

# Flexible Cure Rate Modelling Under Latent Activation Schemes

FREDA COONER, SUDIPTO BANERJEE, BRADLEY P. CARLIN AND  
DEBAJYOTI SINHA <sup>1</sup>

*Division of Biostatistics, School of Public Health, University of Minnesota,  
Mayo Mail Code 303, Minneapolis, Minnesota 55455-0392, U.S.A.*

and

*Department of Biostatistics and Bioinformatics, Medical University of South Carolina,  
135 Cannon Street, Charleston, South Carolina 29425, U.S.A.*

Correspondence author: Suddipto Banerjee  
telephone: (612) 624-0624  
fax: (612) 626-0660  
email: [suddiptob@biostat.umn.edu](mailto:suddiptob@biostat.umn.edu)

December 18, 2006

---

<sup>1</sup>Freda Cooner is Graduate Assistant, Suddipto Banerjee is Assistant Professor of Biostatistics, and Bradley P. Carlin is Professor of Biostatistics and Mayo Professor in Public Health at the Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, 55455. Debajyoti Sinha is Professor, Department of Biostatistics and Bioinformatics, Medical University of South Carolina, Charleston, SC 29425. The work of the first three authors was supported in part by NIH grants 1-R01-CA95995 and 1-R01-CA112444. Work of Dr. Sinha was supported by NCI grant 9-R01-CA69222. The authors thank the Editor, Associate Editor and the three referees for several suggestions and also Professor Minghui Chen of the University of Connecticut, Storrs, for useful discussions.

# Flexible Cure Rate Modelling Under Latent Activation Schemes

## Summary

With rapid improvements in medical treatment and health care, many data sets dealing with time to relapse or death now reveal a substantial portion of patients who are *cured* (that is, who never experience the event). Extended survival models called *cure rate models* account for the probability of a subject being cured and can be broadly classified into the classical mixture models of Berkson and Gage (1952; “BG type”) or the stochastic tumor models pioneered by Yakovlev (1996) and extended to a hierarchical framework by Chen, Ibrahim and Sinha (1999; “YCIS type”). Recent developments in Bayesian hierarchical cure models have evoked significant interest regarding relationships and preferences between these two classes of models. Our present work proposes a unifying class of cure rate models that facilitates flexible hierarchical model-building while including both existing cure model classes as special cases. This unifying class enables robust modelling by accounting for uncertainty in underlying mechanisms leading to cure. Issues such as regressing on the cure fraction and propriety of the associated posterior distributions under different modelling assumptions are also discussed. Finally, we offer a simulation study and also illustrate with two data sets (one on melanoma and the other on breast cancer) that reveal our framework’s ability to distinguish among underlying mechanisms that lead to relapse and cure.

*Key Words:* Survival analysis; Cure rate models; Cure fractions; Latent activation schemes; Moment generating functions; Bayesian hierarchical models; Markov Chain Monte Carlo algorithms.

# 1 Introduction

With significant progress in the medical and health sciences, scientists and health professionals increasingly encounter data sets where some patients are expected to be *cured*. The concept is subtle because a subject surviving the time window of the experiment is considered “censored”, while he/she is cured if he/she will *never* fail from the event under study. We can never “observe” a cure because of a finite monitoring time, yet interest in formulating models accounting for cure is ripe for evaluating prognosis in potentially fatal diseases (e.g. cancers). Here traditional parametric survival models such as Weibull or Gamma (Cox and Oakes, 1984), which do not account for possibility of cure, may prove inadequate.

To address this conceptually challenging problem, *cure rate models* that incorporate a cure fraction have been proposed, starting with the mixture models by Berkson and Gage (1952) (we refer to as the BG model), with subsequent investigations by Farewell (1982, 1986), Goldman (1984) and Ewell and Ibrahim (1997) among others. More recent developments include cancer models (Yakovlev et al., 1993; Yakovlev, 1996; Yakovlev and Tsodikov, 1996) that propose mechanisms that bring about metastasis. Being motivated from mechanistic considerations of the particular disease, these are referred to as *mechanistic models*. Based upon them, Chen, Ibrahim and Sinha (1999) proposed an alternative class of cure models, which we call the YCIS models (see also Ibrahim et al., 2001; Tsodikov et al., 2003; and Banerjee and Carlin, 2004). These models assume that an individual is at risk of failure if he/she is exposed to at least one (perhaps several) of the so-called *latent factors* or *latent risks*. Otherwise, the individual is not at risk and is considered *cured*. Failure is observed when one (or some) of these latent factors become *activated*.

Existing cure models are, almost exclusively, modifications of either the BG (see e.g. Kuk and Chen (1992); Taylor (1995); Sy and Taylor, 2000; Li and Taylor, 2002) or the YCIS models (see e.g. Tsodikov et al., 2003). Although insight into the biological processes leading to disease manifestation often sheds some light upon appropriate model assumptions,

controversy persists on whether to use the BG model or the YCIS model (see, e.g., Yin and Ibrahim, 2006, Tsodikov, 2003, Tucker and Taylor, 1996, and Tucker, Thames and Taylor, 1990). The problem is only exacerbated in realistic settings as survival data sets are unlikely to identify (and hence validate) underlying mechanistic model parameters. In this vein, both the BG and YCIS models are somewhat inflexible as they do not account for the uncertainty in the transition from the underlying process to disease manifestation. This can, in turn, lead to poorer fits to the data.

Our current work proposes a class of hierarchical models that address the above issues by considering stochastically modelling the process of disease manifestation. We achieve greater modelling flexibility by stochastically modelling the ordered sequence of latent events that lead to disease manifestation. We show that this framework leads to identifiable models, of which the BG and YCIS are special instances (Section 2). We also consider the identifiability of regression parameters and the resulting hazard structures (Section 3). While assumptions made on the latent event distributions cannot be verified a priori, we assess their relative merits a posteriori using a posterior predictive criterion (Section 4). Section 5 illustrates model performance with a simulated experiment and then with survival data on two forms of cancer (melanoma and breast), indicating potential advantages of our approach. The melanoma data comes from a clinical trial by the Eastern Cooperative Oncology Group (ECOG; see Kirkwood et al., 1996) and has been analyzed, for instance, in Chen et al. (1999), while the breast cancer data is a sample from the SEER databases ([www.seer.cancer.gov](http://www.seer.cancer.gov)). Finally, Section 6 offers a summary and indicates future research areas.

## 2 Models with Latent Factors and Activation Schemes

Cure models based upon latent activation schemes assume that an *observed* failure time  $T$  (when the individual *fails*) is generated by the *latent* event times (activation times for the  $N$  latent factors)  $Y_1, \dots, Y_N$ . If  $N = 0$  then the individual is not exposed to any of the

latent factors and is considered *cured* (not at risk of failure) and  $T = \infty$ . For a given  $N$ , the  $\{Y_k\}_{k=1}^N$  are assumed to be independently distributed with a common *survival function*  $P(Y > t) = S(t) = 1 - F(t)$  independent of  $N$ .

## 2.1 Hierarchical activation schemes

A crucial issue for further development is relating the latent  $\{Y_k\}_{k=1}^N$  to the observed  $T$ . Generally, if we assume that  $r$  out of  $N$  latent factors need to activate for the subject to fail, then  $T = Y_{(r)}$  for  $1 \leq r \leq N$  where  $Y_{(1)} < \dots < Y_{(N)}$  are the ordered  $Y_k$ . Thus,  $r$  is a *threshold* variable whose biological interpretation will be addressed later. It can be a fixed constant, a function of  $N$ , or even be treated as random by specifying a conditional distribution for  $r$  given  $N$  (denoted as  $r | N$ ).

The conditional distribution of  $T$  given  $N$  and  $r$  can be written as

$$P(T \geq t | N, r) = 1(N = 0) + IB(S(t); N - r + 1, r)1(N \geq r \geq 1), \quad (1)$$

where  $1(\cdot)$  is the indicator function and

$$IB(S(t); N - r + 1, r) = \sum_{j=0}^{r-1} \binom{N}{j} [F(t)]^j [S(t)]^{N-j} = N \binom{N-1}{r-1} \int_0^{S(t)} u^{N-r} (1-u)^{r-1} du$$

is the incomplete Beta function. Derivation of (1) above follows easily from a standard result on order statistics using the binomial theorem (see, e.g., Rao, 1973, p.215). The unconditional survival function of  $T$ , say  $S^*(t)$ , is given in terms of the latent distribution as

$$S^*(t) = E_{N,r}[P(T \geq t | N, r)] = P(N = 0) + E_{N,r}[IB(S(t); N - r + 1, r)1(N \geq r \geq 1)], \quad (2)$$

where the expectation  $E_{N,r}$  is taken over the joint distribution of  $(N, r)$ . Note that  $S^*(t)$  exists and is bounded between 0 and 1 for any valid distribution of  $(N, r)$  restricted to  $N \geq r \geq 1$ . Also, since  $\lim_{t \rightarrow \infty} S(t) = 0$ , we have  $S^*(t)$  is *improper* whenever  $\lim_{t \rightarrow \infty} S^*(t) = P(N = 0) > 0$ . Indeed, then  $P(N = 0)$  is the probability of a person being cured, hence called the *cure fraction*, and depends only upon the distribution of  $N$ , irrespective of what

$r$  is. Although  $S^*(t)$  is improper, we can still consider the hazard  $h^*(t)$  such that  $h^*(t)dt \approx P(T \in [t, t + dt) | T > t)$  and the corresponding *improper* density

$$f^*(t) = f(t)E_{(N,r)} \left[ N \binom{N-1}{r-1} [S(t)]^{N-r} [F(t)]^{r-1} \mathbf{1}(N \geq r \geq 1) \right], \quad (3)$$

where  $f(t)$  is the proper density of  $F(t)$ . The variable  $N$  can never be *observed* and must be modelled using a probabilistic assumption. On the other hand, the scientific context can sometimes suggest a fixed value or a function of  $N$  for  $r$  in (2).

A flexible hierarchical modelling approach specifies the joint distribution of  $r$  and  $N$  through a marginal specification for  $N$  and a conditional distribution for  $r$  given  $N$ . The conditional distribution of  $T$  given  $N$  and  $r$  remains as in (1), and  $S^*(t)$  is easily derived from (2) to yield

$$S^*(t) = P(N = 0) + E_N \left[ \mathbf{1}(N \geq 1) N \int_0^{S(t)} E_{r|N} \left[ \binom{N-1}{r-1} u^{N-r} (1-u)^{r-1} \right] du \right]. \quad (4)$$

Practical modelling further benefits by characterizing cure models that are identifiable from the data (see, e.g., Li, Taylor and Si, 2001). In our hierarchical context, we seek conditions when  $\theta$  is identifiable under a non-informative prior  $g(\theta)$ . This is relevant only when some individuals experience the event (i.e., not all are censored) and amounts to the finiteness of  $I_{f^*}(t) = \int_0^\infty g(\theta) f^*(t) d\theta$ . Identifiability permits regression parameters with flat priors through a link on  $\theta$ . For instance, assuming the marginal specification  $N \sim Po(\theta)$  with mean  $\theta$ , we obtain the cure fraction  $P(N = 0) = \exp(-\theta)$ . The objective (non-informative) scale-invariant prior sets  $g(\theta) = 1/\theta$ , which amounts to a flat prior on  $\log(\theta)$ . Under this setup, we can prove (see item 1 in the Appendix) that  $\theta$  is identifiable whenever  $r = 1$  or  $r = N$ . (In fact, we prove this identifiability more generally whenever  $r$  is fixed, or  $N - r$  is fixed.) Setting  $r = 1$  implies that activating any one of the  $N$  latent events bring about the observed failure, that is,  $T = \min_{1 \leq k \leq N} Y_k$  (*first-activation or FA scheme*). On the other hand, setting  $r = N$  implies  $T = \max_{1 \leq k \leq N} Y_k$  and delivers a different scheme (*last-activation or LA scheme*). A well-identified *Mixture scheme* is obtained immediately with  $r|N$  having

positive mass on  $\{1, N\}$  with probabilities  $1 - \pi$  and  $\pi$  respectively. Indeed, this model is identifiable for any  $\pi \in (0, 1)$  whose posterior distribution will indicate the data’s support for first or last activation.

More generally, with  $r|N$  having positive mass on values other than 1 and  $N$ ,  $\theta$  may no longer be identifiable. A particularly interesting result follows by specifying  $r|N$  as *DiscreteUnif*(1,  $N$ ) (discrete uniform). Then (4) simplifies to (see item 2 in the Appendix)

$$S^*(t) = P(N = 0) + S(t)(1 - P(N = 0)), \quad (5)$$

which is a classical BG-type model with cure fraction  $P(N = 0)$  depending upon the distribution of  $N$ . With  $N \sim Po(\theta)$  and  $g(\theta) = 1/\theta$ , we find  $\theta$  is not identifiable. The same is true for  $\mu$  with  $g(\mu) \propto 1$  in the BG model, where  $N \sim Ber(e^\mu/(1 + e^\mu))$  with cure fraction  $1/(1 + e^\mu)$ .

The identifiability of the cure fraction is determined by the nature of the mixing in (2). The question of characterizing non-degenerate probability distribution for  $r$  given  $N \sim Po(\theta)$  on the support of  $\{1, \dots, N\}$  that leads to an identifiable  $\theta$  is more challenging (see item 3 in the Appendix). For instance, with  $N \sim Po(\theta)$ , we may specify  $r - 1|N$  as *Bin*( $N - 1, \pi$ ) (the HA-Bin model). Although  $\theta$  is weakly identifiable for  $\pi \in (0, 1)$ , the hyperparameter  $\pi$  can, at least conceptually, be assigned a suitable hyperprior (e.g. a uniform or beta distribution) with Markov chain Monte Carlo (MCMC) techniques being employed for their estimation. In fact, this is a clear advantage of the Bayesian mechanism as a richer framework is obtained without impairing model identifiability. Note that the hierarchical model includes the first and last activation schemes as special cases when  $\pi = 0$  and  $\pi = 1$  respectively and its posterior distribution may indicate favoring first or last activation. However, the data often does not inform as much about this hyperparameter and the posterior of  $\pi$  might be inconclusive. A more practical approach fits separate models (say fixing  $\pi = 0, 1$ ) and uses statistical model comparison metrics to assess “better” performance. We discuss such a practical strategy in Section 4. Alternatively, as discussed earlier, the mixing probabilities

in the mixture model are better identified and can be used to test for activation schemes (see Section 5.1).

## 2.2 “First-activation” scheme

The first-activation scheme sets  $r = 1$  in the hierarchical setting assuming a single activation leads to observed failure. A biological model for patients diagnosed with cancer assumes  $N$  to be the number of metastasis-competent cells (clonogens) that are in an irreversible process towards metastasis, and  $Y_k$  is the time for the  $k^{th}$  clonogen to produce “detectable” tumor. This occurs as soon as *any one* of the clonogens metastasize so that  $T = \min_{1 \leq k \leq N} Y_k$  and (1) simplifies to  $P(T \geq t|N) = 1(N = 0) + [S(t)]^N 1(N \geq 1)$ . The YCIS models of Chen et al. (1999) are a special instance of this scheme with  $N \sim Po(\theta)$ . See Hanin et al. (2001) for arguments supporting this assumption for post-radiation clonogens in a patient’s body.

Several authors, including Tucker et al. (1990) and Tucker and Taylor (1996), have questioned the universal validity of the Poisson assumption and this scheme. From a modelling standpoint,  $N$  can have any finite-mean integer-valued distribution with moment generating function  $m(t) = E[\exp(tN)]$  and a cure fraction given by  $P(N = 0) = m(-\infty)$ . The marginal distribution of  $T$  is then obtained as:

$$S^*(t) = E_N[P(T \geq t|N)] = m[\log S(t)] \quad (6)$$

For example, in the YCIS model with  $N \sim Po(\theta)$ , we have  $m(t) = \exp[-\theta(1 - e^t)]$  which yields  $S^*(t) = \exp(-\theta(1 - S(t)))$ , having cure fraction  $\exp(-\theta)$ . Note that (6) offers flexibility to go beyond the YCIS and BG models and model the biology of disease occurrence/relapse suitable for the application in hand.

In the BG-type models  $N$  is binary with only one latent dominant event (say, one dominant metastasis competent tissue-mass), so  $N \sim Ber(\theta)$  with  $\theta$  being the probability of an activation and  $m(t) = 1 - \theta(1 - e^t)$ . Here  $S^*(t) = 1 - \theta(1 - S(t))$ , with cure fraction

$1 - \theta$ . An alternative  $K$ -site cancer model discussed by Gail et al. (1980) suggests that for each patient an unknown  $N$  out of  $K$  dominant mutation sites within a disease location get mutated, with  $K$  fixed a priori from scientific considerations. This implies a  $BG(K)$  model where  $N \sim \text{Bin}(K, \theta)$  and  $m(t) = (1 - \theta(1 - e^t))^K$  yielding  $S^*(t) = (1 - \theta(1 - S(t)))^K$  with cure fraction  $(1 - \theta)^K$ . Ideally, one could stochastically model  $K$  as well although this would be computationally burdensome and would likely entail reversible jump MCMC (see e.g. Carlin and Louis, 2000, Section 6.4.3). Although the choice of  $K$  is subjective, and the model's performance can be sensitive to it, the flexible hierarchical schemes add robustness to the model's performance.

Yet another specification of a mechanistic model of cancer considers a biological model (see Moolgavkar, Luebeck, and De Gunst, 1990), where exposure to genetic damage causes the patient's body to produce  $N$  mutated cells/tissues before activating the immune system. If every new mutation (initiation) produces, with probability  $1 - \theta$ , an effective immune response capable of destroying the last mutated tissue/cell to halt the mutation process, then  $N \sim \text{Geo}(\theta)$ , a geometric distribution with mean  $\theta/(1 - \theta)$  and mgf  $m(t) = (1 - \theta)/(1 - \theta e^t)$ . With  $\{Y_k\}_{k=1}^N$  now being the i.i.d. promotion times of  $N$  mutated cells left in the body, a first-activation scheme results in  $S^*(t) = (1 - \theta)/(1 - \theta S(t))$  with a cure fraction of  $1 - \theta$ .

### 2.3 “Last-activation” scheme

The mechanistic framework of the first activation scheme, originally argued by Yakovlev et al. (1993), can be questioned for certain diseases. Alternatively,  $N$  can be the number of latent factors that must *all* be activated for failure. For instance, for certain situations (e.g. types of cancer), the underlying mechanism involves metastasis competent mutation of the tissue mass that generates the first primary. During the short interval of this mutation, the patient's immune response also gets activated to initiate  $N$  immune responses to this mutation. When there is no mutation, there is no immune response (i.e.,  $N = 0$ ). Each

immune response is a latent factor capable of resisting disease manifestation or death until its promotion (destruction) time  $Y_k$ . Failure (detectable disease/death) occurs after all the  $N$  factors have been activated, so the observed failure time is  $T = \max Y_k$ ,  $k = 1, \dots, N$ , a special case of (1) with  $r = N$ . Here also we discuss Poisson, Bernoulli, binomial and geometric specifications for  $N$ .

The conditional distribution of  $T$  given  $N$  is now easily expressed in terms of  $F(t) = 1 - S(t)$  as

$$S^*(t) = 1 + m(-\infty) - m[\log F(t)] \quad (7)$$

with cure fraction  $S^*(\infty) = m(-\infty)$ . Analogous to (6), (7) characterizes  $S^*(t)$  for last activation in terms of  $m(t)$ , and helps in understanding regression (see Section 3). Also,  $S^*(t)$  in (7), though different from (6), tends to the same cure-fraction,  $m(-\infty) = P(N = 0)$ . In particular, when  $N \sim Po(\theta)$ , we have  $S^*(t) = 1 + \exp(-\theta)(1 - \exp(\theta F(t)))$  which is different from that under first activation, but approaches the same cure fraction,  $\exp(-\theta)$ . For  $N \sim Geo(\theta)$ , using the mgf of the geometric distribution we obtain, after some algebra,  $S^*(t) = (1 - \theta) + [\theta^2 F(t)/(1 - \theta F(t))]$ , with  $1 - \theta$  as the cure fraction for  $0 < \theta < 1$ . The biological motivation for the geometric model is similar to what we described for the first activation, except that now failure occurs after the last promotion time for the mutated cell.

### 3 Regression in cure models

Equation (3) shows  $f^*(t)$  in a generally complex relationship with the latent density function, but the hazard function  $h^*(t)$  for the first and last activation schemes is given in terms of  $m'(t) = \frac{d}{dt}m(t)$  as

$$h_{FA}^*(t) = -\frac{m'[\log\{S(t)\}]}{m[\log\{S(t)\}]}h(t) \quad \text{and} \quad h_{LA}^*(t) = -\frac{m'(\log F(t))f(t)}{[1 + m(-\infty) - m(\log F(t))]F(t)}, \quad \text{respectively.}$$

Forming  $g(t) = h_{FA}^*(t)/h_{LA}^*(t)$ , we see that  $g(t)$  depends upon both  $S(t)$  and  $\theta$ . With  $N \sim Po(\theta)$  we find  $g(t) = (1 + e^{-\theta})e^{\theta S(t)} - 1$ , a *decreasing* function in  $t$  satisfying  $\lim_{t \rightarrow \infty} g(t) = e^{-\theta}$

(the cure fraction) and  $\lim_{t \rightarrow 0} g(t) = e^\theta$ . Indeed, this behavior occurs for any latent survival distribution and suggests that activation schemes are identifiable from data with specific hazard structures.

In the following, we specify the latent survival function  $S(t)$  using a two-parameter Weibull distribution  $Weib(\rho, \eta)$  with survival function  $S(t) = \exp(-t^\rho e^\eta)$ . Cure rate models can incorporate a regressor vector  $\mathbf{x}$  in the Weibull scale parameter  $\eta = \eta(\mathbf{x})$ , ensuring both a proportional hazards structure as well as an accelerated failure time model for the latent activation time. However, this does not necessarily translate to  $S^*(t|\mathbf{x})$ . Whenever the regressor  $\mathbf{x}$  is modelled via the activation time using an accelerated life model structure as  $S(t|\mathbf{x}) = S_0(t\phi(\mathbf{x}))$ , the corresponding cure rate model under the first activation scheme becomes an accelerated life model with  $S^*(t|\mathbf{x}) = m[\log\{S_0(t\phi(\mathbf{x}))\}] = S_0^*(t\phi(\mathbf{x}))$  with  $S_0^*(t) = m[\log\{S_0(t)\}]$  for any  $m(u)$  that is free of  $\mathbf{x}$ . We can show, using the uniqueness of the mgf, that when  $S(t)$  does not depend on  $\mathbf{x}$  the cure model *cannot* have an accelerated failure time structure. Similarly when  $S(t|\mathbf{x}) = S_0(t\phi(\mathbf{x}))$  and  $S^*(t|\mathbf{x})$  has an accelerated failure time distribution, the mgf  $m(u|\mathbf{x}) = E[\exp(uN)|\mathbf{x}]$  is free of  $\mathbf{x}$ . These results characterize the activation time distribution and the distribution of  $N$  when the cure rate model under the first activation scheme follows an accelerated failure time structure.

For the special case of the YCIS model, when  $N$  is Poisson with mean  $\theta = \theta(\mathbf{x})$  and  $S(t)$  is free of  $\mathbf{x}$ , we obtain a proportional hazards structure for  $h^*(t|\mathbf{x}) = \theta(\mathbf{x})f(t)$ . On the other hand a model with either  $N \sim Bin(K, \theta)$  or  $N \sim Geo(\theta)$  does not render a proportional hazards structure, irrespective of whether we regress through  $\theta(\mathbf{x})$  or  $\eta(\mathbf{x})$ . In fact, using the uniqueness property of the Poisson mgf, we can show that when  $S(t)$  is free of  $\mathbf{x}$  and  $h^*(t|\mathbf{x})$  has a proportional hazards structure then  $N | \mathbf{x} \sim Po(\theta(\mathbf{x}))$  with mean  $\theta(\mathbf{x})$  (a function of  $\mathbf{x}$ ). Therefore when the available observed data informs about the form of  $S^*(t|\mathbf{x})$ , one can deduce the distributional structures of the corresponding latent activation times and  $N$ .

Under the last activation scheme, we do not have a proportional hazards structure with

Poisson, binomial or geometric mgf's and the hazards structures are more complex, and perhaps less intuitive, than for the first activation setting. Nevertheless, our earlier discussions on identifiability (see Section 4 and Appendix, item 1) show that with a Poisson mgf we have identifiable regression in  $\log(\theta)$  under fairly general conditions, while for Bernoulli or Binomial mgf's, regressing on  $\theta$  with flat priors on regression coefficients produce improper posteriors. On the other hand, regressing through  $\eta$  in the latent Weibull distribution is always valid, yielding proper posteriors even with improper priors with an appropriate link.

To see how the foregoing assumptions lead to flexible classes of models, consider the setting with  $I$  subjects, where the  $i^{\text{th}}$  subject has a vector of risk factors (covariates)  $\mathbf{x}_i$ . Letting  $\eta_i$  and  $\theta_i$  be the respective analogues of the foregoing  $\eta$  and  $\theta$  for subject  $i$ , we introduce regression either as  $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$  or as  $\theta_i = g(\mathbf{x}_i^T \boldsymbol{\beta})$ , where  $g$  is a suitable link mapping onto the positive real line (recall that  $\theta_i$  has positive support). We investigate cure rate models based upon the activation scheme, the distribution of  $N$ , and the regression, particularly considering each of the following models under different activation schemes:

$$\text{Model 1(a): } \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}; \theta_i = \theta; N \sim Po(\theta);$$

$$\text{Model 1(b): } \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}; \theta_i = \theta; N \sim Ber(\theta);$$

$$\text{Model 1(c): } \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}; \theta_i = \theta; N \sim Bin(K, \theta) \text{ (} K \text{ known);}$$

$$\text{Model 1(d): } \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}; \theta_i = \theta; N \sim Geo(\theta);$$

$$\text{Model 2: } \log(\theta_i) = \mathbf{x}_i^T \boldsymbol{\beta}; \eta_i = \eta; N \sim Po(\theta)$$

Models 1(a)–1(d) regress on the latent Weibull mean, ensuring proper models irrespective of the distribution of  $N$ , but keep a constant cure parameter  $\theta$ . Model 2 regresses on the cure-parameter  $\theta$  and renders proper models for  $N \sim Po(\theta)$ , but restricts the latent survival distribution to a common Weibull mean  $\eta$ . Note that all activation schemes coincide for Model 1(b) ( $N$  binary). In models 1(a)–(d), higher values of  $\eta_i$  indicate longer latent event times and hence prolonged survival, so covariates influence the latent factors that cause failure, while in model 2 they influence the cure fraction directly. In our subsequent analysis,

we will compare performances of these models under first, last and hierarchical activation schemes.

The covariates that affect the latent event-times may also affect probability of cure. In fact, a natural question that arises here is that of feasibility and identification of regressions in *both*  $\eta$  and  $\theta$ . Such settings will not necessarily yield proportional hazards, proportional odds or accelerated failure structures in  $S^*(t)$  and might lead to non-identifiable regression models for the marginal survival function. In a different context, this is a more relevant issue in developing spatial models for geographically referenced cure data (see, e.g., Banerjee and Carlin, 2004) with spatially informative priors and merits separate investigation.

## 4 Bayesian estimation and model comparisons

### 4.1 The likelihood and posterior

For the  $i^{th}$  individual, our observed data consists of  $D_i = (t_i, \nu_i, \mathbf{x}_i)$ , where  $t_i$  is the observed failure time,  $\nu_i$  is the failure indicator, and  $\mathbf{x}_i$  is a set of covariates. We collect the model-specific parameters (and hyperparameters) into  $\Omega_i$ . Generally we write  $\Omega_i = (\eta_i(\boldsymbol{\beta}), \theta_i(\boldsymbol{\beta}), \rho, \psi)$  to denote regression in either  $\eta_i$  or  $\theta_i$  (but not both), and  $\psi$  as the set of other hyperparameters that may arise in specific models. Suppressing the dependence of the latent distributions on  $\eta_i$  and  $\rho$  and implicitly assuming  $N \geq r \geq 1$ , subject  $i$ 's contributes to the data likelihood (in a right-censored setting) is

$$L(D_i; \Omega_i, N_i, r_i) = P(T \geq t_i | N_i, r_i)^{1-\nu_i} \times \left( -\frac{d}{dt_i} P(T \geq t_i | N_i, r_i) \right)^{\nu_i},$$

where  $P(T \geq t_i | N_i, r_i)$  is as in (1) and  $-\frac{d}{dt_i} P(T \geq t_i | N_i, r_i) = N \binom{N-1}{r-1} [S(t_i)]^{N_i-r_i} [F(t_i)]^{r_i-1} f(t_i)$ .

The posterior distribution of  $\{\Omega_i\}$  (marginalized over  $N_i$  and  $r_i$ ) can be now be written down generically as

$$P(\{\Omega_i\} | \{D_i\}) \propto \prod_{i=1}^I \sum_{(N_i, r_i)} P(\Omega_i, N_i, r_i) \times L(D_i; \Omega_i, N_i, r_i), \quad (8)$$

where  $P(\Omega_i, N_i, r_i)$  are the joint prior probabilities. Note that for Models 1(a)–(c) we have  $P(\theta_i(\boldsymbol{\beta})) \equiv P(\theta)$  and  $P(\eta_i(\boldsymbol{\beta})) \equiv P(\boldsymbol{\beta})$ , while the reverse is true for Model 2. Also, for the first- and last-activation schemes  $P(r_i|N_i)$  is degenerate and we need consider  $P(N_i|\theta_i)$  only.

In general the marginalization in (8) is analytically intractable and performed using a Markov chain Monte Carlo algorithm (see e.g. Carlin and Louis, 2000). The implementation of the MCMC may be considerably simplified by using a marginalized likelihood. In fact, marginalizing the  $(N_i, r_i)$ 's out of (8) reduces the estimation space considerably, and amounts to evaluating

$$E_{N_i, r_i | \theta_i} \left[ \left( -\frac{d}{dt} P(T \geq t_i | N_i) \right)^{\nu_i} (P(T \geq t_i))^{1-\nu_i} \right].$$

for each  $i$ . Using the facts that  $\nu_i$  equals 1 or 0, and that the derivative can be interchanged with the expectation, we can rewrite this as  $(h^*(t_i))^{\nu_i} S^*(t_i)$ , yielding the data likelihood as

$$L(\{D_i\}; \{\Omega_i\}) = \prod_{i=1}^I [h^*(t_i)^{\nu_i} S^*(t_i)]. \quad (9)$$

Thus we may sample  $P(\{\Omega_i\}|\{D_i\})$  in (8) using say Metropolis, and avoid sampling the  $(N_i, r_i)$ .

## 4.2 Model comparison

With the broad range of models that the above formulation encompasses, model evaluation and selection become important issues. We believe the model selection issue to involve scientific considerations that might favor certain assumptions and assessing the models with respect to the data at hand. Regarding this, note that the data likelihood in (9) can distinguish between the models and the activation schemes based upon the differing hazard and survival structures  $(h^*(t)$  and  $S^*(t))$  they imply. As a practical guideline for model selection, we adopt a measure that rewards goodness of fit but penalizes complexity. In practice, when several models fit the data adequately well, often scientific considerations (e.g. knowledge about the disease) prompt the final model choice(s). In the absence of such knowledge, our

measure will favor the simpler models unless the complex model offers substantially better fits.

For assessing goodness of fit, we perform posterior predictive comparisons. With  $\Omega$  as the parameter set, we generate (future) data replicates  $\mathbf{t}^* = \{t_i^*\}_{i=1}^I$  from the posterior predictive distribution,

$$P(\mathbf{t}^* | \{D_i\}) = \int P(\mathbf{t}^* | \Omega, \{\nu_i, \mathbf{x}_i\}) P(\Omega | \{D_i\}) d\Omega, \quad (10)$$

where  $P(\mathbf{t}^* | \Omega, \{\nu_i, \mathbf{x}_i\})$  is the underlying probability distribution in the data likelihood. More precisely, given post-convergence posterior samples of size  $M$  from MCMC output, say  $(\Omega_1, \dots, \Omega_M)$  from  $P(\Omega | \{D_i\})$ , for each  $\Omega_j$  we generate a predictive data replicate  $\mathbf{t}_j^*$  (an  $I \times 1$  vector) by drawing from  $P(\mathbf{t}^* | \Omega_j, \{\nu_i, \mathbf{x}_i\})$ . This yields an  $I \times M$  posterior predictive *ensemble* matrix, say  $\mathbf{T}^* = [\mathbf{t}_1^* : \dots : \mathbf{t}_M^*]$ , whose  $i$ -th row is a sample from the posterior predictive distribution of the survival time for the  $i$ -th individual.

Note that, unlike in usual survival models, the impropriety of  $S^*(t)$  in the likelihood somewhat complicates these computations, particularly simulating from  $P(\mathbf{t}^* | \Omega, \{\nu_i, \mathbf{x}_i\})$ . However, by simulating our latent-event schemes, we can generate posterior predictive samples for the  $i$ -th individual by plugging in the posterior sample values  $\Omega_j$ ,  $j = 1, \dots, M$  for the required parameter estimates. For instance, if  $\nu_i = 1$ , then subject  $i$  has failed, so we first generate  $N_i$  from the truncated  $Po(\theta_i(\boldsymbol{\beta}))1(N_i \geq 1)$  distribution, ensuring that  $N_i \geq 1$ , and set  $r_i = 1$  (first-activation),  $r_i = N_i$  (last-activation) or generate  $r_i \sim Bin(N_i, \pi_i)$  (hierarchical-activation) for the respective schemes. Next, we generate  $Y_1, \dots, Y_N$  from the Weibull family and set  $T_i = Y_{(r_i)}$ . On the other hand, if  $\nu_i = 0$ , then the subject has a positive probability of being cured, so we generate  $N_i$  from a  $Po(\theta_i(\boldsymbol{\beta}))$  distribution, admitting the possibility that  $N_i = 0$ . If  $N_i = 0$ , then we set  $T_i = \infty$  (operationally, what is actually assigned as  $\infty$  will depend upon the computing environment, but for practical purposes any number much larger than the observed data suffices). Otherwise, we follow the procedures for  $N_i \geq 1$  as for  $\nu_i = 1$ , but truncate  $T_i$  to exceed the observed censoring time  $t_i$ .

Although several models may provide adequate fit to the data, it is beneficial to have a framework for choosing among them. Theoretically, due to the presence of the point mass at  $T_i = \infty$  the moments of the posterior predictive distribution do not exist for cure models, but in practice we replace  $\infty$  with a large number and effectively build model comparison procedures based upon them. Here we adopt the posterior predictive L-measure (Laud and Ibrahim, 1995; Gelfand and Ghosh, 1998), computed as the sum of a goodness-of-fit measure and a penalty term

$$L = E_{[\mathbf{t}^*|\{D_i\}]} [(\mathbf{t}^* - \boldsymbol{\mu}^*)^T (\mathbf{t}^* - \boldsymbol{\mu}^*)] + \frac{\delta}{\delta + 1} (\mathbf{t} - \boldsymbol{\mu}^*)^T (\mathbf{t} - \boldsymbol{\mu}^*), \quad (11)$$

where  $\boldsymbol{\mu}^* = E_{t^*|\{D_i\}} [\mathbf{t}^*]$  is the posterior predictive mean of the replicated data, and  $\mathbf{t}$  is a vector of the observed time-points  $\{t_i\}$ . Letting  $P = tr(Var(\mathbf{t}^*|\mathbf{t}))$  (which equals the first term in (11)) and  $G = (\mathbf{t} - \boldsymbol{\mu}^*)^T (\mathbf{t} - \boldsymbol{\mu}^*)$ , we see that  $L$  tends to  $P+G$  as  $\delta \rightarrow \infty$ . This approach has decision-theoretic justifications treating model choice as a minimizing decision rule for a squared-error loss function. To compute  $G$ ,  $P$  and  $L$  we first compute  $\boldsymbol{\mu}^* = \frac{1}{M} \sum_{i=1}^M \mathbf{t}_i^*$  and then insert it to compute  $P = \frac{1}{M-1} \sum_{i=1}^M (\mathbf{t}_i^* - \boldsymbol{\mu}^*)^T (\mathbf{t}_i^* - \boldsymbol{\mu}^*)$  and  $G = (\mathbf{t} - \boldsymbol{\mu}^*)^T (\mathbf{t} - \boldsymbol{\mu}^*)$ . Operationally, we often work on a transformed scale, replacing  $t_i$  and  $t_i^*$  by  $\log(t_i)$  and  $\log(t_i^*)$  respectively, to ensure greater numerical stability and a better scale for comparison. Smaller  $L$  scores suggest better models in terms of predictive fit, and simulation draws from different likelihoods can be fairly compared. Alternative model comparison measures such as the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), while computationally simpler, are unsuitable here because our likelihoods are not log-concave. Furthermore, since the L-measure is computed purely from predictive samples, we can not only compare models *within* a given activation scheme, but also *across* different activation schemes – something that is precluded in DIC without computation of normalizing constants that greatly detracts from its computational simplicity.

## 5 Illustrations

We illustrate our methods first through a simulation study for assessing different model performances, and then for data sets from melanoma and breast cancer. The melanoma data, where failure is melanoma *relapse*, has previously been analyzed using YCIS modeling (Chen et al., 1999) to capture a pronounced “plateau” effect in its survival curve. The breast cancer data, where failure is defined as *death* from breast cancer (deaths from many aggressive cancers are attributed to disease *relapse*, and relapse is practically indistinguishable from death), does not reveal a dominant plateau effect in the study time frame. Yet, cure modelling for breast cancer is relevant, especially with prognosis often showing marked improvements where cure fraction estimates are better measures of treatment efficacy.

We always assigned a  $N(0, 10^3)$  prior to the regression slopes, while a relatively weak  $\text{Gamma}(0.001, 0.001)$  prior was assigned to  $\rho$ . For Model 2, we set  $\eta \sim N(0, 100)$ , while  $\theta \sim \text{Gamma}(0.001, 0.001)$  in Model 1a and  $U(0, 1)$  in Models 1b and 1c. These priors were weak enough for the data to drive the posterior inference and yielding acceptable MCMC convergence. Experimenting with different hyper-parameter values revealed very robust posteriors. For the HA-Bin scheme we selected  $r - 1 | N \sim \text{Bin}(N - 1, \pi)$  and experimented with a beta or uniform prior on  $\pi$  and also with  $\pi = 1/2$ . The resulting inferences did not differ much, so we present the  $\pi = 1/2$  cases only. When fitting the mixture model (Mix) where  $r \in \{1, N\}$  with probability  $1 - \pi$  and  $\pi$  respectively, we assigned  $\pi \sim U(0, 1)$ . For each analysis we ran two initially dispersed parallel MCMC chains for 30,000 iterations each. Convergence diagnostics suggested discarding the first 5,000 iterations from each chain as pre-convergence burn-in, yielding  $2 \times 25,000 = 50,000$  samples for posterior analysis.

### 5.1 Simulation study

For assessing performances of the activation schemes, we generated data sets from each set of known model specifications and activation schemes that were subsequently analyzed

by fitting Models 1a, 1c, 1d and Model 2. Model 1b, the classical BG-type model, does not distinguish among the different activation schemes and is not considered here. For each of Models 1(a), 1(c), 1(d) and 2 and each activation scheme, FA, LA, HA-Bin and Mixture (Mix), we generated 10,000 data sets: 2,000 each with cure fractions 20%, 25%, 30%, 35%, and 40%. The regressors comprised an intercept and a continuous covariate generated from a  $N(0, 1)$  distribution. We fixed  $\beta = (-5, 2)$  for Models 1a and 1c and  $(-3, 2)$  for Model 1d,  $\rho = 1.5$  for all the models, and let the censoring rate vary between 20% and 60%. For Model 2 we set  $\eta = -3$  and  $\beta_1 = 1$  while  $\beta_0$  was determined from  $\beta_1$  and the respective cure fraction. MCMC iterations subsequently produced consistent estimates of all the model parameters with the 95% credible interval including the true values 95% – 100% of the time.

The first two columns in Table 1 indicate the model and scheme generating the data with each cell displaying the means of G, P and L over the 10,000 simulations. The bold diagonal entries for the L measure indicate, quite expectedly, better performance of the true model on the average. This pattern is very consistent, barring only the Mixture scheme in Model 1c, where HA-Bin’s L score mean is marginally lower. The table also reveals superior performance of the hierarchical schemes, both HA-Bin and Mix: even when not the truth, they perform very robustly with their  $\bar{G}$ ,  $\bar{P}$  and  $\bar{L}$  appreciably closer to the true model and substantially lower than the misspecified non-hierarchical scheme. Note that the differences between the HA-Bin and the Mixture schemes were usually minor, although the former’s performance under the true Mixture model seems better than the reverse (especially for Models 1c and 1d). Model 1c in Table 1 corresponds to  $K = 10$ ; we also conducted analysis for other values of  $K$  (e.g. 5, 15 and 20) and found performances of the activation schemes extremely consistent with those presented here.

To assess the mixture model’s ability to distinguish between FA and LA we also monitored the mixing probability  $P(r = 1)$ . This parameter appeared to be well-identified in our MCMC computations. For example, with a randomly selected simulated data set, the

Mixture scheme with  $P(r = 1) = P(r = N) = 1/2$  and Model 2 yielded 95% posterior credible interval for  $P(r = 1)$  about  $(0.394, 0.791)$ . With FA and LA as the true underlying schemes they were consistently estimated as  $(0.845, 0.997)$  and  $(0.029, 0.444)$  respectively, while with HA-Bin we found the interval as  $(0.017, 0.557)$ .

These simulations demonstrate the advantages for accounting for the uncertainty in activation schemes. Fixed activation schemes seem to perform inferiorly to our hierarchical schemes, whose comparatively robust performances should make them invaluable modelling tools for more realistic settings where the “true” underlying mechanisms generating the data are rarely known.

## 5.2 Analysis of melanoma data

Melanoma incidence rates are among the highest of all solid tumors, and in spite of earlier detection and screening, high-risk melanoma patients continue to have mortality rates between 60% and 75% (Kirkwood, et al., 2000). The data we consider comes from two recent phase three clinical trials (Kirkwood et al., 1996) where subjects have been administered the post-operative chemotherapy interferon alpha-2b (IFN). The data consists of 284 subjects, of whom 113 are female, with 174 who relapsed and the remaining being censored. In addition, an indicator covariate (fully active, other) representing the performance status (PS) of each subject is also available. Basic descriptive analysis via a Kaplan-Meier curve (see the solid line in Figure 2) reveals a typical “plateau”, indicating a significant proportion of patients who appear cured. Table 2 provides the posterior predictive L-measures for the different models for the melanoma data set. Recall that Model 1b is the BG model, where the activation schemes coincide; hence, the L-measures are same across this row. Model 1d (regressing on the Weibull mean) for the Mixture scheme has the lowest L-measure score among *all* the different models under any activation scheme with a marginal improvement over the YCIS model (Model 2, FA) and more pronounced improvements over the BG-type

model.

Table 3 presents results for the L-best model under the different activation schemes. The covariates included are age at diagnosis, gender (male or female), and performance status (fully active or not). Recall from Section 3 that for Model 2 the parameters impact the cure probability with positive estimates implying lower cure probabilities, whereas for Models 1a, 1b, and 1c they affect the Weibull link, now with positive estimates implying greater hazard of failing. The lack of significance in the covariate estimates is somewhat disappointing, but these findings agree with the results reported by Chen et al. (1999) (though they discuss only the first-activation scheme). The Weibull scale parameter  $\rho$  is significantly greater than one, consistent with the marked slope in the hazard curve. We also remark that the Mixture model 1d estimated of  $P(r = 1)$  as 0.859 with a 95% C.I. (0.651,0.989), indicating strong support for first-activation.

The first column of Figure 1 plots the individual-specific cure fraction estimates (posterior medians) for the different activation schemes as provided by Model 2 for the melanoma data set. In each plot we also show a solid line corresponding the constant cure fraction estimated from Model 1a (regression in the Weibull link). These plots reveal the sensitivity in the cure fractions to individual characteristics, somewhat corroborating the results in Table 2 that indicated Model 2 as one of the best models. Finally, Figure 2 overlays the median of the posterior predictive survival plots (dashed and dotted lines) with Kaplan-Meier plots (solid lines) from the raw data. Almost all of the models seem to provide adequate fit to the empirical curve, capturing the plateau quite effectively. This is noteworthy given that a smooth parametric family (Weibull) is modelling a latent, rather than an observed, survival distribution.

### 5.3 Analysis of breast cancer data

The National Cancer Institute’s Surveillance, Epidemiology, and End Results (SEER) database, available online at <http://seer.cancer.gov>, provides a national cohort of women who have been progressively monitored for assessing breast cancer prognosis. In addition, available individual-level covariates include racial information, age at diagnosis, the number of primary cancers each woman has had diagnosed, and the stage of the disease (local, regional (95 patients) or distant (12 patients), with local as baseline). We consider a sample of 305 patients who were all diagnosed with breast cancer in January 1992 and were monitored until 1998. The response here is time to death *from breast cancer*: only those who have been identified as having died from metastasis of nodes in the breast (there were 102 such deaths) are considered having *failed*, while the rest (including those who might have died from metastasis of other types of cancer or other causes) are considered censored. With the time units being “month”, the longest observation time is 84 months. Note that with such an endpoint definition, a cure rate model is essentially mandatory, since we know that some positive fraction will die from causes other than breast cancer.

Table 4 provides the L-scores for the different models for the breast cancer data set. Clearly, the FA scheme (YCIS model) does not perform well here. Indeed, the hierarchical and the LA frameworks with regression in the cure fraction (Model 2) seem to be more desirable, with the HA-Bin scheme in Model 2 having the lowest L-score. In Table 5, we present parameter inference results from Model 1d with LA along with those from Model 2 for FA and the hierarchical schemes (the best models under each scheme). For all the three models, we see a significant positive impact for age at diagnosis (later diagnosis corresponds to greater hazard). The number of primary nodes has a positive estimate, significantly so for the LA (1d) scheme. Expectedly, regional or distant stages yield significant hazard increases relative to local. The Weibull scale parameter  $\rho$  is robustly estimated across the models. The mixture model 2 still seemed to slightly favor the first activation scheme, though much less

than the melanoma data, estimating  $P(r = 1)$  with 0.745 and a 95% C.I. of (0.495,0.928).

Turning to the median cure fraction estimates in the second column of Figure 1, the variation in the cure fraction estimates appears less tightly centered about the fixed cure fraction estimate than for the melanoma data. Unlike the melanoma data, here the differences in performance across models are quite evident. The lack of a “plateau” in the observed time frame induces the downward shift of cure fraction for the poorly performing models, so Model 1a tends to move the cure fraction bar off the center. This reveals greater sensitivity to the modelling assumptions, and a bias in the models with a constant cure fraction.

This message carries over to the posterior-predictive survival plots in Figure 3, where the Kaplan-Meier plot for the survival data runs only up to 84 months, but we extend our fitted curves up to 200 months to show the “plateau”. Generally, the fit of Models 1(a)–1(d) (including BG-type models) is less satisfactory than for Model 2, showing the inadequacy of a constant cure fraction. In Model 2 the LA and HA-Bin schemes offer better fit than FA (YCIS model), nicely capturing the flattening of the survival curve for the observed data after 50 months. These plots also reveal the differences in the plateaus more distinctly than the melanoma data. Generally, we find that LA reaches the plateau fastest, FA the slowest, while HA-Bin and Mixture have intermediate rates. This is expected due to the more stringent conditions for failure in LA than FA. Finally, we remark that these predictive fits seem to be consistent with the goodness of fit measures (G) in Table 4, although they do not inform about formal model penalty (P), which is accounted for by the L-measure. In any case, the breast cancer data demonstrates the weaker performances of the existing BG-type and YCIS models and the improvements available using the flexible hierarchical methods.

## 6 Summary and Discussion

We have proposed a general class of cure models motivated from latent activation mechanisms and outlined a hierarchical framework with Markov chain Monte Carlo (MCMC) in-

ference. Our models generalize existing ones, offer robust inference and should be especially beneficial in realistic settings where the disease's mechanistic nature can only be surmised. In fact, statistical estimation of *when* the survival function hits a plateau is particularly pertinent in cancer studies, for example of the breast, where cure is believed although short term monitoring data might conceal such a plateau. Our melanoma analysis reveal our models to be performing consistently with existing models, while our breast cancer example shows the need for greater flexibility.

Extensions of our work could explore further possibilities beyond  $N \sim Po(\theta)$ . In fact, interesting hazard structures (e.g. proportional odds) arise with  $N \sim Geo(\theta)$  in Model 2 leading to new classes of identifiable cure models. Also, while the mixing distribution is unidentifiable in semi-parametric formulations, joint modelling of cure-rate survival data and multivariate longitudinal data (e.g. Ibrahim et al., 2001, Taylor, 1995) discuss characterizing the mixing structure via observed multivariate longitudinal data. We envision adapting our richer framework for such data. Along a different route, we could consider spatial cure models extending the simpler BG-type structures (e.g. Banerjee and Carlin, 2004) by incorporating random effects or *frailties* that are correlated. Our framework can certainly accommodate such modelling, but unresolved issues arise regarding associations in both the latent link and the cure fraction.

## APPENDIX

1. We discuss the identifiability of  $\theta$  under the prior  $g(\theta)$  with respect to the density  $f^*(t|\theta)$ , which amounts to  $I_{f^*}(t) = \int_0^\infty g(\theta)f^*(t|\theta)d\theta$  being finite for all  $t > 0$ . One could work with the likelihood  $\prod_{i=1}^n f^*(t_i|\theta)$ , but the general idea of the proof remains the same. Note that if  $N \sim Po(\theta)$ , and  $a(N)$  is any function on the integers then we have the following simple relationship:

$$E_N[Na(N)] = \sum_{n=0}^{\infty} na(n)e^{-\theta}\frac{\theta^n}{n!} = \theta \sum_{n=1}^{\infty} a(n)e^{-\theta}\frac{\theta^{n-1}}{(n-1)!} = \theta E_N[a(N+1)]. \quad (\text{A1.2})$$

First suppose  $r$  is fixed at  $w + 1$  for some non-negative integer  $w$ . Then  $f^*(t|\theta) = f(t)E_N[N1(N \geq w + 1)P(W = w|N)]$ , where  $W | N \sim \text{Bin}(N - 1, F(t))$ , which equals, by (A1.2),  $\theta f(t)E_N[1(N \geq w)P(W^* = w|N)]$ , where  $W^* | N \sim \text{Bin}(N, F(t))$ . This can be further simplified to give

$$f^*(t|\theta) = \theta f(t) \frac{e^{-\theta} [\theta F(t)]^w}{w!} \sum_{n=w}^{\infty} \frac{[\theta S(t)]^{n-w}}{(n-w)!} = \theta f(t) \frac{e^{-\theta F(t)} [\theta F(t)]^w}{w!},$$

which equals  $\theta f(t)P(V = w)$ , where  $V \sim \text{Po}(\theta F(t))$ . Now, we have

$$I_{f^*}(t) = f(t) \int_0^{\infty} g(\theta) \theta \frac{e^{-\theta F(t)} [\theta F(t)]^w}{w!} d\theta = f(t) \frac{[F(t)]^w}{w!} \int_0^{\infty} g(\theta) \theta^{w+1} e^{-\theta F(t)} d\theta.$$

With  $g(\theta) = 1/\theta$ , the above further simplifies to  $I_{f^*}(t) = f(t)/F(t) < \infty$ . This proves the identifiability of  $\theta$  under the scale-invariance prior whenever  $r$  is degenerate at a single point. Even when  $r$  and  $N$  both vary in the above setting but the difference  $N - r$  remains constant, calculations analogous to the above reveal  $I_{f^*}(t) = f(t)/S(t) = h(t) < \infty$  under  $g(\theta) = 1/\theta$ . This proves, in particular, the identifiability of the cure fraction for the last activation models ( $N - r = 0$ ).

2. Here we derive (5), which shows how the hierarchical framework leads to a Berkson-Gage type model. Taking  $r | N \sim \text{DiscreteUnif}(1, N)$  in (4) and setting  $r^* = r - 1$ , we obtain

$$\begin{aligned} S^*(t|\theta) &= P(N = 0) + E_N \left[ 1(N \geq 1) \int_0^{S(t)} \sum_{r=1}^N \binom{N-1}{r-1} u^{N-r} (1-u)^{r-1} du \right] \\ &= P(N = 0) + E_N \left[ 1(N \geq 1) \int_0^{S(t)} \sum_{r^*=0}^{N-1} \binom{N-1}{r^*} u^{N-r^*-1} (1-u)^{r^*} du \right] \\ &= P(N = 0) + E_N [1(N \geq 1)S(t)] \quad (\text{the above binomial expansion equals 1}) \\ &= P(N = 0) + S(t)E_N [1(N \geq 1)] = P(N = 0) + (1 - P(N = 0))S(t). \end{aligned}$$

For  $N \sim \text{Po}(\theta)$ , it is easily seen by differentiating  $S^*(t|\theta)$  above (or directly by using (A1.2)) that  $f^*(t|\theta) = f(t)(1 - e^{-\theta})$ , and hence  $I_{f^*}(t) = f(t) \int_0^{\infty} g(\theta)(1 - e^{-\theta})d\theta$ .

However, with  $g(\theta) = 1/\theta$ , we immediately see that  $I_{f^*}(t) = \infty$  proving the non-identifiability of  $\theta$ . The classical BG-type model assumes  $N \sim Ber(\theta)$ , with  $1 - \theta$  as the cure fraction. With a logistic link we have  $\theta = e^\mu/(1 + e^\mu)$ , which gives  $f^*(t|\mu) = e^\mu f(t)/(1 + e^\mu)$  and  $I_{f^*}(t) = f(t) \int_0^\infty g(\mu)e^\mu/(1 + e^\mu)d\mu$ . With the non-informative flat prior  $g(\mu) \propto 1$ , it is straightforward to show that  $I_{f^*}(t) = \infty$ .

3. Let  $P_N(\cdot)$  denote the conditional probability distribution of  $r - 1$  given  $N$  with support  $\{0, \dots, N - 1\}$ . Then, with  $N \sim Po(\theta)$ , using (A1.2) we can write  $f^*(t|\theta) = \theta f(t)E_{(N, V^*)}[P(W^* = V^*|N)]$ , where  $W^*|N \sim Bin(N, F(t))$  (as in item 1) and  $V^*|N \sim P_{N+1}(\cdot)$ . Therefore, for  $\theta$  to be identifiable under  $g(\theta) = 1/\theta$ , the family  $P_N(\cdot)$  must satisfy

$$\sum_{n=0}^{\infty} \frac{\exp(-\theta)\theta^n}{n!} \sum_{i=0}^n P(W^* = i|N = n)P_{n+1}(i) = O\left(\frac{1}{\theta^{1+\delta}}\right), \quad \text{with } \delta > 0.$$

For general  $P_N(\cdot)$  (not with all its mass at a single point), this condition may be difficult to verify.

## REFERENCES

- Banerjee, S. and Carlin, B.P. (2004), "Parametric Spatial Cure Rate Models for Interval-Censored Time-to-Relapse Data," *Biometrics*, **60**, 268–275.
- Berkson, J. and Gage, R.P. (1952), "Survival Curve for Cancer Patients following Treatment," *Journal of the American Statistical Association*, **47**, 501–515.
- Carlin, B.P. and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC Press.
- Chen, M.-H., Ibrahim, J.G. and Sinha, D. (1999), "A New Bayesian Model for Survival Data with a Surviving Fraction," *Journal of the American Statistical Association*, **94**, 909–919.

- Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.
- Ewell, M. and Ibrahim, J.G. (1997), “The Large Sample Distribution of the Weighted Log-rank Statistic Under General Local Alternatives,” *Lifetime Data Analysis*, **3**, 5-12.
- Farewell, V.T. (1982), “The Use of Mixture Models for the Analysis of Survival Data with Long Term Survivors,” *Biometrics*, **38**, 1041-1046.
- Farewell, V.T. (1986), “Mixture Models in Survival Analysis: Are they worth the risk?” *Canadian Journal of Statistics*, **14**, 257-262.
- Gail, M.H., Santner, T.J. and Brown, C.C. (1980), “An analysis of comparative carcinogenesis experiments based on multiple times to tumor,” *Biometrics*, **36**, 255-266.
- Gelfand, A.E. and Ghosh, S.K. (1998), “Model Choice: A Minimum Posterior Predictive Loss Approach,” *Biometrika*, **85**, 1–11.
- Goldman, A.I. (1984), “Survivorship Analysis when Cure is a Possibility: A Monte Carlo Study,” *Statistics in Medicine* **3**, 153-163.
- Hanin, L., Tsodikov, A. and Yakovlev, A. (2001), “Optimal Schedules of Cancer Surveillance and Tumor Size at Detection,” *Mathematical and Computer Modelling*, **33**, 1419–1430.
- Ibrahim, J.G., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.
- Kirkwood, J.M., Ibrahim, J.G., Sondak, V.K., Richards, J., Flaherty, L.E., Ernstoff, M.S., Smith, T.J., Rao, U., Steele, M. and Blum, R.H. (2000), “High- and Low-dose Interferon Alfa-2b in High-risk Melanoma: First Analysis of Intergroup Trial E1690/S9111/C9190,” *Journal of Clinical Oncology*, **19**, 1226–1228.
- Kirkwood, J.M., Strawderman, M.H., Ernstoff, M.S., Smith, T.J., Borden, E.C. and Blum, R.H. (1996), “Interferon Alfa-2b Adjuvant Therapy of High-risk Resected Cutaneous

- Melanoma: The Eastern Cooperative Oncology Group Trial EST 1684,” *Journal of Clinical Oncology*, **14**, 7–17.
- Kuk, A.Y.C. and Chen, C.H. (1992), “A Mixture Model Combining Logistic Regression with Proportional Hazards Regression,” *Biometrika*, **79**, 531-541.
- Laud, P. and Ibrahim, J. (1995), “Predictive Model Selection,” *J. Roy. Statist. Soc., Ser. B*, **57**, 247–262.
- Li, C.S. and Taylor, J.M.G. (2002), “A Semiparametric Accelerated Failure Time Cure Model,” *Statistics in Medicine*, **21**, 3235–3247.
- Li, C.S., Taylor, J.M.G. and Sy, J.P. (2001), “Identifiability of Cure Models,” *Statistics and Probability Letters*, **54**, 389–395.
- Moolgavkar, S.H., Luebeck, E.G., and De Gunst (1990), “Two Mutation Model for Carcinogenesis: Relative Roles of Somatic Mutations and Cell Proliferation in Determining Risk,” in *Scientific Issues in Quantitative Cancer Risk Assessment*, Editor: S.H.Moolgavkar, Birkhauser: Boston, 136–152.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.
- Spiegelhalter, D.J., Best, N., Carlin, B.P., and van der Linde, A. (2002), “Bayesian Measures of Model Complexity and Fit (with discussion),” *J. Roy. Statist. Soc., Ser. B*, **64**, 583–639.
- Sy, J.P. and Taylor, J.M.G. (2000), “Estimation in a Cox Proportional Hazards Cure Model,” *Biometrics*, **56**, 227–236
- Taylor, J.M.G. (1995). “Semiparametric Estimation in Failure Time Mixture Models,” *Biometrics*, **51**, 899-907.

- Tucker, S.L. and Taylor, J.M.G. (1996), “Improved Models of Tumor Cure,” *International Journal of Radiational Biology* **70**, 539–553.
- Tucker, S.L., Thames, H.D. and Taylor, J.M.G. (1990), “How Well is the Probability of Tumor Cure after Fractionated Irradiation Described By Poisson Statistics?” *Radiation Research* **124**, 273–282.
- Tsodikov, A., Ibrahim, J. and Yakovlev, A. (2003), “Estimating Cure Rates from Survival Data: An Alternative to Two-Component Mixture Models,” *Journal of the American Statistical Association*, **98**, 1063–1078
- Yakovlev, A.Y. (1996), “Threshold models of tumor recurrence,” *Mathematical and Computer Modelling*, **23**, pp. 153–164
- Yakovlev, A.Y., Asselain, B., Bardou, V.J., Fourquet, A., Hoang, T., Rochefordiere, A. and Tsodikov, A.D. (1993), “A Simple Stochastic Model of Tumor Recurrence and its Application to Data on Premenopausal Breast Cancer,” *Biometrie et Analyse de Donnees Spatio-Temporelles*, **12**, eds. B. Asselain, M. Boniface, C. Duby, C. Lopez, J.P. Masson, and J. Tranchefort, Rennes, France: Société Française de Biométrie, 66–82.
- Yakovlev, A.Y. and Tsodikov, A.D. (1996). *Stochastic models of tumor latency and their biostatistical applications*. New Jersey: World Scientific.
- Yin, G.S. and Ibrahim, J.G. (2005), “Cure Rate Models: a Unified Approach,” *The Canadian Journal of Statistics*, **33**, 559–570(12).

Simulated data		$\bar{G}$				$\bar{P}$				$\bar{L}$			
Model	Simulation scheme	FA	LA	HA-Bin	Mix	FA	LA	HA-Bin	Mix	FA	LA	HA-Bin	Mix
1(a)	FA	<b>50.1</b>	57.3	51.4	53.6	<b>55.1</b>	68.2	57.8	57.1	<b>105.2</b>	125.5	109.2	110.7
1(a)	LA	23.1	<b>17.9</b>	18.1	18.9	37.8	<b>20.7</b>	25.8	30.6	60.9	<b>38.6</b>	43.9	49.5
1(a)	HA-Bin	44.2	43.7	<b>29.5</b>	31.5	39.9	38.7	<b>31.6</b>	35.1	84.1	82.4	<b>61.1</b>	66.6
1(a)	Mix	62.7	58.6	53.1	<b>50.2</b>	69.4	68.7	<b>63.8</b>	64.9	132.1	127.3	116.9	<b>115.1</b>
1(c)	FA	<b>51.1</b>	64.1	52.2	51.9	<b>53.9</b>	59.8	55.8	57.4	<b>105.0</b>	123.9	108.0	109.3
1(c)	LA	27.7	<b>19.5</b>	19.8	19.7	27.5	<b>21.6</b>	24.0	24.4	55.2	<b>41.1</b>	43.8	44.1
1(c)	HA-Bin	45.5	43.9	<b>35.8</b>	35.9	48.6	40.0	<b>36.5</b>	39.3	94.1	83.9	<b>72.3</b>	75.2
1(c)	Mix	49.8	49.5	43.7	<b>43.4</b>	62.3	60.8	<b>59.3</b>	60.1	112.1	110.3	<b>103.0</b>	103.5
1(d)	FA	71.6	72.7	72.0	<b>71.2</b>	<b>68.7</b>	83.1	70.7	78.1	<b>140.3</b>	155.8	142.7	149.3
1(d)	LA	26.7	<b>18.5</b>	20.3	20.5	35.5	<b>22.7</b>	26.5	26.1	62.2	<b>41.2</b>	46.8	46.6
1(d)	HA-Bin	43.7	47.8	<b>29.6</b>	33.5	43.5	43.3	<b>31.3</b>	35.6	87.2	91.1	<b>60.9</b>	68.1
1(d)	Mix	67.3	69.8	<b>60.6</b>	61.9	86.9	91.7	<b>82.3</b>	<b>80.6</b>	154.2	161.5	142.9	<b>142.5</b>
2	FA	<b>45.9</b>	53.4	48.0	48.6	<b>55.9</b>	67.4	58.8	61.1	<b>101.8</b>	120.8	106.8	109.7
2	LA	31.6	<b>25.1</b>	26.1	25.2	34.4	<b>29.9</b>	35.1	34.5	66.0	<b>55.0</b>	61.2	59.7
2	HA-Bin	59.7	62.6	<b>47.6</b>	49.5	69.4	69.6	<b>60.6</b>	63.5	129.1	132.2	<b>108.2</b>	113.0
2	Mix	57.8	56.2	<b>46.0</b>	46.2	67.5	65.5	63.4	<b>60.5</b>	125.3	121.7	109.4	<b>106.7</b>

Table 1: Model performances in simulation examples. The first and second columns denote the underlying model and mechanism which generated the data. The remaining columns show means of goodness-of-fit ( $\bar{G}$ ), the penalty ( $\bar{P}$ ) and the L-measure ( $\bar{L}$ ) under first-, last- and hierarchical-activation schemes. The cells with the lowest means for each measure are highlighted

Melanoma	First			Last			HA-Bin			Mixture		
Model	G	P	L	G	P	L	G	P	L	G	P	L
1(a)	206.98	421.93	628.91	206.71	463.99	670.70	210.16	438.00	648.16	209.15	419.40	628.55
1(b)	206.29	438.30	644.59	<b>206.29</b>	<b>438.30</b>	<b>644.59</b>	206.29	438.30	644.59	206.29	438.30	644.59
1(c)	207.93	423.87	631.80	206.87	459.21	666.08	210.09	437.96	648.05	207.78	422.63	630.41
1(d)	<b>209.35</b>	<b>409.10</b>	<b>618.45</b>	206.81	492.75	699.56	206.05	436.08	642.13	<b>210.75</b>	<b>405.57</b>	<b>616.32</b>
2	206.59	414.22	620.81	205.43	459.62	665.05	<b>203.50</b>	<b>425.25</b>	<b>628.75</b>	205.65	413.49	619.14

Table 2: The goodness-of-fit, penalty and L-measure values for various models under first-, last- and hierarchical-activation schemes for the melanoma data set. The model with the best L-score in each column is highlighted.

Parameter	First (Model 1d)		Last (Model 1b)		HA-Bin (Model 2)		Mixture (Model 1d)	
	median	(2.5%, 97.5%)	median	(2.5%, 97.5%)	median	(2.5%, 97.5%)	median	(2.5%, 97.5%)
Intercept	-1.835	(-2.242, -1.479)	-0.924	(-1.196, -0.675)	0.089	(-0.130, 0.255)	-1.665	(-2.156, -1.236)
Age	-0.019	(-0.244, 0.215)	-0.023	(-0.186, 0.144)	0.118	(-0.033, 0.283)	-0.024	(-0.250, 0.220)
Gender (male=0)	-0.041	(-0.568, 0.441)	0.025	(-0.385, 0.396)	-0.164	(-0.472, 0.164)	-0.014	(-0.582, 0.472)
Performance Status	-0.283	(-1.151, 0.472)	-0.246	(-1.023, 0.372)	-0.239	(-0.732, 0.308)	-0.362	(-1.226, 0.407)
$\eta$ (Weibull link)	-	-	-	-	-0.929	(-1.136, -0.749)	-	-
Cure fraction ( $\exp(-\theta)$ )	0.342	(0.270, 0.408)	0.359	(0.297, 0.419)	-	-	0.338	(0.250, 0.404)
$\rho$	1.426	(1.232, 1.622)	1.193	(1.038, 1.353)	1.192	(1.050, 1.356)	1.477	(1.283, 1.670)

Table 3: Posterior quantiles for different models (corresponding to the models with best L-measure) under the three activation schemes for the melanoma data set.

Breast cancer Model	First			Last			HA-Bin			Mixture		
	G	P	L	G	P	L	G	P	L	G	P	L
1(a)	172.28	274.40	446.68	129.42	223.62	353.04	141.42	219.83	361.25	177.80	270.42	448.22
1(b)	149.44	242.15	391.59	149.44	242.15	391.59	149.44	242.15	391.59	149.44	242.15	391.59
1(c)	180.26	283.63	463.89	137.89	235.04	372.93	149.71	246.21	395.92	180.90	288.43	469.33
1(d)	188.19	306.76	494.95	<b>117.49</b>	<b>204.24</b>	<b>321.73</b>	175.87	199.06	374.93	188.20	295.28	483.48
2	<b>98.54</b>	<b>261.03</b>	<b>359.57</b>	98.71	231.43	330.14	<b>91.63</b>	<b>171.11</b>	<b>262.74</b>	<b>98.71</b>	<b>217.25</b>	<b>315.96</b>

Table 4: The goodness-of-fit, penalty and L-measure values for various models under first-, last- and hierarchical-activation schemes for the breast-cancer data set. The model with the best L-score in each column is highlighted.

Parameter	First (Model 2)		Last (Model 1d)		HA-Bin (Model 2)		Mixture (Model 2)	
	median	(2.5%, 97.5%)	median	(2.5%, 97.5%)	median	(2.5%, 97.5%)	median	(2.5%, 97.5%)
Intercept	-2.907	(-4.306,-0.439)	-7.789	(-9.776,-6.018)	-2.650	(-2.705,-2.552)	-3.105	(-4.211,-1.963)
Age	0.034	( 0.018, 0.050)	0.025	( 0.013, 0.037)	0.024	( 0.018, 0.029)	0.032	( 0.018, 0.052)
Primaries	0.196	(-0.191, 0.545)	0.621	( 0.089, 1.119)	0.062	(-0.088, 0.273)	0.351	(-0.172, 1.037)
Stage (local=0)								
Regional	0.538	( 0.103, 0.955)	0.737	( 0.286, 1.201)	0.788	( 0.532, 0.891)	0.700	( 0.142, 1.318)
Distant	2.484	( 1.782, 3.124)	2.531	( 1.882, 3.150)	1.647	( 1.429, 1.763)	3.380	( 2.609, 4.063)
$\eta$ (Weibull link)	-6.403	(-7.835,-5.406)	-	-	-6.002	(-6.061,-5.810)	-6.763	(-8.093,-5.930)
Cure fraction ( $\exp(-\theta)$ )	-	-	0.398	( 0.246, 0.520)	-	-	-	-
$\rho$	1.358	( 1.052, 1.737)	1.262	( 0.992, 1.577)	1.528	( 1.426, 1.612)	1.512	( 1.285, 1.822)

Table 5: Posterior quantiles for different models (corresponding to best L-measures) under the three activation schemes for the breast-cancer data set.

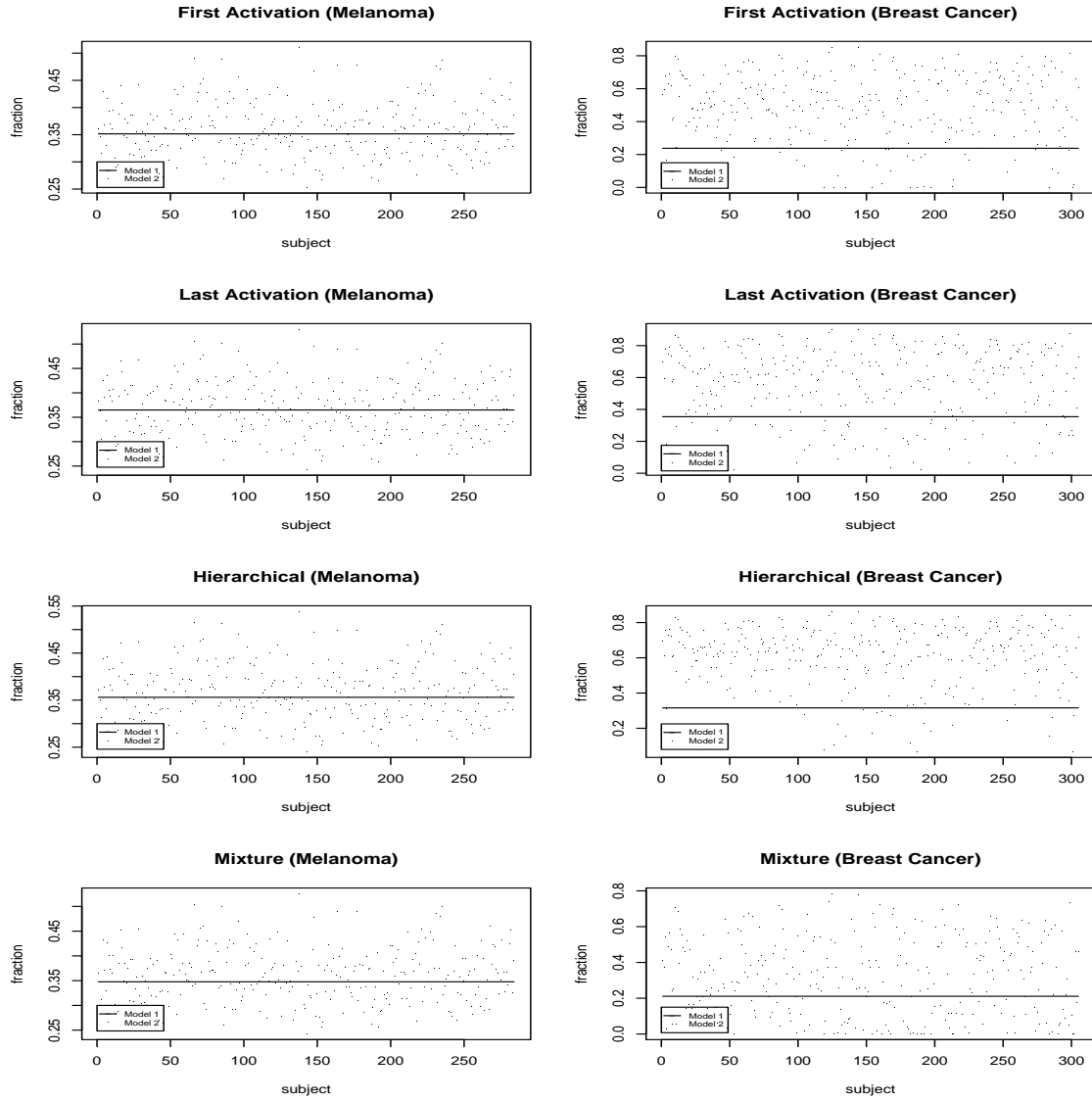


Figure 1: Plots of the median estimates of the cure fractions for the melanoma and breast-cancer data sets under different activation schemes. The horizontal solid line corresponds to the constant cure fraction estimate from Model 1a, while the dots are the subject-specific cure fraction estimates provided by Model 2.

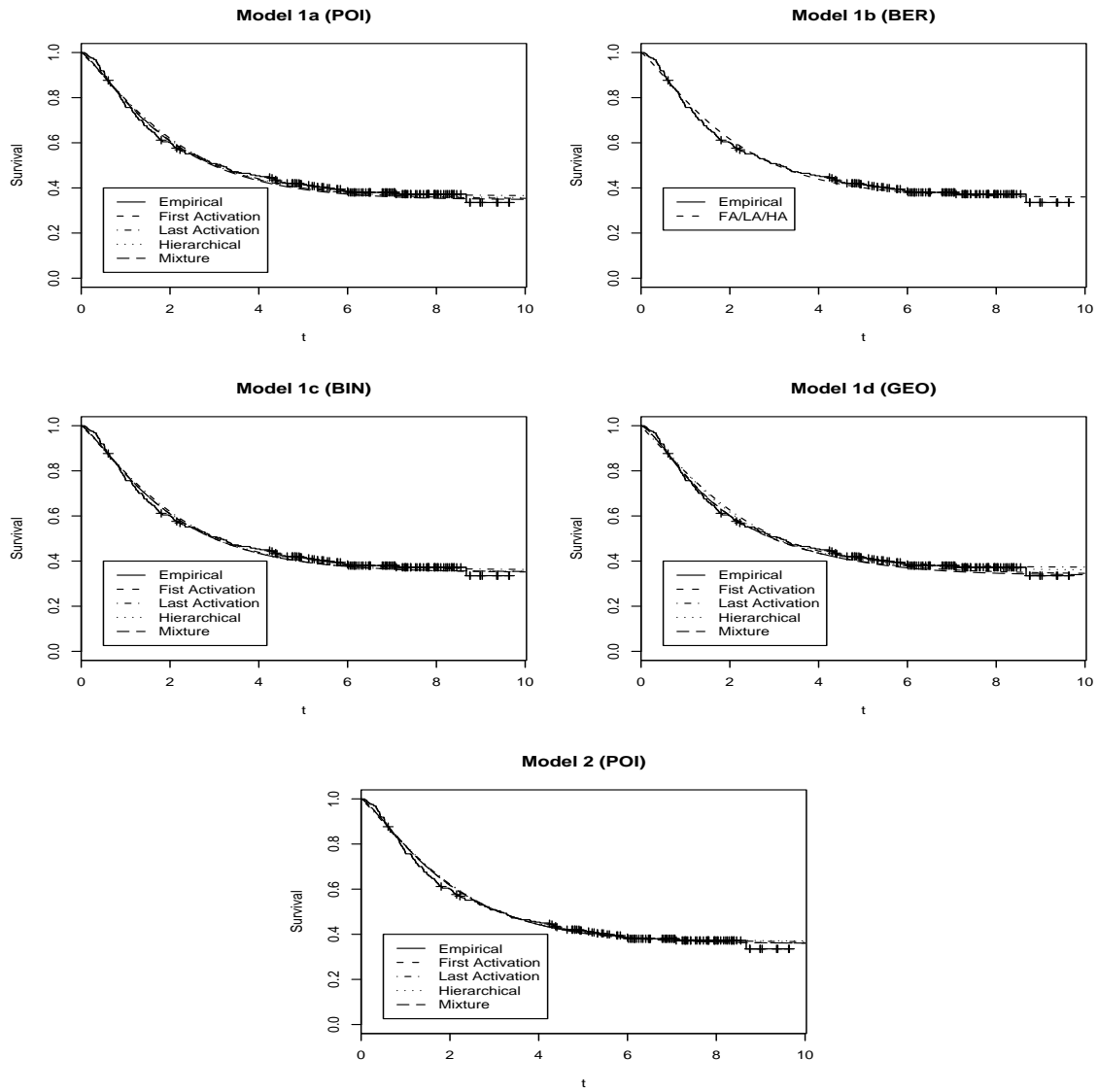


Figure 2: The posterior estimates (median) of the survival plots for the melanoma data under the different activation schemes showing adequate fits under the different schemes.

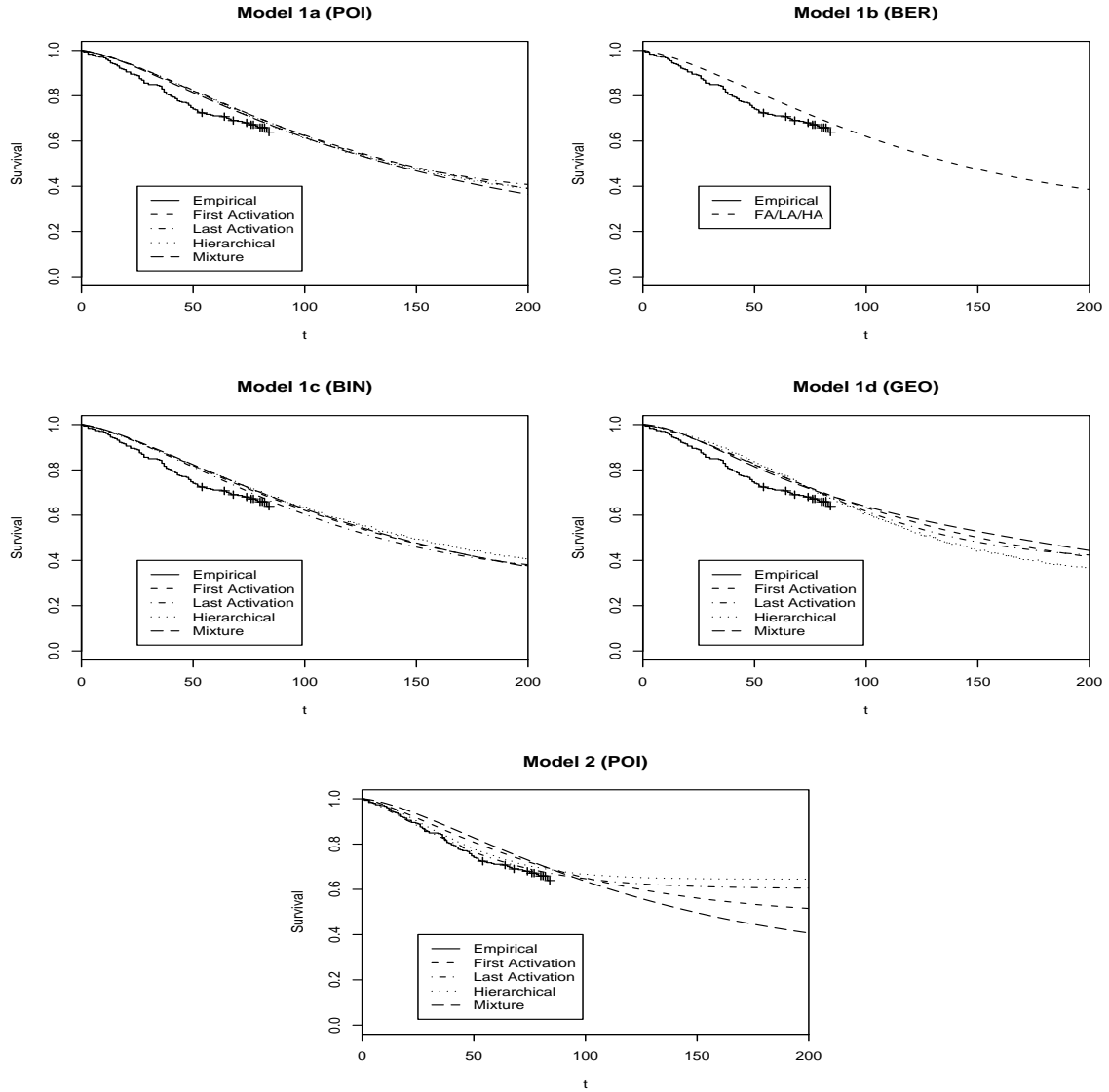


Figure 3: The posterior estimates (median) of the survival plots for the breast-cancer data under the different activation schemes. The Kaplan-Meier plot of the observed data is the solid line up to 84 months, while the fitted curves are extended up to 200 months to better reveal their shape. The estimated curves are much more sensitive to the modelling assumptions than for the melanoma data.