TYPES OF 'EVENT' IN CANCER STUDIES

In many medical studies, time to death is the event of interest. However, in cancer, another important measure is the time between response to treatment and recurrence or relapse-free survival time (also called disease-free survival time). It is important to state what the event is and when the period of observation starts and finishes. For example, we may be interested in relapse in the time period between a confirmed response and the first relapse of cancer.

*Correspondence: Mr TG Clark; E-mail: taane.clark@cancer.org.uk
Received 6 December 2002; accepted 30 April 2003

CENSORING MAKES SURVIVAL ANALYSIS DIFFERENT

The specific difficulties relating to survival analysis arise largely from the fact that only some individuals have experienced the event and, subsequently, survival times will be unknown for a subset of the study group. This phenomenon is called censoring and it may arise in the following ways: (a) a patient has not (yet) experienced the relevant outcome, such as relapse or death, by the time of the close of the study; (b) a patient is lost to follow-up during the study period; (c) a patient experiences a different event that makes further follow-up impossible. Such censored survival times underestimate the true (but unknown) time to event. Visualising the survival process of an individual as a time-line, their event (assuming it were to occur) is beyond the end of the follow-up period. This situation is often called *right censoring*. Censoring can also occur if we observe the presence of a state or condition but do not know where it began. For example, consider a study investigating the time to recurrence of a cancer following surgical removal of the primary tumour. If the patients were examined 3 months after surgery to determine recurrence, then those who had a recurrence would have a survival time that was *left censored* because the actual time of recurrence occurred less than 3 months after surgery. Event time data may also be *interval censored,* meaning that individuals come in and out of observation. If we consider the previous example and patients are also examined at 6 months, then those who are disease free at 3 months and lost to follow-up between 3 and 6 months are considered interval censored. Most survival data include right censored observations, but methods for interval and left censored data are available (Hosmer and Lemeshow, 1999). In the remainder of this paper, we will consider right censored data only.

In general, the feature of censoring means that special methods of analysis are needed, and standard graphical methods of data exploration and presentation, notably scatter diagrams, cannot be used.

ILLUSTRATIVE STUDIES

Ovarian cancer data

This data set relates to 825 patients diagnosed with primary epithelial ovarian carcinoma between January 1990 and December 1999 at the Western General Hospital in Edinburgh. Follow-up data were available up until the end of December 2000, by which time 550 (75.9%) had died (Clark *et al*, 2001). Figure 1 shows data from 10 patients diagnosed in the early 1990s and illustrates how patient profiles in calendar time are converted to time to event

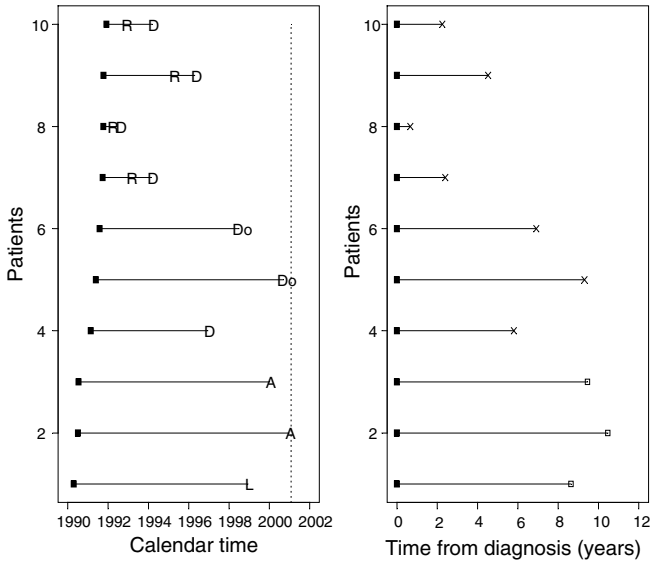


Figure 1 Converting calendar time in the ovarian cancer study to a survival analysis format. Dashed vertical line is the date of the last follow-up, R = relapse, D = death from ovarian cancer, Do = death from other cause, A = attended last clinic visit (alive), L = loss to follow-up, × = death, □ = censored.

(death) data. Figure 1 (left) shows that four patients had a nonfatal relapse, one was lost to follow-up, and seven patients died (five from ovarian cancer). In the other plot, the data are presented in the format for a survival analysis where all-cause mortality is the event of interest. Each patient's 'survival' time has been plotted as the time from diagnosis. It is important to note that because overall mortality is the event of interest, nonfatal relapses are ignored, and those who have not died are considered (right) censored. Figure 1 (right) is specific to the outcome or event of interest. Here, death from any cause, often called overall survival, was the outcome of interest. If we were interested solely in ovarian cancer deaths, then patients 5 and 6 – those who died from nonovarian causes – would be censored. In general, it is good practice to choose an end-point that cannot be misclassified. All-cause mortality is a more robust end-point than a specific cause of death. If we were interested in time to relapse, those who did not have a relapse (fatal or nonfatal) would be censored at either the date of death or the date of last follow-up.

Lung cancer clinical trial data

These data originate from a phase III clinical trial of 164 patients with surgically resected (non-small cell) lung cancer, randomised between 1979 and 1985 to receive radiotherapy either with or without adjuvant combination platinum-based chemotherapy (Lung Cancer Study Group, 1988; Piantadosi, 1997). For the purposes of this series, we will focus on the time to first relapse (including death from lung cancer). Table 1 gives the time of the earliest 15 and latest five relapses for each treatment group, where it can be seen that some patients were alive and relapse-free at the end of the study. The relapse proportions in the radiotherapy and combination arms were 81.4% (70 out of 86) and 69.2% (54 out of 78), respectively. However, these figures are potentially misleading as they ignore the duration spent in remission before these events occurred.

SURVIVAL AND HAZARD

Survival data are generally described and modelled in terms of two related probabilities, namely *survival* and *hazard*. The survival probability (which is also called the survivor function) $S(t)$ is the probability that an individual survives from the time origin (e.g. diagnosis of cancer) to a specified future time t. It is fundamental to a survival analysis because survival probabilities for different values of t provide crucial summary information from time to event data. These values describe directly the survival experience of a study cohort.

Table 1 A sample of times (days) to relapse among patients randomised to receive radiotherapy with or without adjuvant chemotherapy

Radiotherapy ($n = 86$)	18, 23[a], 25, 27, 28, 30, 36, 45, 55, 56, 57, 57, 57, 59, 62, ..., 2252[a], 2286[a], 2305[a], 2318[a], 2940[a]
Radiotherapy+CAP ($n = 78$)	9, 22, 35, 53, 76, 81, 94, 97, 103, 114, 115, 126, 147, 154, ..., 2220[a], 2375, 2566, 2875[b], 3067[b]

CAP = cytoxan, doxorubicin and platinum-based chemotherapy. [a]Lost to follow-up and considered censored. [b]Relapse-free at time of analysis and considered censored.

The hazard is usually denoted by $h(t)$ or $\lambda(t)$ and is the probability that an individual who is under observation at a time t has an event at that time. Put another way, it represents the instantaneous event rate for an individual who has already survived to time t. Note that, in contrast to the survivor function, which focuses on not having an event, the hazard function focuses on the event occurring. It is of interest because it provides insight into the conditional failure rates and provides a vehicle for specifying a survival model. In summary, the hazard relates to the incident (current) event rate, while survival reflects the cumulative non-occurrence.

KAPLAN–MEIER SURVIVAL ESTIMATE

The survival probability can be estimated nonparametrically from observed survival times, both censored and uncensored, using the KM (or product-limit) method (Kaplan and Meier, 1958). Suppose that k patients have events in the period of follow-up at distinct times $t_1 < t_2 < t_3 < t_4 < t_5 < \cdots < t_k$. As events are assumed to occur independently of one another, the probabilities of surviving from one interval to the next may be multiplied together to give the cumulative survival probability. More formally, the probability of being alive at time t_j, $S(t_j)$, is calculated from $S(t_{j-1})$ the probability of being alive at t_{j-1}, n_j the number of patients alive just before t_j, and d_j the number of events at t_j, by

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$$

where $t_0 = 0$ and $S(0) = 1$. The value of $S(t)$ is constant between times of events, and therefore the estimated probability is a step function that changes value only at the time of each event. This estimator allows each patient to contribute information to the calculations for as long as they are known to be event-free. Were every individual to experience the event (i.e. no censoring), this estimator would simply reduce to the ratio of the number of individuals events free at time t divided by the number of people who entered the study.

Confidence intervals for the survival probability can also be calculated. The KM *survival curve*, a plot of the KM survival probability against time, provides a useful summary of the data that can be used to estimate measures such as median survival time. The large skew encountered in the distribution of most survival data is the reason that the mean is not often used.

Survival analysis of the lung cancer trial

Table 2 shows the essential features of the KM survival probability. The estimator at any point in time is obtained by multiplying a sequence of conditional survival probabilities, with the estimate

Table 2 Calculation of the relapse-free survival probability for patients in the lung cancer trial

Radiotherapy ($n=86$)		Radiotherapy+CAP ($n=78$)	
Survival times (days)	**Kaplan–Meier survivor function $S(t)$**	**Survival times (days)**	**Kaplan–Meier survivor function $S(t)$**
18	$1\times(1\text{-}1/86)=0.988$	9	$1\times(1\text{-}1/78)=0.987$
23[a]	$S(18)\times(1\text{-}0/85)=0.988$	22	$S(18)\times(1\text{-}1/77)=0.974$
25	$S(23)\times(1\text{-}1/84)=0.977$	35	$S(22)\times(1\text{-}1/76)=0.962$
27	$S(25)\times(1\text{-}1/83)=0.965$	53	$S(35)\times(1\text{-}1/75)=0.949$
28	$S(27)\times(1\text{-}1/82)=0.953$	76	$S(53)\times(1\text{-}1/74)=0.936$
30	$S(28)\times(1\text{-}1/81)=0.941$	81	$S(76)\times(1\text{-}1/73)=0.923$
36	$S(30)\times(1\text{-}1/80)=0.930$	94	$S(81)\times(1\text{-}1/72)=0.910$
45	$S(36)\times(1\text{-}1/79)=0.918$	97	$S(94)\times(1\text{-}1/71)=0.897$
55	$S(45)\times(1\text{-}1/78)=0.906$	103	$S(97)\times(1\text{-}1/70)=0.885$
56	$S(55)\times(1\text{-}1/77)=0.894$	114	$S(103)\times(1\text{-}1/69)=0.872$
57	$S(56)\times(1\text{-}3/76)=0.859$	115	$S(114)\times(1\text{-}1/68)=0.859$
57	$S(56)\times(1\text{-}3/76)=0.859$	121[a]	$S(115)\times(1\text{-}0/67)=0.859$
57	$S(56)\times(1\text{-}3/76)=0.859$	126	$S(121)\times(1\text{-}1/66)=0.846$
59	$S(57)\times(1\text{-}1/73)=0.847$	147	$S(126)\times(1\text{-}1/65)=0.833$
62	$S(59)\times(1\text{-}1/72)=0.835$	154	$S(147)\times(1\text{-}1/64)=0.820$
⋮	⋮	⋮	⋮
2252[a]	$S(2209)\times(1\text{-}0/5)=0.115$	2220[a]	$S(2218)\times(1\text{-}0/5)=0.273$
2286[a]	$S(2286)\times(1\text{-}0/4)=0.115$	2375	$S(2220)\times(1\text{-}0/4)=0.205$
2305[a]	$S(2305)\times(1\text{-}0/3)=0.115$	2566	$S(2375)\times(1\text{-}0/3)=0.137$
2318[a]	$S(2318)\times(1\text{-}0/2)=0.115$	2875[b]	$S(2566)\times(1\text{-}0/2)=0.137$
2940[a]	$S(2940)\times(1\text{-}0/1)=0.115$	3067[b]	$S(2875)\times(1\text{-}0/1)=0.137$

$S(0)=1$, (CAP = cytoxan, doxorubicin and platinum-based chemotherapy.) [a]Lost to follow-up and considered censored. [b]Relapse-free at time of analysis and considered censored.

being unchanged between subsequent event times. For example, the probability of a member of the radiotherapy alone treatment group surviving (relapse-free) 45 days is the probability of surviving the first 36 days multiplied by the probability of then surviving the interval between 36 and 45 days. The latter is a *conditional* probability as the patient needs to have survived the first period of time in order to remain in the study for the second. The KM estimator utilises this fact by dividing the time axis up according to event times and estimating the event probability in each division, from which the overall estimate of the survivorship is drawn.

Figure 2 shows the survival probabilities for the two treatment groups in the conventional KM graphical display. The median survival times for each group are shown and represent the time at which $S(t)$ is 0.5. The combination group has a median survival time of 402 days (1.10 years), as opposed to 232 days (0.64 years) in the radiotherapy alone arm, providing some evidence of a chemotherapy treatment benefit. Other survival time percentiles may be read directly from the plot or (more accurately) from a full version of Table 2. There appears to be a survival advantage in the combination therapy group, but whether this difference is statistically significant requires a formal statistical test, a subject that is discussed later.

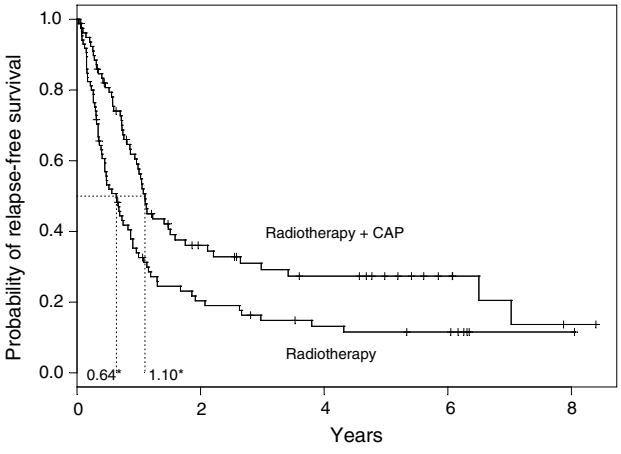


Figure 2 Relapse-free survival curves for the lung cancer trial. * Median relapse-free survival time for each arm, + censoring times, CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

Survival function of the ovarian data

The KM survival curve of the ovarian cancer data is shown in Figure 3A. The steep decline in the early years indicates poor prognosis from the disease. This is also indicated by changes in the cumulative number of events and number at risk. Specifically, of the 825 women diagnosed with ovarian cancer, about a third had died within the first year, accounting for 43% of the total deaths as recorded by the last date of follow-up. The number lost to follow-up can be deduced from the total number in the cohort and the cumulative number of events and number at risk.

The 95% confidence limits of the survivor function are shown. In practice, there are usually patients who are lost to follow-up or alive at the end of follow-up, and confidence limits are often wide at the tail of the curve, making meaningful interpretations difficult. Thus, it may be sensible to curtail plots before the end of follow-up on the x-axis (Pocock *et al*, 2002). Curtailing of the y-axis, a common practice for diseases or events of low incidence, should not be performed. Instead, the incidence of death curve, or $1-S(t)$, (Figure 3B) may be presented (Pocock *et al*, 2002). The cumulative incidence at a time point is simply one minus the survival probability. For example, Figure 3A shows how the 5-year survival of 0.29 (29%) is calculated, and could also be read from Figure 3B as a cumulative incidence of 71% for the first 5 years.

HAZARD AND CUMULATIVE HAZARD

There is a clearly defined relationship between $S(t)$ and $h(t)$, which is given by the calculus formula:

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}[\log S(t)].$$

The formula is unimportant for routine survival analyses as it is incorporated into most statistical computer packages. The point

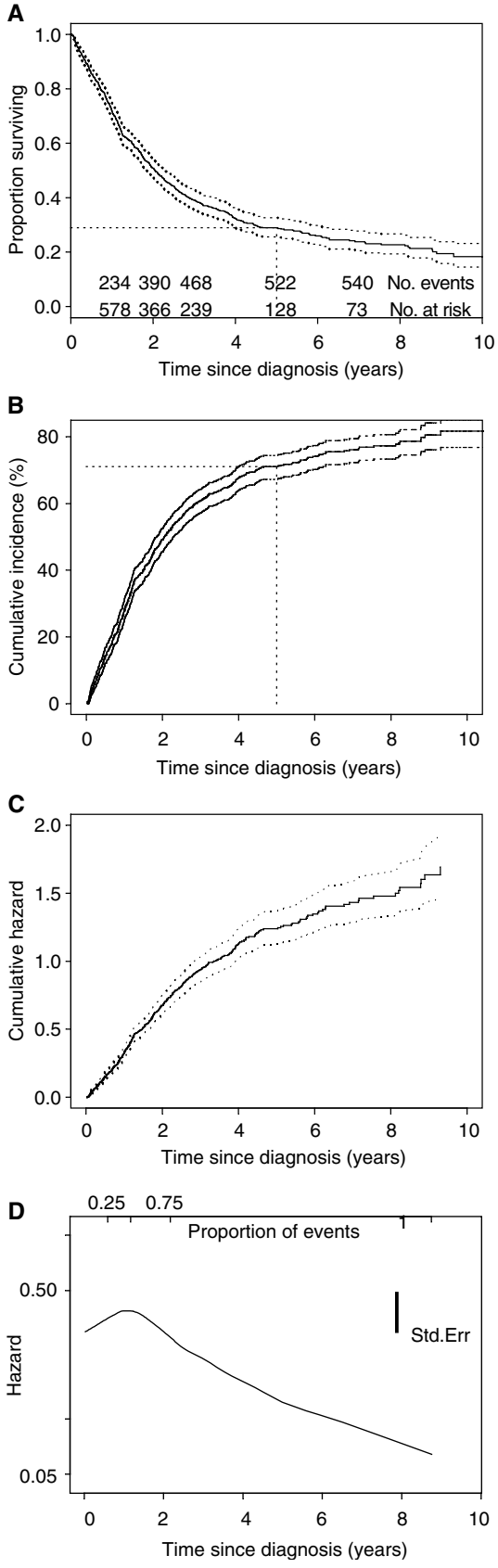


Figure 3 Survival and cumulative hazard curves with 95% CIs for the ovarian cancer study. Std.Err = standard error. (**A**) Kaplan–Meier survivor function, (**B**) cumulative incidence curve, (**C**) cumulative hazard function, (**D**) hazard function (smoothed).

here is simply that if either $S(t)$ or $h(t)$ is known, the other is automatically determined. Consequently, either can be the basis of statistical analysis.

Unfortunately, unlike $S(t)$ there is no simple way to estimate $h(t)$. Instead, a quantity called the *cumulative hazard* $H(t)$ is commonly used. This is defined as the integral of the hazard, or the area under the hazard function between times 0 and t, and differs from the log-survivor curve only by sign, that is $H(t) = -\log[S(t)]$. The interpretation of $H(t)$ is difficult, but perhaps the easiest way to think of $H(t)$ is as the cumulative force of mortality, or the number of events that would be expected for each individual by time t if the event were a repeatable process. $H(t)$ is used an intermediary measure for estimating $h(t)$ and as a diagnostic tool in assessing model validity. A simple nonparametric method for estimating $H(t)$ is the Nelson-Aalen estimator (Hosmer, 1999), from which it is possible to derive an estimate of $h(t)$ by applying a kernel smoother to the increments (Ramlau-Hansen, 1983). Cox (1979) suggests another method to estimate the hazard based on order statistics but similar in spirit to the previous method.

Another approach to estimating the hazard is to assume that the survival times follow a specific mathematical distribution. Figure 4 shows the relationship between four parametrically specified hazards and the corresponding survival probabilities. It illustrates a constant hazard rate over time (which is analogous to an exponential distribution of survival times), strictly increasing/decreasing hazard rates based on a Weibull model, and a combination of decreasing and increasing hazard rates using a log-Normal model. These curves are illustrative examples and other shapes are possible. The specification of hazards using fully parametric distributions is an important and under-utilised modelling technique that will be discussed in subsequent papers.

Hazard function in the ovarian data

Figure 3C shows the cumulative hazard for the ovarian cancer data. The hazard is shown in Figure 3D. As the hazard function is generally very erratic, it is customary to fit a smooth curve to enable the underlying shape to be seen. Figure 3D shows that the (instantaneous) risk of death appears to be high in the first year after diagnosis and decreases afterwards. This observation corresponds to the steeply descending survival probability (Figure 3A) and marked increase in cumulative incidence (Figure 3B) in the first year. The y-axis is difficult to interpret for the hazard and the cumulative hazard, but the decreasing shape of the hazard may be consistent with a decreasing Weibull's model (see Figure 4).

NONPARAMETRIC TESTS COMPARING SURVIVAL

Survival in two or more groups of patients can be compared using a nonparametric test. The logrank test (Peto *et al*, 1977) is the most widely used method of comparing two or more survival curves. The groups may be treatment arms or prognostic groups (e.g. FIGO stage). The method calculates at each event time, for each group, the number of events one would expect since the previous event if there were no difference between the groups. These values are then summed over all event times to give the total expected number of events in each group, say E_i for group i. The logrank test compares observed number of events, say O_i for treatment group i, to the expected number by calculating the test statistic

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}.$$

This value is compared to a χ^2 distribution with $(g-1)$ degrees of freedom, where g is the number of groups. In this manner, a

P-value may be computed to calculate the statistical significance of the differences between the complete survival curves.

If the groups are naturally ordered, a more appropriate test is to consider the possibility that there is a trend in survival across them, for example, age groups or stages of cancer. Calculating O_i and E_i for each group on the basis that survival may increase or decrease across the groups results in a more powerful test. For the new O_i and E_i, the test statistic for trend is compared with the χ^2 distribution with one degree of freedom (Collett, 1994).

When only two groups are compared, the logrank test is testing the null hypothesis that the ratio of the hazard rates in the two groups is equal to 1. The hazard ratio (HR) is a measure of the relative survival experience in the two groups and may be estimated by

$$\mathrm{HR} = \frac{O_1/E_1}{O_2/E_2}$$

where O_i/E_i is the estimated relative (excess) hazard in group i. A confidence interval (CI) for the HR can be calculated (Collett, 1994). The HR has a similar interpretation of the strength of effect as a risk ratio. An HR of 1 indicates no difference in survival. In practice, it is better to estimate HRs using a regression modelling technique, such as Cox regression, as described in the next article.

Other nonparametric tests may be used to compare groups in terms of survival (Collett, 1994). The logrank test is so widely used that the reason for any other method should be stated in the protocol of the study. Alternatives include methods to compare median survival times, but comparing confidence intervals for each group is not recommended (Altman and Bland, 2003). The logrank method is considered more robust (Hosmer and Lemeshow, 1999), but the lack of an accompanying effect size to compliment the *P*-value it provides is a limitation.

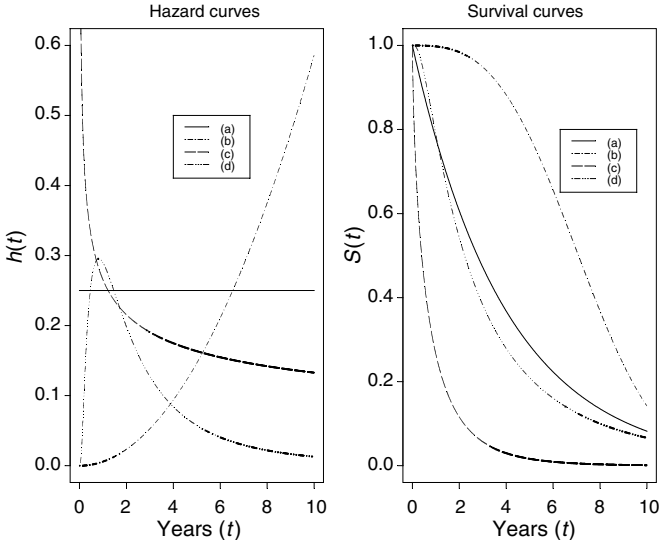


Figure 4 Relationships between (parametric) hazard and survival curves: (a) constant hazard (e.g. healthy persons), (b) increasing Weibull (e.g. leukaemia patients), (c) decreasing Weibull (e.g. patients recovering from surgery), (d) increasing and then decreasing log-normal (e.g. tuberculosis patients).

Survival differences in the lung cancer trial

We have already seen that median survival is greater in the combination treatment arm. Table 3 provides information about (relapse-free) survival differences between the trial arms. A test of differences between median survival times in the groups is indicative of a difference in survival ($P<0.01$). The number of relapses observed among patients treated with radiotherapy + CAP (cytoxan, doxorubin and platinum-based chemotherapy) and radiotherapy alone were 54 and 70, respectively. Using the logrank method, the expected number of relapses for each group were 70.6 and 53.4, respectively. Thus, the logrank test yields a χ^2 value of 9.1 on 1 degree of freedom ($P<0.002$). The HR of 0.58 indicates that there is 42% less risk of relapse at any point in time among patients surviving in the combination treatment group compared with those treated with radiotherapy alone. Overall, there is an indication that the combination treatment is more efficacious than radiotherapy treatment, and may be preventing or delaying relapse.

Survival differences in the ovarian study

In the ovarian study, we wished to compare the survival between patients with different FIGO stages – an ordinal variable. Figure 5 shows overall survival by FIGO stage. A logrank test of trend is statistically significant ($P<0.0001$), and reinforces the visual impression of prognostic separation and a trend towards better survival when the disease is less advanced.

SOME KEY REQUIREMENTS FOR THE ANALYSIS OF SURVIVAL DATA

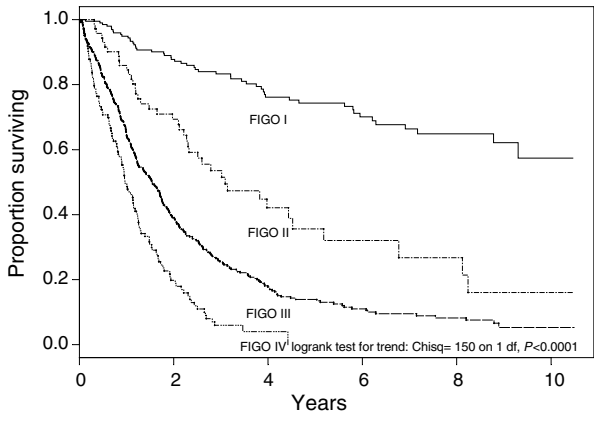
Uninformative censoring

Standard methods used to analyse survival data with censored observations are valid only if the censoring is 'noninformative'. In practical terms, this means that censoring carries no prognostic information about subsequent survival experience; in other words, those who are censored because of loss to follow-up at a given point in time should be as likely to have a subsequent event as those individuals who remain in the study. Informative censoring may occur when patients withdraw from a clinical trial because of drug toxicity or worsening clinical condition. Standard methods for survival analysis are not valid when there is informative censoring. However, when the number of patients lost to follow-up is small, very little bias is likely to result from applying methods based on noninformative censoring.

Table 3 Differences in (relapse-free) survival in the lung cancer trial

	Radiotherapy ($n=86$)	**Radiotherapy+CAP ($n=78$)**
Number of relapses (O_i)	70	54
Median survival time(years) (95% CI)	0.64 (0.45–0.87)	1.10 (0.96–1.59)
Expected number of relapses (E_i)	53.4	70.6
Hazard ratio (95% CI)	0.58 (0.41–0.83)	
Logrank test	$\chi^2=9.1$, 1 df, $P<0.002$	

df = degree of freedom: CAP = cytoxan, doxorubicin and platinum-based chemotherapy.



Number at risk at time (years)	0	2	4	6	8
FIGO I	200	151	93	66	44
FIGO II	72	45	17	10	6
FIGO III	414	142	48	23	14
FIGO IV	123	22	3	1	1

Figure 5 FIGO stage and prognosis in the ovarian study. Chisq = χ^2.

Length of follow-up

In general, the design of a study will influence how it is analysed. Time to event studies must have sufficient follow-up to capture enough events and thereby ensure there is sufficient power to perform appropriate statistical tests. The proposed length of follow-up for a prospective study will be based primarily on the severity of the disease or prognosis of the participants. For example, for a lung cancer trial a 5-year follow-up would be more than adequate, but this follow-up duration will only give a short- to-medium-term indication of survivorship among breast cancer patients.

An indicator of length of follow-up is the median follow-up time. While this could in theory be given as the median follow-up time of all patients, it is better calculated from follow-up among the individuals with censored data. However, both these methods tend to underestimate follow-up, and a more robust measure is based on the reverse KM estimator (Schemper and Smith, 1996), that is the KM method with the event indicator reversed so that the outcome of interest becomes being censored. In the ovarian cohort example, the median follow-up time of all the patients is 1.7 years, although is influenced by the survival times which were early deaths. The median survival of the censored patients was 3.2 years, but the reverse KM estimate of the median follow-up is 5.3 years (95% CI: 4.7 – 6.0 years).

Completeness of follow-up

Each patient who does not have an event can be included in a survival analysis for the period up to the time at which they are censored, but completeness of follow-up is still important. Unequal follow-up between different groups, such as treatment arms, may bias the analysis. A simple count of participants lost to follow-up is one indicator of data incompleteness, but it does not inform us about time lost and another measure has been proposed (Clark *et al*, 2002). In general, disparities in follow-up caused by differential drop-out between arms of a trial or different subgroups in a cohort study need to investigated.

Cohort effect on survival

In survival analysis, there is an assumption of homogeneity of treatment and other factors during the follow-up period. However, in a long-term observational study of patients of cancer, the case mix may change over the period of recruitment, or there may be an innovation in ancillary treatment. The KM method assumes that the survival probabilities are the same for subjects recruited early and late in the study. On average, subjects with longer survival times would have been diagnosed before those with shorter times, and changes in treatments, earlier diagnosis or some other change over time may lead to spurious results. The assumption may be tested, provided we have enough data to estimate survival probabilities in different subsets of the data and, if necessary, adjusted for by further analyses (see next section).

Between-centre differences

In a multicentre study, it is important that there is a consistency between the study methods in each centre. For example, diagnostic instruments, such as staging classification, and treatments should be identical. Heterogeneity in case mix among centres can be adjusted for in an analysis (see next section).

NEED FOR SURVIVAL ANALYSIS ADJUSTING FOR COVARIATES

When comparing treatments in terms of survival, it is often sensible to adjust for patient-related factors, known as covariates or confounders, which could potentially affect the survival time of a patient. For example, suppose that despite the treatment being randomised in the lung cancer trial, older patients were assigned more often to the radiotherapy alone group. This group would have a worse baseline prognosis and so the simple analysis may have underestimated its efficacy compared to the combination treatment, referred to as confounding between treatment and age. Also, we sometimes want to determine the prognostic ability of various factors on overall survival, as in the ovarian study. Figure 5 shows overall survival by FIGO stage, and there is a significant decrease in overall survival with more advanced disease.

Multiple prognostic factors can be adjusted for using multivariate modelling. For example, if those women with early stage disease were younger than those with advanced disease, then the FIGO I and II groups might be surviving longer because of lower age and not because of the effect of FIGO stage. In this case, the FIGO effect is confounded by the effect of age, and a multivariate analysis is required to adjust for the differences in the age distribution. The appropriate analysis is a form of multiple regression, and is the subject of the next paper in this series.

SUMMARY

Survival analysis is a collection of statistical procedures for data analysis where the outcome variable of interest is *time until an event occurs*. Because of censoring – the nonobservation of the event of interest after a period of follow-up – a proportion of the survival times of interest will often be unknown. It is assumed that those patients who are censored have the same survival prospects as those who continue to be followed, that is, the censoring is uninformative. Survival data are generally described and modelled in terms of two related functions, the survivor function and the hazard function. The survivor function represents the probability that an individual survives from the time of origin to some time beyond time *t*. It directly describes the survival experience of a study cohort, and is usually estimated by the KM method. The logrank test may be used to test for differences between survival curves for groups, such as treatment arms. The hazard function gives the instantaneous potential of having an event at a time, given survival up to that time. It is used primarily as a diagnostic tool or for specifying a mathematical model for survival analysis.

In comparing treatments or prognostic groups in terms of survival, it is often necessary to adjust for patient-related factors that could potentially affect the survival time of a patient. Failure to adjust for confounders may result in spurious effects. Multivariate survival analysis, a form of multiple regression, provides a way of doing this adjustment, and is the subject the next paper in this series.

ACKNOWLEDGEMENTS

We thank John Smyth for providing the ovarian cancer data, and Victoria Cornelius and Peter Sasieni for invaluable comments on an earlier manuscript. Cancer Research UK supported all the authors. Taane Clark holds a National Health Service (UK) Research Training Fellowship.

REFERENCES

Altman DG (2003) *Practical statistics for medical research*. London: Chapman & Hall

Altman DG, Bland JM (2003) Statistics notes: interaction revisited: the difference between two estimates. *BMJ* **326:** 219

Clark TG, Altman DG, De Stavola BL (2002) Quantifying the completeness of follow-up. *Lancet* **359:** 1309 – 1310

Clark TG, Stewart ME, Altman DG, Gabra H, Smyth J (2001) A prognostic model for ovarian cancer. *Br J Cancer* **85:** 944 – 952

Collett D (1994) *Modelling Survival Data in Medical Research*. London: Chapman & Hall

Cox DR (1979) A note on the graphical analysis of survival data. *Biometrika* **66:** 188 – 190

Hosmer DW, Lemeshow S (1999) *Applied Survival Analysis: Regression Modelling of Time to Event Data*. New York: Wiley

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* **53:** 457 – 481

Klembaum DG (1996) *Survival analysis*: A self learning text. New York: Springer

Lung Cancer Study Group (1988) The benefit of adjuvant treatment for resected locally advanced non-small cell lung cancer. *J Clin Oncol* **6:** 9 – 17

Parmer M, Machin D (1995) *Survival analysis*. UK: John Wiley and Sons Ltd

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* **35:** 1 – 39

Piantadosi S (1997) *Clinical Trials: A Methodologic Perspective*. New York: Wiley

Pocock S, Clayton TC, Altman DG (2002) Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet* **359:** 1686 – 1689

Ramlau-Hansen H (1983) Smoothing counting process intensities by means of kernel functions. *Ann Statist* **11:** 453 – 466

Schemper M, Smith TL (1996) A note on quantifying follow-up in studies of failure time. *Control Clin Trials* **17:** 343 – 346

British Journal of Cancer (2003) 89, 431 – 436

www.bjcancer.com

Tutorial Paper

Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods

MJ Bradburn*[,1], TG Clark[1], SB Love[1] and DG Altman[1]

[1]*Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Oxford OX3 7LF, UK*

British Journal of Cancer (2003) **89,** 431 – 436. doi:10.1038/sj.bjc.6601119 www.bjcancer.com

Keywords: survival analysis; Cox model; AFT model; model selection

INTRODUCTION

Survival analysis involves the consideration of the time between a fixed starting point (e.g. diagnosis of cancer) and a terminating event (e.g. death). The key feature that distinguishes such data from other types is that the event will not necessarily have occurred in all individuals by the time the study ends, and for these patients, their full survival times are unknown. For instance, in studies that measure the length of survival after diagnosis of cancer, it is common for a proportion of individuals to remain alive and disease-free at the end of the follow-up period, and for these patients, we know only a lower limit on their actual time to event. Thus, special methods are required for these type of data. The explanation and demonstration of some of the methods proposed to analyse such data are the basis of this series.

In the first paper of this series (Clark *et al*, 2003), we described initial methods for analysing and summarising survival data including the definition of hazard and survival functions, and testing for a difference between two groups. We continue here by considering various statistical models and, in particular, how to estimate the effect of one or more factors that may predict survival.

THE NEED FOR MULTIVARIATE STATISTICAL MODELLING

The previous paper demonstrated the construction of (Kaplan–Meier) survival curves for different patient groups, and introduced the logrank test to investigate differences between them. Both these methods are examples of *univariate* analysis; they describe the survival with respect to the factor under investigation, but necessarily ignore the impact of any others. It is more common, at least in clinical investigations, to have a situation where several (known) quantities or *covariates*, potentially affect patient prognosis. For example, suppose two groups of patients are compared: those with and those without a specific genotype. If one of the groups also contains older individuals, any difference in survival may be attributable to genotype or age or indeed both. Hence, when investigating survival in relation to any one factor, it is often desirable to adjust for the impact of others. Moreover, while the logrank test provides a *P*-value for the differences between the groups, it offers no estimate of the actual effect size; in other words, it offers a statistical, but not a clinical, assessment of the factor's impact. The use of a statistical model improves on these methods by allowing survival to be assessed with respect to several factors simultaneously, and in addition, offers estimates of the strength of effect for each constituent factor. Therefore, statistical models are important and frequently used tools which, when constructed appropriately, offer valuable insight into the survival process.

Several statistical methods have been proposed for modelling survival analysis data. We will describe the most important models and illustrate their application using example datasets described in the previous paper (Clark *et al*, 2003). As before, we will assume throughout that all survival times are independent of each other and that censoring occurs solely as right-censoring and is uninformative. The focus is on covariates that are measured at the time of entry to the study, that may be continuous (e.g. the patient age or tumour size), binary (e.g. gender), unordered categorical (e.g. histology) or ordered categorical or ordinal (e.g. performance status or FIGO stage). In the next paper in this series, we will discuss the statistical assumptions made when using statistical models, and provide advice on choosing the appropriate model and covariates therein. We will also consider how to model covariates that change values over time (called 'time-dependent' or 'updated' covariates).

The methods we present here may be divided into two broad categories: proportional hazard approaches (including the semi-parametric Cox model and fully parametric approaches) and accelerated failure time models. These methods have different properties and interpretations, but all may be used to summarise survival data.

THE COX ('SEMI-PARAMETRIC') PROPORTIONAL HAZARDS MODEL

The Cox (proportional hazards or PH) model (Cox, 1972) is the most commonly used multivariate approach for analysing survival time data in medical research. It is a survival analysis regression model, which describes the relation between the event incidence, as expressed by the hazard function and a set of covariates. A fuller explanation of the hazard function was given in the previous article (Clark *et al*, 2003). Put briefly, the hazard is the instantaneous event probability at a given time, or the probability

*Correspondence: Mr M Bradburn; E-mail: mike.bradburn@cancer.org.uk
Received 6 December 2002; accepted 30 April 2003

that an individual under observation experiences the event in a period centred around that point in time.

Mathematically, the Cox model is written as

$$h(t) = h_0(t) \times \exp\{b_1 x_1 + b_2 x_2 + \cdots + b_p x_p\}$$

where the hazard function $h(t)$ is dependent on (or determined by) a set of p covariates (x_1, x_2, …, x_p), whose impact is measured by the size of the respective *coefficients* (b_1, b_2, …, b_p). The term h_0 is called the baseline hazard, and is the value of the hazard if all the x_i are equal to zero (the quantity exp(0) equals 1). The 't' in $h(t)$ reminds us that the hazard may (and probably will) vary over time. An appealing feature of the Cox model is that the baseline hazard function is estimated nonparametrically, and so unlike most other statistical models, the survival times are not assumed to follow a particular statistical distribution.

The Cox model is essentially a multiple linear regression of the logarithm of the hazard on the variables x_i, with the baseline hazard being an 'intercept' term that varies with time. The covariates then act multiplicatively on the hazard at any point in time, and this provides us with the key assumption of the PH model: the hazard of the event in any group is a constant multiple of the hazard in any other. This assumption implies that the hazard curves for the groups should be proportional and cannot cross (see Figure 1 for examples of each). Proportionality implies that the quantities exp(b_i) are called *hazard ratios*. A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the ith covariate increases, the event hazard increases and thus the length of survival decreases. Put another way, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival. This *proportionality assumption* is often appropriate for survival time data but it is important to verify that it holds. We discuss methods for assessing proportionality in the next paper in this series.

The Cox PH model fitted to the ovarian cancer data

This large database, as described in the previous paper of this series (Clark *et al*, 2003), was used to derive a prognostic index for overall survival among ovarian cancer patients in Clark *et al* (2001). Their analysis included 10 variables, but for simplicity we will consider five, all of which were measured at diagnosis: FIGO stage (an ordinal covariate taking values of 1, 2 3 or 4), histology (one of seven subtypes), grade (1, 2 or 3), ascites (yes/no) and patient age.

Table 1 shows the effect sizes (given as hazard ratios), 95% confidence intervals (CI), regression coefficients and statistical significance for each of these in relation to overall survival. Each factor is assessed through separate univariate Cox regressions (left-hand columns). However, the aim of the database is to describe how the factors jointly impact on survival, and so all five factors were incorporated into the multivariate model (right-hand columns). It may be seen that higher FIGO stage, higher grade, presence of ascites and increased age impaired survival to varying (and statistically significant) degrees. The histology was also of importance: the figures describe the survival of patients with each histology type in comparison with the serous type. In principle, any type with a reasonable number of patients could be chosen as the baseline of comparison. On multivariate analysis Mucinous and serous were the tumour types with the best prognosis, whereas undifferentiated and mixed mesodermal were the worst. It is possible to present P-values for the comparisons between each type and serous, but we have given an overall likelihood ratio test for the differences between the categories as a whole. The FIGO stage could be modelled as a categorical variable in the same manner as grade and histology, but assuming it is a continuous variable with a linear trend across the four categories performed sufficiently well.

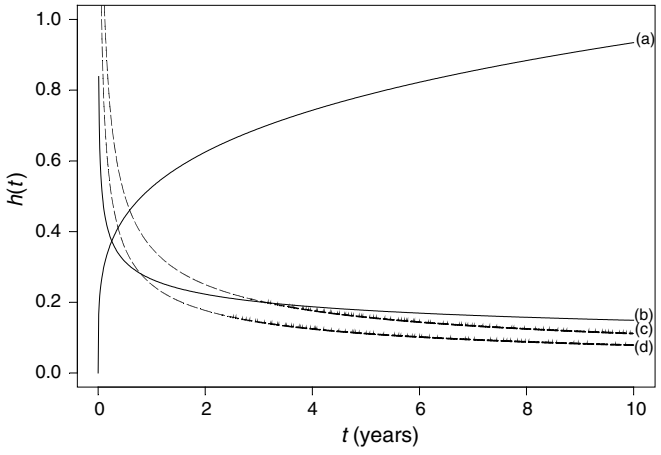


Figure 1 Example of (non-) proportional hazards (groups (c) and (d) only have proportional hazards) using the Weibull distribution. For the Weibull survival model, the hazard function $h(t) = \lambda s(\lambda t)^{s-1}$ for λ, $s > 0$: (a) increasing hazard ($\lambda = 0.5$, $s = 1.25$); (b) decreasing hazard ($\lambda = 0.25$, $s = 0.75$); (c) decreasing hazard ($\lambda = 0.5$, $s = 0.5$); (d) decreasing hazard ($\lambda = 0.25$, $s = 0.5$).

PARAMETRIC PH MODELS

Parametric PH models are a class of models similar in concept and interpretation to the Cox (PH) model. The key difference between the two is that the hazard is assumed to follow a specific statistical distribution when a fully parametric PH model is fitted to the data, whereas the Cox model enforces no such constraint. Other than this, the two model types are equivalent. Hazard ratios have the same interpretation, whether derived from a Cox or a fully parametric regression model, and the proportionality of hazards is still assumed.

A number of different parametric PH models may be derived by choosing different hazard functions. As shown previously, there is a direct link between the survival and hazard, and the choice of hazard distribution determines that of the survival. In fact, the models commonly applied, such as the *Exponential*, *Weibull* or *Gompertz* models, take their names from the distribution that the survival times are assumed to follow, but the most distinguishing features between them are in the hazard function. Examples of survival and hazard functions derived from some of these parametric models were presented in the previous paper (Clark *et al*, 2003). Figure 1 shows increasing and decreasing Weibull hazard functions, as well as two groups with the latter that are proportional to each other.

Parametric models fitted to the ovarian cancer data

The estimated hazard function of the ovarian cancer data as displayed in the previous paper (Clark *et al*, 2003) may be consistent with that derived from a Weibull PH model with decreasing hazard. Fitting this to the ovarian cancer database gives similar results as the Cox model (see Table 2), and may be interpreted in the same manner. Methods to check for the appropriateness of the Weibull distribution will be discussed in the next paper of this series.

COMPARISON OF THE TWO PH APPROACHES

The main drawback of parametric models is the need to specify the distribution that most appropriately mirrors that of the actual

Table 1 Hazard ratios from the Cox PH model for the ovarian dataset

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**
FIGO stage	0.809	2.24	(2.03–2.48)	<0.001	0.731	2.08	(1.82–2.37)	<0.001
Histology				<0.001				<0.001
Serous	(0.000)	(1.00)			(0.000)	(1.00)		
Mucinous	−0.727	0.48	(0.38–0.61)		−0.422	0.66	(0.50–0.85)	
Endometroid	−1.162	0.31	(0.22–0.45)		0.198	1.22	(0.80–1.85)	
Clear cell	−0.343	0.71	(0.52–0.97)		0.342	1.41	(0.99–2.00)	
Adenocarcinoma	0.119	1.13	(0.74–1.72)		0.501	1.65	(0.91–2.99)	
Undifferentiated	0.390	1.48	(0.81–2.70)		0.746	2.11	(1.03–4.29)	
Mixed mesodermal	0.614	1.85	(1.28–2.66)		0.789	2.20	(1.45–3.35)	
Grade				<0.001				<0.001
1	(0.000)	(1.00)			(0.000)	(1.00)		
2	1.116	3.05	(1.90–4.91)		0.885	2.42	(1.40–4.19)	
3	1.650	5.20	(3.31–8.18)		0.885	2.42	(1.40–4.18)	
Absence of ascites	−0.798	0.45	(0.37–0.55)	<0.001	−0.396	0.67	(0.54–0.84)	<0.001
Age (per 5-year increase)	0.153	1.17	(1.12–1.21)	<0.001	0.133	1.14	(1.09–1.19)	<0.001

HR = hazard ratio, CI = confidence interval.

Table 2 Hazard ratios from the Weibull PH model for the ovarian dataset

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**
FIGO stage	0.862	2.37	(2.14–2.62)	<0.001	0.768	2.16	(1.89–2.46)	<0.001
Histology				<0.001				<0.001
Serous	(0.000)	(1.00)			(0.000)	(1.00)		
Mucinous	−0.804	0.45	(0.35–0.57)		−0.496	0.61	(0.47–0.79)	
Endometroid	−1.276	0.28	(0.20–0.40)		0.120	1.13	(0.75–1.70)	
Clear cell	−0.419	0.66	(0.48–0.90)		0.346	1.41	(0.99–2.02)	
Adenocarcinoma	0.113	1.12	(0.73–1.71)		0.499	1.65	(0.91–2.97)	
Undifferentiated	0.397	1.49	(0.82–2.71)		0.765	2.15	(1.06–4.37)	
Mixed Mesodermal	0.638	1.89	(1.31–2.73)		0.804	2.23	(1.47–3.40)	
Grade								<0.001
1	(0.000)	(1.00)		<0.001	(0.000)	(1.00)		
2	1.154	3.17	(1.97–5.10)		0.928	2.53	(1.47–4.36)	
3	1.727	5.62	(3.58–8.84)		0.895	2.45	(1.43–4.20)	
Absence of ascites	−0.840	0.43	(0.36–0.52)	<0.001	−0.404	0.67	(0.54–0.83)	<0.001
Age (per 5-year increase)	0.165	1.18	(1.14–1.22)	<0.001	0.138	1.15	(1.10–1.20)	<0.001

HR = hazard ratio, CI = confidence interval.

survival times. This is an important requirement that needs to be verified and an appropriate distribution may be difficult to identify. Where a suitable distribution can be found, however, the parametric model is more informative than the Cox model. It is straightforward to derive the hazard function and to obtain predicted survival times when using a parametric model, which is not the case in the Cox framework (the use of such quantities is discussed in the next section). Additionally, the appropriate use of these models offers the advantage of being slightly more *efficient*; they yield more precise estimates (i.e. smaller standard errors).

The results from the Cox or parametric PH models may be compared directly, as the model types are merely different approaches to assessing the same quantity. For either method to be valid: (a) the covariate effect needs to be at least approximately constant throughout the duration of the study, and (b) the proportionality assumption must hold. These important issues will be addressed in the subsequent paper in this series.

INTERPRETING THE PH MODEL: BEYOND THE HAZARD RATIO

In addition to the ratio of two hazards, it is possible to obtain other information from a PH regression model. One simple (and possibly underused) quantity that may be derived from a survival model is the predicted survival proportion at any given point in time for a particular risk group. The survival proportion for a given risk group at any time, $S(t)$, is equal to

$$S(t) = S_0(t)^{\exp(\gamma)}$$

where $S_0(t)$ is the baseline survival (the survival proportion when all covariates are equal to zero) and γ is equal to $b_1x_1 + b_2x_2 + \cdots + b_px_p$. Once the value of the baseline survival at a given time is derived, then the predicted survival probabilities for patients with any specified covariate values x_i are easily obtained. This information could then be displayed via tabular or

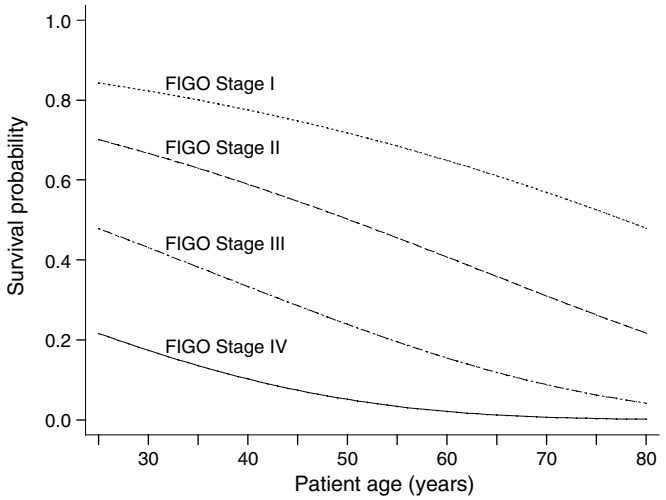


Figure 2 Predicted 5-year survival of ovarian cancer patients by age and FIGO stage.

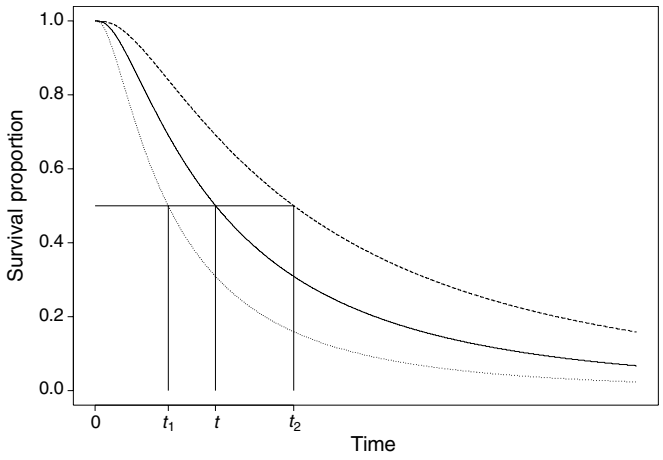


Figure 3 Illustration of the AFT model: (——), $S_0(t)$ the baseline survival function; (·····), $S(t_1) = S_0(\varphi t)$ for $\varphi < 1$; (------), $S(t_2) = S(\varphi t)$ for $\varphi > 1$.

graphical displays. Figure 2 illustrates this by giving predicted 5-year survival according to patient age and FIGO stage. Further examples are demonstrated by Christensen (1987) based on the Cox model, but can also be used when fitting fully parametric models. In a previous analysis that involved some of the patients in the present data, Clark *et al* (2001) produced a nomogram to summarise the impact of these and other covariates, and thus allows the reader to predict the median survival and the 2- and 5-year survival probabilities for patients with given prognostic information.

The advantage of fitting a parametric survival model is that predictions of the event survival, event hazard, mean and median survival times are readily available. For FIGO stages I–IV, the median survival times are estimated to be 7.8, 4.0, 2.0 and 1.0 years, respectively.

ACCELERATED FAILURE TIME MODELS

The accelerated failure time (AFT) model is a different type of model that may be used for the analysis of survival time data. For a group of patients with covariates (x_1, x_2, … x_p), the model is written mathematically as

$$S(t) = S_0(\varphi t)$$

where $S_0(t)$ is the baseline survivor function and φ is an 'acceleration factor' that depends on the covariates according to the formula

$$\varphi = \exp\{(b_1x_1 + b_2x_2 + \cdots + b_px_p)\}.$$

The principle here is that the effect of a covariate is to stretch or shrink the survival curve along the time axis by a constant relative amount φ. Figure 3 demonstrates this for the case of a single covariate (x_1) with two levels, for example, $x_1 = 0$ for a placebo group and $x_1 = 1$ for a new treatment group. The survival probabilities, $S(t)$, for the placebo and new treatment groups are $S_0(t)$ and $S_0(\varphi t)$, respectively. The proportion of patients who are event-free in the placebo group at any time point t_1 is the same as the proportion of those who are event-free in the new treatment group at a time $t_2 = \varphi t_1$. Figure 3 shows the cases where $\varphi > 1$ and $\varphi < 1$, which represent situations where the length of survival is increased and decreased in the new treatment group compared with the placebo, respectively.

The AFT model is commonly rewritten as being log-linear with respect to time, giving

$$\log(T) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \varepsilon$$

where ε is a measure of (residual) variability in the survival times. Thus, the survival times can be seen to be multiplied by a constant effect under this model specification, and the exponentiated coefficients, $\exp(b_i)$, are referred to as *time ratios*. A time ratio above 1 for the covariate implies that this 'slows down', or prolongs the time to the event, while a time ratio below 1 indicates that an earlier event is more likely.

When the survival times follow a Weibull distribution, it can be shown that the AFT and PH models are the same. However, the AFT family of models differs crucially from the PH model types in terms of their interpretation of effect sizes as time ratios as opposed to hazard ratios.

The survival times are usually assumed to follow a specific distributional form in the AFT framework. Distributions such as the *Log-Normal, Log-Logistic, Generalised Gamma* and *Weibull* may be used to represent such survival data. Alternative methods include the method of Buckley and James (1979), which is discussed by Stare *et al* (2000), and semiparametric AFT models, in which the baseline survivor function is estimated nonparametrically (see Wei, 1992, for an overview), but have not yet been widely implemented in statistical software.

As with the PH approach, other quantities such as projected survival probabilities may be derived. Also in keeping with PH models is the fact that AFT models make assumptions; the appropriate choice of statistical distribution needs to be made, and also the covariate effects are assumed to be constant and multiplicative on the timescale, that is, that the covariate impacts on survival by a constant factor.

Parametric AFT models fitted to the lung cancer trial data

We use the non-small cell lung cancer dataset to illustrate the AFT model, focusing on the relapse-free survival (i.e. , the time from diagnosis to the reappearance of cancer, with patients censored at time of death if no recurrence had appeared). Again, we present both the univariate and multivariate effect sizes in Table 3. The specific comparison of interest was the effect of adjuvant (platinum-based) chemotherapy and radiotherapy compared with radiotherapy alone. The unadjusted treatment effect may be

Table 3 Time ratios from the generalised gamma AFT model for the lung cancer trial

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**TR exp(b_i)**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**TR exp(b_i)**	**95% CI**	***P*-value**
Treatment (RT+CAP *vs* RT alone)	0.648	1.91	(1.21–3.01)	0.005	0.718	2.05	(1.29–3.23)	0.002
Cell type (Sq *vs* non-Sq)	0.506	1.66	(1.01–2.71)	0.04	0.511	1.67	(1.04–2.68)	0.03
Performance status (8–10 *vs* 5–7)	0.767	2.15	(1.11–4.19)	0.02	0.729	2.07	(1.00–4.29)	0.05
Tumour status				0.59				0.60
1	(0.000)	(1.00)	—		(0.000)	(1.00)	—	
2	−0.189	0.83	(0.40–1.70)		−0.353	0.70	(0.35–1.41)	
3	−0.388	0.68	(0.31–1.48)		−0.378	0.69	(0.31–1.53)	
Nodal involvement				0.87				0.97
None	(0.000)	(1.00)	—		(0.000)	(1.00)	—	
Limited	0.122	1.13	(0.46–2.79)		−0.059	0.94	(0.36–2.48)	
Extensive	0.206	1.23	(0.55–2.73)		0.029	1.03	(0.42–2.54)	
Age at diagnosis (/years)	−0.013	0.99	(0.96–1.01)	0.34	−0.011	0.99	(0.96–1.01)	0.41
Gender (male *vs* female)	0.032	1.03	(0.62–1.71)	0.90	−0.007	0.99	(0.59–1.67)	0.98
Weight loss ($\geqslant$10 *vs* $<$10%)	−0.477	0.62	(0.29–1.33)	0.22	−0.337	0.71	(0.34–1.51)	0.38
Race (white *vs* non-white)	0.440	1.55	(0.81–2.98)	0.19	0.202	1.22	(0.61–2.46)	0.57

TR = time ratio, CI = confidence interval, RT = radiotherapy, CAP = cytoxan, doxorubicin and platinum-based chemotherapy, Sq = squamous.

summarised by a time ratio of 1.91 (95% CI: 1.21–3.01; $P = 0.005$), which, having allowed for other covariates increased slightly to 2.05. Therefore, we can conclude that the time to recurrence was significantly prolonged (approximately doubled) among patents given adjuvant chemotherapy in comparison with those who were not.

Again, we can derive model-based predictions: overall, patients allocated to receive adjuvant chemotherapy had a predicted median survival time of approximately 16 months, as opposed to 8 months among those treated with radiotherapy alone. Other factors are also significant and would influence these times, but these are of less importance in the context of the comparative trial. We will return to this example in the next paper of this series.

WHICH MODEL SHOULD WE USE: PH OR AFT?

From a statistical viewpoint, an obvious way to choose between the two model types is to fit a type that is in keeping with the data. If the AFT model clearly fits the data better than the PH model, or *vice versa*, this model may be adopted as being the more appropriate. However, in some cases, either type of model may appear to fit the data adequately. In such instances, the choice of model may be influenced by other factors. For instance, if other studies of a similar nature had all used the Cox regression and reported the results as hazard ratios, one may be tempted to follow suit to aid comparability. Against this, the parametric approach offers more in the way of predictions, and the AFT formulation allows the derivation of a time ratio, which is arguably more interpretable than a ratio of two hazards. As yet, however, AFT models are relatively unfamiliar and seen rarely in medical research papers (see Kay and Kinnersley, 2002).

OTHER APPROACHES

Stratified survival analysis

A more straightforward way to incorporate covariates into a survival analysis is to use a stratified survival analysis. For example, suppose the covariate of primary interest is treatment, but we wish to control for the clinical stage of the tumour when making the comparison. Here, the survival in each treatment group can be compared within each stage of disease (the 'strata') by the logrank or some other method, and the differences within each stratum are then combined to give an overall comparison of treatments that has been adjusted for the stage.

The strength of this method is in its simplicity: as the logrank test is nonparametric, few distributional assumptions are made, and its interpretation is straightforward. Its main limitation is that it is only applicable when the covariate is categorical (or with continuous variables that have been arbitrarily categorised). Further, this method does not perform well with several covariates, as the number of individuals in each stratum quickly becomes too small to allow reasonable comparisons. In addition, it does not quantify the strength of effect of each variable, or even offer a *P*-value for factors other than the one of primary interest. This method is not generally regarded as a formal statistical model, but is of use where a very small number of covariates are to be considered, if only as an exploratory method of analysis.

Aalen's additive model

Another approach to modelling the relationship between survival and covariates is to assume that the covariates act additively on the hazard. Aalen's additive hazard model (Aalen, 1989) is one method that has been suggested for this, but its properties are rather unlike any other model described in this paper. The covariates are assumed to impact additively upon a (unknown) baseline hazard, but the effects are not constrained to be constant. The impact is therefore allowed to vary freely over time according to the underlying equation

$$h(t) = h_0(t) + b_1(t)x_1 + b_2(t)x_2 + \cdots + b_p(t)x_p$$

where $h(t)$ is the hazard, $h_0(t)$ is the baseline hazard and the $b_i(t)$ are coefficients, which may change in magnitude and even sign with time. Compare this with the Cox regression, where $h_0(t)$ is also estimated nonparametrically, but the b_i quantify the *multiplicative* effect of covariate i on the hazard and are assumed constant at all times.

As it is not straightforward to estimate $h_0(t)$ nonparametrically, the cumulative baseline hazard is used and the regression coefficients that are actually estimated from the data are also the cumulative (additional) hazard

$$B_i(t) = \int_0^t b_i(u)\mathrm{d}u$$

The usual method of representing these effects is to graph them against time. The further $B_i(t)$ is from zero at time t, the greater the effect the covariate has had on the hazard over the course of the study up to t. The values of $b_i(t)$, the absolute increase in hazard at

time *t*, are not actually observed, but their relative size may be inferred from the slope of the line. These plots are sometimes called Aalen plots, and they are also used to provide an informal assessment of the adequacy of the proportional hazards assumption in the Cox model, although Aalen considered its primary role as an alternative model in its own right (Aalen, 1993).

The flexibility of this approach is tempered by the lack of an easy interpretation. The $B_i(t)$ coefficients are not easy to understand, and as they change repeatedly over time, can offer no single quantifiable effect size. Formal tests of statistically significant covariate effects may be carried out, but Aalen plots are essentially the only manner with which to interpret the effect sizes. These reasons, together with the relative lack of statistical software, are probably the deciding factors in the relatively minimal use of Aalen's model.

Classification trees and artificial neural networks

Two relatively recent developments are classification trees and artificial neural networks. These methods differ substantially in their complexity and interpretation to the methods presented here and to each other. Both approaches are described in more detail in a later paper of this series.

DISCUSSION

The principal strength of statistical models is their ability to assess several covariates simultaneously. The strengths of the stratified logrank test and other such methods are their obvious simplicity and the fact that they make fewer parametric assumptions of the data. Although these reasons are usually insufficient to suggest that the stratified method be used more widely, this second feature is a relevant one, because it needs to be kept in mind that all the models introduced here make certain distributional assumptions of the survival times that will not always be met.

We have focused on the Cox model, the class of parametric PH models and AFT models as tools with which to analyse survival time data. Other models exist (see, e.g., Collett (1994) for a more practical demonstration of some alternatives and Bagdonavičius and Nikulin (2001) for the theoretical background), but many are similar to, if not extensions of, the approaches we have discussed. The use of the Cox model offers greater flexibility than parametric alternatives and, in particular, does not require the direct estimation of the baseline hazard function (i.e. it avoids the need to specify the distribution of the survival times). However, the assumption of proportional hazards is a crucial one that needs to be fulfilled for the results to be meaningful, and will not always be satisfied. Further, while the Cox PH model may be valid, other parametric models will produce more precise estimates where the distribution is specified correctly.

A further concern is that the choice of covariates to include is also far from simple. In the third paper of this series, we will consider ways to choose between the various model types, to identify and assess the importance of covariates, and to verify that the final model is adequate.

ACKNOWLEDGEMENTS

We wish to thank John Smyth for providing the ovarian cancer data, and Victoria Cornelius and Peter Sasieni for invaluable comments on an earlier manuscript. Cancer Research UK supported all authors. Taane Clark holds a National Health Service (UK) Research Training Fellowship.

REFERENCES

Aalen OO (1989) A linear regression model for the analysis of life times. *Stat Med* **8:** 907 – 925

Aalen OO (1993) Further results on the non-parametric linear regression model in survival analysis. *Stat Med* **12:** 1569 – 1588

Bagdonavičius V, Nikulin M (2001) *Accelerated Life Models: Modeling and Statistical Analysis*. London: Chapman & Hall/CRC

Buckley K, James I (1979) Linear regression with censored data. *Biometrics* **66:** 429 – 436

Christensen E (1987) Multivariate survival analysis using Cox's regression model. *Hepatology* **7:** 1346 – 1358

Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis. Part I: basic concepts and first analyses. *Br J Cancer* **89:** 232–238

Clark TG, Stewart ME, Altman DG, Gabra H, Smyth J (2001) A prognostic model for ovarian cancer. *Br J Cancer* **85:** 944 – 952

Collett D (1994) *Modelling Survival Data in Medical Research*. London: Chapman and Hall/CRC

Cox DR (1972) Regression models and life tables (with discussion). *J R Statist Soc B* **34:** 187 – 220

Kay R, Kinnersley N (2002) On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: a case study in influenza. *Drug Inf J* **36:** 571 – 579

Stare J, Heinzl H, Harrell F (2000) On the use of Buckley and James least squares regression for survival data. New approaches in applied statistics: Metodološki zvezki 16 (http://mrvar.fdv.uni-lj.si/pub/mz/mz16/stare.pdf)

Wei LJ (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med* **11:** 1871 – 1879

British Journal of Cancer (2003) 89, 605–611

www.bjcancer.com

Tutorial Paper

Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit

MJ Bradburn*[,1], TG Clark[1], SB Love[1] and DG Altman[1]

[1]*Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Oxford OX3 7LF, UK*

British Journal of Cancer (2003) **89,** 605–611. doi:10.1038/sj.bjc.6601120 www.bjcancer.com

Keywords: survival analysis; Cox model; AFT model; model checking; choice of coavriates; goodness of fit

INTRODUCTION

In this series of papers, we have described a selection of statistical methods used for the initial analysis of survival time data (Clark *et al*, 2003), and introduced a selection of more advanced methods to deal with the situation where several factors impact on the survival process (Bradburn *et al*, 2003). The latter paper focused on proportional hazards (PH) and accelerated failure time (AFT) models, and we continue the series by demonstrating the application of these models in more detail. Whereas the focus of the previous paper was to outline the purpose and interpretation of statistical models for survival analysis, we concentrate here on approaches with which to undertake the actual modelling process. In other words, the aim of this paper is to promote the correct use of the models that have been suggested for the analysis of survival data.

When used inappropriately, statistical models may give rise to misleading conclusions. Checking that a given model is an appropriate representation of the data is therefore an important step. Unfortunately, this is a complicated exercise, and one that has formed the subject of entire books. Here, we aim to present an overview of some of the major issues involved, and to provide general guidance when developing and applying a statistical model. We start by presenting approaches that can be used to ensure that the correct factors have been chosen. Following this, we describe some approaches that will help decide whether the statistical model adequately reflects the survivor patterns observed. Lastly, we describe methods to establish the validity of any assumptions the modelling process makes. We will illustrate each using the two example datasets (a lung cancer trial and an ovarian cancer dataset) that were introduced in the previous papers (Bradburn *et al*, 2003; Clark *et al*, 2003).

CHOICE OF COVARIATES

The covariates that we consider here are fixed, that is, known at baseline or entry to the study. The handling of covariates that change values over time (e.g. white blood cell count as measured at different time points) will be described in the subsequent paper in this series.

*Correspondence: Mr M Bradburn; E-mail: mike.bradburn@cancer.org.uk
Received 6 December 2002; accepted 30 April 2003

Sample size considerations

It is implicitly assumed that the subjects in a study are representative of a wider population to enable the study aims to be addressed. Another important requirement is to have data from an adequate number of subjects. Any estimate based on a small number of individuals will be less reliable than one based on a larger number, and when multivariate models are fitted to small datasets, the estimated impact of the covariates is too imprecise to give reliable answers. The use of variable selection procedures as described below is especially problematic with such data, and often leads to overoptimistic results. Finally, smaller data sets may not have sufficient power to detect a covariate that has a significant impact on survival.

The power (and indeed in some cases validity) of a survival analysis is related to the number of events rather than the number of participants. Simulation work has suggested that at least 10 events need to be observed for each covariate considered, and anything less will lead to problems, for example, the regression coefficients become biased (Peduzzi *et al*, 1995). In the ovarian study, there were 550 deaths and 11 covariates for the five prognostic factors, implying 50 events per covariate. In the liver cancer trial with 114 events, a full model of 11 covariates has approximately 10 events per covariate.

For prospective studies, several books (e.g. Machin *et al*, 1998) and software packages (e.g. nQuery, power and precision) are available to assist the calculation of adequate sample sizes, and many general purpose statistical packages also perform such calculations.

The aim of the study influences the choice of covariates

Before embarking on any statistical modelling, it is helpful to be clear as to why the multivariate model is to be fitted. The models we have presented have the considerable advantage of being able to handle several factors simultaneously, but the choice of which to incorporate lies with the analyst. This choice depends on the study aims. We suggest three possible scenarios as to why a study may use a multivariate model, and deal with each in turn.

(a) A single factor is under investigation for its association with survival, but several other factors exist

The rationale of such a study is to perform a specific test of one factor. This scenario may arise in a randomised controlled trial, such as the lung cancer example, where the aim is to decide

whether a new treatment prolongs survival, but also to adjust for prognostic factors that may or may not be equally matched between treatment groups. Another situation occurs where an association between a marker and patient survival is being assessed. In either case, any terms that are of potential importance could be incorporated whether significant or not, depending on the adequacy of the sample size. All of the covariates (other than the one of primary interest) are essentially 'nuisance' factors that are considered only to ensure they have been taken due account for in assessing the importance of the (prespecified) factor under investigation. Less important covariates may be removed.

(b) A collection of factors of known relevance are under investigation for their ability to predict survival
This arises when one wishes to assess the individual importance of a series of factors, and/or to attempt to build a model that helps predict patient survival. In such cases, the simplest strategy is to attempt to model all covariates, obtain effect sizes and gauge how well the model predicts survival. It may be desirable to remove factors from the model for simplicity, provided this does not compromise the predictive ability of the model. Statistical significance alone is an insufficient measure of assessing the extent to which a covariate can predict survival. Methods that may be used for this evaluation are given in the final paper of this series.

(c) Where a collection of factors are under investigation for their potential association with survival, possibly with additional known factors
Such studies are more 'exploratory' in nature, and the aim is to identify quantities of potential importance for further investigation. Here it is often desired to reduce the number of covariates in the model by excluding those that are not statistically significant and thus concentrate only on 'potentially interesting' ones for future research. Care must be exercised when several covariates are investigated, as the false-positive rate (or the chance of finding a spurious effect) increases with each additional test.

This selection of scenarios is far from exhaustive, and in practice a study may combine all of the above types. The ovarian study is a combination of (b) and (c).

Approaches to adding or removing covariates

Common choices for model building focus on 'semiautomated' methods such as stepwise selection, but other approaches exist. Models that are based purely on statistical significance may not be clinically meaningful. Henderson and Velleman (1981) state this simply: 'The data analyst knows more than the computer', and appropriate use of this knowledge should be incorporated into the analysis. We recommend that the choice of covariates should be verified by a degree of hands-on modelling, where terms are added or removed in a logical order rather than solely according to statistical significance.

We illustrate some straightforward approaches to the choice of covariates in the two example datasets used in this paper. In the final paper of this series, we will outline the rationale behind semiautomated methods (together with their limitations) and give further advice on hands-on modelling.

Selecting covariates for the lung cancer trial

As stated before, the lung cancer trial as presented in the earlier paper is an example of scenario (a). The table of coefficients for the full AFT multivariate model was presented in the previous paper (Bradburn *et al*, 2003). A simpler model would be to consider just the performance status, cell type and treatment covariates. Removing the remaining covariates reduces the model likelihood, but not to a significant degree ($\chi^2 = 3.34$ on 8 degrees of freedom; $P = 0.91$). The new time ratios, confidence intervals and *P*-values are presented in Table 1. They are virtually unchanged from the previous analysis, and thus the earlier conclusions remain the same.

Selecting covariates for the ovarian cancer database

As stated previously, the analysis of the ovarian cancer database (as described in the previous paper) could be considered as a mixture of scenarios (b) and (c). However, as the database is large and the aim is to derive a prognostic model, we will focus on (b). We consider five covariates here, all of which were measured at diagnosis: FIGO stage (an ordinal covariate taking values 1, 2, 3 or 4), histology (with seven possible subtypes), grade, ascites (yes/no) and patient age.

In this analysis, all the covariates were included yielding the model presented in Table 2. Advanced FIGO stage, higher grade, presence of ascites and increased age all impaired survival to varying degrees. The mucinous and serous histology types had a better prognosis, and undifferentiated and mixed mesodermal a lesser one. No grade–histology interactions were included in the final model, either due to insufficient numbers of patients to allow meaningful modelling (e.g. clear cell, mixed mesodermal, adenocarcinoma or undifferentiated), or for statistical insignificance (the remainder). In fact no second-order interaction or nonlinearity was detected.

If this model were to be used for the purpose of predicting future survival patterns, it is appropriate to ensure that the effect sizes are robust. One approach is to use bootstrap sampling, which involves randomly resampling the data and fitting the model to these modified datasets (Clark and Altman, 2002). These produce a series of effect sizes that should be similar to those derived from the original data if the model is sufficiently stable, and indeed do so here.

ASSESSING THE ADEQUACY OF A MODEL

Regardless of which type of model is fitted and how the variables are selected to be in the model, it is important to evaluate how well the model represents the data. A survival model is adequate if it

Table 1 Generalised gamma AFT model applied to the lung cancer data

Covariate	Coefficient (b_i)	TR [exp(b_i)]	95% CI	*P*-value
Treatment (RT+CAP *vs* RT alone)	0.640	1.90	(1.23–2.93)	0.004
Cell type (squamous *vs* nonsquamous)	0.536	1.71	(1.08–2.71)	0.02
Performance status (8–10 *vs* 5–7)	0.765	2.15	(1.09–4.24)	0.03

TR = time ratio; CI = confidence Interval; RT = radiotherapy; CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

Table 2 Cox model applied to the ovarian data

Covariate	Coefficient (b_i)	HR [exp(b_i)]	95% CI	*P*-value
FIGO stage	0.731	2.08	(1.82–2.37)	<0.001
Histology				<0.001
Serous	(0.000)	(1.00)		
Mucinous	−0.422	0.66	(0.50–0.85)	
Endometroid	0.198	1.22	(0.80–1.85)	
Clear cell	0.342	1.41	(0.99–2.00)	
Adenocarcinoma	0.501	1.65	(0.91–2.99)	
Undifferentiated	0.746	2.11	(1.03–4.29)	
Mixed mesodermal	0.789	2.20	(1.45–3.35)	
Grade				<0.001
1	(0.000)	(1.00)		
2	0.885	2.42	(1.40–4.19)	
3	0.885	2.42	(1.40–4.18)	
Absence of ascites	−0.396	0.67	(0.54–0.84)	<0.001
Age (per 5-year increase)	0.133	1.14	(1.09–1.19)	<0.001

HR = hazard ratio; CI = confidence interval.

Table 3 Suggested plots for residual-based diagnostics

Y-axis	*X*-axis	Potential implication	Suggested remedy
Martingale residual	Any omitted covariate	Covariate excluded wrongly	Refit model with covariate included
Martingale residual	Any included covariate	Covariate modelled incorrectly (e.g. nonlinear effect)	Fit nonlinear term (e.g. a squared term)
Martingale residual	Date of enrolment in study	Evidence of temporal effect	Incorporate time of entry as a covariate
Deviance residual	Survival time, log(survival time) or ranks of survival time	Model fails to predict consistently for all survival times	Fit time-dependent PH model or consider using a different (i.e. non-PH) model
Deviance residual	Subject identifier	Individual is an outlier	(1) Check if the data are correct (2) Refit model with individual removed. If effect sizes alter substantially, consider removing individual altogether
Scaled Schoenfeld residual	Survival time, log(survival time) or ranks of survival time	Non-PH	Fit time-dependent PH model or consider using a different (i.e. non-PH) model

PH = proportional hazards. All of the above *X*–*Y* plots should give rise to a plot evenly scattered along a horizontal line that displays no trend. The possible implications where this does not occur and suggested remedies are presented.

represents the survival patterns in the data to an acceptable degree. This aspect of a model is known as *goodness of fit*. For example, if a given group of patients have a poor (or good) prognosis, then the model should predict this group to have that outcome. In practice, the issues in choosing the most appropriate type of model and the most appropriate covariates are heavily related, and the adequacy of a model may be assessed in several ways. In this section, we discuss methods to verify fit that are common across all survival models, before describing approaches specific to different model types. We will use the ovarian database example to demonstrate these checks.

Residuals from survival models

Residuals are a useful method for checking the fit of a statistical model. Essentially, they are the difference between an observed and a model-predicted quantity, with large or systematic differences between the two indicative of a poor model. Several residuals have been proposed, but unfortunately most are rather difficult to understand in the context of survival analyses due to censoring (Collett, 1994). In general, the residuals are skewed and need to have smoothing functions (e.g. Kernel smoother) applied to aid interpretation. Nevertheless, the graphical displays suggested in Table 3 (with appropriate smoothing as required) should all give rise to an evenly scattered horizontal band and display no obvious trend (e.g. no slope). If a trend in these plots is apparent, it should be investigated, perhaps using the method suggested in Table 3. Overall model adequacy may be assessed by use of Cox-Snell residuals (Collett, 1994).

Residual plots for the ovarian cancer data set

Figures 1A illustrates a plot of the Martingale residuals against the patient's age, with a Kernel smoother marked as the dashed line. Figure 1B shows the Martingale residuals plotted against FIGO stage, with the median within each stage represented by the solid bar. Both FIGO and age were modelled as linear effects. If FIGO or age had been wrongly excluded or modelled incorrectly (i.e. nonlinear), the figures should display a trend other than a strictly horizontal line. The age residual plot shows no evidence of a trend. Although there appears to be evidence of a trend in the FIGO plot, the inclusion of this covariate as a categorical covariate fails to

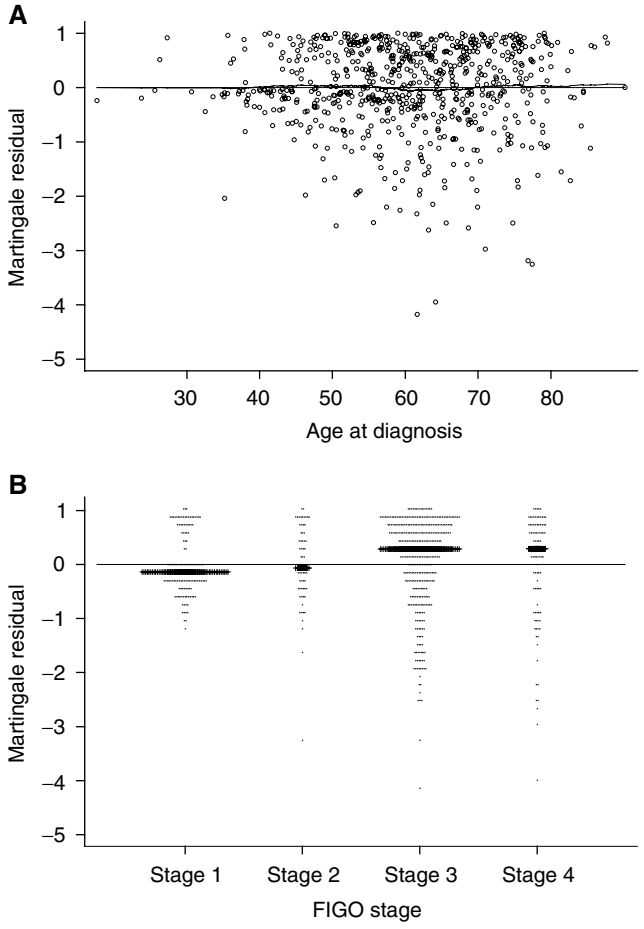


Figure 1 Martingale residuals plotted against (**A**) patient age and (**B**) FIGO stage; median for each stage is denoted by a horizontal line.

improve the fit to a significant degree. Thus, we may be reasonably confident that the model is adequate for both covariates.

Identifying the correct parametric model

When fitting a fully parametric model, the survival times are assumed to follow a statistical distribution. Several different distributions have been proposed, and the identification of a suitable one is a crucial step. The most obvious distinguishing feature between parametric models is in the shape of the hazard they assume the data follow. The Weibull and Gompertz distributions are appropriate when the hazard is always increasing or decreasing; the Log-Logistic may be used where the hazard either rises to a peak and then decreases or always decreases; the Log-Normal and Generalised Gamma models are preferable when the hazard rises to a peak before decreasing. In the Exponential model, the hazard is assumed to be constant over time. The actual shapes of these distributions (e.g. the point in time at which a hazard 'peaks' or the gradient at which it increases/decreases) depend on *ancillary* parameters that are also estimated from the data. For example, when using the Weibull distribution, the hazard function, $h(t)$, is $\lambda s(\lambda t)^{s-1}$. In this case, the shape (s) and scale (λ) parameters are the ancillary parameters to be estimated (see Figure 1 in the previous paper Bradburn *et al*, 2003).

If the shape of the disease hazard is known to be different from that of a particular distribution, then the data should not be analysed with this parametric model. For example, consider the hazard for overall survival after cancer diagnosis. The hazard is rarely constant, thus ruling out an Exponential distribution. In some cases, the hazard rises sharply (due to treatment deaths) before tailing off, which would also rule out the Weibull. An informal assessment of a parametric model's appropriateness may be made via plotting the (smoothed) empirical hazard or cumulative hazard against those estimated by the model, or by log(−log(survival)) plots which are discussed later. Akaike's Information Criterion (AIC) (Akaike, 1974), a statistic that trades off a model's likelihood against its complexity, may also be used when comparing the viability of different parametric models. The AIC of a model may be defined as

$$\mathrm{AIC} = -2\mathrm{LL} + 2(c + a)$$

where LL is the logarithm of the model likelihood (*log-likelihood*), c is the number of *covariates* and s the number of ancillary parameters (e.g. 2 in the case of the Weibull; λ and s). A *lower* value of the AIC suggests a better model. Note, however, that the likelihood computed in a Cox model is a partial likelihood, and so it is not possible to compare Cox PH models to fully parametric ones in this manner.

In the PH framework, it may be clear that none of the parametric models suggested here or elsewhere adequately capture the distributional form of the data. In such cases, the more flexible Cox model is the obvious choice. Commonly used parametric models in the AFT framework are arguably more flexible than those available in the PH framework, and so fitting a parametric AFT model is another option.

Overall goodness-of-fit tests

A simple test for the model adequacy is to compare the overall (Kaplan–Meier) survival curve to the model-based predicted survival and, ideally, for any group of patients the two should be close, if not identical. Hosmer and Lemeshow (1999) suggest using a more formal measure of fit based on comparing observed and expected events in different *risk groups* as defined by the model. Specifically, the predicted risk or prognostic index (PI) from a model consisting of covariates x_1, x_2, …, x_p with estimated coefficients b_1, b_2, …, b_p, respectively, is

$$\mathrm{PI} = b_1x_1 + b_2x_2 + \cdots + b_px_p$$

PI is calculated for each patient. Risk groups are constructed by categorising the (ranked) PIs, for example, three risk groups can be created using the highest, middle and lowest tertiles of PI. A score test is then applied to the differences between the observed and expected events in the risk groups. A simple approximation to this calculation may be obtained by adding the risk groups as a series of covariates to the survival model itself. A significant improvement in the model likelihood suggests that the original covariates form an insufficient model for the data.

Assessing overall goodness of fit on the ovarian cancer data

The predicted survival curves for the ovarian cancer model are potentially misleading. Several factors are associated with length of survival, and some are also related to (or correlated with) each other (e.g. histology and stage). Predicted survival curves for each histological group may be estimated by fixing all other covariates at their mean values. However, this approach will give an estimate of survival that is different to those observed in the data because correlations are ignored. The test of Hosmer and Lemeshow (1999) is more useful here. The patients are split into ten risk groups, with the proportion of deaths in each ranging from 10% in the best prognosis group to 94% in the worst. The approximate score test, derived from adding nine covariates to the model, produced no evidence of a poor fit (likelihood ratio test $\chi^2 = 7.84$ on 9 degrees of freedom, $P = 0.55$).

ASSESSING WHETHER PH IS APPROPRIATE

The PH assumption, that is, the hazards are proportional (and not overlapping) at all points in time, should be verified. An obvious approach is to plot the hazard in each group, but this is of limited use. The empirical hazard function is generally not well estimated, and instead the cumulative hazard is generally preferred to assess the PH assumption. If a PH model is valid, a plot of the logarithm of the cumulative hazard function in each group against the logarithm of time should give rise to lines that are parallel. Continuous variables need to be categorised into groups. The plot described is also known as the log(−log(survival)) plot, as the cumulative hazard is equal to the negative logarithm of the survival proportion. This approach requires a subjective assessment. Unfortunately, convergent or divergent lines may be due to either a lack of proportionality or to the omission of an important covariate. In practice, it is not known which, but this phenomenon suggests an inadequate model. On the other hand, parallel lines suggest that models assuming PHs may be suitable. In the case of fitting a Weibull or an Exponential parametric model, the lines should be parallel and straight.

Several formal statistical tests have been proposed for assessment of proportionality of hazards. A simulation study by Ng'andu (1997) described and compared several tests in the Cox PH framework, and concluded that the (weighted) scaled Schoenfeld residuals test (Grambsch and Therneau, 1994), the linear correlation test (Harrell, 1986) and the time-dependent covariate test (Cox, 1972) were the most powerful diagnostic tools for proportionality. The first two of these test for an association between residuals and time (evidence of which indicates a bad fit), and the third tests whether the effect (coefficient) of a covariate changes with time (i.e. nonconstant hazard ratio). This latter method is appealing as it not only detects nonproportionality, but allows it to be modelled validly. An alternative is to fit a *stratified* model, wherein a covariate that displays nonproportionality is modelled without the constraint of proportionality. Such a covariate must obviously be categorical (or be categorised), but more importantly has no estimated effect size provided when forming the strata of a stratified model, and thus is suitable only for covariates that are not of primary interest. Abandoning the PH approach in favour of some other model is clearly another option.

Assessing the appropriateness of PH for the ovarian cancer data

The Kaplan–Meier survival curves and log(−log(survival)) *vs* log(time) plots are shown for FIGO stage and histology in Figure 2A–D. The log(−log(survival)) plot for FIGO stage gave rise to reasonably parallel lines and therefore suggests proportionality. However, in the case of the histology, this appears to be violated. In particular, the prognosis for the endometroid group sits in the middle of all the groups in the first year but improves thereafter. A similar feature was apparent for the presence of ascites, where the initial detrimental effect becomes less important with time (data not shown). The (weighted) scaled Schoenfeld residuals test suggested significant overall nonproportionality ($P = 0.05$), as did the time-dependent covariate tests for these terms. Therefore, despite other aspects of this model appearing adequate, the assumption of proportionality appears to be violated.

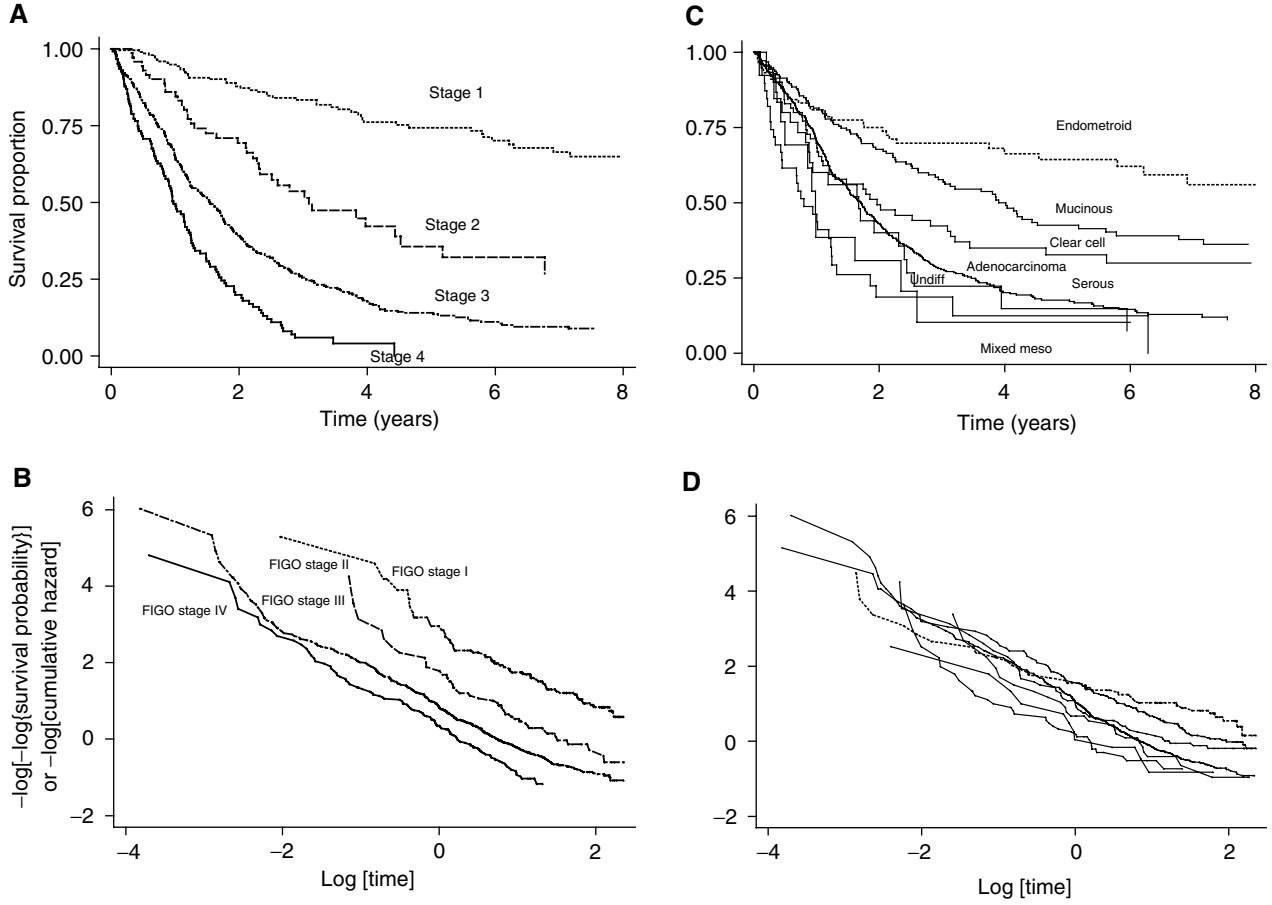


Figure 2 (**A**) Survival according to FIGO stage. (**B**) Log(−log(survival)) for FIGO stage. (**c**) Survival according to histology. (**D**) Log(−log(survival)) for histology. The endometroid group is shown by the dotted line.

Table 4 The Cox model applied to the ovarian data, with a time dependency added to ascites and endometroid terms

Covariate	Coefficient (b_i)	HR [exp(b_i)]	95% CI	*P*-value
FIGO	0.734	2.09	(1.83–2.38)	<0.001
Histology				<0.001
Serous	(0.000)	(1.00)		
Mucinous	−0.432	0.65	(0.50–0.85)	
Clear cell	0.344	1.41	(0.99–2.01)	
Adenocarcinoma	0.494	1.64	(0.91–2.96)	
Undifferentiated	0.769	2.16	(1.06–4.40)	
Mixed mesodermal	0.825	2.28	(1.50–3.47)	
Endometroid	0.312	1.37	(0.90–2.07)	
Endometroid × log(time)	−0.500	0.61	(0.45–0.82)	0.001
Grade				<0.001
1	(0.000)	(1.00)		
2	0.826	2.28	(1.32–3.95)	
3	0.843	2.32	(1.35–4.00)	
Absence of ascites	−0.466	0.63	(0.50–0.80)	<0.001
Ascites × log(time)	0.233	1.26	(1.01–1.58)	0.04
Age (per 5-year increase)	0.134	1.14	(1.09–1.20)	<0.001

HR = hazard ratio; CI = confidence interval.

Table 5 Akaike Information Criterion (AIC) of five different distributions fitted to the full model

Model	Log likelihood (LL)	No. of covariates (*c*)	No. of ancillary parameters (*a*)	AIC
Exponential	−259.28	11	1	542.55
Weibull	−253.21	11	2	532.41
Log-Normal	−238.22	11	2	502.44
Log-Logistic	−236.33	11	2	498.65
Generalised Gamma	−235.79	11	3	499.58

AIC = −2LL+2(*c*+*a*).

Nevertheless, we can still use a Cox PH model with time-dependent covariates implemented, which is a model that includes interaction terms between the covariates and (log) time, and thus allows the effect of the relevant covariates to change with time. Table 4 shows the amended model that now allows the effects to vary with time. The time-dependent terms suggest that the absence of ascites and endometroid histology have effects that diminish (the hazard ratios tend towards 1) with time. For example, the absence of ascites is judged to have a hazard ratio of $\exp(-0.466 + 0.233 \times \log(2)) = 0.74$ at 2 years but $\exp(-0.466 + 0.233 \times \log(5)) = 0.91$ at 5 years.

ASSESSING WHETHER AN AFT MODEL IS ADEQUATE

In the AFT model, the survival proportion in one group at any time t is equal to the survival proportion in the second at time φt, where φ is constant. Therefore, a Quantile–Quantile (Q–Q) plot of the times of survival percentiles should lie on a straight line of slope φ that passes through (0, 0). As with the log(−log(survival)) plot in PH models, this is a useful but limited approach as departures from linearity could be due to the AFT model being inappropriate or that one or more important covariates have been omitted. The methods of stratification or modelling with time-dependent covariates suggested in the PH section may be applied here as well.

The lung cancer trial data

We assessed the adequacy of the Generalised Gamma and four other parametric models (each with all covariates included) and present their AIC values in Table 5. The Generalised Gamma model has a higher log-likelihood than the other models and a lower AIC, indicating that this distribution may be the most accurate. To check for excluded covariates, the Martingale residuals were plotted against potential model terms as before. None of these plots suggested that a covariate was incorrectly omitted. Figure 3

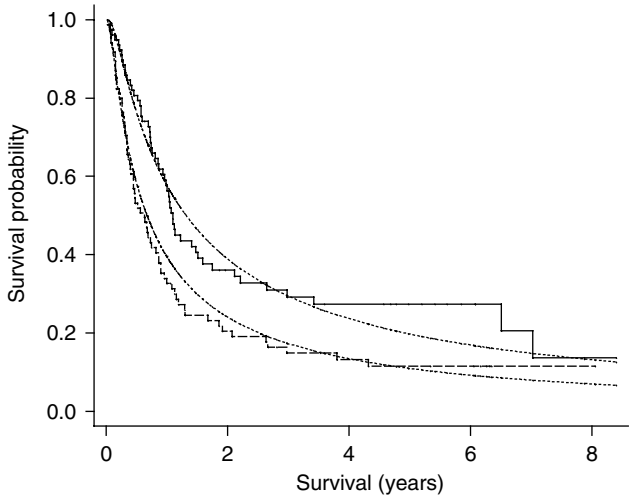


Figure 3 Kaplan–Meier survival probabilities for patients treated by RT + CAP (solid line) and RT alone (dashed line). The respective predicted survival proportions of a generalised gamma multivariate model are given by the faint dotted lines for grouped mean covariates. RT = radiotherapy, CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

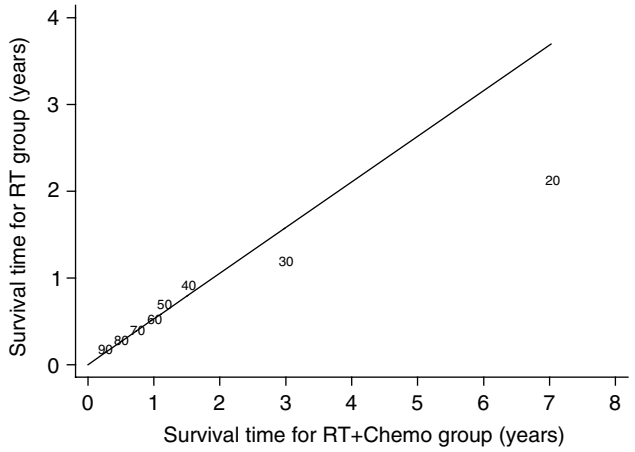


Figure 4 Q–Q plot (percentiles of survival distribution) for patients RT + CAP against those with RT only. The plot symbols are the survival percentiles* and the slope corresponds to the value of the time ratio (= 1/1.90). * The 10th percentile is omitted: 4.9 and 11.4 years for RT and RT + CAP respectively, RT = radiotherapy, CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

gives the predicted observed survival curves together with the predicted survival under a Generalised Gamma model; for each treatment group, the Karnofsky performance status and cell type were fixed at their mean values. The medium-term survival is not as well fitted by the model but is tolerably close. The long-term survival is also less well estimated, but because few patients survive this length of time the estimated survival is imprecise and so this does not cause grounds for concern. The survival times for the 10th, 20th, …, 90th survival percentiles for each treatment group are plotted as a Q–Q plot in Figure 4 and, again, apart from the later times seem to fit adequately.

DISCUSSION

This paper has sought to demonstrate the models introduced in the previous paper in this series (Bradburn *et al*, 2003), to offer practical advice on how to select a method that represents the data fairly, and how to present and interpret it. Good modelling of survival data is not a straightforward exercise, and it is not possible to suggest an 'off the peg' solution. Before starting the process of deciding which (if any) of the models suggested is most suitable for an individual dataset, the important question of why the model should be fitted needs to be considered. The answer should inform the modelling process. Although it is possible to choose a model from those suggested that is optimal from a purely statistical point of view (e.g. goodness-of-fit measures), nonstatistical considerations should to be taken into account. The choice of model and of covariates therein should, in general, be suggested from experience and based on the specific question under investigation. However, good nonstatistical reasons informing model choice should not override good statistical reasons for not choosing that model. The diagnostics (e.g. residuals) for the different models may be difficult to interpret, but they will give an indication of whether modelling assumptions hold and, ultimately, should be considered when model building.

In some cases, all of the models mentioned above may not be wholly appropriate either for modelling the data or answering the relevant question. Consider an example where the time between treatment and possible multiple cancer relapse is to be investigated. The methods introduced assume one survival time (culminating in one type of event), but we may be dealing with patients who have one or more relapses of different type or levels. In the final paper of this series, we introduce models that extend the types of models described here to incorporate recurrent events. We also present approaches to modelling continuous covariates in a nonlinear fashion, validating models and discuss alternatives when fundamental censoring assumptions do not hold.

ACKNOWLEDGEMENTS

We wish to thank John Smyth for providing the ovarian cancer data, and Victoria Cornelius and Peter Sasieni for invaluable comments on an earlier manuscript. Cancer Research UK supported all authors. Taane Clark holds a National Health Service (UK) Research Training Fellowship.

REFERENCES

Akaike H (1974) A new look at the statistical model identification. *IEEE Transaction and Automatic Control* **AC-19:** 716–723

Bradburn MJ, Clark TG, Love S, Altman DG (2003) Survival analysis. Part II: multivariate data analysis – an introduction to concepts and methods. *Br J Cancer* **89**(3)**:** 431–436

Clark TG, Altman DG (2002) Developing a prognostic model in the presence of missing data: an ovarian cancer case-study. *J Clin Epidemiol* **56:** 28–37

Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis. Part I: basic concepts and first analyses. *Br J Cancer* **89**(2)**:** 232–238

Collett D (1994) *Modelling Survival Data in Medical Research*. London: Chapman & Hall

Cox DR (1972) Regression models and life tables (with discussion). *J R Stat Soc B* **34:** 187–220

Grambsch PM, Therneau TM (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81:** 515–526

Harrell FE (1986) The PHGLM procedure. *SAS Supplemental Library Users Guide*, Version 5 edition. Cary, NC: SAS Institute

Henderson HV, Velleman PF (1981) Building multiple regression models interactively. *Biometrics* **37:** 391–411

Hosmer DW, Lemeshow S (1999) *Applied Survival Analysis: Regression Modelling of Time to Event Data*. New York: Wiley

Machin D, Campbell MJ, Fayers PM, Pinol APY (1998) *Sample Size Tables for Clinical Studies*. Oxford: Blackwell Science

Ng'andu NH (1997) An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Stat Med* **16:** 611–626

Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis II: accuracy and precision of regression estimates. *J Clin Epidemiol* **48:** 1503–1510

British Journal of Cancer (2003) 89, 781 – 786

www.bjcancer.com

Tutorial Paper

Survival Analysis Part IV: Further concepts and methods in survival analysis

TG Clark*,[1], MJ Bradburn[1], SB Love[1] and DG Altman[1]

[1]*Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Oxford OX3 7LF, UK*

British Journal of Cancer (2003) **89,** 781 – 786. doi:10.1038/sj.bjc.6601117 www.bjcancer.com

Keywords: survival analysis; missing data; validation; repeated events

INTRODUCTION

In the previous papers in this series (Bradburn *et al*, 2003a, b; Clark *et al*, 2003), we discussed methods for analysing survival time data, both univariate and multivariate. We have dealt with only a portion of the methods available for analysing survival time data, and in many cases, useful alternatives to (or extensions of) these methods exist. We have also left unanswered other questions regarding the design and analysis of studies that measure survival time and, in particular, dealing with situations where some standard modelling assumptions do not hold. We conclude this series by tackling these issues. These ideas are described in a question and answer format, and introductory references are provided for the reader to investigate further.

IN A SURVIVAL ANALYSIS, CONTINUOUS VARIABLES ARE SOMETIMES CATEGORISED. SHOULD WE DO THIS (AND IF SO, HOW)?

In medical research, it is common to see continuous measures grouped into categories to simplify a covariate's relationship with survival and its interpretation. There is no statistical reason for grouping and it can lead to as many problems as it seeks to avoid. The categorisation of a continuous covariate by definition discards data and can be seen as introducing measurement error. It also leads to biased estimates and a reduced ability to detect real relationships (Schmoor and Schumacher, 1997; Altman, 1998). Nevertheless, there are sometimes good reasons to categorise a continuous covariate in the analysis of survival (and indeed any) data. When doing so, it is wise to note the following points:

1. Use cut-points that have been predetermined rather than testing multiple values. A common choice of boundaries is fixed centiles such as quartiles. It is preferable though to use established cut-points that have clinical meaning, and therefore provide consistent groupings between studies. Examples include dividing oestrogen receptor level at 10 fmol, and age into 5- or 10-year intervals.
2. Do not choose cut-points based on minimising *P*-values, as this method gives biased results (Altman *et al*, 1994; Altman, 1998).
3. If possible, use more than two categories to reduce the loss of information and allow some assessment of the linearity of any trend.
4. Ensure that each group contains an adequate number of individuals (and events).

Grouping is sometimes used because there are concerns with mismodelling the relationship when there is a nonlinear relationship between the variable and log hazard. The simplest approach is to evaluate the effect of adding a quadratic term to the model, but better approaches to use are smoothing splines (Therneau and Grambsch, 2000) or fractional polynomials (Royston *et al*, 1999). Figure 1 shows the result of modelling a new covariate, (log) CA125, in the previously used ovarian cancer data, by the method of smoothing splines (with 11 degrees of freedom). There is evidence of nonlinearity ($P = 0.002$) and the plot suggests that CA125 might be modelled as a cubic effect. It is clear that modelling the data using a binary or linear variable would be inappropriate here (see Figure 1). Knorr *et al* (1992) discussed these issues in the context of prognostic studies in cancer.

IN OUR CLINICAL TRIAL, WE COLLECTED MEASUREMENTS AT PREARRANGED VISITS. CAN WE INCLUDE MULTIPLE MEASUREMENTS FOR THE SAME COVARIATE IN OUR SURVIVAL ANALYSIS?

If variables measured after entry into the study are to be included in the survival model, special methods are required. Such methods are called *time-dependent* (or *updated*) *covariate methods*, as the variables they incorporate may change value over time. For example, if a longitudinal study seeks to assess the effects of smoking on cancer, a variable for each patient may be defined, being equal to 0 (nonsmoker) or 1 (smoker) at any time. If a nonsmoker begins smoking after entering the study, then this covariate is updated (from '0' to '1') at the time that smoking begins. This covariate contributes more information than using smoking status at time of entry alone. It is important to note that post-entry measurements cannot be validly incorporated into a survival model without using these methods.

Recall that for the proportional hazards model, the formula relating a covariate x_1 to the hazard $h(t)$ at time t is

$$h(t) = h_0(t) \exp[b_1 x_1]$$

where $h_0(t)$ is the baseline hazard. If repeated measurements of a covariate x_1 are available, the formula changes to

$$h(t) = h_0(t) \exp[b_1 x_1(t)]$$

*Correspondence: Mr TG Clark; E-mail: taane.clark@cancer.org.uk
Received 6 December 2002; accepted 30 April 2003

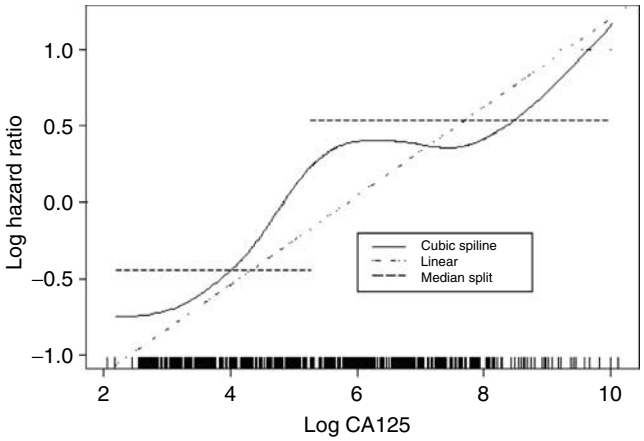


Figure 1 Modelling log CA125 using spline functions: | corresponds to measurements.

where $x_1(t)$ is the value of x_1 at time t. (It is also possible to use, but harder to interpret, an accelerated failure time model here.) The covariate x_1 may be continuous or categorical, and may change freely or at fixed time intervals. The coefficient b_1 represents the additional relative hazard for each unit increase in x_1 at any given time. This model is different from models with *time-dependent coefficients* (Bradburn *et al*, 2003b), in which the *effect* of a covariate changes rather than the value of the covariate itself, that is, $h(t) = h_0(t) \exp[b_1(t)\ x_1]$.

The time dependent method can be applied in many standard statistical software packages. However, the approach described requires a large amount of data and is therefore rarely seen. One also has to be confident that the collection process is not *itself* dependent on clinical progress, perhaps by using scheduled assessments. Further details of the method, and some precautions, are noted in Altman and De Stavola (1994).

MOST SURVIVAL ANALYSIS METHODS ASSUME THE CENSORING IS NONINFORMATIVE. WHAT IF THE CENSORING IS INFORMATIVE?

Informative censoring occurs when individuals are lost to follow-up for reasons that may relate to their (unknown) outcome. For example, in a randomised trial in which the main outcome is time to cancer recurrence, a patient who is lost to follow-up may be more likely to have experienced drug toxicity or ill health and thus may also be more susceptible to (earlier) relapse. Informative censoring introduces bias into the standard methods discussed previously. Unfortunately, it is difficult both to identify informative censoring and to assess its impact. It is helpful though to know what proportion of censored individuals were lost to follow-up before the end of the study (Clark *et al*, 2002).

A simple, *ad hoc* approach to the problem is to perform sensitivity analyses, to assess the impact of assigning different survival times to those patients whose observed (censored) survival times may have been affected in this manner. For example, if a patient suspected to be in ill health exits the study at 4 weeks, a first analysis may be performed with this patient censored at 4 weeks and a second where the patient is assumed to have relapsed at 4 weeks (i.e. a 'best case – worst case' scenario). This approach works best when there are few such patients, but in that situation, the possible bias will be very small. Another possibility is to decide *a priori* that all such patients will be treated in a particular way. The issue has been of particular concern in randomised trials of nicotine replacement therapy, in which losses to follow-up are considerable. In a systematic review of randomised trials, patients who were lost to follow-up were regarded as being continuing smokers (Silagy *et al*, 2002).

More formal approaches have been proposed (e.g. Robins, 1995a, b; Scharfstein *et al*, 2001). In general, they assume that a relationship exists (and can be modelled) between censoring times and baseline covariates and perhaps also post-treatment patient data. It is difficult to evaluate the assumptions of these complex methods, and implementation in statistical software is limited.

If follow-up stops because the patient has experienced a different defined event, the problem may be viewed as a competing risk scenario (see below), or handled via a mixture model (or 'cure' model), where the differing event types are explicitly modelled. The latter method makes particular sense if the two events are quite dissimilar, such as patient recovery and patient death.

In practice, if there is little informative censoring, the bias introduced to standard methods is minimal, and in general using these along with simply reporting loss to follow-up (perhaps with a basic sensitivity analysis) will suffice. Good patient follow-up and avoidance of unnecessary drop-out is by far the best solution, and when and why drop-out occurs should always be reported (Moher *et al*, 2001).

SOME COVARIATE DATA ARE MISSING IN OUR ANALYSIS. WHAT SHOULD WE DO?

Missing data are a common problem when developing survival models in cancer. Individuals without complete covariate data are usually omitted, but the resulting analysis has reduced power and may be an unrepresentative subset of patients. Often many covariates have missing data, and the absence of a small percentage of data points for each variable can lead to a greatly depleted sample. Unless only a few values are missing, some investigation of the missing data and methods that accommodate it should be considered. In the ovarian cancer data set presented previously, a small number of important factors containing little or no missing data were used. The database contains several other factors in which missing data were frequently encountered, and a more definitive analysis (Clark *et al*, 2001) was able to incorporate these factors, while retaining all patients by applying multiple imputation methods (Van Buuren *et al*, 1999). Multiple imputation is a framework in which missing data are imputed or replaced with a set of plausible values. Several data sets are then constructed, each being analysed separately, and their results are combined while allowing for the uncertainty introduced in the imputation. Other approaches exist (e.g. Lipsitz and Ibrahim, 1998), but imputation approaches have more software available (Horton and Lipsitz, 2001). Further details, discussion and references are given in another analysis of the ovarian data found in Clark and Altman (2002).

We recommend that authors of research papers are explicit about the amount of missing data for each variable and indicate how many patients did not have complete data. Imputation techniques are powerful tools and are increasingly available in software, but are not a panacea. Inherent in the method is the assumption that a model relating data absence to other measured covariates (and possibly survival too) exists and can be specified. This has much in common with the situation where informative censoring is suspected, and similarly, their practical experience is limited at the present time. Researchers should be aware of the assumptions, most of which are untestable, and use sensitivity analysis to assess the robustness of results. Ultimately, these problems are best avoided by minimising missing data.

HOW SHOULD WE CHOOSE WHICH VARIABLES TO INCLUDE IN OUR SURVIVAL MODEL?

In some cases, the factors to be included in the model will be predetermined. In many others, there will be several possible covariates from which only a handful are to be chosen. This is

often because there are a large number of covariates of which some are unimportant, but the identification and elimination of these is not always easy. As a starting point, it is good practice to include known prognostic factors and any that are specifically required by the study aims (e.g. the treatment identifier in the analysis of a clinical trial). It is then the burden of new factors to add significant additional predictive ability (Simon and Altman, 1994).

If there are a large number of factors of interest and there is relatively little information about their prognostic influence, automated selection techniques such as stepwise methods can be used. There are variations on these that start either with all covariates (backward elimination) or none (forward selection), adding or removing covariates according to statistical significance at some predecided level. A disadvantage of both is that they only evaluate a small number of the set of possible models. Instead, each possible model could be fitted, with the best being picked on the basis of a goodness-of-fit measure such as Mallow's C (Hosmer and Lemeshow, 1999). However, this may be time-consuming with many covariates, multiple testing is problematic, and is seldom used due to its noninclusion in many software packages.

Unfortunately, all these methods are problematic. The 'best' model is derived solely on statistical grounds (and indeed may lack any clinical meaning), the regression coefficients produced are biased (too large) and standard errors and *P*-values are too small, especially for smaller sample sizes and when few events occur. Backward elimination is possibly the best of the above methods for identifying the important variables, and it allows one to examine the full model, which is the only fit providing accurate standard errors and *P*-values (Harrell, 2001). An alternative, the lasso method (Harrell, 2001) attempts to force some regression coefficient estimates to be exactly zero, thus achieving variable selection while shrinking the remaining coefficients toward zero to reflect the overfitting and overestimation caused by data-based model selection.

If one cannot completely prespecify a model, it may be best to apply backward elimination or lasso to a full model of prespecified covariates of interest, and use bootstrap methods to compare the stability and predictive accuracy of the full model with that of a reduced one (see next question for further details).

WE HAVE DEVELOPED A PROGNOSTIC MODEL FOR OVERALL SURVIVAL. HOW CAN WE MEASURE ITS PREDICTIVE ABILITY? HOW CAN THE MODEL BE VALIDATED?

In survival analysis, statistical models are employed to identify or propose combinations of risk factors that might predict patient survival. It follows that to be of use, the model must be able to: (1) make unbiased predictions, that is, give predicted probabilities that match closely those observed, and (2) distinguish higher and lower risk patients. These are the two components of predictive ability, and are called *calibration* and *discrimination,* respectively. Importantly, models rarely perform as well on either basis when used to predict survival in patients other than those used to derive the model. A model that closely mirrors the survival patterns of the present data is said to have *internal validity*, but to be of wider use should do so for other groups of patients as well (be *externally valid*). Before a model is applied routinely in clinical practice, it should have been shown to meet both criteria.

Measures of discrimination include the *c*-index and Nagelkerke's $R^2(R^2_N)$ (Harrell, 2001). The *c*-index, a generalisation of the area under the receiver operating characteristic (ROC) curve, is the probability of concordance between observed and predicted survival based on pairs of individuals, with $c=0.5$ for random predictions and $c=1$ for a perfectly discriminating model. Similarly, $R^2_N=0$ indicates no predictive ability and $R^2_N=1$ indicates perfect predictions. Calibration may be quantified using an estimate of slope shrinkage (Harrell, 2001). Each quantity may be evaluated for the data used in the modelling by randomly splitting the patients into two samples, one to derive the model and the other to validate it. The proportion of data to include in each sample is, however, arbitrary and although estimates of predictive accuracy from this approach are unbiased, they also tend to be imprecise. Bootstrapping, a method that involves analysing subsamples from a data set, or 'leave-one out' cross-validation may be more beneficial. For these analyses, an alternative is to estimate shrinkage factors and apply these to regression coefficients to counter overoptimism. These techniques allow evaluation on multiple data sets. Once the internal validity of a model has been established, it can be tested for its generalisability by applying the model to other patients, and using the above methods to assess the adequacy of the predictions.

A good summary of important issues can be found in Justice *et al* (1999) and Wyatt and Altman (1995), and more details on the statistical methods are given in Altman and Royston (2000). In summary, internal validation is necessary before a model is proposed, and external validation is highly recommended before it is to used in clinical practice.

CAN WE PERFORM AN ANALYSIS WHERE THERE ARE UNMEASURED FACTORS THAT MAY AFFECT SURVIVAL TIME?

In practice, one cannot be sure that all important prognostic variables have been measured. In general, omitting variables will simply reduce the predictive ability of a model, so that patients with similar measured covariates will exhibit large variability in their survival. When a strongly prognostic variable is omitted, however, the model may be biased. In particular, the estimated treatment effect in a randomised trial may be biased if an important prognostic variable is not adjusted for, even when that variable is balanced between the treatment groups (Schmoor and Schumacher, 1997; Chastang *et al*, 1988). It is inappropriate to proceed at all if vital information such as clinical stage in breast cancer patients is unavailable.

Another form of missing covariate is when some individuals have a shared exposure that is unmeasured. For example, members of the same family will have shared dietary and other environmental exposures, so that their outcomes cannot be considered to be independent. A similar situation arises in cluster randomised trials and multicentre trials in general (Yamaguchi *et al*, 2002). Such data can also be considered as being 'multilevel', with variation both between and within groups. Random effects (or 'frailty') models can be used to allow covariate effects to vary across groups (O'Quigley and Stare, 2002). Such models are widely used in other contexts, in particular, in meta-analysis. Frailty can also be considered to apply to individuals, relating to the idea of unmeasured variables as a possible explanation for observed heterogeneity of outcome. Use of such models depends on precise knowledge of the frailty distribution, which is generally not available (Keiding *et al*, 1997).

Lack of fit of a Cox model may be better explained by other modelling approaches (O'Quigley and Stare, 2002), such as the accelerated failure time model (Keiding *et al*, 1997).

SEVERAL PAPERS IN OUR RESEARCH AREA HAVE APPLIED (ARTIFICIAL) NEURAL NETWORKS AND REGRESSION TREES AS AN ALTERNATIVE TO THE COX MODEL. WHAT ARE THESE METHODS?

Artificial neural networks

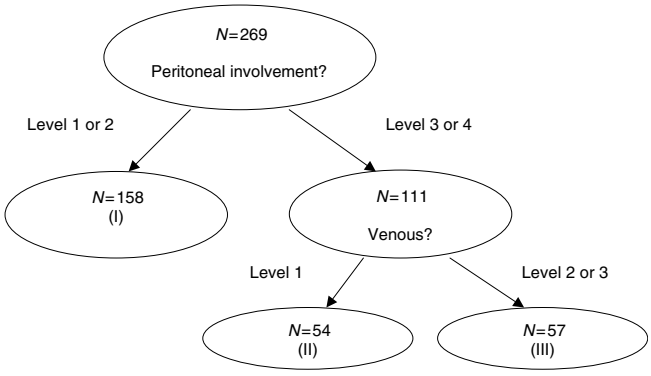
Artificial neural networks (ANNs) are a relatively new method for assessing the extent to which a series of covariates explain patient

outcomes. The key feature of the ANN methodology is to assume that there are some latent, or 'hidden', intermediary variables in the input (covariate) and output (survival probability) processes. The most common model is the three-layer model shown in Figure 2. Under this model, the covariates (input) do not act directly on the response variable (output), but channel their influence into a series of latent (hidden) variables. It is the relative importance of these unobservable variables which determines the survival. For a more detailed introduction to these methods, see Cross *et al* (1995).

This methodology is appealing in that it can incorporate complex relationships between covariates and survival more easily than standard approaches such as Cox regression, which may be too simplistic. However, there have been several major criticisms of the method: (a) the high chance of overfitting the data, (b) the lack of easy interpretation of the model and of the impact of individual covariates, (c) the perceived 'black box' methodology involved, and (d) the difficulty in handling censored survival times. The last issue arises because it is usually the status of the individuals (i.e. alive or dead) at a given point (or points) in time that is taken to be the response. Biganzoli *et al* (1998) and others have modelled the hazard functions directly, in a promising attempt to extend this method. Reviews comparing the examples where both ANN and regression methods had been used to derive prognostic models have found that overall ANNs are little better than classical statistical modelling approaches (Sargent, 2001), and misuses of ANNs in oncology are common (Schwarzer *et al*, 2000). We therefore advise caution in their use, and the involvement of an experienced statistician.

Classification and regression trees

The classification and regression tree (C&RT) approach is based on dividing the cohort into groups of similar response patterns, using covariates (Lausen *et al*, 1994). The partitioning algorithm starts with the covariate that best discriminates the survival outcome between two subgroups. For continuous or multicategory variables, the method thus needs to determine the threshold that best dichotomises the variable. This process is repeated for each subgroup in turn using all the available covariates. The same covariate may be used more than once, and the process stops eventually with either no covariate adequately dividing the subgroups further or when the subgroups have reached a specified minimum size. Figure 3 shows an unpublished C&RT analysis in a Dukes' B colonic cancer study, in which four categorical variables (perforation, peritoneal involvement, venous and margin) were assessed for their prognostic value in overall survival. Using a logrank test at each step, it was found that peritoneal involvement (levels 1, 2 *vs* 3, 4) discriminated best between good and bad survival, and level 1 venous subdivided patients with high levels of peritoneal involvement. The stopping rule employed was the first occurrence of either (a) the maximum logrank statistic is not statistically significant at the 1% level or (b) when any subgroup contains less than 25 patients. The latter condition ceased the partitioning algorithm in the example, yielding the three groups of patients described in Figure 3.



	5-year survival (95% CI)	*N*	Deaths
(I) Peritoneal 1 or 2	87.1 (80.2, 91.7)	158	21
(II) Peritoneal 3 or 4 and venous 1	73.7 (58.8, 83.9)	54	13
(III) Peritoneal 3 or 4 and venous 2 or 3	45.7 (31.0, 59.3)	57	29

Figure 3 A CART for Dukes' B colonic cancer study.

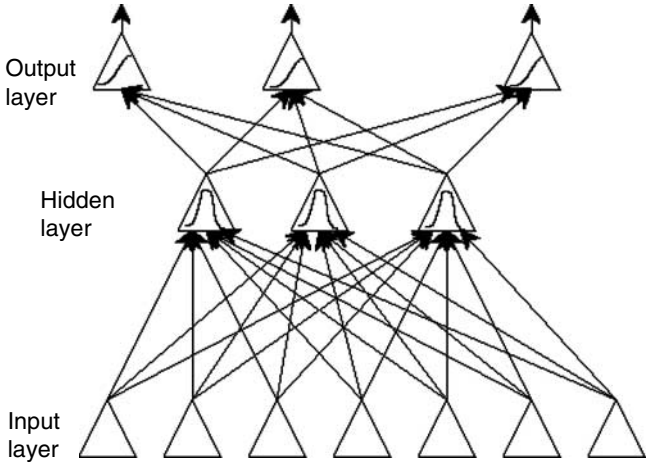


Figure 2 An example of an ANN.

The major advantage of C&RT is its ease of interpretability – it reflects how many decisions are made. It also relies on fewer distributional assumptions (Schmoor *et al*, 1993) and is particularly useful in situations where there are interactions. The disadvantages of C&RT lie in having to decide what threshold to use for continuous covariates, and to correct for multiple testing and overfitting. The automated covariate selection is similar to forward stepwise methods in regression, and hence shares their problems (see the choice of covariate section). Finally, as C&RT seeks to classify patients into groups, it offers little in the way of estimated effect of risk factors. Nevertheless, C&RT is a useful complement to other methods, in particular as an exploratory tool that can inform future research.

CAN WE ANALYSE DIFFERENT TYPES OF EVENTS OR REPEATED EVENTS?

Traditional survival analysis methods (including all those discussed so far) assume that only one type of event of interest occurs, and at most once. More advanced methods exist to allow the investigation of several types of events (e.g. cancer death, vascular death, other), or an event that may occur repeatedly (e.g. cancer recurrence). We will describe methods for each in turn.

Where the survival duration is ended by the first of several events, it is referred to as *competing risks analyses*. Analysing the time to each event separately can be misleading, and in this context the Kaplan–Meier method, in particular, tends to overestimate the proportion of subjects experiencing each event. The cumulative

incidence method, in which the overall event probability at any time is the sum of the event-specific probabilities, may be used to address this. Univariate tests and statistical models also exist, and an overview of several of the methods proposed can be found in Tai *et al* (2001). Models are generally implemented by entering each patient several times – one per event type – and for each patient, the time to any event is censored on the time at which the patient experienced another event.

Where multiple events of the same type occur, it is common practice to use the first event only, but this ignores information. Three approaches to use this extra information are demonstrated using artificial patient data in Table 1. In a *conditional* model, follow-up time is broken up into segments defined by events, with each patient being at risk for an ith event once the $(i-1)$th has occurred. Patient 1 in Table 1 is therefore assumed not to be at risk of a second event until the first has occurred, and so is at risk of experiencing this from time 8 until time 12. This model comes in two types: using either the time since the beginning of the study (type A) or the time since the previous event (type B). The *marginal* model, on the other hand, considers each event to be a separate process and, by definition, the time for each event starts at the beginning of follow-up for each patient. Here, all patients are considered to be at risk for all events, regardless of how many events they have previously had, and so patient 2, for example, was considered at risk of events 3 and 4 despite being lost to follow-up at the second. A third approach, called the *independent increment* model, is closest in spirit to a conditional model but takes no account of the number of previous events experienced by a patient, and for this reason the conditional and marginal models are often preferable. For each model, the data should be entered in the form of one patient record per event number as illustrated in Table 1.

Table 1 Data layout under four recurrent event models with patient 1 having three events (at times 8, 12 and 26) and patient 2 having two events (at times 10, 18)

Model	Patient i.d.	Time interval	Event[a]	Stratum[b]
Conditional A	1	(0,8]	1	1
	1	(8,12]	1	2
	1	(12,26]	1	3
	1	(26,31]	0	4
	2	(0,10]	1	1
	2	(10,18]	1	2
Conditional B[c]	1	(0,8]	1	1
	1	(0,4]	1	2
	1	(0,14]	1	3
	1	(0,5]	0	4
	2	(0,10]	1	1
	2	(0,8]	1	2
Marginal model	1	(0,8]	1	1
	1	(0,12]	1	2
	1	(0,26]	1	3
	1	(0,31]	0	4
	2	(0,10]	1	1
	2	(0,18]	1	2
	2	(0,18]	0	3
	2	(0,18]	0	4
Independent increment	1	(0,8]	1	1
	1	(8,12]	1	1
	1	(12,26]	1	1
	1	(26,31]	0	1
	2	(0,10]	1	1
	2	(10,18]	1	1

[a]1 = had event of interest, 0 = censored. [b]Relates to the number of events and is used in the fitting of the model as the strata variable. [c]Gap time model.

All of the above models are usually applied within a Cox model framework, although accelerated failure time methods may equally be used. These models are fitted using the same basis as standard approaches, with two exceptions: (1) a cluster effect is used to adjust the standard errors because patients are repeated in the study, and (2) the analysis is stratified – with the exception of the independent increment method – with the event type (for competing risks) or number (for recurrent events) defining the strata. Interaction effects between covariates and strata may be used to assess whether covariate effects vary across competing outcomes or event number. For example, Kay (1986) presents an example of a treatment that reduces the risk of death from one cause, but increases the risk of death from another.

More thorough reviews of the above (and other related) methods can be found in Hosmer and Lemeshow (1999), and Therneau and Grambsch (2000). These modelling procedures are generally only a little more difficult than for single-event data, and software is widely available. As with any statistical model though, it is still important to assess its adequacy and fit. In each case, the choice of the best method of analysis will depend on the disease in question and the goals of the analysis. However, the aims such as those described here can often be highly relevant, and where this is the case these methods should be strongly considered.

SUMMARY

Most analyses of survival data use primarily Kaplan–Meier plots, logrank tests and Cox models. We have described the rationale and interpretation of each method in previous papers of this series, but here we have sought to highlight some of their limitations. We have also suggested alternative methods that can be applied when either the data or a given model is deficient, or when more difficult or specific problems are to be addressed. For example, analysis of recurrent events can make an important contribution to the understanding of the survival process, and so investigating repeat cancer relapses may be more informative than concentrating only on the time until the first. More fundamentally, missing data are a common issue in data collection that in some cases can seriously flaw a proposed analysis. Such considerations may be highly relevant to the analysis of a data set, but are rarely mentioned in the analysis of survival data. One possible reason for this is a perceived lack of computer software, but many of the approaches discussed here are currently incorporated into existing commercial statistical packages (e.g. SAS, S-Plus, Stata) and freeware (e.g. R). On the other hand, the desire may be to 'keep things simple for the readership'. This view is reasonable, but is valid only where a simple analysis adequately represents the survival experience of patients in the study. Ensuring the analyses are appropriate is therefore crucial. More advanced survival methods can derive more information from the collected data; their use may admittedly convey a less straightforward message, but at the same time could allow a better understanding of the survival process.

The aim of this series has been to aid awareness, understanding and interpretation of the many and varied methods that constitute the analysis of survival data. It is paramount that analyses are performed in the knowledge of the assumptions that are made therein, and the more complex methods, in particular, are best applied by a statistician.

ACKNOWLEDGEMENTS

We thank Victoria Cornelius and Peter Sasieni for providing comments on an earlier draft. Taane Clark holds a National Health Service (UK) Research Training Fellowship. Cancer Research UK supported Douglas Altman, Mike Bradburn and Sharon Love.

REFERENCES

Altman DG (1998) Suboptimal analysis using 'optimal' cutpoints. *Br J Cancer* **78:** 556 – 557

Altman DG, De Stavola BL (1994) Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Stat Med* **13:** 301 – 341

Altman DG, Lausen B, Sauerbrei W, Schumacher M (1994) Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* **86:** 829 – 835

Altman DG, Royston P (2000) What do we mean by validating a prognostic model? *Stat Med* **19:** 453 – 473

Biganzoli E, Boracchi P, Mariani L, Marubini E (1998) Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat Med* **17:** 1169 – 1186

Bradburn MJ, Clark TG, Love S, Altman DG (2003a) Survival analysis. Part II: multivariate data analysis – an introduction to concepts and methods. *Br J Cancer*

Bradburn MJ, Clark TG, Love S, Altman DG (2003b) Survival analysis. Part III: multivariate data analysis – choosing a model and assessing its adequacy and fit. *Br J Cancer* (submitted)

Chastang C, Byar D, Piantadosi S (1988) A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. *Stat Med* **7:** 1243 – 1255

Clark TG, Altman DG (2002) Developing a prognostic model in the presence of missing data: an ovarian cancer case-study. *J Clin Epidemiol* **56:** 28 – 37

Clark TG, Altman DG, De Stavola BL (2002) Quantifying the completeness of follow-up. *Lancet* **359:** 1309 – 1310

Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis. Part I: basic concepts and first analyses. *Br J Cancer* (submitted)

Clark TG, Stewart ME, Altman DG, Gabra H, Smyth J (2001) A prognostic model for ovarian cancer. *Br J Cancer* **85:** 944 – 952

Cross SS, Harrison RF, Kennedy RL (1995) Introduction to neural networks. *Lancet* **346:** 1075 – 1079

Harrell FE (2001) *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer-Verlag

Horton NJ, Lipsitz SR (2001) Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am Stat* **55:** 244 – 254

Hosmer DW, Lemeshow S (1999) *Applied Survival Analysis: Regression Modelling of Time to Event Data*. New York: Wiley

Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalisability of prognostic information. *Ann Int Med* **130:** 515 – 524

Kay R (1986) Treatment effects in competing-risks analysis of prostate cancer data. *Biometrics* **42:** 203 – 211

Keiding N, Andersen PK, Klein JP (1997) The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Stat Med* **16:** 215 – 224

Knorr KL, Hilsenbeck SG, Wenger CR, Pounds G, Oldaker T, Vendely P, Pandian MR, Harrington D, Clark GM (1992) Making the most of your prognostic factors: presenting a more accurate survival model for breast cancer patients. *Breast Cancer Res Treat* **22:** 251 – 262

Lausen B, Sauerbrei W, Schumacher M (1994) Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In *Computational Statistics*, Dirschedl P, Osermann R (eds) Heidelberg/New York: Physica-Verlag

Lipsitz SR, Ibrahim JG (1998) Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics* **54:** 1002 – 1013

Moher D, Schulz KF, Altman DG for the CONSORT Group (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomised trials. *Lancet* **357:** 1191 – 1194

O'Quigley J, Stare J (2002) Proportional hazards models with frailties and random effects. *Stat Med* **21:** 3219 – 3233

Robins JM (1995a) An analytic method for randomized trials with informative censoring: Part I. *Lifetime Data Anal* **1:** 241 – 254

Robins JM (1995b) An analytic method for randomized trials with informative censoring: Part II. *Lifetime Data Anal* **1:** 417 – 434

Royston P, Ambler G, Sauerbrei W (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* **28:** 964 – 974

Sargent DJ (2001) Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer* **91:** 1636 – 1642

Scharfstein D, Robins JM, Eddings W, Rotnitzky A (2001) Inference in randomised studies with informative censoring and discrete time-to-event endpoints. *Biometrics* **57:** 404 – 413

Schmoor C, Schumacher M (1997) Effects of covariate omission and categorization when analysing randomized trials with the Cox model. *Stat Med* **16:** 225 – 237

Schmoor C, Ulm K, Schumacher M (1993) Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial. *Stat Med* **12:** 2351 – 2366

Schwarzer G, Vach W, Schumacher M (2000) On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* **19:** 541 – 561

Silagy C, Lancaster T, Stead L, Mant D, Fowler G (2002) Nicotine replacement therapy for smoking cessation (Cochrane Review). In: *The Cochrane Library*, Issue 4. Oxford: Update Software

Simon R, Altman DG (1994) Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* **69:** 979 – 985

Tai B, Machin D, White I, Gebski V (2001) Competing risks analysis of patients with osteosarcoma: a comparison of four different approaches. *Stat Med* **20:** 661 – 684

Therneau TM, Grambsch PM (2000) *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag

Van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* **18:** 681 – 694

Wyatt JC, Altman DG (1995) Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ* **311:** 1539 – 1541

Yamaguchi T, Ohashi Y, Matsuyama Y (2002) Proportional hazards models with random effects to examine centre effects in multicentre cancer clinical trials. *Stat Methods Med Res* **11:** 221 – 236

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 10 November 2017
doi: 10.3389/fvets.2017.00191

Risk Factors for Culling, Sales and Deaths in New Zealand Dairy Goat Herds, 2000–2009

Milan Gautam[1], Mark A. Stevenson[2], Nicolas Lopez-Villalobos[1] and Victoria McLean[3]*

[1] Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand, [2] Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC, Australia, [3] Dairy Goat Co-Operative, Hamilton, New Zealand

The aim of this study was to identify risk factors for culling, sales and deaths in intensively managed dairy goat herds in New Zealand. A data set provided by the New Zealand Dairy Goat Cooperative (n = 13,197 does) was analyzed using a Cox proportional hazard model. The outcome of interest was length of productive life (LPL), defined as the number of days from the date of second kidding to the date of removal from the herd or the date on which follow-up was terminated, whichever occurred first. Milk solids yield in the first lactation (MSL1) as a predictor of LPL was parameterized in the model as a penalized spline term. To account for MSL1 violating the proportional hazards assumption of the Cox model, LPL was divided into two intervals: T1 (less than or equal to 730 days from the date of second kidding) and T2 (greater than 730 days from the date of second kidding). MSL1 was then included in the model as a time-dependent covariate. A frailty term was included in the model to account for unmeasured, herd-level effects on LPL. During T1, the daily hazard of removal for does that produced 80 kg milk solids in the first lactation was 0.84 (95% CI 0.58–1.23) times the daily hazard of removal for does that produced 30 kg milk solids in the first lactation. During T2, the daily hazard of removal for does that produced 80 kg milk solids in the first lactation was 1.44 (95% CI 0.79–2.65) times the daily hazard of removal for does that produced 30 kg milk solids in the first lactation. We conclude that involuntary losses may be avoided if high MSL1 yielding does are preferentially managed from 2 years beyond the date of second kidding.

Keywords: epidemiology, dairy goats, length of productive life, survival analysis, Cox proportional hazards regression

OPEN ACCESS

Edited by:
Moh A. Alkhamis, Kuwait Institute for Scientific Research, Kuwait

Reviewed by:
Ane Nødtvedt, Norwegian University of Life Sciences, Norway
Catalina Picasso, University of Minnesota, United States

****Correspondence:***
Milan Gautam
m.gautam@massey.ac.nz

Specialty section:
This article was submitted to Veterinary Epidemiology and Economics, a section of the journal Frontiers in Veterinary Science

Received: *15 May 2017*
Accepted: *23 October 2017*
Published: *10 November 2017*

Citation:
Gautam M, Stevenson MA, Lopez-Villalobos N and McLean V (2017) Risk Factors for Culling, Sales and Deaths in New Zealand Dairy Goat Herds, 2000–2009. Front. Vet. Sci. 4:191. doi: 10.3389/fvets.2017.00191

INTRODUCTION

In farmed animal production systems (e.g., dairy, beef cattle, pig, and dairy goat farms) a long, productive life of individual production units is an essential prerequisite for economic efficiency (1). In dairy systems, longevity is defined as the interval between delivery of the first offspring and the date of removal from the herd (2). Increasing the longevity of dairy animals is desirable because it means that the cost of rearing replacements is amortized over a longer period of income production. Since longevity is a desirable quality in production animals (3), it is important to have an understanding of factors influencing the same. Very little work has been done in this area of the dairy goat industry, and an understanding of risk factors for culling, sales and deaths in dairy goats is limited.

In New Zealand, the number of dairy goat herds is small relative to the number of dairy cow herds, and a key industry focus is on the production of infant formula (4). Typically, does are housed indoors in open sided free stall barns and are fed fresh-cut pasture. Approximately two-thirds of the commercial dairy goat farms are concentrated in the Waikato region, in the upper North Island. Purebred and crossbred Saanens are the predominant breeds, but other breeds such as Toggenburgs and Alpines are common (4). At the time of writing, there were 69 herds registered with the New Zealand Dairy Goat Cooperative (NZDGC), a farmer-owned cooperative, each with around 700 milking does per herd, on average.

A better understanding of the various risk factors for removal can be used to enhance longevity in dairy animals. With this knowledge, it is possible to identify characteristics that can serve as early indicators of culling and, depending upon how strong the effect of a particular risk factor on removal is, it is possible to plan in advance the best time to remove an animal from the herd when it is still profitable to do so or at least incur minimal loss. Survival analysis is a commonly used technique to quantify longevity in domestic animals (5, 6). Using this technique, the association between risk factors and culling can be examined in relation to their effect on the length of productive life (LPL) instead of simply describing the relationship in terms of risk (5). In survival analysis, a quantity termed "hazard" is modeled instead of longevity itself (7). Hazard represents the instantaneous probability that an animal is removed at a given time, given that it is still present up to that time. Since it is the hazard that is modeled and not longevity, it is possible to use data from animals that have not yet been removed from the herd (as censored observations) as well as those that have been removed (7).

Although a number of studies have been carried out to identify risk factors for removal in dairy cows (5, 8, 9), the number of similar studies in dairy goats is limited (1, 10) and, to the best of our knowledge, none have been conducted in a New Zealand context. To address this knowledge gap, the aim of this study was to identify factors that influence the risk of removal in commercial dairy goat herds in New Zealand (11). This knowledge will allow managers of dairy goat herds take a more planned approach to culling: either to remove does at higher risk of removal at a time when it is economic to do so, or to preferentially manage profitable animals if it is known that they are at greater risk of removal compared to their herd mates.

MATERIALS AND METHODS

Study Population and Data Collection

The data for this study were obtained from the NZDGC. Since the total number of dairy goat herds in New Zealand is relatively small, we assumed the dairy goat herds affiliated with NZDGC provided an accurate reflection of commercial dairy goat farming in New Zealand. Although the complete data set was comprised of records for a total of 48,699 animals (including those with birth dates as early as August 1983 and production records up to December 2009), only those born on or after 1st January 2000 were used in the analyses presented in this paper. This restriction was applied because a large proportion of animals born prior to 1st January 2000 had missing observations, particularly those related to total lactation length and milk, fat, and protein yields.

Several exclusion criteria were applied to the NZDGC data (**Figure 1**). Bucks were excluded from the analyses. A doe had to complete her first lactation and then kid for a second time to be included in the data set so that the correct temporal sequence between first lactation milk solids yield (MSL1) and LPL was ensured. Finally, records were screened and limited to does having a first lactation length between 0 and 305 days and/or a first lactation total milk solids yield of less than or equal to 1,800 kg. Lactations of greater than 305 days and total lactation yields of more than 1,800 kg milk solids were deemed implausible. Finally, does for which the first lactation fat and protein yields

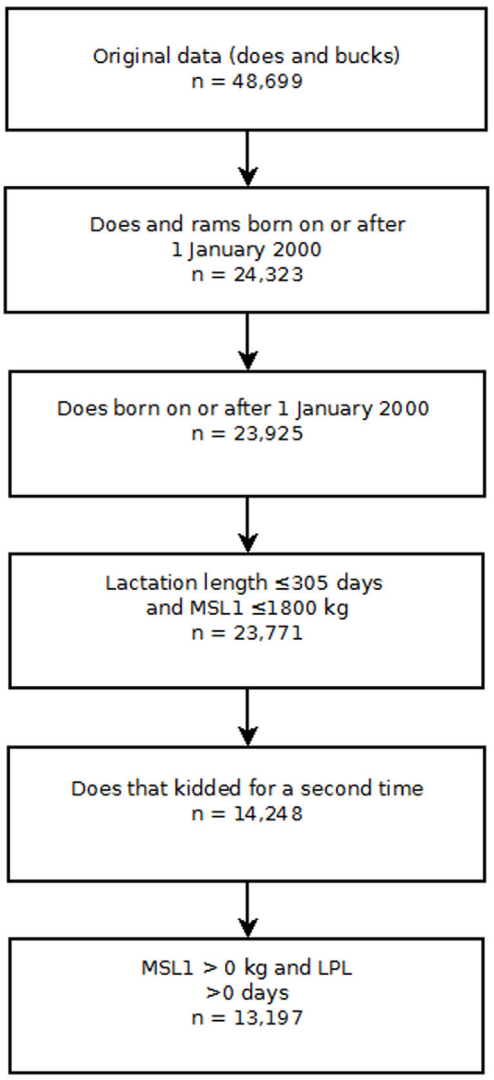


FIGURE 1 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009. Flow chart showing the exclusion criteria used to select individual doe records for analysis in this study. Key: MSL1 first lactation milk solids yield (kilograms); LPL, length of productive life.

were recorded as 0 were excluded from the analyses. Does were followed until 31st December 2009 or the date on which they were removed from the herd, whichever occurred first.

Herds registered with the NZDGC record data for individual animals including the date of birth, the unique animal identifier, breed, parity date(s), and the date and reasons for removal from the herd (if applicable). Herd managers record details of individual animals into paper diaries or, more rarely in the case of dairy goats, into dedicated herd health software. This information is then sent to the national milk recording authority, Livestock Improvement Corporation (LIC) who merge these details with test day milk yields measured at roughly 60-day intervals throughout the lactation. Animal biographical and production data recorded in the central database of LIC are then transferred to NZDGC in digital format. This information is used by NZDGC for genetic evaluation of individual animal (12). Estimated breeding values for milk, fat, protein, and milk solids (fat and protein) obtained from genetic evaluations are reported to the NZDGC and each herd manager receives an individual report with the genetic evaluation of his/her animals.

The outcome of interest in this study was LPL, defined as the difference in time (days) between the date of second kidding and the date of removal from the herd. In the context of this study, we use the term "removal" to refer to animals that leave the herd as either culled animals, sales, or deaths. For does that were still in the herd at the termination of the study (censored observations), LPL was quantified as the time between the date of second kidding and 31st December 2009.

Model Building

Selection of Explanatory Variables

The total yields of milk protein and milk fat from each animal in the first lactation were added to create a single variable called first lactation milk solids yield (MSL1).

Based on the reported breed composition of the sire and dam the breed of each animal was recorded in 16th for the following breeds: Saanen, Toggenburg, Nubian, Alpine, and "unknown." From these fractions (the total of which sum to one), the proportion of each breed was calculated. For instance, the breed composition of a doe with pedigree values 8, 4, 0, 0, 4 for Saanen, Toggenburg, Nubian, Alpine, and unknown (respectively) would be 50% Saanen, 25% Toggenburg, 0% Nubian, 0% Alpine, and 25% unknown. Given the several possible combinations of crossbreds, it was decided that the percentage of each breed would be forced into the model as a series of continuous variables to avoid any ambiguity created by breed defined as a categorical variable. The recorded parentage details for all does were not available. Where parentage details were not available, breed fractions were estimated by the herd manager.

Bivariate Analyses

Since all the explanatory variables in our study were continuously distributed, they were categorized into quartiles. The Kaplan–Meier technique (13) was then used to quantify LPL of does within each quartile. The log rank statistic was used to test the homogeneity of survivorship between quartile groups. Those explanatory variables that showed an association with LPL (that is, a difference in the Kaplan–Meier survival curves that was significant at $P < 0.20$) were selected for inclusion in the multivariate analyses.

Multivariable Analyses

Factors influencing LPL were quantified using a Cox proportional hazard model (14). Here, the hazard of removal at time t can be expressed as:

$$H(t,x) = h_0(t)\exp^{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}. \quad (1)$$

Equation 1 shows the hazard of an event at time t is the product of $h_0(t)$ and $\exp^{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}$. The first of these quantities, $h_0(t)$, is called the baseline hazard function and includes a time component t, representing how the hazard of removal changes as a function of time. The remaining quantity $exp^{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}$ is the exponential of the linear sum of a series of k explanatory variables. This quantity represents how the baseline hazard function is modified in response to a given set of explanatory variables. In contrast to the baseline hazard function, the set of explanatory variables does not involve a time component (15).

A key assumption of the Cox model is that of proportionality of hazards. According to this assumption, the effect of an explanatory variable on the outcome of interest does not change over time, i.e., the hazards for each level of an explanatory variable must be proportional at all times. In situations where this assumption is violated, modifications such as stratified analyses or inclusion of time-dependent covariates are necessary (16).

Model development was carried out using the contributed survival package (17) implemented in R version 3.3.3 (18). To start, a saturated Cox model was run including all explanatory variables identified as influencing LPL at the bivariate level. Explanatory variables that were not statistically significant were removed from the model one at a time, beginning with the least significant, until the estimated regression coefficients for all explanatory variables retained were significant at an alpha level of less than 0.05. Explanatory variables that were excluded at the initial screening stage were tested for inclusion in the final model and were retained in the model if their inclusion changed any of the estimated regression coefficients by more than 20%. Biologically plausible two-way interactions were between explanatory variables were assessed.

Checking the Scale of Continuous Covariates

A key assumption in including MSL1 into the model as a continuous variable was that the relationship between MSL1 and log hazard was linear. To test this assumption, MSL1 was categorized into quartiles and the regression coefficient for each quartile plotted as function of the midpoint of each quartile group. Since the line connecting the four midpoints was not linear, we concluded that MSL1 was not linear in its log hazard. Based on these findings, a penalized spline term was used to account for the non-linear association between MSL1 and LPL.

Testing the Proportional Hazard Assumption

To verify that the proportional hazards assumption of the Cox model was valid a plot of the scaled Schoenfeld residuals from the model as a function of time was constructed. In a model

where the proportional hazards assumption holds the Schoenfeld residuals should be scattered around 0. We calculated the Pearson product–moment correlation between the scaled Schoenfeld residuals and time and the hypothesis of no correlation between the two variables was assessed using a χ^2 test statistic. From these analyses, we concluded that MSL1 violated the proportional hazards assumption. To account for non-proportionality of hazards, we divided LPL into two intervals: less than or equal to 730 days (referred to as T1 in the remainder of this paper) and greater than 730 days (T2). The decision to use 730 days was semi-arbitrary and was selected because, being equivalent to 2 years, it approximated median LPL in this population. This division allowed us to quantify the effect of MSL1 separately for each period [less than or equal to 730 days (T1) and greater than 730 days (T2)]. The technique of dividing the time component into intervals to investigate the time-dependent effect of covariates is called a piecewise Cox proportional hazards model or a step function proportional hazards model.

Final Model

In addition to the terms to allow for the interaction between time and penalized MSL1, our final model included herd as a random effect, otherwise known as a frailty term.

RESULTS

The final data set was comprised of 23,771 does with a birth date greater than or equal to 1st January 2000. Of this group, 14,248 does completed their first lactation and kidded for the second time. Further screening of the production data and removal of implausible records reduced the final data set to comprised 13,197 does from 38 herds (**Figure 1**). Of this group, 5,386 animals were removed during the follow-up period and the remaining 7,811 animals that were recorded as being alive in the herd on 31st December 2009 were treated as censored observations. Descriptive statistics of the study population are presented in **Table 1**.

Inclusion of terms for breed in the Cox proportional hazards model was not statistically significant. Biologically plausible two-way interactions were tested and none were significant at an alpha level of 0.05.

As shown in **Table 2**, the interaction between MSL1 and time was significant for T1, but was not statistically significant for T2. During T1, the hazard of removal for does that produced 80 kg milk solids in the first lactation was 0.84 (95% CI 0.58–1.23) times the daily hazard of removal for does that produced 30 kg milk solids in the first lactation (**Figure 2**). During T2 (730 days after the date of second kidding), high producing MSL1 does had a higher daily hazard of removal compared to average producing herd mates: a doe producing 80 kg milk solids in the first lactation had 1.44 (95% CI 0.79–2.65) times the daily hazard of removal compared with does that produced 30 kg milk solids in the first lactation (**Figure 3**). These results show that relatively high levels of MSL1 production had no strong association with daily hazard of removal during the early phase of productive life, however, as LPL progressed, does with higher MSL1 yields were at greater risk of removal.

DISCUSSION

We used a piece-wise Cox proportional hazards model, to quantify the effect of MSL1 on LPL in dairy goats that completed their first lactation and kidded a second time. To the best of our knowledge, this is the first study of its kind to evaluate the effect of a time-dependent covariate on longevity in dairy goats.

Although the results presented in this study are based on data which were not originally collected for the purpose of this study, consent to use and analyze the data was obtained from NZDGC before the start of the study and results were presented to NZDGC stakeholders. A possible limitation of our study was selection bias in that the herds used for these analyses were those that participated in herd testing programs and were, therefore, likely to be a more intensively managed subset of dairy goat herds compared with the general population of New Zealand dairy goat herds. A second limitation was that we could not investigate the effect of specific diseases or disease categories on longevity. There were two reasons for this: (1) we had no reassurance that disease case definitions were used consistently over time and across each of the herds that took part in the study; and (2) does were removed for a wide range of reasons resulting in relatively low numbers of animals in each category. When studying factors influencing LPL in production animals, it is desirable to identify

TABLE 1 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009.

Outcome	*n*	Mean	SD	Median	Q1; Q3
L1 fat yield (kg)	13,197	16	8	16	10; 21
L1 protein yield (kg)	13,197	14	7	14	9; 18
L1 milk solids (kg)	13,197	30	15	29	19; 40
Age at first kidding (days)	13,197	580	421	390	369; 669
LPL (days)	5,386[a]	763	547	663	327; 1,084
Age at removal (days)	5,386[a]	1,644	596	1,500	1,142; 2,026
Number of lactations	5,386[a]	3	1.4	3	2; 4

Descriptive statistics of first lactation production outcomes, age at first kidding, LPL, age at removal and total number of lactations.
L1, lactation 1; LPL, length of productive life; Q1, first quantile; Q3, third quantile.
[a]*Uncensored does only.*

TABLE 2 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009.

Variable	Coefficient (SE)	Chi square	df	P
MSL1 × T1				
Linear	–0.0033 (0.0014)	5.31	1.0	0.021
Non-linear	–	1.68	3.0	0.650
MSL1 × T2				
Linear	0.0014 (0.0016)	1.00	1.0	0.360
Non-linear		3.05	3.0	0.030
Herd-level random effect	–	2,358.74	13.60	0.000

Regression coefficients of factors influencing risk of culling in dairy goats from the final piecewise Cox model.
MSL1, first lactation milk solids yield (kilogram); T1, 0–730 days from the date of second kidding; T2, greater than 730 days from the date of second kidding.

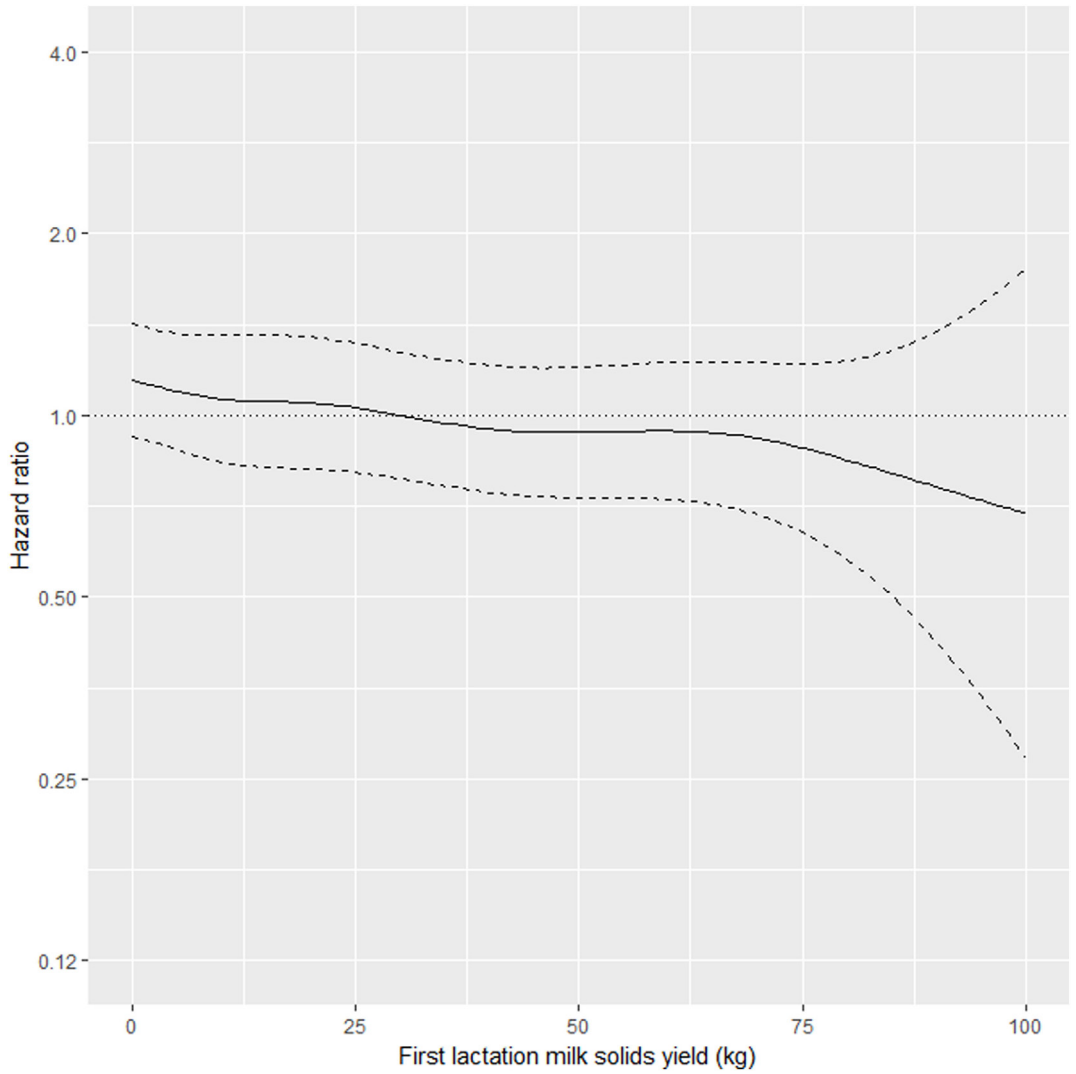


FIGURE 2 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009. Line plot showing, for the interval 0–730 days from the date of second kidding, the hazard ratio for removal as a function of first lactation milk solids yield (based on the model presented in **Table 2**). The dashed lines represent 95% confidence intervals around the point estimates of the hazard ratio. In the above plot, the reference category was a doe producing 30 kg milk solids in the first lactation. A doe producing 80 kg milk solids in the first lactation had 0.84 (95% CI 0.58–1.23) times the daily hazard of removal compared with a doe that produced 30 kg milk solids in the first lactation.

risk factors for specific removal reasons (e.g., reproductive failure, udder health, lameness) as opposed to considering all removals as a single group. Failure to do so is likely to mask some of the more subtle influences on longevity. As a prerequisite for being able to examine specific reasons for removal, it is necessary that removal reasons are recorded accurately and consistently across herds and over time.

Our results show that in the first 2 years after the date of second kidding, there was an inverse association between MSL1 yields and the daily hazard of removal (**Figure 2**). Does with higher MSL1 yields had lower daily hazards of removal compared with average producing herd mates. This trend reversed beyond 2 years from the date of second kidding (**Figure 3**) with high MSL1 yields having a higher daily hazard of removal compared with average producing herd mates. We believe these results provide useful information for the management of dairy goat herds. As high producers get older, herd managers need to take special steps to ensure that this group of animals is managed in such a way to minimize the impact of factors that could influence removal risk. For example, a herd manager might elect to run his/her high MSL1 producers as a separate mob and to provide preferential feeding, housing, and milking management.

A search of the literature did not identify any previous studies that investigated the association between first lactation milk solids yield and longevity in dairy goats. Even in dairy cattle, the number of studies that have examined the association between first lactation milk yield and longevity is limited (19–22). It has been shown that mean daily yield of milk in the first lactation of

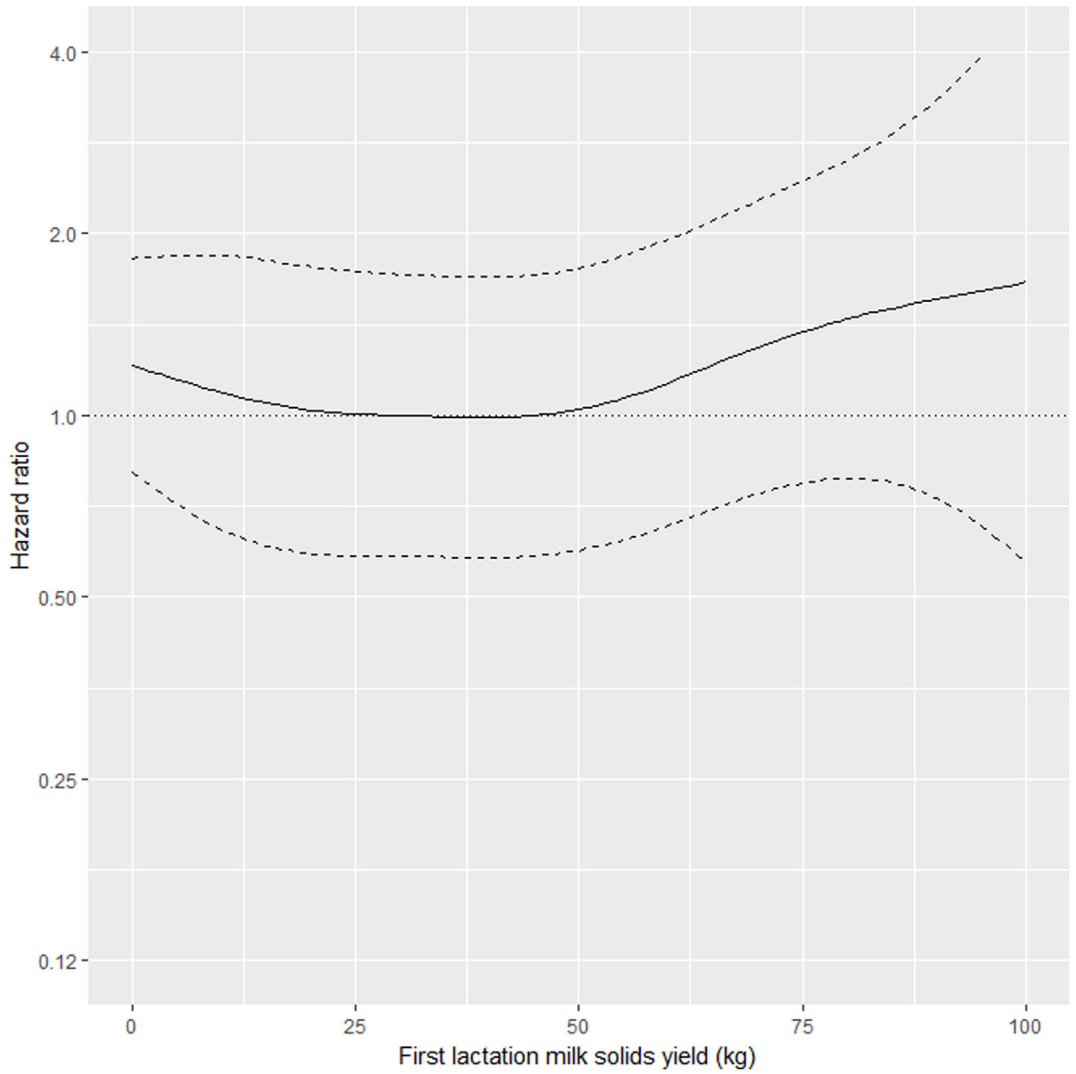


FIGURE 3 | Risk factors for culling, sales and deaths in New Zealand dairy goat herds, 2000–2009. Line plot showing, for the interval greater than 730 days from the date of second kidding, the hazard ratio for removal as a function of MSL1 (based on the model presented in **Table 2**). The dashed lines represent 95% confidence intervals around the point estimates of the hazard ratio. In the above plot, the reference category was a doe producing 30 kg milk solids in the first lactation. A doe producing 80 kg milk solids in the first lactation had 1.44 (95% CI 0.79–2.65) times the daily hazard of removal compared with a doe that produced 30 kg milk solids in the first lactation.

a cow is an early indicator of lifetime yield (21–23). While the total lifetime yield or daily milk yield in animals in subsequent lactations can be expected to be high in animals that produce more milk in the first lactation, overall reproductive performance decreases (22). Animals producing high amounts of milk in the first lactation are subject to a greater level of metabolic stress as a result of negative energy balance (20), which consequently leads to impaired fertility (24). Since we investigated the effect of MSL1 on LPL instead of first lactation milk yield and our study involved dairy goats, it is not possible to directly extrapolate the results of the above cow-based research to our study. Nevertheless, it is biologically plausible to assume that high yields of milk solids in first lactation would have a negative impact on the energy balance of dairy animals regardless of species. However, with good management, this negative effect may be unapparent for a reasonable period of time which, in our case, was approximately 2 years after the date of second kidding.

In this study, the effect of MSL1 on LPL was investigated using a model that included herd as a random effect (frailty) term. A frailty term is a continuous variable that quantifies the unobserved heterogeneity for groups of individuals such as those in families, classes, schools, or herds (25). Frailty terms are important because they provide a means for accounting for heterogeneity (i.e., "clustering") in outcome risk that arises from individuals within a cluster being more similar than individuals selected at random from the general population. Since variations in management practices among herds can be expected, the use of herd level effect as a frailty term is a standard practice

in epidemiological studies that quantify risk factors for given outcomes in domestic, farmed animal populations (26). The significance of the herd-level effect term in the model indicates that the hazard of removal as a function of LPL varied across herds. We propose that studies comparing herds with upper quartile frailty terms with those with lower quartile frailty terms may be useful to identify specific herd-level factors that are influential determinants of LPL. For example, a cross-sectional questionnaire survey can be designed to investigate various aspects of management such as nutrition, veterinary care, breeding practices, and milking practices in these two categories of farms and the data used to analyze differences between "low risk" and "high risk" herds in terms of survival.

In general, where heterogeneity is an unavoidable feature of the population under investigation, researchers should take into account the existence of dissimilarities among groups to avoid errors during analysis. By failing to acknowledge such heterogeneity, a researcher is more likely to make Type I error, which means he/she is likely to report a false association between explanatory and outcome variables when there is none. Interestingly, the protective effect of high MSL1 on the hazard of removal during T1 was evident only after the effect of herd was accounted-for in the model as a frailty term. When herd-level effects were not controlled-for, high MSL1 in L1 was positively associated with an increase in the risk of removal.

Several studies conducted on dairy cows have studied animal traits affecting LPL. Since longevity usually refers to the time between the first parity of an animal and its removal from the herd, it is not possible to get a direct measure of longevity for all animals, particularly those that are younger (6). However, with the use of survival analysis, such issues can be accounted-for because the technique uses information from all animals used in the study regardless of their culling status at the end of the study. Since we were interested to find out if MSL1 was associated with longevity, we defined longevity as the number of days between the date of second kidding and the date of removal from the herd. In this way, we could be sure that the explanatory variable (MSL1) preceded the study outcome (LPL), ensuring the correct temporal sequence between cause and effect.

CONCLUSION

This study identified a time varying effect of MSL1 on removal in New Zealand dairy goats. We found that does with high MSL1 yields had a lower risk of removal during the first 2 years following the second kidding compared with compared with their average producing herd mates. Beyond 2 years following the second kidding, does with high MSL1 yields had a relatively high hazard of removal compared with their average producing herd mates. We conclude that involuntary losses may be avoided if high MSL1 yielding does are preferentially managed from 2 years beyond the date of second kidding.

The data and analyses presented in this paper are based on the first author's thesis presented as partial fulfillment of the requirements for the degree of Master of Veterinary Studies at Massey University, New Zealand.

AUTHOR CONTRIBUTIONS

Study conception and design and critical revision: MG, MS, NL-V, and VM. Acquisition of data: NL-V and VM. Analysis and interpretation of data, and drafting of manuscript: MG and MS.

REFERENCES

1. Pérez-Razo M, Sánchez F, Torres-Hernández G, Becerril-Pérez C, Gallegos-Sánchez J, González-Cosío F, et al. Risk factors associated with dairy goats stayability. *Livest Prod Sci* (2004) 89:139–46. doi:10.1016/j.livprodsci.2004.02.008
2. Essl A. Longevity in dairy cattle breeding: a review. *Livest Prod Sci* (1998) 57:79–89. doi:10.1016/S0301-6226(98)00160-2
3. Jovanovac S, Raguž N, Sölkner J, Mészáros G. Genetic evaluation for longevity of Croatian Simmental bulls using a piecewise Weibull model. *Arch Anim Breed* (2013) 56:89–101. doi:10.7482/0003-9438-56-009
4. Solis-Ramirez J, Lopez-Villalobos N, Blair HT. Dairy goat production systems in Waikato, New Zealand. *Proceedings of the New Zealand Society of Animal Production*. Invercargill (2011). p. 86–91.
5. Stevenson M, Lean I. Risk factors for culling and deaths in eight dairy herds. *Aust Vet J* (1998) 76:489–94. doi:10.1111/j.1751-0813.1998.tb10188.x
6. Szabó F, Dákay I. Estimation of some productive and reproductive effects on longevity of beef cows using survival analysis. *Livest Sci* (2009) 122:271–5. doi:10.1016/j.livsci.2008.09.024
7. Forabosco F. *Breeding for Longevity in Italian Chianina Cattle [Doctor of Philosophy Dissertation]*. Wageningen, The Netherlands: Department of Animal Science, University of Wageningen (2005).
8. Seegers H, Beaudeau F, Fourichon C, Bareille N. Reason for culling in French Holstein cows. *Prev Vet Med* (1998) 36:257–71. doi:10.1016/S0167-5877(98)00093-2
9. Bell M, Wall E, Russell G, Roberts D, Simm G. Risk factors for culling in Holstein-Friesian dairy cows. *Vet Rec* (2010) 167:238–40. doi:10.1136/vr.c4267
10. Malher X, Seegers H, Beaudeau F. Culling and mortality in large dairy goat herds managed under intensive conditions in western France. *Livest Prod Sci* (2001) 71:75–86. doi:10.1016/S0301-6226(01)00242-1
11. Gautam M. *Epidemiological Study of Removals in New Zealand Dairy Goat Herds [Master of Veterinary Science Dissertation]*. Palmerston North, New Zealand: Institute of Veterinary, Animal and Biological Sciences, Massey University (2012).
12. Singireddy SR, Lopez-Villalobos N, Garrick DJ. Across-breed genetic evaluation of New Zealand dairy goats. *Proceedings of the New Zealand Society of Animal Production*. Lincoln: New Zealand Society of Animal Production (1997). p. 43–5.
13. Efron B. The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc* (1977) 76:312–9. doi:10.1080/01621459.1981.10477650
14. Cox D. Regression models and life tables. *J R Stat Soc* (1972) 34(B):187–220.
15. Kleinbaum D, Klein M. *Survival Analysis: A Self Learning Text*. New York, USA: Springer-Verlag (2012).
16. Ata N, Sözer M. Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe J Math Stat* (2007) 36:157–67.
17. Therneau T, Grambsch P. *Modeling Survival Data: Extending the Cox Model*. New York: Springer (2000).
18. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (2017).
19. Robertson A, Barker JSF. The correlation between first lactation milk production and longevity in dairy cattle. *Anim Prod* (1966) 8:241–52. doi:10.1017/S0003356100034619
20. Pasman E, Otte M, Esslemont R. Influences of milk yield, fertility and health in the first lactation on the length of productive life of dairy cows in Great Britain. *Prev Vet Med* (1995) 24:55–63. doi:10.1016/0167-5877(94)00457-T
21. Haworth G, Tranter W, Chuck J, Cheng Z, Wathes D. Relationships between age at first calving and first lactation milk yield, and lifetime productivity and longevity in dairy cows. *Vet Rec* (2008) 162:1–6. doi:10.1136/vr.162.20.643

22. Sawa A, Krężel-Czopek S. Effect of first lactation milk yield on efficiency of cows in herds with different production levels. *Arch Anim Breed* (2009) 52:7–14.
23. Jairath L, Hayes J, Cue R. Correlations between first lactation and lifetime performance traits of Canadian Holsteins. *J Dairy Sci* (1995) 78:438–48. doi:10.3168/jds.S0022-0302(95)76653-X
24. Pryce J, Royal M, Garnsworthy P, Mao I. Fertility in the high-producing dairy cow. *Livest Prod Sci* (2004) 86:125–35. doi:10.1016/S0301-6226(03)00145-3
25. Wienke A. *Frailty Models. MPIDR Working Paper WP 2003-032*. Max Planck Institute for Demographic Research (2003). Available from: https://www.demogr.mpg.de/papers/working/wp-2003-032.pdf
26. Dohoo I, Martin S, Stryhn H. *Veterinary Epidemiologic Research*. Prince Edward Island, Canada: AVC Inc Charlottetown (2009).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer CP and handling editor declared their shared affiliation.

EQUINE VETERINARY JOURNAL
Equine vet. J. (2002) **34** (5) 438-443

Long-term survival of equine surgical colic cases. Part 2: Modelling postoperative survival

C. J. PROUDMAN*, J. E. SMITH, G. B. EDWARDS and N. P. FRENCH

Faculty of Veterinary Science, University of Liverpool, Leahurst, Neston, Wirral CH64 7TE, UK.

Keywords: horse; colic; survival analysis; penalised Cox regression; random effects modelling

Summary

Colic surgery is a frequently performed operation with high postoperative mortality. This study was undertaken to identify variables associated with decreased postoperative survival. We used data from 321 horse years of postoperative survival time to model the probability of survival following recovery from colic surgery. Continuous variables were modelled using a 6 variable, penalised Cox regression model. This demonstrated approximately linear relationships between survival and the following variables: increase in packed cell volume (PCV), intestinal resection length, time to surgery (interval between onset of colic and surgery) and duration of surgery. No significant decrease in survival was demonstrated with increasing age of the patient or with heart rate. The only categorical variable to be significantly associated with decreased survival was epiploic foramen entrapment. The final, fixed effects Cox proportional hazards model of postoperative survival included the variables epiploic foramen entrapment, PCV, resection length and duration of surgery, each variable adjusted for the nonlinear relationship with time to surgery. Residual variation in postoperative survival attributable to professional personnel (referring veterinary surgeon, anaesthetist and surgeon) was explored by fitting each as a random effects term in the model. Little of the residual variation could be attributed to any category of personnel. Model diagnostics indicated little influence by individual outliers on model parameters and little evidence of subjects poorly predicted by the final model. The study highlights factors influencing the long-term survival of horses recovering from colic surgery and proposes a model that can be used to inform prognosis.

Introduction

The management of horses with colic is a major challenge to equine veterinary surgeons. The term 'colic' encompasses a wide range of disease entities, all of which have a similar clinical presentation. Prognosis however, varies greatly between the different diseases and according to the treatment regimen selected. A further complication in the assessment of colic cases is the large number of clinical parameters that can be measured and the variability of each parameter. Recent advances in epidemiology and statistical modelling enable the detailed exploration of complex relationships between explanatory variables and specific outcomes (e.g. survival, hernia formation, postoperative ileus). Furthermore, the co-relationships between explanatory variables can be studied and accounted for (Reeves and Curtis 1989).

Multivariable modelling has been used in a number of studies of equine colic (reviewed by Reeves and Curtis 1989). The most common application has been the development of prognostic models (Parry *et al.* 1983; Puotunen-Reinert 1986; Orsini *et al.* 1988; Reeves *et al.* 1989, 1990, 1992; Pascoe *et al.* 1990; Furr *et al.* 1995). None of these models have gained widespread acceptance in clinical practice. Reasons for this include the apparent complexity of the models (Reeves and Curtis 1989), but also difficulties in validating models developed on data from one hospital for use in a different hospital with different prevalences of surgery, death and different colic types.

Estimating prognosis in surgical cases prior to surgery necessarily precludes information gained at the time of surgery e.g. nature of the lesion, length of ischaemic bowel, duration of surgery. All of these factors may have a profound influence on the probability of survival. Exploratory laparotomy is a diagnostic procedure as much as a therapeutic one. For this reason we wanted to evaluate the prognostic importance of intra-operative variables.

The present study uses data derived from a 3 year, prospective study of postoperative survival of colic cases (Proudman *et al.* 2002). Specific objectives of data modelling are the development of simple models of postoperative survival, hypothesis generation and the description of the functional form of the relationships between continuous variables and the probability of survival. Random effects models (Aalen 1988) are used to evaluate, in a nondivisive manner, the influence of professional personnel on the outcome of colic surgery.

Materials and methods

Study population and data collection

Data from the study described by Proudman *et al.* (2002) were used. In brief, the clinical details from 341 horses that recovered from colic surgery were recorded on a computer database. All horses eligible were recruited between March 1998 and August

*Author to whom correspondence should be addressed.

TABLE 1: Comparison of age, heart rate, PCV and time to surgery between horses surviving colic surgery and those dying or undergoing euthanasia prior to recovery

	Survivors		Nonsurvivors	
	Median	s.d.	Median	s.d.
Age (years)	10	6.4	13	5.6
Heart rate (beats/min)*	50	19.0	78	22.0
PCV (l/l)*	0.38	0.08	0.50	0.11
Time to surgery (h)	15.3	25.3	18.5	70.6

*IP<0.05 for Wilcoxon two-sample test.

2000. Postoperative progress was rigorously followed during hospitalisation and by means of periodic telephone and postal questionnaire, after discharge. A total of 321 horse years of postoperative survival were documented. The study population was a subpopulation of all surgical colic cases as horses were only recruited onto the study upon successful recovery from anaesthesia following surgery. Table 1 provides summary data for some clinical parameters for the study population (n = 311, excluding grass sickness cases), and for the 32 colic cases, that were anaesthetised for surgery during this period, that failed to recover from anaesthesia and were, therefore, ineligible for recruitment into the study.

Data analysis

Descriptive data for surgical survivors and nonsurvivors was compared with the Wilcoxon 2 sample test. Critical probability was set at P<0.05. The shape of the relationship between continuous variables (e.g. heart rate at admission, age, length of resection) and mortality was explored using penalised Cox regression models[1] (Therneau and Grambsh 2000). These are extensions of Cox regression models that fit nonparametric functions (p-spline smoothers) to estimate the relationships between outcome and explanatory variables (Anon 2001). The results can be displayed graphically to illustrate the multivariable functional form of these relationships (e.g. linear, quadratic or cubic). Penalised Cox regression modelling in S-Plus[1] has the additional advantage of testing fitted functions for linearity and the significance of nonlinearity. Final models were constructed using backwards elimination procedures and an assessment of the effect of variable inclusion on parameter estimates. A critical probability of 0.05 was used to assess effect. Model diagnostics explored the influence of individual observations on regression coefficients using plots of the scaled change in regression coefficient for each observation. Changes in coefficient greater than 0.4 of s.e. can be interpreted as exerting disproportional influence (Therneau 1994). A deviance plot was also generated, indicating the ability of the model to differentiate survivors and non-survivors. The combined effect of pairs of variables on the probability of mortality up to 100 days was explored by plotting 3-dimensional graphs of the smoothed relationships generated by generalised additive models (Hastie and Tibshirani 1990).

Residual variability in survival that could be attributed to random effects was tested by including variables such as referring clinician, surgeon and anaesthetist as frailty (gamma) terms in the model (Aalen 1988). A critical probability of 0.05 was used to determine significant effects.

TABLE 2: Variables explored as potential explanatory variables for postoperative survival

Continuous variables	Categorical variables
Age	Breed
Heart rate at admission	Gender
Duration of colic prior to surgery	Laparotomy diagnosis
Packed cell volume at admission	Resection (yes/no)
Resection length	Anastomosis type (jejuno-caecal vs. jejuno-jejunal)
Duration of surgery	Anastomosis method (stapled vs. handsewn)
	Anaesthetic induction agent
	Anaesthetic gaseous agent

Results

Continuous variables

Six continuous variables thought *a priori* to influence postoperative survival were modelled in a 6 variable, penalised Cox regression model using p-spline smoothers. Figure 1 illustrates the multivariable smoothed relationships between these variables and mortality risk (log hazard). It is apparent that age and heart rate at admission show no marked or consistent association with mortality. However, packed cell volume at admission (PCV), resection length and duration of surgery all show reasonably linear increases in mortality with increasing values. P values for linearity for PCV, resection length, duration of surgery and time to surgery (time between onset of colic and surgery) are all less than 0.05 indicating a significant linear component to their relationship with mortality. Of these 4 variables, only time to surgery has a P value for non linearity approaching 0.05, suggesting a significant non linear component to this p-spline fit.

Figure 2 illustrates the combined effect of PCV and duration of surgery and of PCV and resection length on the probability of death before 100 days. It is apparent that horses with high values of 2 variables had a significantly greater probability of death. For example, a horse with a PCV of 0.45 l/l at admission and no intestine resected had a probability of death of approximately 0.2 (20% of horses at risk). Whereas one with a similar PCV and 30 feet of intestine resected had a probability of approximately 0.45 (45% of horses at risk), rising to 0.6 if 35 feet of intestine were resected. No significant multiplicative interaction between variables was detected. Table 1 provides summary data for 4 continuous variables in survivors and nonsurvivors of colic surgery. Heart rate and PCV for these 2 groups are significantly different.

Categorical variables

Survival was categorised by a number of variables (listed in Table 2) and the influence of each on postoperative survival examined using Cox proportional-hazards modelling. Only epiploic foramen entrapment emerged as a significant variable. Probability of survival of epiploic foramen entrapment cases is significantly different to that of ileal impaction and pedunculated lipoma cases (RR = 2.1, 95% CI 1.4, 2.8, P = 0.033). Kaplan-Meier plots comparing the cumulative probability of survival after surgery for different small intestinal lesions are shown in Proudman *et al.* (2002).

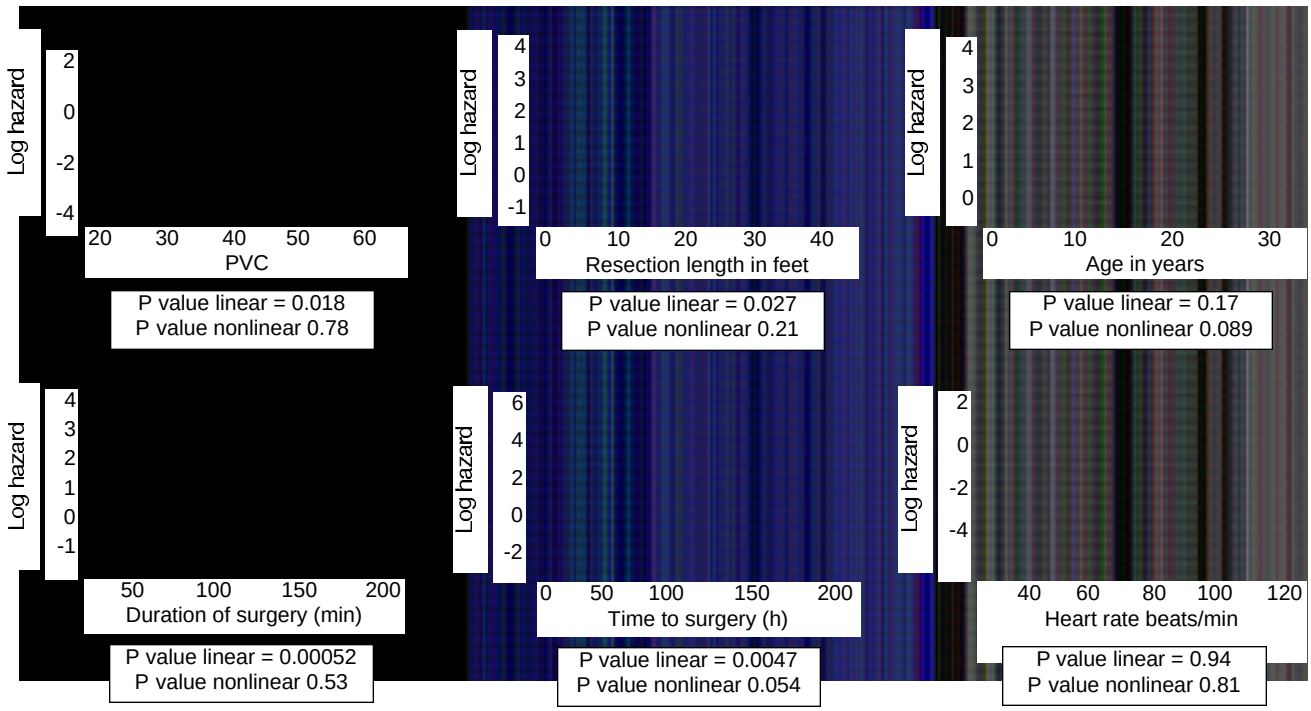


Fig 1: Plot of p-spline smoothers for the 6 continuous variables considered for inclusion in the fixed effects model. Results of significance tests for linearity and nonlinearity shown for each variable.

Model construction

Based on the results described above, a Cox proportional-hazards model for postoperative survival was constructed including PCV, resection length, time to surgery, duration of surgery and epiploic foramen entrapment. Due to the nonlinear relationship between time to surgery and mortality, this variable was fitted as a p-spline with 5 degrees of freedom (the most parsimonious fit). Table 3 gives details of parameter values for the model, adjusted for the nonlinear relationship with time to surgery.

Random effects

The residual variation due to referring clinician, surgeon and anaesthetist was explored by inclusion of each as a random effect (frailty) term in the fixed effects model. The *a priori* hypothesis being tested was that some referring clinicians, surgeons and anaesthetists were associated with better long-term survival. The variance estimates for these effects are given in Table 4 and it is apparent that residual variation attributable to anaesthetist, surgeon or referring veterinary surgeon in this hospital was nonsignificant.

TABLE 3: Parameter values for a fixed effects model of postoperative survival (adjusted for nonlinear relationship with time to surgery)

Variable	Coefficient (ß)	s.d.	P value
Fixed effects model			
Resection length (increment per foot)	0.029	0.014	0.0310
Duration of surgery (increment per min)	0.012	0.004	0.0011
PCV (increment per 1%)	0.046	0.018	0.0088
Epiploic foramen entrapment (y/n)	0.75 (RR = 2.11; 1.42, 2.79)	0.350	0.0330

y/n = yes/no.

Model diagnostics

Apart from 4 observations on the regression coefficient for the variable 'time to surgery,' the scaled residuals are less than 0.4 of s.e. (Fig 3) indicating that individual observations have relatively little influence on the parameters in the final model. The plot of deviance residuals from the final model shows little evidence of poorly predicted subjects.

Discussion

The model described above differs radically from previous models of colic survival because it deals with the long-term survival of horses undergoing colic surgery. Previous prognostic models (Parry *et al.* 1983; Puotunen-Reinert 1986; Orsini *et al.* 1988; Reeves *et al.* 1989, 1990, 1992; Pascoe *et al.* 1990; Furr *et al.* 1995) have sought to use pre-operative clinical data to predict short-term outcome. The

TABLE 4: Variance estimates for random effects terms in the fixed effects model

Variable	Variance estimate	P value
Random effects		
Referring veterinary surgeon	0.00	0.93
Surgeon	0.06	0.15
Anaesthetist	0.14	0.25

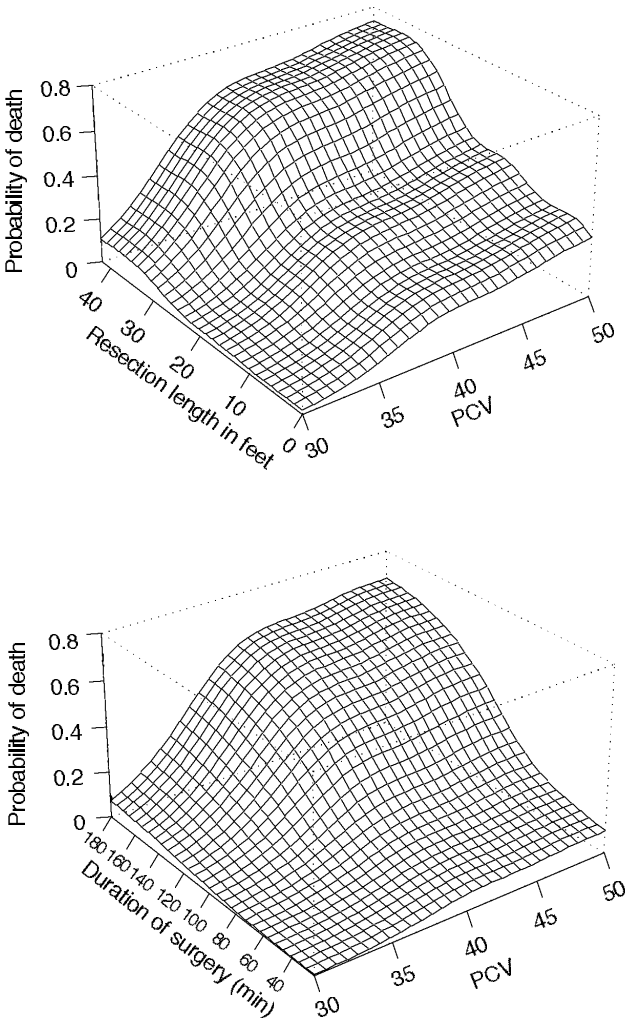


Fig 2: Three-dimensional graphs illustrating the combined effect of PCV and resection length (above), and PCVand duration of surgery (below) on the probability of death by 100 days.

modelling techniques used in our study also differ from those employed previously. In particular, evaluation of the functional form of relationships between continuous variables and mortality, and the use of penalised Cox models, have not previously been applied to postoperative colic data.

There are clear differences between our study population (horses surviving colic surgery) and the population of surgical nonsurvivors not eligible for inclusion in this study (Table 1). Median heart rate and PCV are higher than for the surgical survivor group. This suggests that our case definition (recovery from surgery) excludes many severely endotoxaemic horses which are unlikely to survive surgery. Clinical parameters indicating endotoxaemia have been identified previously as important prognostic indicators in horses prior to colic surgery (Orsini *et al.* 1988; Reeves *et al.* 1989, 1990, 1992; Pascoe *et al.* 1990; Furr *et al.* 1995; Thoefner *et al.* 2001).

The current approach acknowledges that much of the variability in survival is due to the different disease processes that cause colic. Of the 5 significant variables, only one is a pre-operative clinical parameter. The other 4 variables (time to surgery, epiploic foramen entrapment, duration of surgery and resection length) are contingent upon surgery having taken place. This study suggests that trying to predict postoperative survival without these data would exclude much useful information. It is also questionable how necessary it is to use models to predict survival preoperatively. Blikslager and Roberts (1995) demonstrated that clinicians were reasonably accurate at predicting survival on the basis of clinical examination. They cite positive predictive values of 83–91%. A more important prediction might be the need for surgical intervention in colic cases. This has been addressed by Reeves *et al.* (1992) who developed a logistic regression model to predict the need for surgery. This model included seven clinical variables. Validation of the model indicated that the model fitted the data poorly. In the light of these previous studies and the reluctance of clinicians to use predictive models incorporating preoperative data only, our study was designed to evaluate data derived at surgery as well as preoperatively, and to identify variables associated with long-term survival of those horses successfully recovering from surgery.

The relationships between continuous variables and survival are illustrated by the output from the 6 variable, penalised Cox regression model. This suggests that PCV, resection length and duration of surgery have an approximately linear association with decreased survival (mortality). Increasing values of ‘time to surgery’are associated with a significant increase in mortality but this relationship is significantly nonlinear. It is of interest that age shows no marked association with survival. Although the smoothing spline indicates increased mortality in horses over age 20 years, the confidence interval is wide due to the small number of observations and suggests that any difference is non-significant. This study offers little evidence that older horses recovering from colic surgery have a worse prognosis than younger ones.

The observed relationship between heart rate and survival indicates no increase in mortality with increasing heart rate. This is apparently contrary to the findings of Reeves *et al.* (1990, 1992) who found that a number of variables relating to cardiovascular compromise (PCV, heart rate, capillary refill time, pulse quality) were significantly associated with survival. This apparent difference can be explained by the different study populations. Our study focussed on surgical survivors only, largely excluding severely endotoxaemic horses (with high heart rate, PCV, poor pulse quality) as they were unlikely to survive surgery. Freeman *et al.* (2000), in a study of exclusively small intestinal colic cases, also reported no effect of heart rate.

The 3 dimensional plots, illustrating the combined effect of changes in 2 variables on the probability of death, highlight the need for a multivariable approach to modelling survival in postoperative colic cases. Single variables, considered in isolation, will not give an accurate estimate of the probability of survival in individual cases. Although the 3 dimensional plots consider the influence of 2 variables, the final model contains 5 variables that should be considered in combination when estimating the probability of death.

Modelling postoperative survival serves to highlight the variables with the greatest influence on postoperative survival. The results of our study suggest that both duration of surgery and length of intestinal resection are surgical variables with considerable prognostic value. Efforts to improve long-term survival should be focussed on decreasing surgery time and on understanding the mechanisms underlying the association

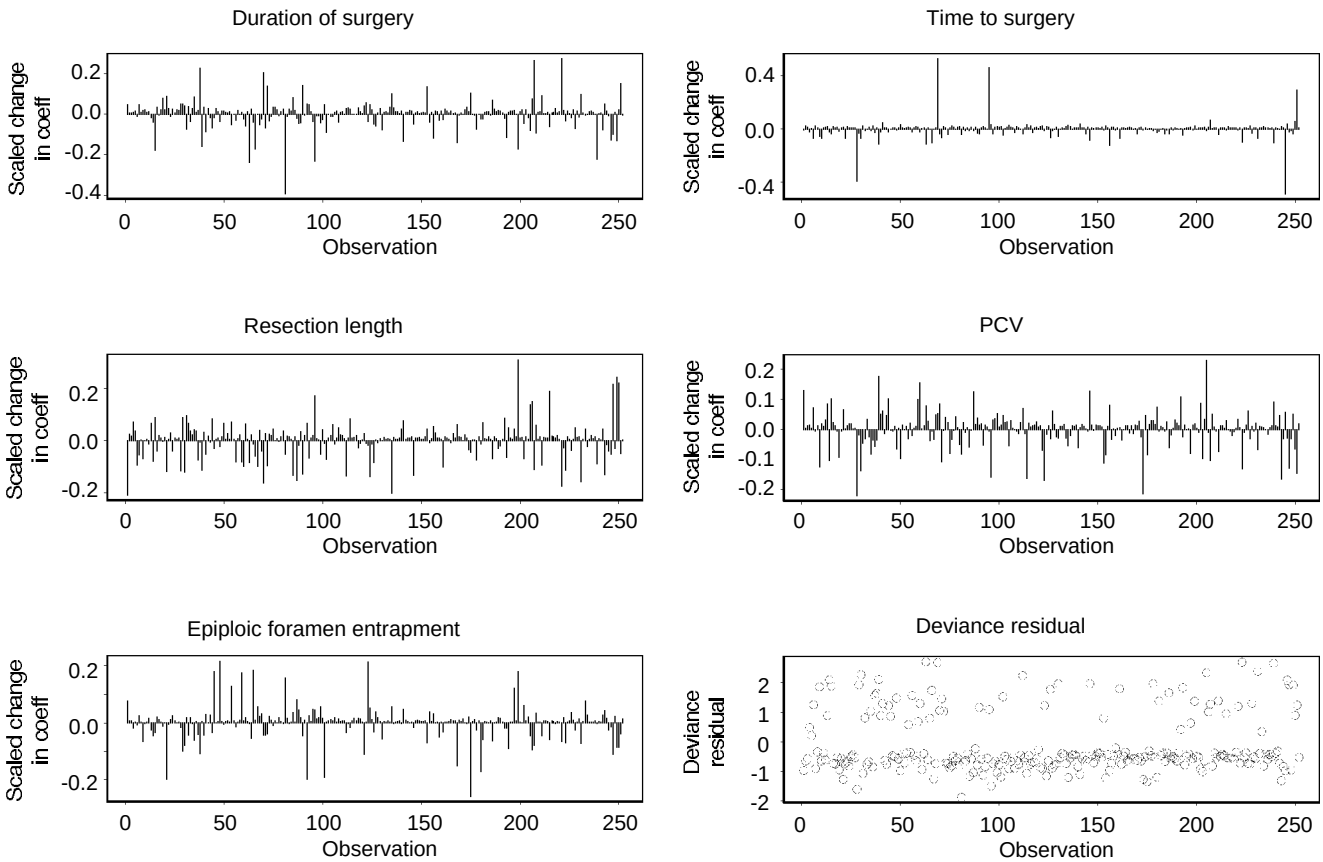


Fig 3: Plots of scaled residuals for each variable in the final model. Values of >0.4 for individual observations indicate outliers that may exert disproportional influence on the regression coefficient. The deviance residual plot indicates the ability of the model to differentiate survivors from nonsurvivors.

between length of resected bowel and risk of mortality. While some of the causes of prolonged surgery time are beyond the control of the surgeon (e.g. length of bowel involved, accessibility of diseased bowel), others are within his/her control. Examples of such include quick identification of the lesion and rapid surgical decision making. Whether it is possible to reduce the length of bowel resected in specific cases needs further investigation. At present, intestinal viability following strangulation is assessed subjectively. The advent of more objective measures of intestinal viability may allow resection to be minimised, or avoided altogether, in order to maximise the probability of long-term survival.

The influence of professional personnel on the postoperative survival of our colic cases was one that the investigators were keen to assess. However, we were also keen to perform the analysis in a non divisive manner in order to prevent direct comparisons between individuals. This was achieved by investigating the influence of our personnel as a random effect term in the model. Therefore a summary statistic, representing the contribution to residual variation (i.e. variation not explained by the terms in the model) made by professional personnel, was generated. This variation was small in magnitude and no further analysis was therefore necessary. Had this model suggested that differences in personnel made a major contribution to variation in survival, then the investigators would have been justified in exploring further. Freeman *et al.* (2000) reported a reduced probability of survival in horses operated on by surgeons with experience of less than 9 colic operations. Our database included 3 surgeons that met this criterion but no effect was demonstrated. This may reflect a genuine lack of effect or may be due to low statistical power. The authors suggest that random effects modelling is an effective screening method for conducting non divisive surgical audit amongst groups of professional personnel. If a significant effect is identified, further analysis might be justified to identify individuals associated with worse postoperative prognosis.

Surgery, involving both human and animal patients, is notoriously difficult to evaluate scientifically. Surgeons have traditionally taken an anecdotal approach to evaluating their own performance, with case series predominating (Horton 1996). A further problem associated with evaluation of surgical success is the artificial categorisation of follow-up as 'during hospitalisation' or 'after discharge from the hospital'. In the veterinary literature this has lead to an over-emphasis on immediate postoperative success and a failure to account for mortality or postoperative complications that occurred after discharge. It is hoped that further prospective studies such as ours, with rigorous monitoring of animals after discharge, will generate high quality data about a range of surgical procedures. This will allow accurate descriptions of postoperative progress to be made and the development of models describing the risk of survival and the risk of developing postoperative complications. The results described in this study relate specifically to one hospital, they should not be taken as representative of all equine hospitals.

Variables not recorded in this study may also be of prognostic value. Differences in prognosis and in significant prognostic variables may arise from different populations of horse, different professional personnel and different surgical procedures.

This study has described the relationship between certain continuous variables and long-term postoperative survival following recovery from colic surgery. It has highlighted the association of epiploic foramen entrapment with reduced long-term survival and has described a multivariable model for postoperative survival. The influence of professional personnel (referring clinician, anaesthetist and surgeon) on the probability of long-term survival was found to be small.

Acknowledegements

This study is funded entirely by The Home of Rest for Horses. The authors thank their anaesthetist colleagues and residents in the Phillip Leverhulme Large Animal Hospital for their assistance with data collection. The willing co-operation of referring veterinary surgeons and horse owners is gratefully acknowledged. Some of the data presented in this paper were first presented at the Annual Conference of The Society for Veterinary Epidemiology and Preventive Medicine, 2001.

Manufacturer's address

[1]Insightful Corporation, Seattle, USA.

References

Aalen, O.O. (1988) Heterogeneity in survival analysis. *Stats. in Med.* **7**, 1121-1137.

Anon (2001) Penalised Cox models. In: *S-Plus 6 Guide to Statistics*. **2**. Insightful Corporation, Washington. pp 373-383.

Blikslager, A.T. and Roberts, M.C. (1995) Accuracy of clinicians in predicting site and type of lesion as well as outcome in horses with colic. *J. Am. vet. med. Ass.* **207**, 1444-1447.

Freeman, D.E., Hammock, P., Baker, G.J., Goetz, Foreman, J.H., Schaeffer, D.J., Richter, R.-A., Inoue, O. and Magid, J.H. (2000) Short- and long-term survival and prevalence of postoperative ileus after small intestinal surgery in the horse. *Equine vet. J., Suppl.* **32**, 42-51.

Furr, M.O., Lessard, P. and White, N.A. (1995) Development of a colic severity score for predicting the outcome of equine colic. *Vet Surg.* **24**, 97-101.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*, Chapman and Hall, London.

Horton, R. (1996) Surgical research or comic opera: questions, but few answers. *Lancet* **347**, 984-985.

Orsini, J.A., Elser, A.H., Galligan, D.T., Donawick, W.J. and Kronfeld, D.S. (1988) Prognostic index for acute abdominal crisis (colic) in horses. *Am. J. vet. Res.* **49**, 1969-1971.

Parry, B.W., Anderson, G.A. and Gay, C.C. (1983) Prognosis in equine colic: a comparative study of variables used to assess individual cases. *Equine vet. J.* **15**, 211-215.

Pascoe, P.J., Ducharme, N.G., Ducharme, G.R. and Lumsden, J.H. (1990) A computer-derived protocol using recursive partitioning to aid in estimating prognosis of horses with abdominal pain in referral hospitals. *Can. J. vet. Res.* **54**, 373-378.

Proudman, C.J., Smith, J.E., Edwards, G.B. and French, N.P. (2002) Long-term survival of equine surgical colic cases. Part 1: Mortality and morbidity. *Equine vet. J.* **34**, 432-437.

Puotunen-Reinert, A. (1986) Study of variables commonly used in examination of equine colic cases to assess prognostic value. *Equine vet. J.* **18**, 275-277.

Reeves, M.J. and Curtis, C.R. (1989) "By the seat of your pants" or multivariable predictive modelling. *Equine vet. J.* **21**, 83-84.

Reeves, M.J., Curtis, C.R., Salman, M.D. and Hilbert, B.J. (1989) Prognosis in equine colic patients using multivariable analysis. *Can. J. vet. Res.* **53**, 87-94.

Reeves, M.J., Curtis, C.R., Salman, M.D. Reif, J.S. and Stashak, T.S. (1990) A multivariable model for equine colic patients. *Prev. vet. Med.* **9**, 241-257.

Reeves, M.J., Curtis, C.R., Salman, M.D., Stashak, T.S. and Reif, J.S. (1992) Validation of logistic regression models used in the assessment of prognosis and the need for surgery in equine colic patients. *Prev. vet. Med.* **13**, 155-172.

Therneau, T.M. (1994) *A Package for Survival Analysis in S*, Technical Report, Mayo Clinic.

Therneau, T.M. and Grambsh, P.M. (1998) Penalised models. In: *Modeling Survival Data*, Springer-Verlag, New York. pp 120-126.

Thoefner, M.B., Ersboll, A.K., Jensen, A.L. and Hesselholt, M. (2001) Factor analysis of the interrelationships between clinical variables in horses with colic. *Prev. vet. Med.* **48**, 201-214

Received for publication: 17.10.01
Accepted: 7.2.02

Available online at www.sciencedirect.com
SCIENCE DIRECT®
The Veterinary Journal 171 (2006) 89–97
The Veterinary Journal
www.elsevier.com/locate/tvjl

Pre-operative and anaesthesia-related risk factors for mortality in equine colic cases

C.J. Proudman [a,*], A.H.A. Dugdale [b], J.M. Senior [b], G.B. Edwards [a], J.E. Smith [a], M.L. Leuwer [b], N.P. French [a]

[a] *Faculty of Veterinary Science, University of Liverpool, Leahurst, Neston, Wirral CH64 7TE, UK*
[b] *Faculty of Medicine, Department of Anaesthesia, The Duncan Building, Daulby Street, Liverpool L69 3GA, UK*

Accepted 24 September 2004

Abstract

Mortality rates for horses that have undergone emergency abdominal surgery are higher than for other procedures. Here, multivariable modelling of data from 774 surgical colic cases is used to identify pre-operative and anaesthesia-related variables associated with intra- and post-operative mortality.

Intra-operative mortality was significantly ($P < 0.05$), and positively associated with heart rate and packed cell volume (PCV) at admission, and negatively associated with the severity of pain. Post-operative mortality increased with increasing age and PCV at admission. Draught horses, Thoroughbreds and Thoroughbred-cross horses carried a significantly worse prognosis. We detected a small but significant variability in the risk of intra-operative death amongst referring veterinary surgeons. Different anaesthetic induction agents, inhalation maintenance agents and the use, or not, of intermittent positive pressure ventilation had no significant effect on risk of death. We conclude that cardiovascular compromise, level of pain, age, and breed are all associated with the risk of mortality in equine surgical colic cases.

Keywords: Horse; Anaesthesia; Colic surgery; Survival; Multivariable models

1. Introduction

Anaesthesia of healthy horses for elective surgery carries a greater risk of mortality than human anaesthesia or anaesthesia of other companion animals (Clarke and Hall, 1990; Johnston et al., 1995; Jones, 2001a; Lunn and Mushin, 1982). Surgery for equine colic carries an even higher risk of peri-operative mortality (Johnston et al., 1995; Johnston et al., 2002; Mee et al., 1998a,b).

The largest study of mortality rates in equine anaesthesia reported a mortality prevalence of 0.9% from 35,978 non-colic anaesthetics, and 8.0% from 5330 anaesthetics for colic cases (Johnston et al., 2002). A retrospective study carried out at our hospital previously reported a mortality rate for colic cases of 5.4% from 635 cases (Mee et al., 1998b). Both studies used different inclusion/exclusion criteria making comparisons difficult. Neither study attempted to identify risk factors for mortality in colic cases. Here we use data from horses that have undergone general anaesthesia for colic surgery in multivariable models to determine risk factors for intra-operative and post-operative mortality. In particular, we focus on the influence of pre-operative and anaesthesia-related variables.

* Corresponding author. Tel.: +44 151 6041; fax: +44 151 6048.
E-mail address: c.j.proudman@liv.ac.uk (C.J. Proudman).

1090-0233/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tvjl.2004.09.005

2. Materials and methods

Between March 1998 and the end of May 2003, the clinical details of horses admitted to the Philip Leverhulme Large Animal Hospital, University of Liverpool, for treatment of acute abdominal pain (''colic'') were recorded on a computer database. Analysis of data from the first 341 horses recruited has been reported previously (Proudman et al., 2002a,b). For inclusion in the study horses had to be demonstrating behavioural signs of abdominal pain, the cause of which was confirmed as gastrointestinal in origin at laparotomy.

Data were extracted from clinical records. This study focussed on pre-operative and anaesthesia-related variables which included historical details e.g. signalment, duration of colic signs and treatment prior to referral; pre-operative clinical findings e.g. degree of pain, heart rate, packed cell volume (PCV); and anaesthesia-related variables e.g. induction protocol used, inhalation agent used, use of intermittent positive pressure ventilation, use of hypertonic saline. The choice of anaesthetic technique, use of hypertonic saline or other fluids and use of IPPV were all matters of individual clinical preference.

The post-operative progress of horses that recovered from anaesthesia during this period was recorded on the database. Long-term survival data were acquired by owner-initiated reporting and by telephone questionnaires administered every three months for the first year following discharge and every six months thereafter (Proudman et al., 2002a). Owners were questioned in order to determine precise dates of events of interest (e.g. death of the horse, change of owner). This allowed event-time data to be recorded for use in survival analysis.

Failure time or outcome of interest was often euthanasia rather than death. Euthanasia was usually performed when (a) the horse's condition was such that the treating veterinary surgeon considered that there was no reasonable hope of the animal recovering, or (b) the horse's clinical condition demanded further surgical intervention that was declined by the owner. In most cases this resulted in a horse undergoing euthanasia shortly before it would have died. Throughout this paper the term ''death'' is used although in most cases death was the result of euthanasia.

2.1. Data analysis

Two distinct risk periods were examined: (1) The intra-operative period, from administration of anaesthetic pre-medication, to recovery from anaesthesia indicated by the patient walking out of the anaesthetic recovery box; (2) the post-operative period which commenced when the patient walked from the recovery box.

Hypotheses about the influence of anaesthetic induction agents and inhalation agents on intra-operative mortality of colic cases were evaluated. The functional form of relationships between continuous variables and mortality was evaluated using generalised additive models (Hastie and Tibshirani, 1990). Variables showing evidence of association with the outcome ($P < 0.20$) were carried forward for evaluation in multivariable generalised linear (logistic regression) models. If the nature of the association between variable and outcome was significantly non-linear, continuous variables were centred (each value subtracted from the mean) and polynomial transformations considered for inclusion in models. A final model was built using backward elimination procedures. Variables remained in models if they significantly reduced the residual deviance (likelihood ratio chi-squared statistic $P < 0.05$).

Post-operative survival time was modelled using Cox proportional hazards models. Kaplan–Meier plots of survival time were generated and differences in mortality rate were tested for significance in univariable Cox proportional hazards models. The functional form of relationships between continuous variables and post-operative mortality was modelled with a penalised Cox regression model that fits non-parametric functions (p-spline smoothers) to the data (Therneau and Grambsh, 1998). Categorical and continuous variables were evaluated for inclusion in the final multivariable model using a backwards elimination procedure and a likelihood ratio test statistic (LRTS) critical probability of $P < 0.05$. Cases anaesthetised between the start of the study and the end of September 1999 were part of a randomised clinical trial of isoflurane and halothane (CEPEF 3, Johnston et al., 2004). Colic cases anaesthetised during this period were treated as a separate subset as they were randomly allocated either halothane or isoflurane for the maintenance of anaesthesia. The fit of our survival model was assessed by global likelihood ratio test statistic P-value and the assumption of proportional hazards evaluated by examination of Schoenfeld residuals (Therneau and Grambsh, 1998).

The influence of anaesthetist, surgeon and referring veterinary surgeon on the risk of intra-operative and post-operative death was explored using random effects (frailty) terms in the final, fixed effects model (Therneau and Grambsh, 1998). Each random effects term in the model estimates the degree of variability attributable to differences in anaesthetist, surgeon or referring veterinary surgeon but without identifying individuals or making direct comparisons between individuals.

3. Results

Data were available from 774 horses that underwent general anaesthesia for exploratory laparotomy because

Table 1
Univariable analysis of pre-operative and anaesthesia-related variables potentially associated with intra-operative death (including euthanasia) in 774 horses undergoing exploratory laparotomy for acute abdominal pain

	Intra-operative death		Chi-squared
	No	Yes	*P*-value
Age (years)			
0–4	153	16	0.66
5–9	169	20	
0–12	103	18	
13–16	126	15	
>16	135	19	
Breed			
Pony	139	27	0.28
Thoroughbred/Tb cross	281	28	
Cob	69	5	
Draught	21	4	
Warmblood	43	6	
Arab	36	3	
Other	46	7	
Gender			
Female	284	31	0.39
Male (neutered)	369	54	
Male (entire)	33	3	
Hypertonic saline			
No	572	52	<0.001
Yes	114	36	
Induction agents			
GGE + thiopentone	272	44	<0.001
Alpha 2 agonist + Diazepam + ketamine	367	30	
Other	42	13	
Inhalation agent			
Halothane	257	26	0.34
Isoflurane	428	55	
Pain score			
None/Mild	208	43	<0.001
Moderate	347	33	
Severe/violent	123	10	
Heart rate (bpm)			
27–48	286	19	<0.001
49–69	209	17	
70–90	130	26	
90–111	40	19	
>111	11	2	
Packed cell volume (L/L)			
<0.35	210	16	<0.001
0.35–0.39	138	7	
0.40–0.46	187	11	
>0.46	124	44	
Duration of colic prior to surgery (h)			
<11	165	14	0.047
11–17	151	14	
18–26	156	20	
>26	137	27	

of acute abdominal pain. A total of 88 horses died or underwent euthanasia intra-operatively. A further 49 horses diagnosed with equine grass sickness (a usually fatal dysautonomia) were excluded from further analyses, leaving 637 horses that recovered from general anaesthesia and were subjected to long-term follow-up post-operatively. In excess of 1100 horse years of survival were documented.

3.1. Intra-operative mortality

Table 1 lists some of the variables that were screened for univariable association with intra-operative death. The following variables were not significantly associated with outcome: pre-operative total protein, peritoneal fluid colour, presence of nasogastric reflux, or laparotomy diagnosis (e.g. colon torsion, pedunculated lipoma strangulation). The categorical variable "resection (yes/no)" was highly correlated with outcome so was not considered further. A generalised additive model of four continuous variables (age, heart rate, PCV and duration of colic prior to surgery) was used to generate the plots shown in Fig. 1, illustrating the form of the relationship between each variable and the risk of intra-operative death. The variable "PCV" demonstrated significant non-linearity ($P < 0.001$) suggesting that a quadratic term may fit the model best. The following variables were carried forward into a multivariable model: Induction agent, duration prior to surgery, pain score, PCV, $(\text{PCV centred})^2$, HR, and use of hypertonic saline. Variables that remained significantly associated with outcome are listed (with coefficient values) in Table 2. $(\text{PCV centred})^2$ fitted the final model better than the non-transformed variable PCV.

3.2. Post-operative mortality

The continuous variables age and PCV showed a linear association with the rate of mortality (Fig. 2) and were carried forward for evaluation in the final multivariable model. Plasma total protein at admission showed no evidence of association and was not considered further. The univariable association between the following variables and post-operative mortality was explored with Cox proportional hazards models: Age, breed, heart rate at admission, PCV at admission, induction agent, pain score, duration of colic prior to surgery, inhalation agent, IPPV and gender. The first five variables listed showed some evidence of association so they were evaluated in a multivariable model. No association was found between the use of IPPV and post-operative mortality (univariable Cox proportional hazards model LRTS $P = 0.3$). Significant variables in the final model were age, breed and PCV at admission (Table 2). Schoenfeld residual plots indicated that the hazard associated with PCV did decline with time ($P < 0.001$), but the hazards attributable to age and breed did not.

The influence of inhalation anaesthetic agent on survival for the first 100 days post-operatively is illustrated by the Kaplan–Meier plots in Fig. 3. When the whole dataset was represented, horses receiving isoflurane appear to have a lower probability of survival, due largely

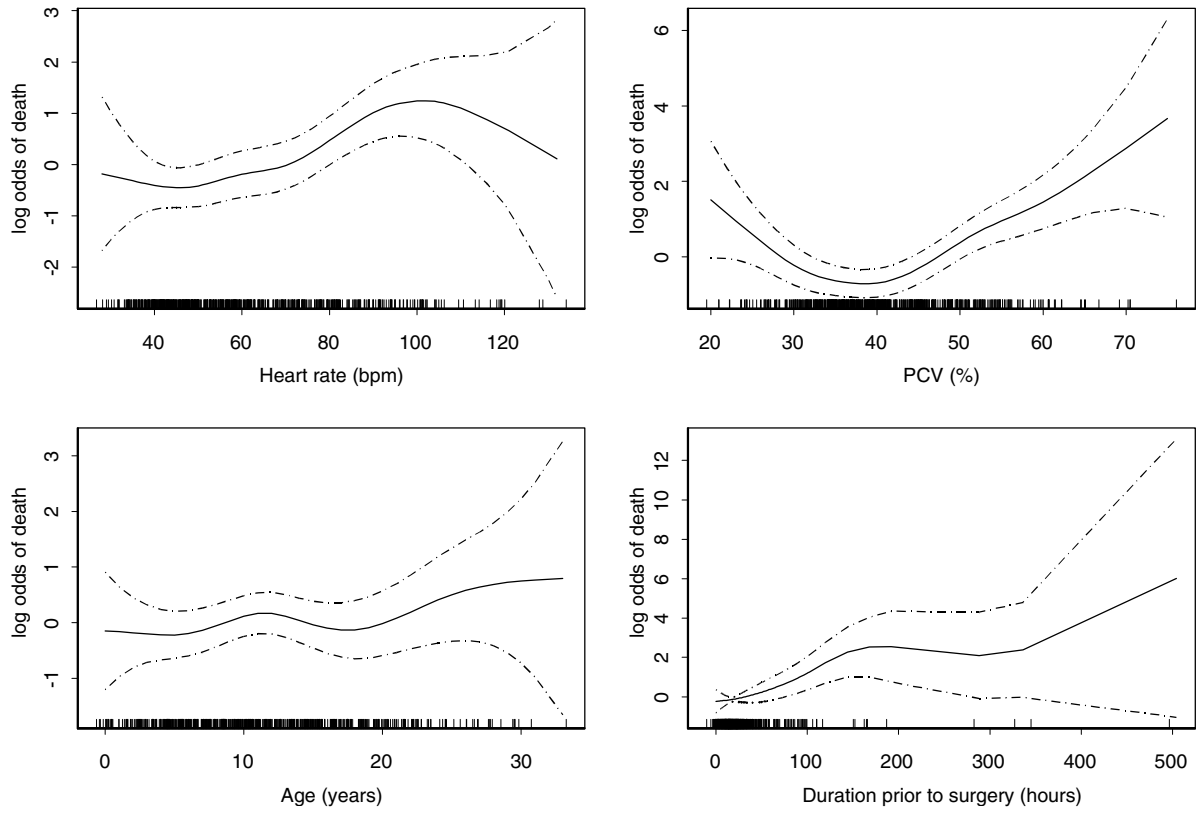


Fig. 1. Graphs from a four-variable generalised additive model illustrating the relationship between continuous variables and the risk of intra-operative death. Rug plot on the *X* axis indicates number of observations, dashed lines indicate 95% confidence intervals.

to an increased rate of death between 10 and 25 days (Fig. 3(a)). However, when only horses that were randomly allocated inhalation anaesthetic agent were considered (Fig. 3(b)), there is no significant difference in mortality (univariable Cox proportional hazards model LRTS $P = 0.7$).

3.3. Influence of professional personnel on outcome

The inclusion of random effects terms in the final fixed effects model was used to explore the influence of professional staff on intra-operative and post-operative death. Variance estimates for these effects are displayed

Table 2
Variables significant in the final, fixed effects, multivariable models for risk of death in equine colic patients

	Coefficient	SE	OR/HR	95% CI	*P*-value
Intra-operative death model					
PCV centred	0.02	0.02	1.017	0.98–1.05	<0.001
(PCV centred)^2	0.005	0.001	1.005	1.003–1.006	<0.001
Heart rate	0.02	0.01	1.02	1.00–1.04	0.006
Pain score None/mild	Referent				
Moderate	−0.70	0.28	0.50	0.29–0.86	0.019
Severe	−0.79	0.40	0.45	0.21–0.99	
Post-operative death model[a]					
PCV	0.05	0.01	1.05	1.03–1.07	<0.001
Age (per year)	0.04	0.01	1.04	1.01–1.06	<0.01
Breed					
Pony	Referent				
Tb/Tb cross	0.52	0.22	1.70	1.09–2.59	0.02
Cob	0.41	0.31	1.51	0.82–2.76	0.18
Draught	1.70	0.42	5.45	2.40–12.47	<0.001
WmBl	0.43	0.38	1.53	0.73–3.24	0.26
Arab/Arab cross	0.39	0.35	1.48	0.74–2.93	0.26
Other	−0.37	0.43	0.69	0.29–1.60	0.39

[a] Global LRTS < 0.001.

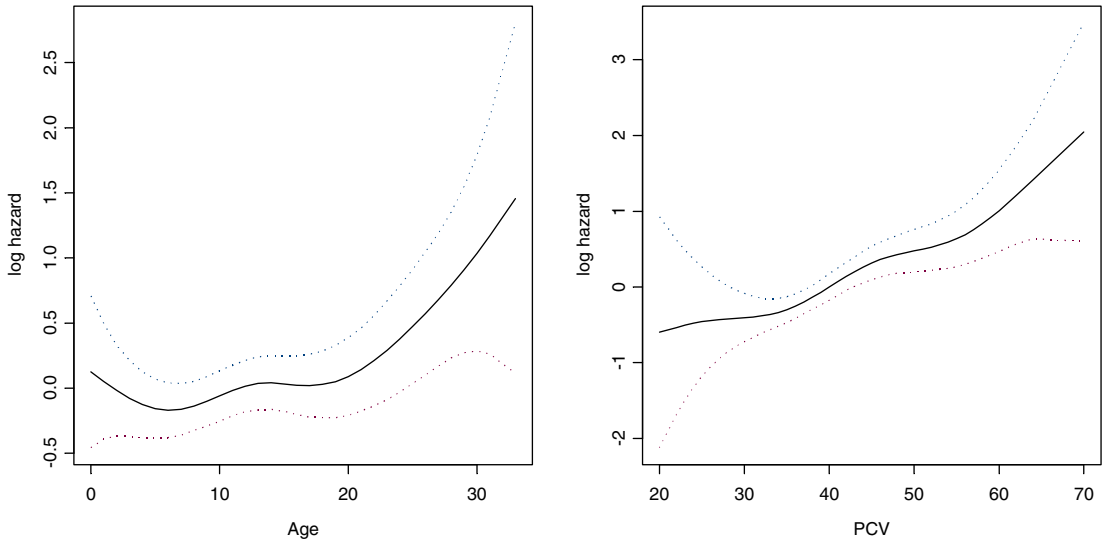


Fig. 2. Plots of p-spline smoothers for age and PCV illustrating the functional form of their relationship with the rate of post-operative mortality. Neither age nor PCV had a significantly non-linear relationship with mortality ($P = 0.18$ and 0.53 respectively).

in Table 3. Referring veterinary surgeon is significantly associated with intra-operative death. Similar results were obtained with random effects terms in an intercept-only model with no fixed effects. These results suggest that some referring vets are associated with surgical colic cases that have a higher risk of intra-operative death.

4. Discussion

A number of studies have reported various referral-, patient- and surgery-related factors to be associated with a poorer prognosis for survival following equine colic surgery (Freeman et al., 2000; Morton and Blikslager, 2002; Parry et al., 1983a,b; Phillips and Walmsley, 1993; Proudman et al., 2002b; Puotunen-Reinert, 1986; Pascoe et al., 1983). In particular these studies have highlighted the increased risk of death associated with cardiovascular compromise (resulting from endotoxaemia), duration of surgery and nature of the disease (Proudman et al., 2002a,b; Freeman et al., 2000; Morton and Blikslager, 2002; Phillips and Walmsley, 1993; Pascoe et al., 1983).

In the human literature there have been several reports documenting post-operative survival following abdominal surgery, many focusing on the elderly patient (>65 years old). These have similarly concluded that factors such as poor pre-operative ASA grade, prolonged pre-operative duration of the problem, greater emergency of the surgery and in some studies increasing age, were associated with increased risk of both intra- and post-operative complications, and mortality (Cohen et al., 1988; Cook and Day, 1998; Edwards et al., 1996; El-Haddawi et al., 2002; Pedersen et al., 1990; Tiret et al., 1988). Prolonged duration of anaesthesia, has also been shown to increase the risk of peri-operative mortality in man, but has been difficult to separate from the extent of surgical intervention required (Pedersen et al., 1990).

In our study, PCV and heart rate at presentation were associated with poorer intra-operative prognosis, and PCV was also associated with poor long-term prognosis. Both variables reflect the degree of cardiovascular compromise, predominantly due to endotoxaemia, and would affect the patient's ASA score in analogous human studies. The functional form of the relationship between PCV and risk of intra-operative mortality is non-linear (Fig. 1). This was further confirmed by the best model fit being achieved with the quadratic term $(\text{PCV centred})^2$. The parabolic form of a quadratic function is consistent with the clinical observation that there is an ''optimum'' PCV for survival. Values of this variable higher than the optimum or lower than the optimum result in a poorer prognosis for survival.

It is not current practice in our hospital to attempt extensive pre-surgical cardiovascular resuscitation. Although univariable analysis indicates that the use of hypertonic saline is associated with increased intra-operative mortality, this likely reflects the fact that its use was reserved for the most hypovolaemic cases. In this study the pre-induction PCV would have differed from PCV at admission only in those horses given hypertonic saline, but was not recorded. Data on blood pressure and intra-operative vasopressor use were collected during this study. Interpretation of the relationship between these variables and survival is complex and will be the subject of a future publication.

Our results raise the issue of whether fluid resuscitation to an optimum PCV should always be performed

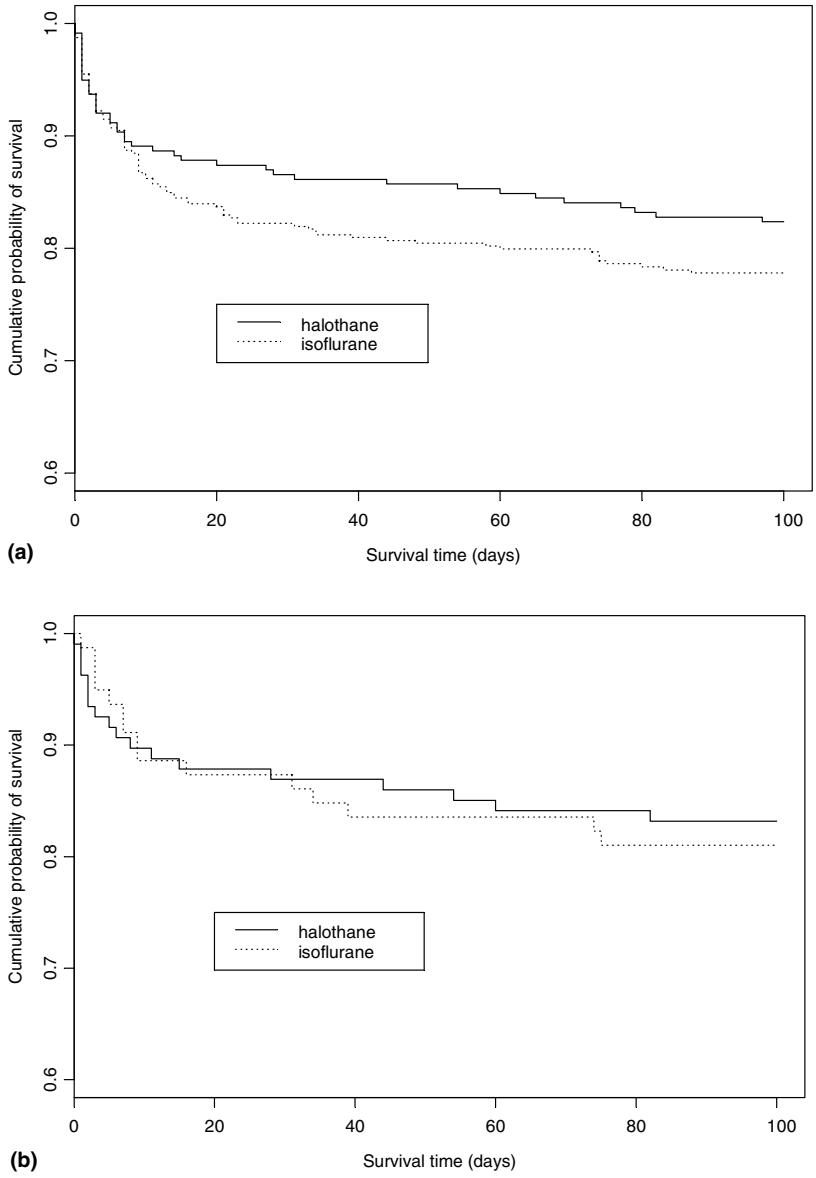


Fig. 3. (a) Probability of survival of 637 horses recovering from anaesthesia following colic surgery. All cases represented. (b) Probability of survival of 186 horses recovering from anaesthesia following colic surgery where inhalation agent was randomly allocated.

pre-operatively? Although hypertonic saline can dramatically reduce PCV and improve cardiovascular status, its effects are very transient (approximately 30 min) in the horse (Bertone et al., 1990; Bertone and Shoemaker, 1992). The extensive use of pre-operative fluid therapy also carries the risk of sequestration of fluid into obstructed bowel and into bowel wall (Chan et al., 1983; Allen et al., 1986; Prien et al., 1990). The use of colloids for pre-operative fluid therapy, where relatively small volumes can be administered with longer lasting effects and improved intra-vascular retention, warrants further investigation.

Table 3
Variance estimates for random effects terms in final fixed effects models to assess the influence of professional personnel on outcome of equine colic surgery

	Variance estimate	*P*-value
Intra-operative death		
Anaesthetist	−0.044	0.18
Surgeon	0.061	0.24
Referring veterinary surgeon	0.004	0.01
Post-operative death		
Anaesthetist	<0.001	0.66
Surgeon	0.023	0.20
Referring veterinary surgeon	<0.001	0.91

Our study did reveal an association between increasing age and the risk of post-operative mortality, but not intra-operative mortality. This association was still significant after adjusting for duration of colic and pain score. Two previous studies of equine colic cases found no such association (Proudman et al., 2002b; Phillips and Walmsley, 1993). In contrast, following non-colic

surgeries, mortality was found to be highest in very young horses, lowest for young adults (age range 12 months to 5 years old), and risk of mortality then gradually increased with increasing age (Johnston et al., 2002).

Human studies have also disagreed about the importance of age, with some studies demonstrating a clear effect (Cohen et al., 1988; Cook and Day, 1998; Edwards et al., 1996; Tiret et al., 1988), and others not (El-Haddawi et al., 2002). However, many of these studies were performed in an already geriatric population (Cook and Day, 1998; Edwards et al., 1996; El-Haddawi et al., 2002). Increasing age is accompanied by both physiological changes that result in decreased functional reserves and the increasing likelihood of coexisting systemic diseases (Jones, 2001b; Severn, 2001). Both of these may contribute to a worse ASA status, and also to the overall risk of mortality.

Thoroughbreds, Thoroughbred crosses and draught horses have a higher rate of post-operative mortality than other breeds. These breeds are heavier than the reference breed (pony) and it may be that larger animals are at increased risk rather than breed specific susceptibilities. Intra-operatively the larger horse breeds tend to suffer worse compression atelectasis and ventilation/perfusion mismatching than smaller Equidae, resulting in greater degrees of intra-operative hypoxaemia (Stegmann and Littlejohn, 1987; Nyman et al., 1990). In addition to body size, the abdominal shape may also play a role (Moens et al., 1995). It is therefore possible that the greater degree of hypoxaemia in larger horses might be responsible for the worse post-operative prognosis.

Although IPPV is less successful for the treatment of established hypoxaemia in horses than in man (Shawley and Mandsager, 1990), its early instigation may help to prevent severe hypoxaemia from developing (Day et al., 1995). IPPV can be employed to treat hypercapnia, and/or prevent its development (Gasthuys et al., 1991), and can be used to aid the intra-operative management of acid–base disturbances. Eighty-five percent of horses recovering from anaesthesia in our study were subject to IPPV but no significant difference in prognosis was detected compared to horses breathing spontaneously. The CEPEF (Confidential Enquiry into Perioperative Equine Fatalities) phase 3 study, however, suggested an association between the use of IPPV and increased risk of mortality (Johnston et al., 2004). IPPV can be detrimental to the cardiovascular system (Gasthuys et al., 1991), especially where there is pre-existing cardiovascular compromise. However, if haemodynamic monitoring is performed, and appropriate cardiovascular support provided, as in our study, these effects can be offset (Wilson and McFeely, 1991), perhaps explaining our disagreement with the CEPEF 3 study results. Johnston et al. (2004) made no mention of whether most horses subjected to IPPV were closely monitored under anaesthesia, but they did conclude that the risk of mortality was reduced where arterial blood pressure was monitored.

Our results demonstrated an increased likelihood of intra-operative mortality in horses that showed less severe signs of abdominal pain on admission to the hospital. This may reflect the extent of devitalisation of bowel (pain reduces as ischaemia becomes more advanced) and thus the severity and/or duration of vascular compromise and endotoxaemia. An alternative explanation is that horses showing moderate or severe pain generally undergo immediate surgery. Surgery may be delayed in horses showing less obvious signs of pain.

Although several studies of post-operative survival have been reported, only the CEPEF studies have attempted to separate the influence of certain anaesthetic agents on mortality (Johnston et al., 1995, 2002, 2004). CEPEF phase 3, was a randomised comparison trial of isoflurane and halothane and showed no overall difference in mortality between these two agents (Johnston et al., 2004). However, there was a decreased incidence of death from "cardiac causes" in horses anaesthetised with isoflurane compared with halothane, which was balanced by an increased incidence of death due to other (predominantly central nervous system) causes. Data from CEPEF 3 suggested that isoflurane might be a more appropriate choice in animals with pre-existing myocardial depression. Our data, which includes a high proportion of horses with myocardial depression due to endotoxaemia, offers no support for the hypothesis that isoflurane is safer under such circumstances.

The same study (CEPEF 3) failed to detect any differences in mortality between different injectable anaesthetic induction agents but much variation in protocols between hospitals was observed. Our results are consistent with this finding, probably because the impact of induction agents is greatly reduced if anaesthetic maintenance is conducted with inhalation agents for any length of time (Taylor and Young, 1993).

We have previously reported variables significantly associated with long-term survival of equine colic cases from this population (Proudman et al., 2002b), but not variables associated with intra-operative mortality. The focus of the present study was the influence of anaesthetic-related variables on outcome so it was appropriate to consider both intra- and post-operative mortality. The model previously reported to describe post-operative mortality included the variables epiploic foramen entrapment (yes/no), duration of surgery and resection length. Surgical and post-operative management variables were not considered in the present study although epiploic foramen entrapment and duration of surgery remained significantly associated with post-operative outcome if added to the final model in Table 2.

We conclude that cardiovascular parameters (PCV and heart rate) and degree of pain are associated with

intra-operative death in our equine colic population. Reduced long-term prognosis is associated with increasing age, larger breeds of horse, and elevated PCV on admission to the hospital. We found no association between short-term or long-term prognosis and choice of anaesthetic induction protocol, inhalation maintenance agent or the use of IPPV.

Acknowledgements

We thank the many veterinary surgeons involved in managing the cases in this study, and horse owners and veterinary surgeons that provided information on long-term survival. The Liverpool colic survival study is funded by The Home of Rest for Horses and PetPlan Charitable Trust.

References

Allen, D., Kvietys, P.R., Granger, D.N., 1986. Crystalloids versus colloids: Implications in fluid therapy of dogs with intestinal obstruction. American Journal of Veterinary Research 47, 1751–1755.

Bertone, J.J., Shoemaker, K.E., 1992. Effects of hypertonic and isotonic saline solutions on plasma constituents of conscious horses. American Journal of Veterinary Research 53, 1844–1849.

Bertone, J.J., Gossett, K.A., Shoemaker, K.E., Bertone, A.L., Schneiter, H.L., 1990. Effect of hypertonic vs. isotonic saline solution on response to sublethal *Escherichia coli* endotoxaemia in horses. American Journal of Veterinary Research 51, 999–1007.

Chan, S.T., Kapadia, C.R., Johnson, A.W., Radcliffe, A.G., Dudley, H.A., 1983. Extracellular fluid volume expansion and third space sequestration at the site of small bowel anastomoses. British Journal of Surgery 70, 36–39.

Clarke, K.C., Hall, L.W., 1990. A survey of anaesthesia in small animal practice. AVA/BSAVA report. Journal of Veterinary Anaesthesia 17, 4–10.

Cohen, M.M., Duncan, P.G., Tate, R.B., 1988. Does anaesthesia contribute to operative mortality. Journal of the American Medical Association 260, 2859–2863.

Cook, T.M., Day, C.J.E., 1998. Hospital mortality after urgent and emergency laparotomy in patients aged 65 yr and over. Risk and prediction of risk using multiple logistic regression analysis. British Journal of Anaesthesia 80, 776–781.

Day, T.K., Gaynor, J.S., Muir, W.W., Bednarski, R.M., Mason, D.E., 1995. Blood gas values during intermittent positive pressure ventilation and spontaneous ventilation in 160 anesthetized horses positioned in lateral or dorsal recumbency. Veterinary Surgery 24, 266–276.

Edwards, A.E., Seymour, D.G., McCarthy, J.M., Crumplin, M.K.H., 1996. A 5-year survival study of general surgical patients aged 65 years and over. Anaesthesia 51, 3–10.

El-Haddawi, F., Abu-Zidan, F.M., Jones, W., 2002. Factors affecting surgical outcome in the elderly at Auckland hospital. Australian and New Zealand Journal of Surgery 72, 537–541.

Freeman, D.E., Hammock, P., Baker, G.J., Goetz, T., Foreman, T.H., Schaeffer, D.J., Richter, R.A., Inoue, O., Magid, J.H., 2000. Short- and long-term survival and prevalence of post-operative ileus after small intestinal surgery in the horse. Equine Veterinary Journal 32, 42–51.

Gasthuys, F., De Moor, A., Parmenter, D., 1991. Haemodynamic effects of change in position and respiration mode during a standard halothane anaesthesia in ponies. Journal of Veterinary Medicine Series A 38, 203–211.

Hastie, T., Tibshirani, R., 1990. Generalised Additive Models. Chapman and Hall, London.

Johnston, G.M., Taylor, P.M., Holmes, M.A., Wood, J.L.N., 1995. Confidential enquiry into perioperative equine fatalities (CEPEF-1): preliminary results. Equine Veterinary Journal 27, 193–200.

Johnston, G.M., Eastment, J.K., Wood, J.L.N., Taylor, P.M., 2002. The confidential enquiry into perioperative equine fatalities (CEPEF): mortality results of Phases 1 and 2. Veterinary Anaesthesia and Analgesia 29, 159–170.

Johnston, G.M., Eastment, J.K., Taylor, P.M., Wood, J.L.N., 2004. Is isoflurane safer than halothane in equine anaesthesia. Results from a prospective multicentre randomised controlled trial. Equine Veterinary Journal 36, 64–71.

Jones, R.M., 2001a. Anaesthesia in old age. Editorial. Anaesthesia 44, 377–378.

Jones, R.S., 2001b. Comparative mortality in anaesthesia. British Journal of Anaesthesia 87, 813–815.

Lunn, J.N., Mushin, W.W., 1982. Mortality associated with anaesthesia. Anaesthesia 37, 856.

Mee, A.M., Cripps, P.J., Jones, R.S., 1998a. A retrospective study of mortality associated with general anaesthesia in horses: elective procedures. Veterinary Record 142, 275–276.

Mee, A.M., Cripps, P.J., Jones, R.S., 1998b. A retrospective study of mortality associated with general anaesthesia in horses: emergency procedures. Veterinary Record 142, 307–309.

Moens, Y., Lagerweij, E., Gootjes, P., Poortman, J., 1995. Distribution of inspired gas to each lung in the anaesthetised horse and influence of body shape. Equine Veterinary Journal 27, 110–116.

Morton, A.J., Blikslager, A.T., 2002. Surgical and postoperative factors influencing short-term survival of horses following small intestinal resection: 92 cases (1994–2001). Equine Veterinary Journal 34, 450–454.

Nyman, G., Funkquist, B., Kvart, C., Frostell, C., Tokics, L., Strandberg, A., Lundquist, H., Lundh, B., Brismar, B., Hedenstierna, G., 1990. Atelectasis causes gas exchange impairment in the anaesthetised horse. Equine Veterinary Journal 22, 317–324.

Parry, B.W., Anderson, G.A., Gay, C.C., 1983a. Prognosis in equine colic: A comparative study of variables used to assess individual cases. Equine Veterinary Journal 15, 211–215.

Parry, B.W., Anderson, G.A., Gay, C.C., 1983b. Prognosis in equine colic: A study of individual variables used in case assessment. Equine Veterinary Journal 15, 337–344.

Pascoe, P.J., McDonnell, W.N., Trim, C.M., Van Gorder, J., 1983. Mortality rates and associated factors in equine colic operations – a retrospective study of 341 operations. Canadian Journal of Veterinary Research 24, 76–85.

Pedersen, T., Eliasen, K., Henriksen, E., 1990. A prospective study of mortality associated with anaesthesia and surgery: risk indicators of mortality in hospital. Acta Anaesthesiologica Scandinavica 34, 176–182.

Phillips, T.J., Walmsley, J.P., 1993. Retrospective analysis of the results of 151 exploratory laparotomies in horses with gastrointestinal disease. Equine Veterinary Journal 25, 427–431.

Prien, T., Backhaus, N., Pelster, F., Pircher, W., Bunte, H., Lawin, P., 1990. Effect of intraoperative fluid administration and colloid osmotic pressure on the formation of intestinal edema during gastrointestinal surgery. Journal of Clinical Anesthesia 2, 317–323.

Proudman, C.J., Smith, J.E., Edwards, G.B., French, N.P., 2002a. Long term survival of equine surgical colic cases. Part 1: Patterns of mortality and morbidity. Equine Veterinary Journal 34, 432–437.

Proudman, C.J., Smith, J.E., Edwards, G.B., French, N.P., 2002b. Long-term survival of equine surgical colic cases. Part 2: Modelling postoperative survival. Equine Veterinary Journal 34, 438–443.

Puotunen-Reinert, A., 1986. Study of variables commonly used on examination of equine colic cases to assess prognostic value. Equine Veterinary Journal 18, 275–277.

Severn, A.M., 2001. Time to light the grey touchpaper! The challenge of anaesthesia in the elderly. Editorial. British Journal of Anaesthesia 87, 533–536.

Shawley, R.V., Mandsager, R.E., 1990. Clinical use of positive-pressure ventilation in the horse. The Veterinary Clinics of North America; Equine Practice 6, 575–585.

Stegmann, G.F., Littlejohn, A., 1987. The effect of lateral and dorsal recumbency on cardiopulmonary function in the anaesthetised horse. Journal of the South African Veterinary Association 58, 21–27.

Taylor, P.M., Young, S.S., 1993. Does the induction agent affect the course of halothane anaesthesia in horses. Journal of Veterinary Anaesthesia 20, 84–91.

Therneau, T.M., Grambsh, P.M., 1998. Modelling Survival Data. Springer-Verlag, New York.

Tiret, L., Hatton, F., Desmonts, J.M., Vourc'h, G., 1988. Prediction of outcome of anaesthesia in patients over 40 years: a multifactorial risk index. Statistics in Medicine 7, 947–954.

Wilson, D.V., McFeely, A.M., 1991. Positive end-expiratory pressure during colic surgery in horses: 74 cases (1986–88). Journal of the American Veterinary Medical Association 199, 917–921.

ELSEVIER

Preventive Veterinary Medicine 61 (2003) 157–170

PREVENTIVE VETERINARY MEDICINE

www.elsevier.com/locate/prevetmed

Spatio-temporal epidemiology of foot-and-mouth disease in two counties of Great Britain in 2001

J.W. Wilesmith [a,1], M.A. Stevenson [b,*], C.B. King [c], R.S. Morris [b]

[a] *Epidemiology Department, Veterinary Laboratories Agency, New Haw, Addlestone, Surrey KT15 3NB, UK*
[b] *EpiCentre, Institute of Veterinary, Animal, and Biomedical Sciences, Massey University, Private Bag 11-222, Palmerston North, New Zealand*
[c] *National Centre for Disease Investigation, P.O. Box 40742, Upper Hutt, New Zealand*

Received 17 September 2002; accepted 15 August 2003

Abstract

The spatial, temporal, and spatio-temporal features of the 2001 British foot-and-mouth disease epidemic in selected areas within the counties of Cumbria and Devon, which experienced the greatest incidence of disease, are described using hazard functions, extraction mapping and the space-time *K*-function.

In Cumbria, the hazard of foot-and-mouth disease infection peaked at 2.8% in the week commencing 8 March 2001 and farm holdings in this area continued to be identified with disease to 12 September 2001. In contrast, peak infection hazard in Devon was 0.7% in the week commencing 15 March 2001 and eradication of the disease was achieved in this area by 31 May 2001. Persistence of the disease in Cumbria was consistent with: (1) many cattle holdings infected early in the epidemic (creating a high environmental viral load), and (2) a relatively large amount of medium-to-long-distance spread of the virus associated with seasonal farming activities—compounded to some extent by the movement of people and vehicles between disaggregated farm land parcels. The interaction of disease risk in Cumbria showed that premises remained infectious for longer throughout May, June and July, consistent with delays in disease detection during this period.

Keywords: Foot-and-mouth disease; Epidemiology; Spatial epidemiology; Spatio-temporal epidemiology; Space-time *K*-function

* Corresponding author. Tel.: +64-6-3505915/3506149; fax: +64-6-3505716.
E-mail address: m.stevenson@massey.ac.nz (M.A. Stevenson).
[1] Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, University of London, Keppel Street, London WC1E 7HT, UK.

0167-5877/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.prevetmed.2003.08.002

1. Introduction

The 2001 foot-and-mouth disease (FMD) epidemic in Great Britain has been a timely reminder of the value of geo-referenced farm data in the management of animal disease outbreaks. Where many personnel are deployed to carry out control activities, accurate and up-to-date maps showing farm locations and herd disease status are key (Morris et al., 2002). In addition, geo-referenced data are essential—firstly for monitoring progress (Sanson et al., 1991) and secondly for predictive modelling of alternative control strategies (Howard and Donnelly, 2000; Ferguson et al., 2001; Kao, 2001; Keeling et al., 2001; Morris et al., 2001).

In the absence of vaccination, strategies for dealing with a FMD epidemic involve four main elements: (1) rapid slaughter of all stock on premises identified as infected, (2) pre-emptive slaughtering of stock on premises on the basis that they had been exposed to infection (termed 'pre-emptive' culling in this paper), (3) imposition of bans on the movement of stock and/or people within defined infected areas, and (4) surveillance to detect infected premises (Sanson, 1993; Radostits et al., 2000). The removal of stock on infected premises and those potentially exposed in the immediate vicinity is to limit local spread, and restrictions on animal and animal-product movement is to reduce the spread of virus across larger distances.

The magnitude of FMD epidemics vary in response to the scale of the original infection challenge, the geographical distribution of the animal population at risk and the effectiveness of control efforts during eradication. In dynamic outbreak situations (where, as a result of pre-emptive culling, the animal population-at-risk is constantly changing and control measures vary in their effectiveness and intensity of application), the spatial and temporal components of disease risk can change markedly throughout an epidemic's course. Within the epidemic of FMD in Great Britain in 2001, there were apparent differences between geographical areas in response to control measures applied. In this paper, we considered the two counties which experienced the greatest incidence of disease (Cumbria and Devon). Our first aim was to describe the spatial and temporal features of the incidence of FMD among farm holdings in these areas. The second aim was to describe the spatio–temporal interaction of infection risk and to describe changes in the components of this interaction that occurred over the course of the epidemic. Better appreciation of these aspects of an epidemic means that decisions concerning practical issues (such as pre-emptive culling radii, surveillance periods, the resources required and the appropriate effort to trace the spread of infection from infected premises) can be made with greater objectivity and modified appropriately for individual outbreaks.

2. Materials and methods

The areas investigated are shown in Fig. 1. Both were 2500 km^2 (50 km $\times$ 50 km): the first in the west of the county of Cumbria and the second in the north of the county of Devon. These areas were selected because they were regions where the behaviour of the disease (that is, the spread and ability to be controlled) showed marked differences (Gibbens et al., 2001). The period of interest was from 20 February 2001 to 30 September 2001.

Fig. 1. Map of Great Britain showing the location of the study areas described in this study: (a) Cumbria, (b) Devon.

The population of interest included all farm holdings containing at least one of the five FMD-susceptible domestic species (cattle, sheep, goats, deer or pigs) that were located within each of the defined areas. Holding-associated data were retrieved from the agricultural-census data collected by the Ministry of Agriculture, Fisheries and Food (MAFF, 2000) for the 12 months to June 2000 and subsequently revised throughout the course of the 2001 FMD epidemic. Data recorded for each holding included the county-parish-holding identifier, easting and northing coordinates of the main farm building and the number of adults present of each of the FMD-susceptible species. On the basis of adult-stock counts, a holding-level type classification was calculated as follows. Holdings where 80% of the total animal population on the day of census were dairy or beef suckler animals were designated as 'cattle' enterprise types. Holdings where 80% of the total animal population were a single species were designated as enterprises of that type. Holdings unable to be classified by this method were designated as 'mixed'.

Cases were holdings located within the boundaries of each defined study area that were declared as FMD-infected premises under the foot-and-mouth Disease Order 1983 (HMSO, 1983). These were holdings where the clinical signs of FMD were observed by a Department for Environment, Food and Rural Affairs (DEFRA) investigating officer or holdings where there was laboratory confirmation of infection within the herd or flock (typically the case

for holdings depopulated as direct contacts or holdings that were culled on suspicion of infection). Details of case holdings (county-parish-holding identifier, estimated infection date, confirmation date and cull date) were retrieved from the Disease Control System Database (Gibbens et al., 2001) and merged with the agricultural-census database.

We used three independent methods to analyse these data. Firstly, the temporal evolution of the epidemic was described using survival analyses. Secondly, the spatial distribution of infected premises was described using kernel density estimation methods. Thirdly, the spatio–temporal interaction among cases was described using the space-time *K*-function.

For the survival analyses, the outcome was the estimated FMD infection date for case holdings. Estimated infection date was recorded at the time of diagnosis, and was determined either on the basis of the history provided by the holding manager or on the estimated age of lesions observed at the time of examination. For the later case, estimated infection date was the date of examination minus the age (in days) of the oldest lesions identified minus an additional 5 days to represent an incubation period for the disease (Gibbens and Wilesmith, 2002). Holdings that were culled pre-emptively or culled on suspicion of infection but not confirmed as infected were right-censored on the date of cull. Holdings that remained free of infection throughout the period of interest were right-censored at the end of the observation period on 30 September 2001. The weekly hazard of FMD (representing the weekly probability of becoming FMD infected, given that a holding remained free of infection to at least the specified point in time) was computed using the LOCFIT library (Loader, 1999) implemented in the *R* statistical package (Ihaka and Gentleman, 1996). We compared the hazard of FMD for two groups: cattle holdings versus sheep, goat, deer, pig and mixed holdings (considered as a single group under the title of 'other'). Instantaneous hazard functions of this type allowed high-risk and low-risk periods for infection to be readily identified.

To describe the spatial pattern of infection among holdings in each area, four time periods were defined starting from 20 February 2001. These periods were thought broadly to represent the natural phases of the epidemic: the first (20 February to 28 March 2001) was the period of rapid spread of the disease; the second (29 March to 23 May 2001) a period in which there was a sharp reduction in incidence; the third (24 May to 18 July 2001) a period of relatively constant incidence, and the fourth (19 July to 30 September 2001) the period when eradication was achieved. For each period, two density surfaces representing the number of holdings per km^2 were constructed using a Gaussian-kernel smoothing function: the first (numerator) was based on all holdings identified as FMD-positive during the period and the second (denominator) on holdings that were present at the start of the period and considered at risk. The ratio of the density surface of FMD-positive holdings to the density surface of the population of holdings at risk at the start of each period provided a relief map of the distribution of the proportion of holdings per km^2 which became FMD-affected (Bithell, 1990; Lawson and Williams, 1994; Bowman and Azzalini, 1997). These plots (termed 'extraction maps' by Lawson and Williams, 1993) provided an objective means for identifying areas where there were relatively high rates of infection. Bandwidth parameters for the kernel functions (used to control for the degree of smoothing of the estimated density surface) were calculated by cross validation (Bowman and Azzalini, 1997) and were based on the holding population at risk at the start of each time interval. To account for edge effects (Lawson, 2001), details of holdings located within a 5 km guard area around the periphery

of each study area were included in each data set. Kernel density surfaces were computed on the basis of holdings located in both the study area and the guard area; only the density estimates for each defined 2500 km^2 area are reported.

The spatio–temporal interaction of infection risk was described using the space-time *K*-function (Diggle et al., 1995) implemented in the SPLANCS library (Rowlingson and Diggle, 1993; Bivand and Gebhardt, 2000) in R. Here, the respective data sets were restricted to include case holdings only. The space-time *K*-function $K(s, t)$ was calculated as the cumulative number of cases that were expected within distance s and time interval t of an arbitrarily-selected case divided by the intensity λ (the expected number of events per unit space and per unit time). Letting $K_S(s)$ define the *K*-function in space and $K_T(t)$ define the *K*-function in time, the *K*-function difference $D(s, t)$ was computed as:

$$D(s, t) = K(s, t) - K_S(s)K_T(t) \quad (1)$$

which estimates the cumulative number of cases expected within distance s and time interval t of an arbitrarily-selected case that were attributable to the interaction between space and time, scaled by λ. To facilitate comparison among time periods, we computed $D_0(s, t)$:

$$D_0(s, t) = \frac{D(s, t)}{K_S(s)K_T(t)} \quad (2)$$

which estimates, for given distance and time separations, the proportional increase in cases attributable to space-time interaction (Diggle et al., 1995). Separate space-time *K*-function analyses were conducted for each of the four time periods described using maximum distance and time separations of 10 km and 21 days, respectively. The use of a small maximum distance separation (relative to the overall dimensions of the study area) reduced the likely influence of first-order, trend effects on each *K*-function that was computed. Space-time *K*-functions were computed using details of holdings within each 2500 km^2 study area: the *K*-function implementation within SPLANCS providing a correction term for edge effects. Crude estimates of the confidence limits for $D_0(s, t)$ were determined by estimating the lower and upper limits of $D(s, t) \times \lambda$, assuming that $D(s, t) \times \lambda$ (being an estimate of count data) followed a Poisson distribution.

A formal test for the presence of space-time interaction was performed by conducting m Monte Carlo simulations in which each of the n case events were labelled with the observed n time 'markers'. A total of m estimates of $D(s, t)$ were obtained and for each simulation, the sum of $D(s, t)$ over all s and t were obtained. The sum of $D(s, t)$ for the observed data then was ranked among the empirical frequency of m sums. If the observed sum ranked *kth* largest (or smallest) the one-sided attained significance level was k/m.

3. Results

Tables 1 and 2 provide counts of holdings present in each area at the start of the study period stratified by holding type and FMD status on 30 September 2001. The density of holdings in the Cumbria study area was 0.6 km^2, approximately half that of Devon (1.3 holdings per km^2). In Cumbria, there was one holding pre-emptively culled for each holding

Table 1
Counts of holdings present in the Cumbria study area (Great Britain) on 24 February 2001 stratified by holding type classification and foot-and-mouth disease status on 30 September 2001

Type	FMD-positive	FMD-negative and culled		FMD-negative	Total
		Pre-emptively	On suspicion		
Cattle	76	36	0	148	260
Other					
Deer	0	1	0	0	1
Goat	0	1	0	11	12
Mixed	11	1	0	2	14
Pig	3	2	0	6	11
Sheep	256	288	5	708	1257

diagnosed FMD-positive. In Devon, 3.2 holdings were culled pre-emptively for each holding diagnosed FMD-positive.

In Cumbria, the hazard of FMD peaked in the week commencing 8 March 2001 (Fig. 2a). During this period, the infection hazard was greater for cattle holdings (6.0%) than for other holding types (2.5%). Infection hazard rapidly declined from 28 March as additional control measures were applied. From April through to August, the hazard of FMD for cattle holdings remained static at 0.8%—whereas for other holding types there was a steady rise in hazard (reaching 1.4% by the end of July). In Devon, the peak infection hazard (0.7%) occurred in the week commencing 15 March followed by a rapid decline for both holding classifications.

There were marked changes in the spatial pattern of incident holdings in Cumbria for the four time periods. From 20 February to 28 March, the highest density of incident holdings was in an area 570 km^2 north west of (and including) the town of Penrith (Fig. 3a). The next period was characterised by incident holdings appearing further to the north west and south east of this primary focus—indicative of the infection moving into a population of susceptible holdings (Fig. 3b). In the third period, a secondary focus of incident holdings of 500 km^2 developed within an area bounded by the towns of Windermere, Penrith, and Brough (Fig. 3c). In Devon, a single, relatively high-density focus of incident holdings

Table 2
Counts of holdings present in the Devon study area (Great Britain) on 24 February 2001 stratified by holding type classification and foot-and-mouth disease status on 30 September 2001

Type	FMD-positive	FMD-negative and culled		FMD-negative	Total
		Pre-emptively	On suspicion		
Cattle	45	162	2	1024	1233
Other					
Deer	0	0	1	10	11
Goat	1	7	0	81	89
Mixed	9	4	0	34	47
Pig	2	15	0	56	73
Sheep	85	267	14	1475	1841

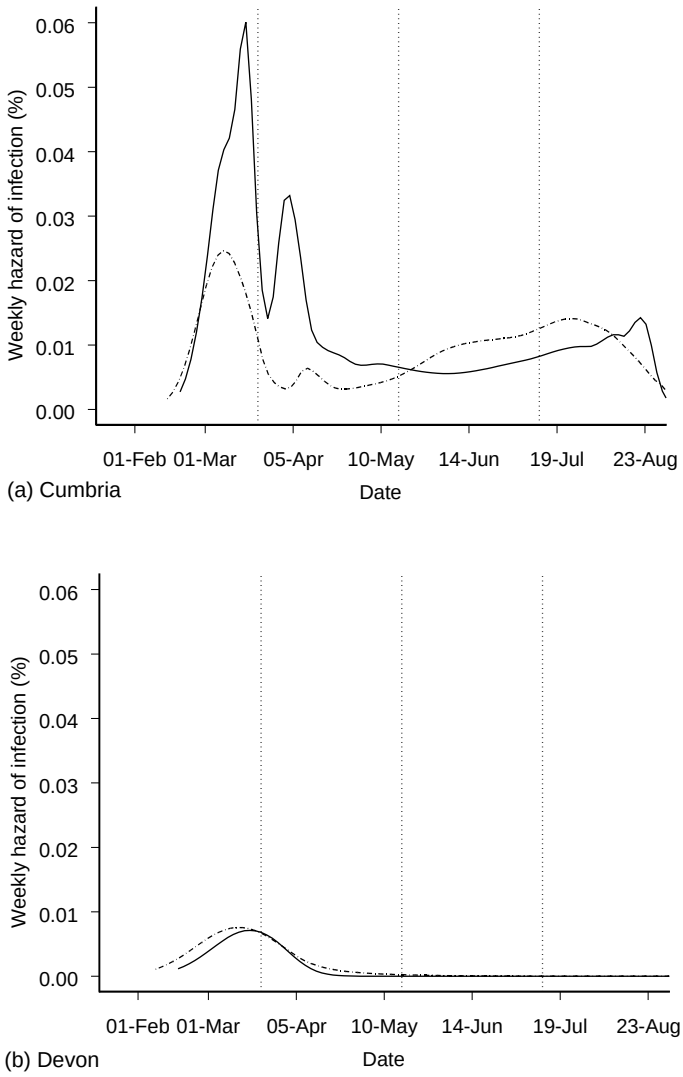


Fig. 2. Weekly hazard of foot-and-mouth disease infection for cattle holdings (solid lines) and 'other' holdings (dashed lines) in Cumbria and Devon (Great Britain). Vertical lines mark the four time periods described (20 February to 28 March 2001, 29 March to 23 May 2001, 24 May to 18 July 2001, and 19 July to 30 September 2001, respectively).

was present in the first period (Fig. 4a). In the second period, there were no high density zones identifiable in Devon: incident holdings during this time were randomly distributed. The widely-dispersed pattern of FMD-affected holdings in Fig. 4b and the absence of new infections after 31 May 2001 reflects the effectiveness of the control measures applied in this area.

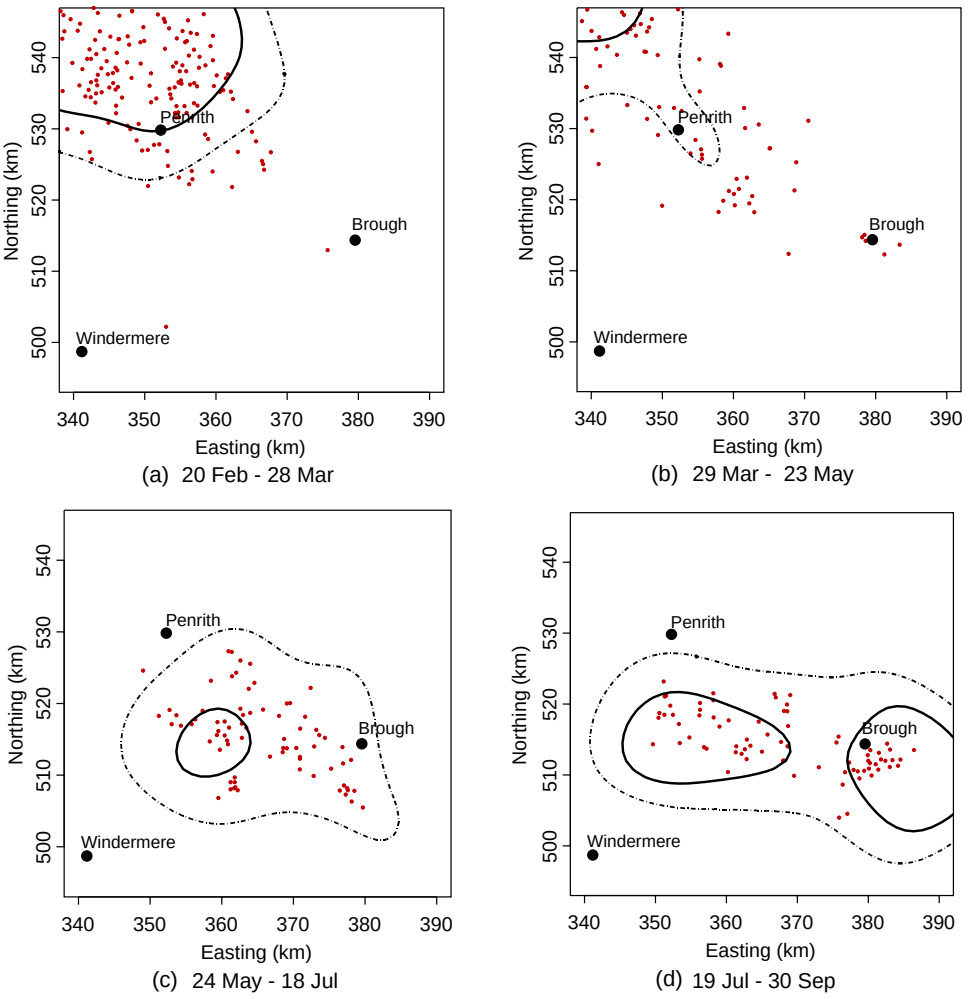


Fig. 3. Contour plots showing locations in Cumbria (Great Britain) where >10% (dashed lines) and >20% (solid lines) of holdings were diagnosed with foot-and-mouth disease. Point locations of incident holdings in each period have been superimposed, for reference.

In the surface plots of $D_0(s, t)$ for Cumbria (Fig. 5a–d), surface values, where $D_0(s, t)$ exceeds 1.0, are marked (showing the distance and time separation from an arbitrarily-selected case where the observed number of cases exceeded that which was expected, assuming that space-time interaction did not exist). A Monte Carlo test for space-time interaction was based on values of (s, t) in a 30×30 grid running in each coordinate direction. The observed values of the test statistic for each period were 0.852, 2.42, 3.34, and 3.89×10^6, whereas the values from 99 Monte Carlo permutations of the times ranged from -0.433 to 0.815, -1.31 to 1.30, -0.714 to 1.51, and -1.02 to 1.67×10^6—corresponding to a one-sided $P < 0.01$ in each instance. For the Devon study area (Fig. 6a and b), the test statistics for the two periods were 0.935 and 0.124×10^6. The values from 99 Monte Carlo

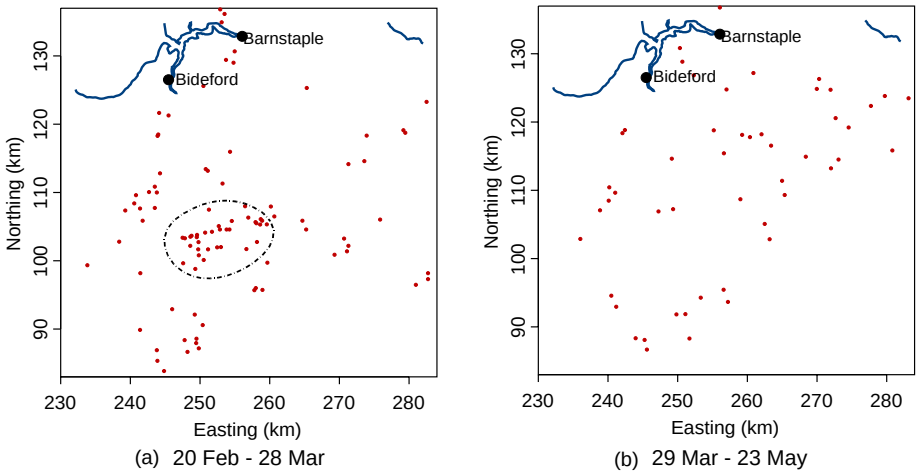


Fig. 4. Contour plots showing locations in Devon (Great Britain) where >10% (dashed lines) of holdings were diagnosed with foot-and-mouth disease. Point locations of incident holdings in each period have been superimposed, for reference.

permutations of the times ranged from −0.817 to 1.62 and −0.58 to 0.56 × 10^6, corresponding to a one-sided P of 0.04 and 0.39 for the earlier and later periods, respectively.

The median days between estimated infection and confirmation date in the Cumbria study area was 6 (95%, CI 5–12) (Fig. 7); median days between confirmation and cull date was 2 (95%, CI 1–3) (Fig. 8). For the Devon study area, the median days between estimated infection and confirmation date was 7 (95%, CI 5–11). Median days between confirmation and cull date was 2 (95%, CI 1–3).

4. Discussion

The observed start of the 2001 FMD epidemic in Great Britain was 19 February when the disease was identified in a group of cull sows sent for slaughter to an abattoir located in the south east of the country. By the end of the epidemic on 30 September 2001, 2026 holdings throughout Great Britain had been confirmed as infected and had been culled and an additional 8481 holdings had been culled pre-emptively as part of the control measures introduced after 24 February 2001, making this the largest and most-expensive FMD epidemic in a temperate country in recent years. Our analyses relate to two areas of Great Britain where there were relatively many farm holdings affected.

Several factors explain the striking differences in the behaviour of infection in the two areas investigated. In Cumbria in the early phase of the epidemic, many cattle holdings were infected—probably resulting in a large environmental viral load (Sellers, 1971). Although control activities resulted in a rapid decrease in infection hazard after 28 March, the epidemic remained incompletely controlled. Persistence of the disease—accompanied by the onset of seasonal farming activities such as contract hay and silage making—facilitated medium-to-long distance spread of virus and is thought to be at least partly responsible for

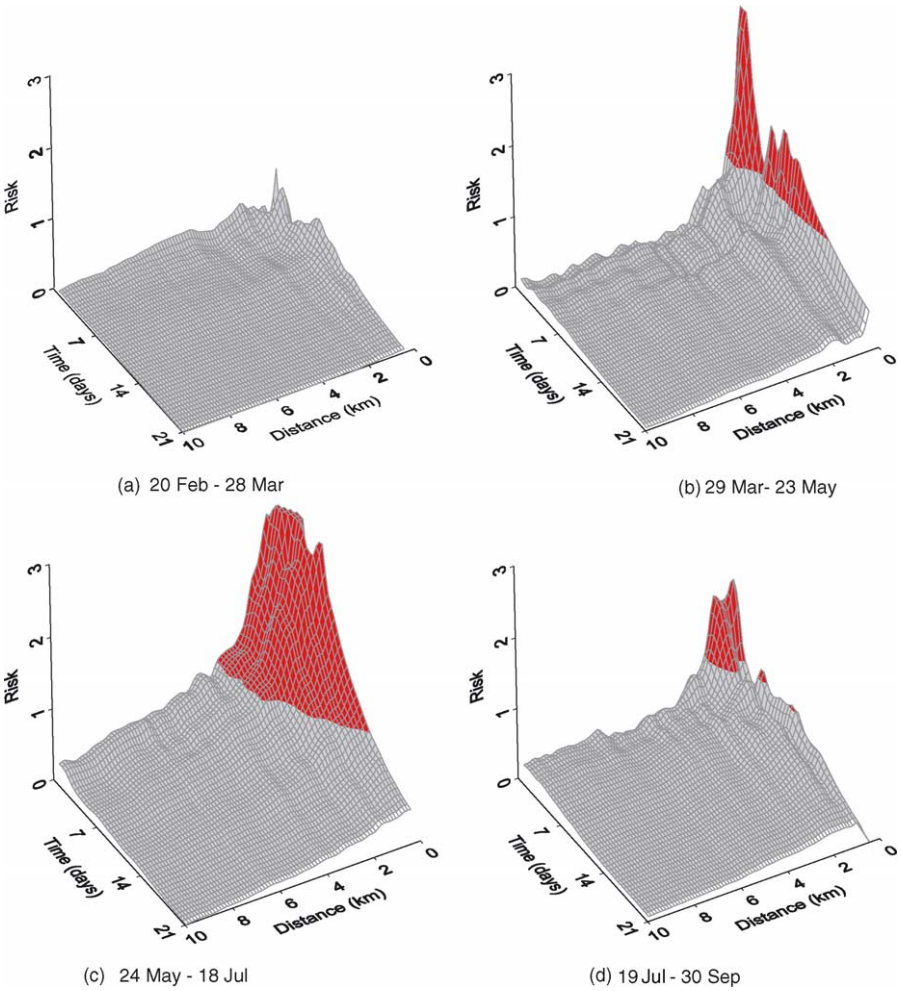


Fig. 5. Spatio–temporal interaction of foot-and-mouth disease risk among infected premises in Cumbria (Great Britain) On each surface, the dark-shaded area shows the distance-time separations where the observed number of cases exceeded that expected.

the steady increase in infection hazard that occurred from 19 April to 15 August (Fig. 2a). Although not evident from these analyses (where farm holdings have been represented by a single point in space), holdings made up of disaggregated land parcels and the unavoidable movement of animals and animal products between these parcels also would have complicated control efforts (Ferguson et al., 2001; Kao, 2001). In contrast, in Devon, we speculate that a lower initial environmental viral load and an already-mobilised control effort resulted in relatively quick eradication (Fig. 2b).

Fig. 5a shows that in the Cumbria study area for the period from 20 February to 28 March 2000 there was (in comparison with other periods) relatively low spatio–temporal

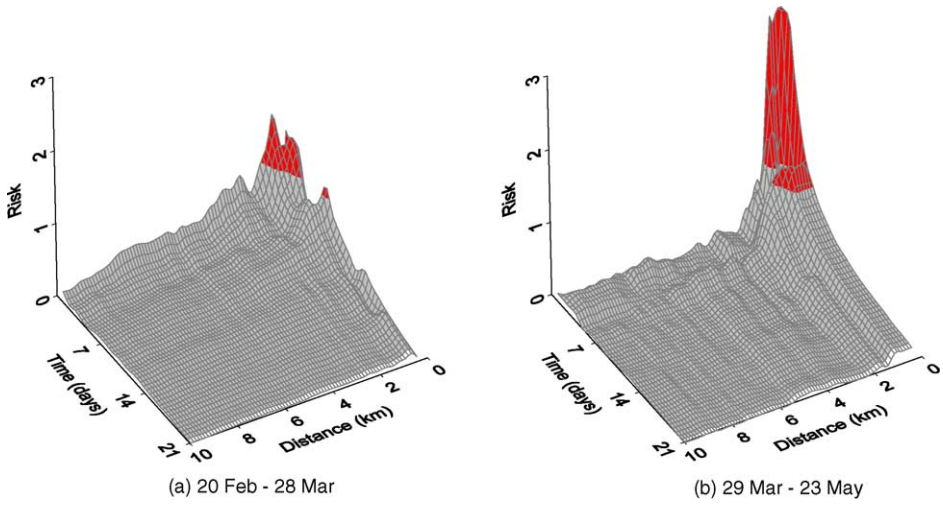


Fig. 6. Spatio–temporal interaction of foot-and-mouth disease risk among infected premises in Devon (Great Britain) On each surface, the dark-shaded area shows the distance-time separations where the observed number of cases exceeded that expected.

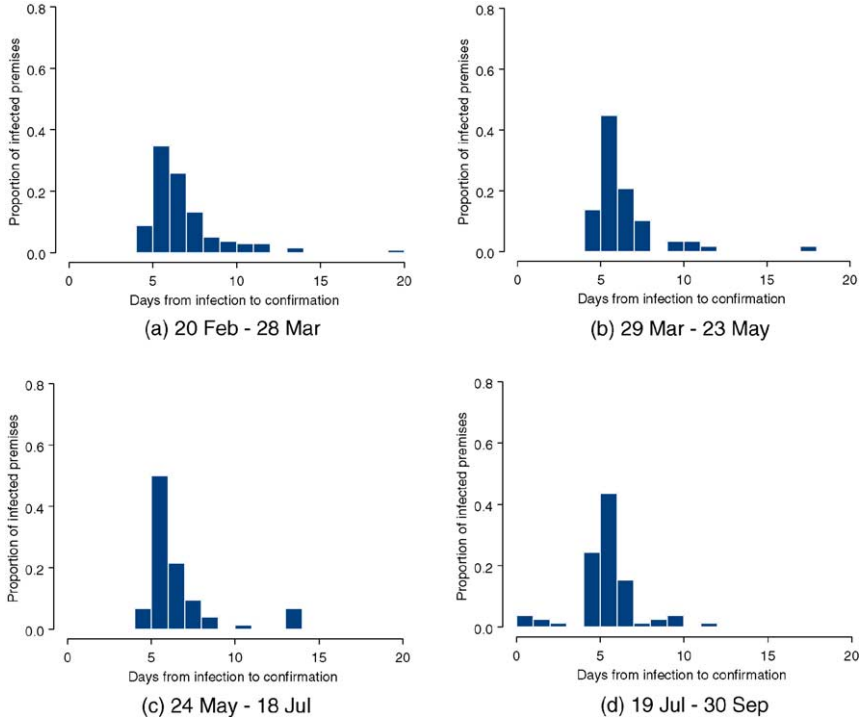


Fig. 7. Frequency distribution of the interval (in days) between foot-and-mouth disease infection date and confirmation date in Cumbria (Great Britain).

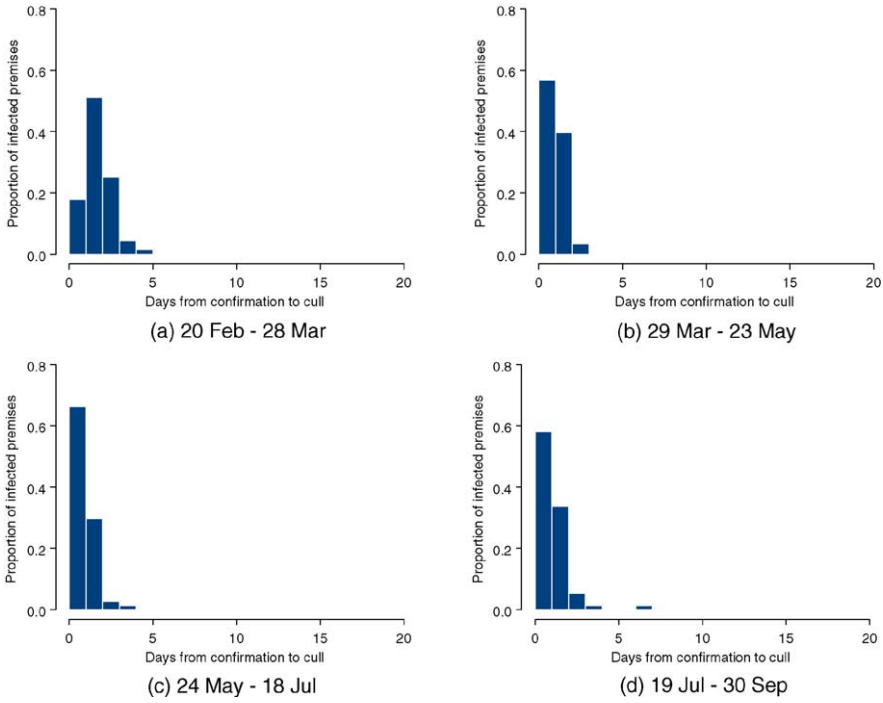


Fig. 8. Frequency distribution of the interval (in days) between foot-and-mouth disease confirmation date and cull date in Cumbria (Great Britain).

interaction of infection. Given the relatively high infection hazard evident at this time (Fig. 2a) this would suggest heavy and recent 'seeding' of virus throughout this area. As the epidemic progressed, spatio–temporal interaction became more prominent—evidenced by $D_0(s, t)$ values of >1 at relatively small distance-time separations (Fig. 5a–d). In Devon, spatio–temporal interaction of infection risk was evident in both the 20 February to 28 March and 29 March to 23 May periods (Fig. 6a and b)—though compared with the equivalent series for Cumbria, the space-time separations where $D_0(s, t) > 1$ were relatively small.

The distance component of the $D_0(s, t) = 1$ risk contour identifies the extent of spatio–temporal interaction of infection risk (colloquially termed 'contagiousness') in space and the temporal component of the $D_0(s, t) = 1$ risk contour identifies the extent of contagiousness in time. Figs. 5 and 6 show that these two components are not static throughout an epidemic; increases in the maximum distance component of the $D_0(s, t) = 1$ contour in sequential analyses is indicative of increases in local-spread distances—implying a need to re-assess the effectiveness of control measures aimed at reducing local spread of the infection. Increases in the maximum temporal position of the $D_0(s, t) = 1$ risk contour suggest that infected holdings are remaining infectious for longer—implying a need to enhance detection, the speed of depopulation and/or cleaning and disinfection procedures. Fig. 5b shows that for holdings infected between 29 March and 23 May 2001 in Cumbria, the risk that an arbitrarily-selected case holding posed to others extended no greater than 1 km (95% CI

0 km, 3 km). For holdings infected between 24 May and 18 July, the shape of the contour differed: the $D_0(s, t) = 1$ risk contour persisted at 3 km (95% CI 0 km, 5 km) for 7 days after infection date, then declined to 0 km by 14 days (Fig. 5c). For the period 19 July to 30 September 2001, the $D_0(s, t) = 1$ risk contour (Fig. 5d) declined from 2 km (95% CI 0 km, 5 km) at $t = 0$ days to 0 km at 7 days. Since the process of culling, disposal and disinfection of infected premises was well-established in Cumbria early in the epidemic (from 29 March 2001 to the end of the epidemic >50% of infected premises were culled within 1 day of confirmation; Fig. 8b–d) we therefore conclude that the major reason for the increased time infected premises posed a risk to others during May to July was delayed detection of disease. The movement of the disease into 'other' (predominantly sheep) holding types from May to July—where the identification of clinical signs was more difficult (Kitching and Mackay, 1995) probably influenced this pattern. This is important because the epidemic in the studied area Cumbria was predominantly a cattle-based outbreak with infection spilling over into sheep flocks.

Although this has been a retrospective analysis of two purposively-selected areas we propose that the analytical approach described could be applied on a routine basis during future FMD epidemics to describe and better understand the temporal, spatial and spatio-temporal features of infection risk. Instantaneous hazard functions provide a description of the probability of infection per unit time, accounting for temporal changes in the size of the holding population at risk. Extraction maps account for the spatial distribution of the farm holding population at risk and provide a means for defining locations where there are high proportions of infected premises per unit area. Whereas these techniques consider time and space independently, the ability to describe the risk of infection attributable to spatio–temporal interaction provides further insight—identifying the extent of 'contagiousness' firstly in space (and therefore providing an objective means for defining suitable pre-emptive culling distances) and secondly in time (indicating how quickly infection risk is 'extinguished' after a holding becomes infected). The space-time *K*-function appears well-suited for this purpose, and in the example provided in these analyses we have shown that the level of spatio–temporal interaction itself can change over time, indicative of important changes in the epidemiological behaviour of the disease.

5. Conclusion

Our analyses showed that in Cumbria (where many cattle holdings initially were infected and eradication took several months) the distance over which contagiousness operated and the temporal duration of contagiousness from source farms changed throughout the course of the epidemic. In contrast, in Devon (where the epidemic was controlled quickly), there was less evidence of space-time interaction and the relevant spatial and temporal effects were much shorter.

Acknowledgements

We are grateful to DEFRA for funding for this study. We thank Professor Peter Diggle and Dr Jim Della-Vedova for helpful comments provided during the preparation of this paper.

References

Bithell, J.F., 1990. An application of density estimation to geographical epidemiology. Stat. Med. 9, 691–701.
Bivand, R., Gebhardt, A., 2000. Implementing functions for spatial statistical analysis using the R language. J. Geograph. Syst. 2, 307–317.
Bowman, A.W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-PLUS Illustrations. Oxford University Press, London, pp. 112–117.
Diggle, P.J., Chetwynd, A.G., Häggkvist, R., Morris, S., 1995. Second order analysis of space-time clustering. Stat. Methods Med. Res. 4, 124–136.
Ferguson, N.M., Donnelly, C.A., Anderson, R.M., 2001. The foot-and-mouth disease epidemic in Great Britain: pattern of spread and impact of interventions. Science 292, 1155–1160.
Gibbens, J.C., Sharpe, C.E., Wilesmith, J.W., Mansley, L.M., Michalopoulou, E., Ryan, J.B.M., Hudson, M., 2001. Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. Vet. Rec. 149, 729–743.
Gibbens, J.C., Wilesmith, J.W., 2002. Temporal and geographical distribution of cases of foot-and-mouth disease during the early weeks of the 2001 epidemic in Great Britain. Vet. Rec. 151, 407–412.
HMSO, 1983. Foot-and-Mouth Disease Order. Statutory Instrument 1983, Number 1950. Her Majesty's Stationary Office, London.
Howard, S.C., Donnelly, C.A., 2000. The importance of immediate destruction in epidemics of foot-and-mouth disease. Res. Vet. Sci. 69, 189–196.
Ihaka, R., Gentleman, R., 1996. R: A language for data analysis and graphics. J. Comput. Graph. Stat. 5, 299–314.
Kao, R.R., 2001. Landscape fragmentation and foot-and-mouth disease transmission. Vet. Rec. 148, 746–747.
Keeling, M.J., Woolhouse, M.E.J., Shaw, D.J., Matthews, L., Chase-Topping, M., Haydon, D.T., Cornell, S.J., Kappey, J., Wilesmith, J.W., Grenfell, B.T., 2001. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science 294, 813–817.
Kitching, R., Mackay, D., 1995. Foot-and-mouth disease. State Vet. J. 5, 4–8.
Lawson, A.B., 2001. Statistical Methods in Spatial Epidemiology. Wiley, London, pp. 91–99.
Lawson, A.B., Williams, F.L.R., 1994. Armadale: a case study in environmental epidemiology. J. R. Stat. Soc. Ser. A: Gen. 157, 285–298.
Loader, C., 1999. Local Regression and Likelihood. Springer-Verlag, New York.
MAFF, 2000. Agricultural statistics—United Kingdom. Her Majesty's Stationary Office, London.
Morris, R.S., Wilesmith, J.W., Stern, M.W., Sanson, R.L., Stevenson, M.A., 2001. Predictive spatial modelling of alternative control strategies for the foot-and-mouth disease epidemic in Great Britain, 2001. Vet. Rec. 149, 137–144.
Morris, R.S., Sanson, R.L., Stern, M.W., Stevenson, M.A., Wilesmith, J.W., 2002. Decision-support tools for foot and mouth disease control. Rev. Off. Int. Epizoot. 21, 557–567.
Radostits, O.M., Gay, C.C., Blood, D.C., Hinchcliff, K.W., 2000. Veterinary Medicine, ninth ed. W.B. Saunders, London, pp. 1059–1069.
Rowlingson, B.S., Diggle, P.J., 1993. SPLANCS: Spatial point pattern analysis code in S-PLUS. Comput. Geosci. UK 19, 627–655.
Sanson, R.L., 1993. The development of a decision support system for an animal disease emergency. Unpublished PhD thesis Massey University, Palmerston North, New Zealand.
Sanson, R.L., Liberona, H., Morris, R.S., 1991. The use of a geographical information system in the management of a foot-and-mouth disease epidemic. Prev. Vet. Med. 11, 309–313.
Sellers, R.F., 1971. Quantitative aspects of the spread of foot-and-mouth disease. Vet. Bull. 41, 431–439.

References

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's Proporional Hazards Model. *American Statistical Association*, *83*, 204 - 212.

Black, D., & French, N. (2004). Effects of three types of trace element supplementation on the fertility of three commercial dairy herds. *Veterinary Record*, *154*, 652 - 658.

Bradburn, M., Clark, T., Love, S., & Altman, D. (2003a). Survival analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer*, *89*, 431 - 436.

Bradburn, M., Clark, T., Love, S., & Altman, D. (2003b). Survival analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit. *British Journal of Cancer*, *89*, 605 - 611.

Caplehorn, J., & Bell, J. (1991). Methadone dosage and retention of patients in maintenance treatment. *Medical Journal of Australia*, *154*(3), 195 - 199.

Clark, T., Bradburn, M., Love, S., & Altman, D. (2003a). Survival analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*, *89*, 232 - 238.

Clark, T., Bradburn, M., Love, S., & Altman, D. (2003b). Survival Analysis Part IV: Further concepts and methods in survival analysis. *British Journal of Cancer*, *89*, 781 - 786.

Collett, D. (1994). *Modelling Survival Data in Medical Research*. London: Chapman and Hall.

Crowley, J., & Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, *72*(27 - 36).

Dawson, K., Stevenson, M., Sinclair, J., & Bosson, M. (2014). Recurrent bovine tuberculosis in New Zealand cattle and deer herds, 2006 – 2010. *Epidemiology and Infection*. doi: 10.1017/S0950268814000910

Dohoo, I., Martin, S., & Stryhn, H. (2003). *Veterinary Epidemiologic Research*. Charlottetown, Prince Edward Island, Canada: AVC Inc.

Fisher, L., & Lin, D. (1999). Time-dependent covariates in the Cox proportional hazards regression model. *Annual Reviews in Public Health*, *20*, 145 - 157.

Fleming, T., & Harrington, D. (1991). Counting processes and survival analysis. *Communication in Statistics Theory, Methods*, *13*, 2469 - 2486.

Gautam, M. (2012). *Epidemiological Study of Removals in New Zealand Dairy Goats* (Unpublished master's thesis). Massey University.

Gautam, M., Stevenson, M., Lopez-Villalobos, N., & McLean, V. (2017). Risk Factors for Culling, Sales and Deaths in New Zealand Dairy Goat Herds, 2000-2009. *Frontiers in Veterinary Science*, *4*, 191. doi: 10.3389/fvets.2017.00191

Haerting, J., Mansmann, U., & Duchateau, L. (2007). *Frailty Models in Survival Analysis* (Unpublished doctoral dissertation). Martin-Luther-Universität Halle-Wittenberg.

Hosmer, D., & Lemeshow, S. (1999). *Applied Survival Analysis Regression Modeling of Time to Event Data*. London: Jon Wiley and Sons Inc.

Kleinbaum, D. (1996). *Survival Analysis: A Self-Learning Text*. New York: Springer-Verlag.

Kyle, R. (1993). 'Benign' monoclonal gammopathy — after 20 to 35 years of follow-up. *Mayo Clinic Proceedings*, *68*, 26 - 36.

Lee, E. (1992). *Statistical Methods for Survival Analysis*. London: John Wiley and Sons Inc.

Lee, E., & Go, O. (1997). Survival analysis in public health research. *Annual Reviews in Public Health*, *18*, 105 - 134.

Leung, K., Elashoff, R., & Afifi, A. (1997). Censoring issues in survival analysis. *Annual Reviews in Public Health*, *18*, 83 - 104.

Loprinzi, C., Laurie, J., Wieand, H., Krook, J., Novotny, P., Kugler, J., ... Moertel, C. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. *Journal of Clinical Oncology*, *12*, 601 - 607.

More, S. (1996). The performance of farmed ostrich eggs in eastern Australia. *Preventive Veterinary Medicine*, *29*, 121 - 134.

Prentice, R., Williams, B., & Peterson, A. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, *68*, 373 - 379.

Proudman, C., Dugdale, A., Senior, J., Edwards, G., Smith, J., Leuwer, M., & French, N. (2006). Pre-operative and anaesthesia-related risk factors for mortality in equine colic cases. *The Veterinary Journal*, *171*(1), 89 - 97.

Proudman, C., Pinchbeck, G., Clegg, P., & French, N. (2004). Risk of horses falling in the Grand National. *Nature*, *428*, 385 - 386.

Schemper, M., & Stare, J. (1996). Explained variation in survival analysis. *Statistics in Medicine*, *15*, 1999 - 2012.

Stevenson, M., Wilesmith, J., Ryan, J., Morris, R., Lockhart, J., Lin, D., & Jackson, R. (2000). Temporal aspects of the bovine spongiform encephalopathy epidemic in Great Britain: Individual animal-associated risk factors for disease. *Veterinary Record*, *147*(13), 349 - 354.

Tableman, M., & Kim, J. (2004). *Survival Analysis Using S*. New York: Chapman Hall/CRC.

The Diabetes Control and Complications Trial Research Group. (1996). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal Of Medicine*, *329*(14), 977 - 986.

Therneau, T., & Grambsch, P. (2001). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Venables, W., & Ripley, B. (2002). *Modern Applied Statistics with S*. New York: Springer-Verlag.

Wei, L., Lin, D., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association*, *84*, 1065 - 1073.

Wilesmith, J., Ryan, J., Stevenson, M., Morris, R., Pfeiffer, D., Lin, D., ... Sanson, R. (2000). Temporal aspects of the bovine spongiform encephalopathy epidemic in Great Britain: Holding-associated risk factors for disease. *Veterinary Record*, *147*(12), 319 - 325.

Wilesmith, J., Stevenson, M., King, C., & Morris, R. (2003). Spatio-temporal epidemiology of foot-and-mouth disease in two counties of Great Britain in 2001. *Preventive Veterinary Medicine*, *61*(3), 157 - 170.