The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 34

Survival Models in Health Economic Evaluations: Balancing Fit and Parsimony to Improve Prediction

Christopher H. Jackson, *MRC Biostatistics Unit*

Linda D. Sharples, *MRC Biostatistics Unit*

Simon G. Thompson, *MRC Biostatistics Unit*

Recommended Citation:

Jackson, Christopher H.; Sharples, Linda D.; and Thompson, Simon G. (2010) "Survival Models in Health Economic Evaluations: Balancing Fit and Parsimony to Improve Prediction," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 34.

DOI: 10.2202/1557-4679.1269

Survival Models in Health Economic Evaluations: Balancing Fit and Parsimony to Improve Prediction

Christopher H. Jackson, Linda D. Sharples, and Simon G. Thompson

Abstract

Health economic decision models compare costs and health effects of different interventions over the long term and usually incorporate survival data. Since survival is often extrapolated beyond the range of the data, inaccurate model specification can result in very different policy decisions. However, in this area, flexible survival models are rarely considered, and model uncertainty is rarely accounted for. In this article, various survival distributions are applied in a decision model for oral cancer screening. Flexible parametric models are compared with Bayesian semiparametric models, in which the baseline hazard can be made arbitrarily complex while still enabling survival to be extrapolated. A fully Bayesian framework is used for all models so that uncertainties can be easily incorporated in estimates of long-term costs and effects. The fit and predictive ability of both parametric and semiparametric models are compared using the deviance information criterion in order to account for model uncertainty in the cost-effectiveness analysis. Under the Bayesian semiparametric models, some smoothing of the hazard function is required to obtain adequate predictive ability and avoid sensitivity to the choice of prior. We determine that one flexible parametric survival model fits substantially better than the others considered in the oral cancer example.

KEYWORDS: cost-effectiveness, model uncertainty, model comparison, Bayesian semiparametric, generalized F

Author Notes: Please send correspondence to: Christopher Jackson, MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 0SR, U.K. Email: chris.jackson@mrc-bsu.cam.ac.uk. We thank Stephen Palmer at the University of York for permission to use the oral cancer example and providing the economic model implementation in Excel. The authors are funded by the UK Medical Research Council (grants U.1052.00.008 and U.1052.00.001).

1 Introduction

Decision analytic models (Claxton *et al.*, 2002) are the standard framework for economic evaluations of health technologies in the UK (National Institute for Health and Clinical Excellence, 2008) and elsewhere. These are commonly Markov models for transitions between important clinical states. Expected long-term costs and health effects of each intervention are estimated from the cost and effect of occupying each state, and the probability that an individual occupies each state over time. Modelling enables all available evidence on alternative interventions to be synthesised, clinical evidence to be extrapolated over the long term, and the need for further research to be assessed (Claxton, 1999). Since health effects are usually measured by quality-adjusted life years (QALYs), the modelling of survival is an important component.

The principal source of evidence on the effect of an intervention is usually a randomised controlled trial, but long-term forecasts require extrapolating beyond the follow-up period of the trial. Parametric survival models, most commonly using the Weibull distribution, are employed to extrapolate survival and construct transition probabilities for economic models (Briggs *et al.*, 2006), but the adequacy of these models is rarely assessed. Effects of treatment and other covariates on survival are usually modelled through proportional hazards, but effects may not continue after the observed follow-up. Population mortality data from national registries can sometimes be used to estimate long-term survival (Demiris and Sharples, 2006), but the patients receiving the interventions of interest may differ systematically from the general population.

If no relevant long-term survival data are available, it is particularly important that the short-term data are modelled as accurately as possible. A suitable compromise is desired between semiparametric methods which allow flexible survival distributions, and parametric models which allow extrapolation. Nelson *et al.* (2008) partitioned follow-up time into early and late periods, using Kaplan-Meier estimation for the early survival data, and parametric models to extrapolate the later data. Demiris and Sharples (2006) used Bayesian semiparametric and Weibull parametric models for extrapolating mean survival, with covariates entering through proportional and additive hazards. The UK National Institute for Health and Clinical Excellence (2008) recommend that different plausible extrapolation assumptions be compared in sensitivity analysis. This is usually performed by presenting cost-effectiveness under alternative models, but with no formal measures of the plausibility of different assumptions. Jackson *et al.* (2009, 2010) described how statistical models used in economic evaluations can be formally compared based on pre-

dictive criteria, and used model averaging to account for uncertainty about model selection, although they did not explicitly discuss survival modelling.

In this paper, we show how parametric and semiparametric survival models can be compared by putting all models in a Bayesian setting. This allows survival to be extrapolated to inform the long-term model, and allows the uncertainty surrounding survival to be straightforwardly included in estimates of lifetime cost-effectiveness. The models are evaluated using criteria which assess expected predictive ability based on the observed data. Uncertainty about model selection can be accounted for in the conclusions using a Bayesian bootstrap technique. We use flexible three and four-parameter survival models for the first time in health economic modelling and show how they can form a compromise between the lower bias of semiparametric methods and the lower predictive variance of a two-parameter model. These are compared with a variety of Bayesian semiparametric survival models with piecewise-constant hazards. We discuss how to choose the baseline hazard function in semiparametric models and its variation between covariate categories in order to compromise between flexibility and parsimony and avoid sensitivity to prior assumptions. All models are implemented in the WinBUGS (Lunn *et al.*, 2000) software for generic Bayesian modelling.

These methods are illustrated in an economic model for screening for oral cancer, described in Section 2. Section 3 details the alternative parametric and semiparametric models for oral cancer survival, and Section 4 describes how they are compared using informal plots and formal criteria. In Section 5 the inferences about screening cost-effectiveness under the alternative models are compared. We conclude that one single flexible parametric model fits substantially better than the others considered, so that it is reasonable to base conclusions on this model alone. Among the semiparametric models, those with hazard functions smoothed to adapt to the data showed the best predictive ability and were not sensitive to the choice of prior. We finish by discussing potential extensions of these methods, particularly those which model continuous variations in the hazard, and other problems in extrapolating survival for health economic models.

2 Oral cancer screening model

2.1 Markov model

A decision-analytic model has previously been developed to compare the cost-effectiveness of eight screening strategies for oral cancer in various age and sex

groups. This is described in detail elsewhere (Speight *et al.*, 2006), a summary is presented here. In this paper, for simplicity, we consider only two screening strategies:

1. no screen,
2. opportunistic screening of patients attending a general medical practitioner, by visual examination.

The results (Speight *et al.*, 2006) indicated that one or the other of these two strategies were likely to be cost-effective for the majority of health budgets.

A discrete-time Markov model is used with 7 states representing no cancer, detectable preclinical cancer, four stages of cancer and death, and a yearly time unit. Individuals may progress to the next more severe disease state or die from any state, and disease status is assumed not to improve. After screening and subsequent diagnosis of cancer by biopsy, only transitions to death are modelled. The probabilities of death depend on whether cancer has been diagnosed at screen and on the cancer stage at diagnosis, and then change in each of the first five years following diagnosis. The cost associated with a period spent in each state is assumed to be zero before diagnosis. After diagnosis, the cost varies according to the stage at diagnosis and the time since diagnosis. A year spent in each state is quality-adjusted by utilities which depend on the age and current cancer stage.

The principal source of data for the Markov model was the 6093 registered diagnoses of oral cancer between 1980 and 1995 from the Thames Cancer Registry. Cases were followed up until the year 2000, and 4295 were observed to die. Survival with diagnosed or undiagnosed cancer was estimated from this dataset, with undiagnosed patients assumed to face an additional risk of disease progression which was elicited from a panel of experts. Other parameters of the decision model, including progression rates between stages of cancer and the sensitivity and specificity of screening, were estimated from meta-analyses of summary results of trials or by expert elicitation. These parameters and their sources are listed in greater detail by Speight *et al.* (2006).

2.2 Cost-effectiveness estimation

The Markov model allows the estimation of the expected cost $E(C_u|\mathbf{x})$ and QALY $E(Q_u|\mathbf{x})$ for a patient with given characteristics $\mathbf{x}$, under each screening strategy $u = 1$ or 2 , as described by Spiegelhalter and Best (2003) and Briggs *et al.* (2006). This is now the standard framework for health economic policy-making in the UK (Claxton *et al.*, 2002, National Institute for Health

and Clinical Excellence, 2008) and elsewhere. There is a vector of costs $\mathbf{c}_{ut}$ associated with spending year t in each of the 7 states. Let $P_t(\mathbf{x})$ be the transition probability matrix of the Markov model at year t for this patient. The vector of probabilities $\boldsymbol{\pi}_t(\mathbf{x})$ that a patient is in each state during year t follows the recursive relationship $\boldsymbol{\pi}_t(\mathbf{x}) = \boldsymbol{\pi}_{t-1}(\mathbf{x})P_t(\mathbf{x})$. The initial prevalence $\boldsymbol{\pi}_1$ of undiagnosed states is determined from survey data (Speight *et al.*, 2006). Future costs and QALYs are discounted at a rate of $\delta = 3.5\%$ per year, as recommended by the UK National Institute for Health and Clinical Excellence (2008). The expected cost over a lifetime of $T = 60$ years is then

$$E(C_u|\mathbf{x}) = \sum_{t=1}^T \frac{\boldsymbol{\pi}_t(\mathbf{x})\mathbf{c}'_{ut}}{(1 + \delta)^{t-1}}. \quad (1)$$

(Spiegelhalter and Best, 2003, Briggs *et al.*, 2006). The expected QALY $E(Q_u|\mathbf{x})$ is defined similarly in terms of state-specific quality-of-life weights $\mathbf{q}_{ut}$.

The uncertainty surrounding expected costs and effects is calculated by *probabilistic sensitivity analysis*. All parameters influencing $P_t(\mathbf{x})$, $\mathbf{c}_{ut}$ and $\mathbf{q}_{ut}$ are entered into the model as probability distributions, obtained either as expert judgements of the typical location and spread of values, or posterior distributions estimated from data. Monte Carlo simulation is then used to propagate parameter uncertainty and obtain a joint posterior distribution for the expected costs and effects.

Note that the Markov economic model differs from the methods for analysing individual-level cost-effectiveness data discussed by Lin *et al.* (1997), Lin (2000) and others, which cannot be used for extrapolating beyond the time limit of the data, and require adjustment for informative censoring on the cost and quality-adjusted survival scale. Since $\mathbf{c}_{ut}$, $\mathbf{q}_{ut}$ are not individual-level observations, but the assumed cost and quality of spending year t in each state, for any patient receiving intervention u , they are not subject to any censoring. The survival registry data which inform the model, however, are subject to non-informative, administrative censoring. If individual-level cost and quality-of-life data had been available, these could have been used in Equation 1 to inform the distributions assigned to $\mathbf{c}_{ut}$ and $\mathbf{q}_{ut}$ in probabilistic sensitivity analysis.

2.3 Use of survival data in the Markov model

Although there are many other parameters in the Markov model, we focus our investigation of model uncertainty on the choice of survival model, since

the large individual-patient dataset which informed this choice is available to us. A Weibull survival model was used in the original analysis of these data (Speight *et al.*, 2006) to obtain annual probabilities of death in the Markov model, but the adequacy of this model was only assessed against a simpler exponential alternative. A Kaplan-Meier estimate of the survivor function for male patients aged 50-59 diagnosed in cancer stage 1, and a corresponding kernel-based nonparametric hazard estimate (Mueller and Wang, 1994), are illustrated in Figures 1 and 2 respectively. The hazard is initially high after diagnosis, decreasing sharply in the first five years, decreasing slightly up to 10 years, and increasing thereafter, though the increase is estimated from only 13 deaths in this subgroup which occurred after 10 years. Figure 2 shows that a Weibull model underestimates the hazard for this subgroup in the first five years after diagnosis and overestimates from then on. In this paper we fit a range of different survival models, and assess the impact of model choice on the estimates of screening cost-effectiveness.

Up to 5 years after diagnosis, the Markov model probability $p_t(\mathbf{x})$ of death by year $t + 1$, given survival up to t , for an individual with covariates $\mathbf{x}$, is obtained as $p_t(\mathbf{x}) = (\hat{S}(t|\mathbf{x}) - \hat{S}(t+1|\mathbf{x}))/\hat{S}(t|\mathbf{x})$, where $\hat{S}(t|\mathbf{x})$ is the fitted survival function. However, since the Markov model is used to extrapolate over patient lifetimes, with up to 60 years of follow-up, and the registry data provide only an average of around 10 years of follow-up, extra assumptions are required to model the increase in the hazard of death with age. We use the same assumptions as Speight *et al.* (2006), specifically, after 5 years after diagnosis, the baseline hazard was assumed to stabilise at the fitted five-year value (as suggested by Figure 2). In addition, the effect of increasing age as *the same* patient grows older is assumed to be the same as the effect of age at diagnosis between *different* patients. That is, a hazard $h(t|G, \mathbf{x})$ is estimated from the entire data, using either parametric or semiparametric methods, including a covariate G representing age group at diagnosis (40–49 years, 50–59 years, 60–69 years, 70–79 years or 80+ years). Then when calculating the Markov transition probability, a person diagnosed at age 60, for example, is assumed to have hazard $h(t|G = 60\text{--}69, \mathbf{x})$ for t in the first five years after diagnosis, $h(5|G = 60\text{--}69, \mathbf{x})$ for years 6–10, $h(5|G = 70\text{--}79, \mathbf{x})$ for years 11–20 and $h(5|G = 80+, \mathbf{x})$ thereafter. There are no suitable long-term data to assess these assumptions formally; in this paper we focus on different possible choices for the survival function.

The remaining transition probabilities in the Markov model, describing progression between stages of cancer and the additional mortality risk for undiagnosed patients, were informed by expert elicitation.

Figure 1: Oral cancer empirical (Kaplan-Meier) survival curve and fitted parametric models for male, age 50-59, stage 1. (The fitted generalized gamma model is indistinguishable from the log-normal.)

Figure 2: Nonparametric estimate of hazard of death from oral cancer and fitted parametric models for male, age 50-59, stage 1.

3 Alternative survival models

Several alternative parametric and nonparametric survival models are fitted to the registry data. All models include covariates representing age (five groups: age 40–49, age 50–59, age 60–69, age 70–79 and age 80+), sex and cancer stage at diagnosis (four categories), which are well-established risk factors for oral cancer survival, as in Speight *et al.* (2006). These are fitted in a Bayesian framework, obtaining posterior distributions for the annual survival probabilities to be employed in the probabilistic sensitivity analysis of the Markov decision model (Section 2.2). The Bayesian approach is well suited to economic modelling, due to the need to express uncertainty surrounding expected costs and effects which are complex functions of the model parameters. Commonly in health economic probabilistic sensitivity analyses, independent distributions are used for parameters, informed by their estimates and standard errors, an approach which has been criticised (Ades *et al.*, 2006). A fully Bayesian joint posterior distribution accounts properly for correlations between parameters.

3.1 Parametric models

We compare five fully parametric survival models, with the standard likelihood:

$$l(\boldsymbol{\theta}) = \prod_i f(T_i|\boldsymbol{\theta})^{D_i} (1 - F(T_i^*|\boldsymbol{\theta}))^{1-D_i}, \quad (2)$$

where T_i is the survival time, T_i^* is the censoring time, and D_i is the indicator for an observed death for individual i , $f(.|\boldsymbol{\theta})$ is the probability density and $F(.|\boldsymbol{\theta})$ is the cumulative density of the survival model with parameters $\boldsymbol{\theta}$. In each model, a “location” parameter μ is modelled on covariates via a linear model $\mu_i = \boldsymbol{\beta}'\mathbf{x}_i$. An accelerated failure time parameterisation is used (with the exception of the Gompertz, as detailed below) where the survival time is a log-linear function of covariates: $\log(T_i) = \mu_i + \epsilon_i$, and the ϵ_i are independent and identically distributed. The five models, together with the priors used, are:

1. Weibull (base case, 2 parameters $\boldsymbol{\theta} = (\sigma, \lambda)$)

$$f(t|\boldsymbol{\theta}) = a\lambda^a t^{a-1} \exp(-(\lambda t)^a)$$

$$\mu = \log(\lambda). \text{ Prior: } \sigma = 1/a, \log(\sigma) \sim N(0, 100000).$$

2. Log-normal (2 parameters $\boldsymbol{\theta} = (\mu, \sigma)$)

$$f(t|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{1}{2\sigma^2}(\log(t) - \mu)^2\right)$$

Prior: $\log(\sigma) \sim N(0, 100000)$.

3. Gompertz (2 parameters $\boldsymbol{\theta} = (\lambda, \gamma)$), often used to model human population survival due to its exponentially increasing hazard.

$$f(t|\boldsymbol{\theta}) = \lambda e^{\gamma t} \exp(-\lambda/\gamma(e^{\gamma t} - 1))$$

for $\gamma \neq 0$, or exponential $f(t|\boldsymbol{\theta}) = \lambda e^{\lambda t}$ for $\gamma = 0$. This does not have a convenient accelerated failure time parameterisation unless γ is restricted to be positive. Therefore covariates are included under a proportional hazards model, with $\mu = \log(\lambda)$. Prior: $\gamma \sim N(0, 1000)$.

4. Generalized gamma (using the parameterisation by Prentice (1974) which allows $q \leq 0$). With 3 parameters ($\boldsymbol{\theta} = (\mu, \sigma, q)$), this can represent a variety of hazard trajectories (Cox *et al.*, 2007).

$$f(t|\boldsymbol{\theta}) = \frac{|q|}{\sigma t \Gamma(q^{-2})} \exp(q^{-2}(q\mu - e^{q\mu}))(q^{-2})^{q^{-2}}$$

where $w = (\log(t) - \mu)/\sigma$, for $q \neq 0$. For $q = 0$, $f(t|\boldsymbol{\theta})$ is defined as the above log-normal density with parameters μ, σ . $f(t|\boldsymbol{\theta})$ also reduces to a Weibull distribution with parameters $a = 1/\sigma$ and $\lambda = \exp(-\mu)$ for $q = 1$, and a Gamma distribution for $\sigma = 1, q > 0$.

Priors: $\log(\sigma) \sim N(0, 100000)$, $q \sim N(0, 1000)$ independently.

5. Generalized F (Cox, 2008), an even more flexible distribution with 4 parameters ($\boldsymbol{\theta} = (\mu, \sigma, q, p)$), using the parameterisation described by Prentice (1975). The limiting case for $p \rightarrow 0$ is the above generalized gamma density.

$$f(t|\boldsymbol{\theta}) = \frac{\delta(m_1/m_2)^{m_1} e^{m_1 w}}{\sigma t (1 + m_1 e^w/m_2)^{(m_1+m_2)} B(m_1, m_2)}$$

where $w = (\log(t) - \mu)\delta/\sigma$, $\delta = (q^2 + 2p)^{1/2}$, $m_1 = 2(q^2 + 2p + q\delta)^{-1}$, $m_2 = 2(q^2 + 2p - q\delta)^{-1}$, and $B()$ is the beta function.

Priors: $\log(\sigma) \sim N(0, 100000)$, $q \sim N(0, 1000)$, $\log(p) \sim N(0, 100000)$ independently.

The covariate effects comprising $\boldsymbol{\beta}$ are given independent $N(0, 100000)$ priors.

3.2 Bayesian semiparametric model

A Bayesian semiparametric version of the Cox model (Kalbfleisch, 1978) was also fitted. This allows an arbitrarily flexible baseline hazard. Covariates can be included under a proportional hazards assumption or by stratifying the baseline hazard. This is implemented using a counting process formulation (Clayton, 1991). For individual i , $N_i(t)$ counts the number of deaths up to time t . The increment of the counting process $dN_i(t)$ over the time interval $[t, t + dt]$ takes the value 1 if individual i dies in this interval and 0 otherwise. The hazard function of the process is $\lambda_i(t) = \lim_{dt \rightarrow 0} E(dN_i(t)|\mathcal{F}_t)/dt$, where $\mathcal{F}_t$ is the history of the process up to t . Let $t_j : j = 1, 2, \dots$ represent all unique death or censoring times ($t_0 = 0$). The likelihood of these data is then computed by multiplying probabilities of survival or death over intervals up to each individual's observed death or censoring time:

$$\prod_i \prod_{j: t_j \leq \min(T_i, T_i^*)} \lambda_i(t_j)^{dN_i(t_j)} \exp(-\lambda_i(t_j)(t_j - t_{j-1})),$$

equivalent to a Poisson model for $dN_i(t_j)$ with mean $\lambda_i(t_j)(t_j - t_{j-1})$. The hazard $\lambda_i(t)$ is assumed to be piecewise-constant, changing only at observed death times.

For computational efficiency, the $dN_i(t_j)$ are aggregated over individuals i with identical age, sex and stage at diagnosis, and so identical hazard functions $\lambda_i(t)$. The data are then supplied as dN_{jk} , the number of deaths at each t_j in each of $5 \times 2 \times 4$ covariate categories k , and modelled as

$$dN_{jk} \sim \text{Poisson}(\mu_{jk}), \quad \mu_{jk} = n_{jk} \lambda_{0j} \exp(\beta_k)(t_j - t_{j-1}) \quad (3)$$

where n_{jk} is the number of individuals in category k at risk at time t_{j-1} , λ_{0j} is the baseline hazard, and $\exp(\beta_k)$ is the hazard ratio for covariate category k , equal to the product of the relevant age, sex and stage hazard ratios, with no interactions. We consider three more flexible models for covariate effects: for each alternative, one of the covariates (age, sex or stage) defines stratum-specific baseline hazards λ_{0jk} , and the remaining two covariates are assumed to confer proportional hazards. In a fourth, even more flexible model, $5 \times 2 \times 4$ strata are used representing an interaction between all three covariates.

3.2.1 Predicting from the Bayesian semiparametric model

The model yields posterior distributions of survival probabilities within the period of the observed data. Extrapolations are made by assuming the esti-

mated hazard at the final observed death time remains constant in the future. While a classical Cox regression with a nonparametric baseline hazard estimate could also be used to extrapolate estimated survival with similar assumptions, expressing uncertainty surrounding functions of predicted survival probabilities is more straightforward under a Bayesian semiparametric model. Note that a constant hazard is unrealistic for extrapolating mortality risk to older ages, therefore additional data or expert judgements are required in practice to inform survival in long-term economic models. Recall that the assumptions described in Section 2.3 were adopted when using survival estimates to populate the Markov economic model in this example.

3.2.2 Prior distributions for the baseline hazard

The elements of β are given independent $N(0, 10000)$ priors, while we vary the prior for the baseline hazard λ_{0j} . Although the dataset is large, with 6093 individuals of whom 4295 were observed to die, the semiparametric model requires estimating a hazard for each of 1861 unique times t_j (in days) at which an observed death occurs. In the absence of substantive prior information, we assess the sensitivity of the results to different choices of diffuse prior. We choose independent log-normal or conjugate Gamma priors (Kalbfleisch, 1978) for each λ_{0j} . Two different prior variances, one very large and one mildly informative, are used in each case. Each is essentially uniform in the area supported by the likelihood, but with different degrees of skewness at higher values. A smaller prior variance may yield a better-predicting model, while still giving a reasonable reflection of the prior belief. Figure 3 illustrates the prior distributions for the 5-year survival probability implied by the following four priors which we use:

- (i) $\log(\lambda_{0j}) \sim N(0, 10000)$.
- (ii) $\log(\lambda_{0j}) \sim N(-8, 10)$, so that the 5 year survival probability has a prior mean of 0.5 with a 95% credible interval of (0.01, 0.99).
- (iii) $\lambda_{0j} \sim \text{Gamma}(\mu c, c)$, with prior mean $\mu = 0.00038$ equivalent to a 5 year survival probability of 0.5, and low precision $c = 0.001$.
- (iv) $\lambda_{0j} \sim \text{Gamma}(\mu c, c)$, $\mu = 0.00038$, with $c = 100$, giving a 95% credible interval of (0.000354, 1) for the 5 year survival probability.

Rather than successive hazards being independent, a more realistic reflection of prior opinion is that hazards closer together in time are more highly correlated. Sinha and Dey (1997) discussed priors for hazard functions with

smoothly-varying correlation. Posterior estimation of such models would be computationally difficult in this example, with many distinct survival times, where there would be very long chains of dependence between successive hazards. Instead, we estimated four further models in which λ_{0j} is constrained to be piecewise-constant in time, changing either:

- (i) every month containing an observed death,
- (ii) every year containing an observed death,
- (iii) after every 10 distinct observed death times, or
- (iv) after every 100 distinct observed death times.

The $N(0, 10000)$ prior is used for $\log(\lambda_{0j})$ in each of these four models. In (iii) and (iv), more complex hazard variations can be estimated in periods where the data are more abundant.

3.3 Implementation

The models are fitted by MCMC posterior sampling, implemented in WinBUGS (Lunn *et al.*, 2000). The generalized F and the Prentice (1974) parameterisation of the generalized gamma are not available as built-in distributions in WinBUGS. Therefore these were programmed using the WBDev interface (Lunn, 2003), which allows users to “hard-wire” arbitrary univariate distributions as compiled code. The code to implement these distributions is made available as an appendix to this paper on the publisher’s web site. The simpler Gompertz model is implemented using the “zeroes trick” described in the WinBUGS 1.4 manual.

The semiparametric models are fitted using an MCMC algorithm (Clayton, 1991) implemented in WinBUGS, as described by Demiris and Sharples (2006) and the `leuk` example from the WinBUGS 1.4 Examples Vol. 1. The Markov economic decision model is adapted from the original model in Microsoft Excel used in the published analysis (Speight *et al.*, 2006).

Figure 3: Prior distributions for the baseline hazard, parameterised in terms of the 5 year survival probability. Only a small region is displayed, since they all appear uniform on the remaining region.

4 Model comparison

4.1 Illustration

Figure 1 illustrates the survival curves at the posterior mean baseline log hazard under the parametric models for male patients aged between 50 and 59 diagnosed in cancer stage 1, compared with a Kaplan-Meier estimate, and Figure 2 illustrates corresponding hazard curves for the same patient group. In the first five years, the Weibull model shows the worst fit, while the Gompertz and the generalized F estimates appear the closest to the observed hazard.

Lack of fit could be caused by misspecification of the baseline hazard, by assumptions about how the covariate effects enter each model in Section 3, or by a combination of these. The fit in the first five years is particularly important – while the shape of the survival curves are informed by the entire survival data, only fitted survival probabilities from the first five years, and covariate effects, are used directly for cost-effectiveness prediction. The fitted age effects are used as explained in Section 2.3 for extrapolating survival probabilities as modelled individuals grow older.

Although the semiparametric models we consider have arbitrarily flexible baseline hazards, they are not expected to fit exactly to the Kaplan-Meier estimates for particular subgroups unless the baseline hazard is allowed to be different for every covariate category. Allowing the hazard to change with every observed death provides a very good fit to the early survival data for males, age 50-59 diagnosed in stage 1 (Figure 4), but the simpler model, with hazards changing every year, appeared to fit better to the later survival data. Allowing the hazard to change after every 100 deaths allows more flexible hazard variations in the first five years when the majority of deaths occurred, while not over-fitting the sparse later data. The models with monthly hazards and hazards changing every 10 deaths, not shown in Figure 4, had qualitatively similar fit to this model. The model with the first Gamma prior produces consistently lower estimates of survival than the log-normal prior, and reducing the prior variance gives similar underestimates (not illustrated).

The plot of fitted hazards (Figure 5) may clarify how the fitted survival probabilities are sensitive to the smoothness in the assumed hazard and the prior. The step functions show the posterior mean hazard trajectories under the model with hazard changing every 100 deaths and the model with monthly hazards, compared to a smooth empirical kernel-based hazard estimate (Mueller and Wang, 1994). The grey points are the posterior means under the model with hazard changing at every distinct death time and a $N(0,10000)$ prior. Each point is based on updating the prior with a very small amount of data – between 1 and 17 observed deaths (mean 2.3) at each distinct time. Each “stripe” of points is based on a common number of observed deaths. The hazard increases within a “stripe” since the number at risk decreases through time, while the number of observed deaths remains the same. The influence of the prior seems to lead to underestimates of the hazard under this model between 3 and 6 years, which accumulate to give an overestimate of the survival probability at later times (Figure 4). The survival probabilities at later times under the gamma prior are similarly underestimated.

Figure 6 compares the semiparametric models stratified by covariate categories. In order to retain flexibility while limiting the effective number of parameters, these models were implemented with the baseline hazards changing every 10 observed deaths within a stratum. As expected, the “saturated” model with a different baseline hazard for each age×sex×stage group fits essentially exactly to the Kaplan-Meier estimate. The three models stratified by only one of these covariates each deviate from the data in certain time periods, although these estimates are smoother since there are more observed deaths in each stratum, so that the hazard changes more frequently.

4.2 Formal criteria

To compare the parametric and semiparametric models formally we assess their predictive utility using the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002). This estimates the expected deviance ($-2 \times \log$ -likelihood) for a replicate dataset $\mathbf{y}$ at the expected parameter values $E(\boldsymbol{\theta})$ for a model M_k fitted to data $\mathbf{x}$.

$$E \{-2 \log f(\mathbf{y}|E(\boldsymbol{\theta}), \mathbf{x}, M_k)\} \approx DIC(\mathbf{x}|M_k) = D(\mathbf{x}|\hat{\boldsymbol{\theta}}) + 2p_D$$

where

$$p_D = \overline{D(\mathbf{x}|\boldsymbol{\theta})} - D(\mathbf{x}|\hat{\boldsymbol{\theta}})$$

is the “effective number of parameters”. $\overline{D(\mathbf{x}|\boldsymbol{\theta})}$ is the posterior mean deviance and $D(\mathbf{x}|\hat{\boldsymbol{\theta}}) = \hat{D}$ is the deviance evaluated at the posterior means, an estimate of model fit unpenalized for complexity.

Comparison of DIC between the parametric and semiparametric models is complicated by the different likelihoods used (equations 2 and 3). To elucidate the difference, suppose that in the semiparametric model (3), $\lambda_{0j} = \lambda$, a constant hazard at all times t . The model is then equivalent to an exponential survival model $f()$ in (2) – but while the posterior distributions of λ and β under (2) and (3) are then identical, the log-likelihoods (2) and (3), thus the DIC values, differ by a constant which depends only on the data, not on the parameters. Thus the DIC values obtained by MCMC estimation of the semiparametric model using (3) must be transformed, using this constant, to the scale of the parametric models, to allow model comparison (see the Appendix).

Figure 4: Kaplan-Meier estimate of survival from oral cancer for male, age 50-59, cancer stage 1, and fitted Bayesian semiparametric models

Figure 5: Log hazard of survival from oral cancer for male, age 50-59, cancer stage 1, estimated by an empirical kernel-based method and Bayesian semi-parametric models under three different hazard functions.

Figure 6: Kaplan-Meier estimate of survival from oral cancer for male, age 50-59, cancer stage 1, and fitted Bayesian semiparametric models stratified by covariate categories

For comparison, we computed the $-2\log$ pseudo-marginal likelihood (Geisser and Eddy, 1979, Jackson *et al.*, 2010) using importance sampling (Gelfand and Dey, 1994). This gave very similar model assessments to DIC, though failed to converge for the more complex semiparametric models. The criteria we use assess the expected ability of models to predict beyond the data, in contrast with methods such as posterior model probabilities, Bayes factors and BIC, which aim to identify a relatively simple true model (Jackson *et al.*, 2010).

4.2.1 Accounting for model uncertainty

If different models lead to different inferences, and there is doubt about which model is the most adequate, then model uncertainty should be accounted for in the results. We use a technique described by Jackson *et al.* (2010) to account for model uncertainty between Bayesian models which are evaluated by their predictive ability. A Bayesian bootstrap method yields a probability, under sampling uncertainty, that each model gives the best predictive criterion among the models being compared. These probabilities can be used to form a weighted average of model-specific posterior distributions, which can be used to make inferences which account for uncertainty about model selection.

4.2.2 Results of formal model comparison

Table 1 compares the DIC for all models on the common scale of the parametric likelihood. The generalized F has a substantially better DIC than any of the other parametric and semiparametric survival models. All other parametric models, apart from the Gompertz, are nested within the generalized F.

The semiparametric models fit the observed data better than the parametric models (low $\hat{D}$), as expected, but their complexity (p_D), and hence their predictive ability, varies widely. The semiparametric models with hazard changing at each of 1861 unique survival times have the highest DIC of all models, due to the very high effective number of parameters. By reducing the prior variance of either the log-normal or gamma priors, predictive precision is increased and the DIC is somewhat reduced. However, by reducing the number of change points for the nonparametric baseline hazard, the complexity

of the model, and thus the DIC, is vastly reduced. The model with hazard changing after every 100 deaths has the lowest DIC among the semiparametric models. Allowing the baseline hazard to vary between covariate strata gives small improvements in predictive fit, comparing DIC between the stratified models labelled (c) in Table 1 and the proportional hazard model labelled (b) with hazard changing every 10 death times.

The Bayesian bootstrap algorithm gives a probability of over 0.99 that the generalized F model has the greatest predictive utility, measured by DIC. This indicates that the cost-effectiveness analysis could reasonably be based on this model alone. Note that the dataset included over 6000 cases – with a smaller dataset we would expect more uncertainty about the most adequate model.

5 Results of cost-effectiveness analyses

The Markov economic decision model was evaluated with transition probabilities derived from each of the alternative survival models. The incremental cost $\Delta_C = E(C_2) - E(C_1)$ and incremental QALY $\Delta_Q = E(Q_2) - E(Q_1)$ of screening, compared to no screen, for men aged 50, is presented in Table 1 for all alternative survival models, along with the incremental cost per QALY (incremental cost-effectiveness ratio, or ICER).

Amongst the parametric models, the ICER is lowest (around £22,000) under the Weibull model, which fits poorly. Under the more flexible parametric models (generalized gamma and generalized F), the ICERs are increased to around £24,000. The generalized F is the best-fitting model according to the formal criteria. Elaborating the model further to include a nonparametric baseline hazard does not change the results further, judging from the two models with less informative priors for the baseline hazard (log normal, $\sigma^2 = 10000$, gamma, $c = 0.001$). Reducing this prior variance changes parameter estimates, hence reduces the ICERs by around 10% (log normal, $\sigma^2 = 10$) or around 20% (gamma, $c = 100$). However, when the prior variance is reduced to $\sigma^2 = 10$ under the more parsimonious semiparametric model with hazards changing every 100 deaths, the ICER only changes by less than 3%. These results suggest that to avoid sensitivity to choices of weakly informative priors, some smoothing of the baseline hazard is advised.

The incremental QALYs are lower (thus the ICERs are higher) under the semiparametric models with hazards changing every month or year, compared to the model with hazard changing at each observed death time, probably due to the poorer fit of these models in the first few years after diagnosis when the majority of deaths occurred (Figure 4). When the change points are chosen according to the abundance of data, changing every 10 or 100 deaths, this apparent bias is alleviated and ICERs are closer to those under the most flexible models with hazard changing every death. Furthermore, these models with fewer change points have better predictive fit. Relaxing the proportional hazard assumption for age, sex or stage alone reduces estimated ICERs to under £20,000. However, under a better-fitting model which stratifies by all three covariates, the ICER remains above £20,000.

The decision to adopt a health technology is taken on the basis of comparing the ICER to the maximum amount a provider is willing to pay for one QALY (Claxton *et al.*, 2002), conventionally around £20–30,000 in the UK (National Institute for Health and Clinical Excellence, 2008). Therefore oral cancer screening would be judged borderline cost-effective for 50-year old men from these results. While accounting for model uncertainty would probably not have changed this decision, there are differences in the ICERs of up to 20% between plausible models.

The ICER used for this decision is the ratio of posterior mean expected costs and effects, but the uncertainty about the expected costs and effects can be encapsulated in the *probability of cost-effectiveness* $PCE(30)$, which is the posterior probability that the ICER is less than a willing-to-pay amount of £30,000 (Table 1). This probability is generally higher under models where the ICER is lower, though it also depends on the uncertainty surrounding the estimates. The increased uncertainty in the more complex survival models has a negligible effect on the $PCE(30)$, since these uncertainties are dominated by the uncertainties in the other Markov model parameters described in Section 2.1, and these are identical for different survival models.

Table 1: Measures of overall predictive ability under alternative survival models, with cost-effectiveness of oral cancer screening for 50-year old men.

<i>Bayesian parametric models</i>	Overall measures of adequacy			Cost-effectiveness for 50-year old men			
	$\hat{D}$	p_D	DIC	Incremental cost (£)	Incremental QALY	ICER	PCE(30)
Weibull	21510	9.8	21530	87.7	0.00398	22035	0.510
Log-normal	21078	9.8	21098	87.7	0.00379	23134	0.491
Gompertz	21241	9.8	21260	86.8	0.00341	25455	0.449
Generalized gamma	21077	10.9	21099	87.8	0.00366	23989	0.481
Generalized F	21033	12	21057	87.6	0.00369	23727	0.481
<i>Bayesian semiparametric models</i>							
<i>a) Models with hazard per death: alternative priors on hazard</i>							
Log-normal, $\sigma^2=10000$	20783	1714.1	24211	87.5	0.00360	24308	0.469
Log-normal, $\sigma^2=10$	20637	1600.5	23838	87.7	0.00393	22327	0.501
Gamma, c=0.001	20785	1715.8	24217	87.5	0.00354	24730	0.463
Gamma, c=100	20746	1695.4	24137	88.2	0.00434	20323	0.528
<i>b) Alternative baseline hazards, all under log-normal, $\sigma^2=10000$ prior</i>							
Yearly hazard	22047	25.7	22099	87.0	0.00306	28430	0.404
Monthly hazard	21843	182.5	22208	86.9	0.00314	27672	0.422
Hazard per 10 death times	21681	378	22437	87.0	0.00393	22128	0.501
Hazard per 100 death times	21967	50.9	22069	87.7	0.00406	21610	0.516
<i>c) Stratified hazards, all under log-normal, $\sigma^2=10000$ prior, changing every 10 death times within strata</i>							
By age	21513	413.6	22340	88.3	0.00500	17654	0.590
By sex	21649	400.6	22450	88.0	0.00465	18926	0.567
By stage	21513	416.4	22346	88.3	0.00456	19362	0.558
By age, sex and stage	21418	436	22290	87.4	0.00420	20798	0.528

6 Discussion

The accurate modelling of survival is important in health economic evaluations which estimate cost-effectiveness of interventions over a long-term horizon using short-term data. Diverse parametric and semiparametric models should be considered and their expected predictive ability compared. This can be done by placing all models in a Bayesian framework, which also allows parameter uncertainty to be easily propagated to the estimates of cost-effectiveness. In the example of oral cancer screening, a single four-parameter survival model fitted substantially better than the other candidates, so that cost-effectiveness estimates could reasonably be based on this model alone. In other examples there may be model uncertainty, which can be accounted for by model averaging, in order to improve predictive ability.

Bayesian semiparametric models allow fine variations in the hazard through time and between covariate categories to be estimated, but two problems result if insufficient data are available to capture the detail demanded. Firstly, the resulting estimation uncertainty (Figure 5) leads to poorer predictive ability, and secondly, inferences are sensitive to the choice of apparently-uninformative prior. To address these problems, some smoothing of the hazard is recommended. Our favoured approach of fitting a piecewise-constant hazard function, with change points chosen to ensure a constant number of deaths in each interval, is easy to implement. A more realistic reflection of prior opinion would be to allow hazards to be correlated in time, with the correlation decreasing continuously with increasing time difference (Sinha and Dey, 1997). Standard Gibbs sampling algorithms are inefficient for estimating such autocorrelated processes. Sequential Monte Carlo methods (Doucet *et al.*, 2000) would be more suitable, though these are not currently implemented in general-purpose software. Ideally, the prior should represent substantive previous information, but this is not always possible or practical to obtain, and the process of eliciting such priors would often raise more uncertainties.

6.1 Alternative flexible survival models

Other families of prior used for the baseline survival distribution in Bayesian semiparametric models include Dirichlet processes (Susarla and Van Ryzin, 1976, Ferguson and Phadia, 1979), mixtures of Dirichlet processes (Kuo and Mallick, 1997) and Polya trees (Walker and Mallick, 1999). We expect similar issues of tradeoff between flexibility and parsimony to arise in practice with all these priors, and these can be addressed by the predictive criteria we use.

In the oral cancer application, we only presented cost-effectiveness results for one patient group. In general, decision-making requires assessing heterogeneity in effects and costs between patients, so that the choice of covariates and accurately modelling their effects can be important. In this paper, flexibly-modelled baseline hazards were stratified to account for non-proportionality between covariate groups and potential interactions. While stratified models are straightforward to interpret and implement, flexible models which borrow strength between covariate groups may be more efficient. For example, De Iorio *et al.* (2009) used linear dependent Dirichlet processes to flexibly model both baseline survival and variations between individuals. Regression coefficients could also be explicitly modelled as functions of time. Royston and Parmar (2002) and Abrahamowicz and MacKenzie (2007) described spline-based models including time-varying covariate effects on survival, and Binquet *et al.* (2008) compared covariate selection strategies for these models.

In a Bayesian setting, reversible jump MCMC (Green, 1995, Lunn *et al.*, 2009) allows model uncertainties, such as covariate selection or change-point selection for time-varying parameters, to be accounted for using a single MCMC sampler. Certain reversible jump procedures may be implemented in the WinBUGS software (Lunn *et al.*, 2009). The Bayesian MCMC framework is also suited to implementing frailty models (Clayton, 1991) to account for unexplained heterogeneity in survival between patients.

Further analysis of the oral cancer data might also explore the impact of missing covariates. Age at diagnosis was missing for 3% of the registered cancer cases, which were omitted in the analysis. These cases have mean 5 year survival 0.65, compared to 0.38 for the complete cases. This may be because 50% of these, versus 44% of the complete cases, were diagnosed in cancer stage 1, and 42% were female, compared to 37% of the complete cases. A Bayesian multiple imputation may easily be constructed based on these covariates, but further sensitivity analysis may be required to assess the potential consequence of patients with missing age being younger.

6.2 Other extensions

Extrapolation of survival to inform long-term economic models often requires more than a single source of data. The case registry used in this example had up to 20 years of follow up, but economic models are often informed by randomised controlled trials with typical follow-up of up to 5 years. Mortality data from the general population might then be required to obtain good estimates of longer-term survival. One problem with incorporating population

data is that the patients of interest to the economic decision may have the same survival as the general population only for causes unrelated to the treatment of interest. Data on cause of death may not be available; for example this information was not in the oral cancer dataset available to us. Strong assumptions may then be necessary to characterise the difference between the patients of interest and the general population (Demiris and Sharples, 2006). Expert belief might also be elicited as prior distributions (O’Hagan *et al.*, 2006) to inform either these assumptions, or other assumptions about extrapolation beyond the observed data. Survival data arising from latent multiple causes of death may also, in principle, be modelled flexibly using polyhazard models (Louzada-Neto, 1999) though exploratory use of these models in the oral cancer application did not lead to improvements in fit.

Statistical model uncertainty in this example was addressed by comparing predictive criteria, and model averaging was not judged necessary. DIC is an easily-calculated though approximate criterion, while the alternative pseudo-marginal likelihood relies on fewer approximations though is difficult to compute for complex models. Improvements to the approximation used by DIC have been suggested (Plummer, 2008) for situations when p_D is large compared to the sample size, as in the semiparametric models we used. Jackson *et al.* (2009, 2010) reviewed methods for accounting for model uncertainty among frequentist and Bayesian models in a health economic setting. One particular principle identified, model averaging based on *focused* criteria (Claeskens and Hjort, 2003, Hjort and Claeskens, 2003), might be of benefit in the oral cancer application. Only certain aspects of the survival models, the survival probabilities in the first five years and the covariate effects, are used to populate the oral cancer economic model. An overall model adequacy criterion may not therefore be the most appropriate basis for comparing models. A focused criterion compares the ability of the models to estimate particular parameters of interest, though this may be sensitive to the choice of focus. Further work would be required to extend the deviance-based Bayesian model comparison measures we used to focus on specific parameters (Spiegelhalter *et al.*, 2002, Plummer, 2008).

Appendix

In Section 4.2, we mentioned that the likelihoods used for the parametric and semiparametric models are on different scales. Therefore their DIC values must be transformed onto a common scale for model comparison, so that they represent the same predictive loss function. Here the required scaling factor

is derived. For clarity and without loss of generality, we compare the log-likelihoods under an exponential survival model with a constant hazard.

The log-likelihood corresponding to the Poisson model for survival data formulated as a counting process (equation 3), assuming a constant hazard λ , is

$$\sum_{j,k} dN_{jk} \log(n_{jk}(t_j - t_{j-1})) + \sum_{j,k} dN_{jk} \log(\lambda) - \lambda \sum_{j,k} n_{jk}(t_j - t_{j-1}) - \sum_{j,k} \log(dN_{jk}!)$$

Since $dN_{jk} = 0$ or 1 , $\log(dN_{jk}!) = 0$. The standard survival log-likelihood (equation 2) under an exponential model with $f(t|\lambda) = \lambda \exp(-\lambda t)$ and $1 - F(t|\lambda) = \exp(-\lambda t)$ is

$$\sum D_i \log(\lambda) - \lambda(\sum T_i + \sum T_i^*)$$

Since the number of observed deaths $\sum D_i = \sum_{j,k} dN_{jk}$, and $\sum T_i + \sum T_i^* = \sum_{j,k} n_{jk}(t_j - t_{j-1})$, the total time that all individuals are under observation, the standard log-likelihood differs from the counting process log-likelihood by a constant $\sum_{j,k} dN_{jk} \log(n_{jk}(t_j - t_{j-1}))$ which depends only on the data, not the parameters. Therefore $-2 \sum_{j,k} dN_{jk} \log(n_{jk}(t_j - t_{j-1}))$ should be subtracted from the $\hat{D}$, hence the DIC, calculated for the semiparametric model to allow comparison with the parametric model.

References

- Abrahamowicz, M. and MacKenzie, T. A. (2007) Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine*, **26**, 392–408.
- Ades, A. E., Claxton, K. and Sculpher, M. (2006) Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. *Health Economics*, **15**, 373–381.
- Binquet, C., Abrahamowicz, M., Mahboubi, A., Jooste, V., Faivre, J., Bonithon-Kopp, C. and Quantin, C. (2008) Empirical study of the dependence of the results of multivariable flexible survival analyses on model selection strategy. *Statistics in Medicine* **27**, 6470–6488.
- Briggs, A., Sculpher, M. and Claxton, K. (2006) *Decision modelling for health economic evaluation*. Handbooks in Health Economic Evaluation. Oxford: Oxford University Press.

- Claeskens, G. and Hjort, N. (2003) The focused information criterion (with discussion). *Journal of the American Statistical Association*, **98**, 900–945.
- Claxton, K. (1999) The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics* **18**, 341–364.
- Claxton, K., Sculpher, M. and Drummond, M. (2002) A rational framework for decision-making by the National Institute for Clinical Excellence (NICE). *Lancet*, **360**, 711–715.
- Clayton, D. G. (1991) A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, **47**, 467–485.
- Cox, C. (2008) The generalized F distribution: An umbrella for parametric survival analysis. *Statistics in Medicine*, **27**, 4301–4312.
- Cox, C., Chu, H., Schneider, M. F. and Muoz, A. (2007) Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, **26**, 4352–4374.
- De Iorio, M., Johnson, W. O., Müller, P. and Rosner, G. L. (2009) Bayesian nonparametric nonproportional hazards survival modelling. *Biometrics*, **65**, 762–771.
- Demiris, N. and Sharples, L. D. (2006) Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. *Statistics in Medicine*, **25**, 1960–1975.
- Doucet, A., Godsill, S. and Andrieu, C. (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, **10**, 197–208.
- Ferguson, T. S. and Phadia, E. G. (1979) Bayesian nonparametric estimation based on censored data. *Annals of Statistics*, **7**, 163–186.
- Geisser, S. and Eddy, W. (1979) A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.
- Gelfand, A. and Dey, D. (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 501–514.
- Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

- Hjort, N. and Claeskens, G. (2003) Frequentist model average estimators (with discussion). *Journal of the American Statistical Association*, **98**, 879–899.
- Jackson, C. H., Sharples, L. D. and Thompson, S. G. (2010) Structural and parameter uncertainty in Bayesian cost-effectiveness models. *Applied Statistics*, **59**, 233–253.
- Jackson, C. H., Thompson, S. G. and Sharples, L. D. (2009) Accounting for uncertainty in health economic decision models by using model averaging. *Journal of the Royal Statistical Society, Series A*, **172**, 383–404.
- Kalbfleisch, J. D. (1978) Bayesian analysis of survival distributions. *Journal of the Royal Statistical Society, Series B*, **40**, 214–221.
- Kuo, L. and Mallick, B. (1997) Bayesian semiparametric inference for the accelerated failure-time model. *Canadian Journal of Statistics*, **25**, 457–472.
- Lin, D. Y. (2000) Linear regression analysis of censored medical costs. *Biostatistics* **1**, 35–47.
- Lin, D. Y., Feuer, E. J., Etzioni, R. and Wax (1997) Estimating medical costs from incomplete follow-up data. *Biometrics* **53**, 419–434.
- Louzada-Neto, F. (1999) Polyhazard models for lifetime data. *Biometrics*, **55**, 1281–1285.
- Lunn, D. J. (2003) WinBUGS Development Interface (WBDev). *ISBA Bulletin* **10**
- Lunn, D. J., Best, N. and Whittaker, J. C. (2009) Generic reversible jump MCMC using graphical models. *Statistics and Computing*, **19**, 395–408.
- Lunn, D. J., Thomas, A., Best, N. G. and Spiegelhalter, D. J. (2000) WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.
- Mueller, H. G. and Wang, J. L. (1994) Hazard rates estimation under random censoring with varying kernels and bandwidths. *Biometrics*, **50**, 61–76.
- National Institute for Health and Clinical Excellence (2008) *Guide to the methods of technology appraisal*. London: National Institute for Health and Clinical Excellence.

- Nelson, C. L., Sun, J. L., Tsiatis, A. A. and Mark, D. B. (2008) Empirical estimation of life expectancy from large clinical trials: use of left-truncated, right-censored survival analysis methodology. *Statistics in Medicine*, **27**, 5525–5555.
- O’Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J. and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts’ Probabilities*. Statistics in Practice. Wiley.
- Plummer, M. (2008) Penalized loss functions for Bayesian model comparison. *Biostatistics* **9**, 523–539.
- Prentice, R. L. (1974) A log gamma model and its maximum likelihood estimation. *Biometrika*, **61**, 539–544.
- (1975) Discrimination among some parametric models. *Biometrika*, **62**, 607–614.
- Royston, P. and Parmar, M. (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* **21**, 2175–2197.
- Sinha, D. and Dey, D. K. (1997) Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, **92**, 1195–1212.
- Speight, P. M., Palmer, S., Moles, D. R., Downer, M. C., Smith, D. H., Henriksson, M. and Augustovski, F. (2006) The cost-effectiveness of screening for oral cancer in primary care. *Health Technology Assessment*, **10**.
- Spiegelhalter, D. J. and Best, N. G. (2003) Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine*, **22**, 3687–3709.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.
- Susarla, V. and Van Ryzin, J. (1976) Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71**, 897–902.
- Walker, S. and Mallick, B. K. (1999) A Bayesian semiparametric accelerated failure time model. *Biometrics*, **55**, 477–483.