Flexible Cure Rate Modelling Under Latent Activation Schemes

FREDA COONER, SUDIPTO BANERJEE, BRADLEY P. CARLIN AND DEBAJYOTI SINHA

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: Sudipto Banerjee
telephone: (612) 624-0624
fax: (612) 626-0660
email: sudiptob@biostat.umn.edu

March 4, 2005

1Freda Cooner is Graduate Assistant, Sudipto 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 grant 1–R01–CA95995. Work of Dr. Sinha was supported by NCI grant 9–R01-CA69222.
Flexible Cure Rate Modelling Under Latent Activation Schemes

Summary

Survival models have been and continue to be extremely popular in analyzing time-to-failure data, where failure is disease relapse or death. With rapid improvements in medical treatment and health care, many survival data sets now reveal a substantial portion of patients who are cured (that is, who never experience the endpoint). Extended survival models called cure rate models account for the probability of a subject being cured. Popular cure models can be broadly classified into the classical mixture models of Berkson and Gage (1952; “BG type”) or the hierarchical classes of Chen, Ibrahim and Sinha (1999; “CIS type”). Recent developments in formulating 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 also elucidates the relationship between classical and hierarchical cure models. 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 model’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

As significant progress continues to be made in medical and health sciences, scientists and health professionals increasingly encounter data sets where patients are expected to be cured. Formulation and estimation of models that account for cure are important for understanding prognosis in potentially terminal diseases. Traditional parametric survival models such as Weibull or Gamma (see, e.g., Cox and Oakes, 1984) do not account for cure, assuming instead that individuals who do not experience the event are
The subtle distinction between censoring and cure is worth noting: a subject who does not fail within the time window of the experiment is considered “censored”, while a subject is cured if he will never fail. Clearly the latter is a more abstract concept in that we are never able to “observe” a cure (because of a finite monitoring time), yet there is interest in estimating the probability of such an outcome, especially in various cancer-relapse settings.

In order to address this conceptually challenging problem, models incorporating a cure fraction, called cure rate models, are formulated where the probability of a cure is parametrized. One of the earliest such models was a class of mixture models by Berkson and Gage (1952) (which we refer to as the BG model) which generated subsequent investigations by Farewell (1982, 1986), Goldman (1984) and Ewell and Ibrahim (1997) among others. Recently Yakovlev et al. (1993) and Chen, Ibrahim and Sinha (1999) offered an alternative approach to formulating cure models which we refer to as the CIS model (see also Ibrahim et al., 2001; Tsodikov et al., 2003; and Banerjee and Carlin, 2004). These models assume that a latent biological process is generating the observed failure (say, cancer relapse). More generally, there may be (perhaps several) latent factors or latent risks corresponding to each patient. For an individual to be at risk of failure, he/she must be exposed to at least one of these latent factors. If not, then the individual is not at risk and is considered cured. Failure is observed when one (or some) of these latent factors become activated. These models are scientifically appealing as they are motivated from mechanistic considerations of the disease under study, and are thus sometimes referred to as mechanistic models.

While mechanistic assumptions often lead to reasonable classes of models, such modelling is likely to vary with the experimental setting; see recent discussions by Tsodikov et al. (2003), Tucker and Taylor (1996) and Tucker, Thames and Taylor (1990) regarding the appropriateness of competing model assumptions for different biological/clinical settings. In certain situations, say where there is a single latent factor or dominant tumor driving the relapse, the BG model may be more appropriate, whereas when there are several latent factors, such as metastasis-competent clonogenic tumor cells generating cancer relapse, the CIS models appear more sensible. Most of the existing cure models in the literature are modifications of either the BG (see e.g. Sy and Taylor, 2000; Li and Taylor, 2002) or the CIS models (see e.g. Tsodikov et al., 2003). While scientific insight into the biological processes leading to disease manifestation can shed some light upon the appropriate model assumptions, controversy persists. In traditional survival analysis,
often the disease or biological phenomenon under study suggests a preferred assumption. In cure rate models, however, assumptions have to be made on the latent event distributions that are almost impossible to verify a priori, but whose relative merits can be evaluated a posteriori using a model assessment criterion. Therefore, what is needed is a versatile and unifying class of cure models that would not only clarify the relationship between the BG and CIS models, but also admit competing frameworks for different scientific hypotheses.

In this article we develop a general class of cure models based upon latent factors and their activation schemes that lead to observed failure. The key components in this formulation include the number of latent events/factors (e.g., metastasis-competent tumor cells) and the latent event times that bring about observed failure. We use a Bayesian hierarchical framework that models the distributions of the number of latent events and the latent event times, generating flexible classes that include the BG and CIS models as special cases, and investigate their characterizations and properties. We compare modelling covariates or risk factors in the latent mean versus in the cure fraction, and discuss subsequent identifiability issues and simulation-based model estimation techniques.

The remainder of the paper evolves as follows. In Section 2 we outline our general framework and discuss flexible cure rate modelling under different latent factor activation schemes. Section 3 considers the incorporation of covariates into our models. Section 4 discusses model fitting and comparison. Section 5 illustrates the performance of our models using 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 (http://www.seer.cancer.gov/). Finally, Section 6 summarizes and indicates future areas for research.

2 Flexible Cure Rate Models with Latent Factors

Cure models based upon latent activation schemes will involve failure times at two different levels: an observed failure time, say $T$, corresponding to the time when the individual fails, and the latent event times, $Y_k, k = 1, \ldots, N$, the activation times for the $N$ latent factors that generate the observed failure at
time $T$. Note that if $N = 0$ then the individual is not exposed to any of the latent factors and is considered immune from failure; thus he or she is cured and $T = \infty$. For a given $N$, the $\{Y_k\}_{k=1}^N$ are assumed to be independently and identically distributed with a survival distribution $P(Y > t) = S(t)$ that does not depend upon $N$. We call this the latent survival function and denote the corresponding distribution function by $F(t) = 1 - S(t)$.

A crucial issue for further development is relating the latent activation times to the observed failure times. Generally, if we assume that $r$ out of $N$ latent factors need to activate for the subject to fail, then we have $T = Y_{(r)}$, $r = 1, \ldots, N$ where $Y_{(1)} < \cdots < Y_{(N)}$ 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 $r|N$. In any case, note that an exposed subject ($N > 0$) at any time point will not experience detectable failure if the number of latent event occurrences at that time is less than $r$.

The conditional distribution of $T$ given $N$ and $r$ can be written down in terms of the incomplete beta function (or beta cdf, denoted by $IB(S(t); N - r + 1, r)$) as

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

(1)

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

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

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

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

(2)

where the expectation 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 $N \geq r \geq 1$. This restriction is implicit in our modelling since $IB(S(t); N - r + 1, r)$ is otherwise 0. Thus, when discussing the distribution of $(N, r)$ we are assuming a distribution (perhaps) truncated to $N \geq r \geq 1$. Also, since $\lim_{t \to \infty} S(t) = 0$, we have that $\lim_{t \to \infty} S^*(t) = P(N = 0)$, showing that $S^*(t)$ is improper whenever $P(N = 0) > 0$. Indeed,
then $P(N = 0)$ is the probability of a person being cured or immune, 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, say $h^*(t)dt \approx P(T \in [t, t + dt] | T > t)$, so that $h^*(t)$ is evaluated as $-\frac{d}{dt} \log S^*(t)$, or $f^*(t)/S^*(t)$, where $f^*(t)$ is the corresponding *improper* density. In fact, if $f(t)$ is the proper latent density corresponding to $F(t)$, then

$$f^*(t) = f(t)E_{(N,r)} \left[ N \left( \begin{array}{c} N - 1 \\ r - 1 \end{array} \right) [S(t)]^{N-r}[F(t)]^{r-1}1(N \geq r \geq 1) \right].$$

(3)

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 $r$, whence we take the expectations with respect to $N$ in (2). For example, the CIS models assume that $N$ has a Poisson distribution and *any one* of these latent events bring about the observed failure, so $r = 1$ and, conditional upon $N$, $T = \min_{1 \leq k \leq N} Y_k$. We call this the *first-activation* scheme. A different scheme assumes that an individual is able to “resist” up to $N - 1$ activations and fails with the last activation. This implies $r = N$ and $T = \max_{1 \leq k \leq N} Y_k$, which we call the *last-activation scheme*. In general $r$ will be unknown and can be modelled hierarchically with $N$ in a *hierarchical activation scheme*, where we specify a probability distribution for $r$ given $N$. We formulate classes of cure models under each of the above schemes, generalizing the BG and CIS models using the moment generating function (mgf) of the distribution of $N$ to characterize these classes.

### 2.1 “First-activation” scheme

The first-activation scheme assumes that activation of a single latent factor will lead to observed failure. The CIS models of Chen et al. (2003) can be considered as a special instance of this scheme that models metastatic tumor cells in a cancer relapse setting. According to a biological model for patients diagnosed with cancer, $N$ is the number of metastasis-competent clonogenic cells that are in an irreversible process towards metastasis, and $Y_k$ is the time for the $k^{th}$ clonogenic cell to produce “detectable” tumor. Detectable metastasis occurs as soon as *any one* of the clonogens metastasize so that $T = \min_{1 \leq k \leq N} Y_k$. This arises as a special case of (1) with $r = 1$ so that $P(T \geq t | N) = 1(N = 0) + IB(S(t); N, 1)1(N \geq 1)$, which simplifies to $P(T \geq t | N) = 1(N = 0) + [S(t)]^N1(N \geq 1)$. As mentioned above, the CIS model assumes $N \sim Po(\theta)$. The arguments for using this assumption to model number of clonogenic cells left in a patient’s body after
radiation are put forward by Hanin et al. (2001) among others.

According to several authors including Tucker et al. (1990) and Tucker and Taylor (1996), it is at best debatable whether the Poisson assumption is valid irrespective of any particular cancer cure scenario. As a generalization of the CIS model, the number of possible latent events \( N \) can have any finite-mean integer-valued distribution (e.g. binary, Poisson, negative binomial, etc.) with moment generating function \( m(t) = E[\exp(tN)] \) and a cure fraction defined as \( P(N = 0) = m(-\infty) \). The marginal distribution of \( T \) is given in terms of the moment generating function of \( N \) as:

\[
S^*(t) = E_N[P(T \geq t|N)] = m[\log S(t)] \tag{4}
\]

For example, in the CIS model with \( N \sim \text{Po} (\theta) \), we have \( m(t) = \exp[-\theta(1-e^t)] \) and the marginal cure rate model is available in closed form as: \( S^*(t) = \exp(-\theta(1 - S(t))) \), an improper distribution with cure fraction \( \exp(-\theta) \). In traditional cure rate models (e.g., the BG model and in Farewell (1986)), \( N \) is binary with only one latent dominant event (e.g., a dominant metastasis competent tissue-mass in breast cancer), so \( N \sim \text{Ber} (\theta) \) (with \( \theta \) being the probability of an activation) and \( m(t) = 1 - \theta(1 - e^t) \). Analogously, in the BG model we have \( S^*(t) = 1 - \theta(1 - S(t)) \), approaching the cure fraction \( 1 - \theta \). Note that the support for \( \theta \) is different in the two settings: in the former it is a proportion between \((0, 1)\), while in the latter it is a Poisson mean taking any positive real value. Also, the formulation in (4) provides us with a class of cure-rate models that includes both the CIS and BG models as special cases. An important advantage of (4) lies in its flexibility to go beyond the CIS and BG models and help understand the biology of disease occurrence/relapse suitable for the application in hand.

An extension of the BG model, say the \( \text{BG}(K) \) model, follows by assuming \( N \) to be the number of latent factors, out of a possibly unknown but fixed number of latent factors (the same for each patient), say \( K \), that the individual is subjected to. This is similar to the \( m \)-site cancer model discussed by Gail et al. (1980) among others where \( K \) represents the number of dominant mutation sites within a disease location and we assume that for each patient an unobservable (random) \( N \) of these sites get mutated. Again, a single mutation (activation) of any one of these sites results in the observed failure/occurrence, so that \( T = \min_k Y_k \) as above. Here \( N \sim \text{Bin}(K, \theta) \), so that \( m(t) = (1 - \theta(1 - e^t))^K \) implying the marginal failure time distribution \( S^*(t) = (1 - \theta(1 - S(t)))^K \), with cure fraction \((1 - \theta)^K\).

Another specification we consider here is \( N \sim \text{Geo}(\theta) \), a geometric distribution with mean \( \theta/(1 - \theta) \)
\((P(N = n) = \theta^n(1 - \theta))\). A biological motivation, different from the clonogen motivation in Chen et al. (2003), can be offered as follows (also see Moolgavkar, Luebeck, and De Gunst, 1990). Assuming a short time-interval of the mutation (initiation) period (due to exposure to genetic damage), the patient’s body produces a sequence of \(N\) mutated cells/tissues before activating the immune system, but each of which are at risk of activation. Every mutation (initiation) may give rise to an effective immune response from the body with probability \(1 - \theta\), which is capable of destroying the last mutated tissue/cell and halting the mutation process. With \(\{Y_k\}_{k=1}^N\) now being the promotion times of the mutated cells, a first-activation scheme with \(T = \min_k Y_k\) models failure with any one activation. Also, here \(m(t) = (1 - \theta)/(1 - \theta e^t)\) which results in \(S^*(t) = (1 - \theta)/(1 - \theta S(t))\) with a cure fraction of \(1 - \theta\).

A major advantage of (4) is that it allows us to investigate the identifiability of the model parameters in the presence of covariates and observed data when we use any particular hierarchical model for \(N\). To address the identifiability of the parameters within a model, we only need to consider the marginal survival function via (4) and assess whether available data from the survival function can identify each parameter (see Section 3). The properties of the subclasses of cure rate models based upon the first activation scheme can also be addressed via equation (4); see Section 3.

2.2 “Last-activation” scheme

While certainly reasonable, the physical framework of the first activation scheme is not unique in assuming that only one latent factor needs to be activated for failure. Alternatively \(N\) can be the number of latent factors that must all be activated for failure. For instance, in certain types of cancer such as breast cancer, a patient’s immune response gets activated after the initiation of \(N\) cell mutations. This immune response may be able to resist up to \(N - 1\) promotions of mutated cells before disease manifestation or death. Note that \(Y_k\) now is the time to promotion of the \(k^{th}\) latent factor and failure occurs after the \(N^{th}\) factor is activated, so the observed failure time is \(T = \max Y_k, k = 1, \ldots, N\). Again, we have a special case of (1) with \(r = N\) and \(P(T \geq t|N = 1(N = 0) + IB(S(t); 1, N)1(N \geq 1))\). The distribution of \(N\) can again be chosen as any integer-valued distribution. Here we persist with the same choices as for the first activation scheme: Poisson, Bernoulli, binomial and geometric.

The conditional distribution of \(T\) given \(N\) is now easily expressed in terms of the latent distribution
function \( F(t) = 1 - S(t) \) as \( P(T \leq t|N) = [F(t)]^N 1(N \geq 1) \). Letting \( F^*(t) = E_N[P(T \leq t|N)] \), we obtain the marginal distribution of the failure time as \( F^*(t) = m[\log F(t)] - m(-\infty) \). Since \( S^*(t) = 1 - F^*(t) \), we have
\[
S^*(t) = 1 + m(-\infty) - m[\log F(t)]
\]
which approaches \( m(-\infty) \), the cure fraction. Note that although \( S^*(t) \) in (5) is different from the one in (4), they tend to the same limit, \( m(-\infty) = P(N = 0) \), resulting in the same cure fraction. 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, \( \exp(-\theta) \).

In the \( BG(K) \) setup, recall \( m(t) = (1 - \theta (1 - e^t))^K \), as before, producing \( S^*(t) = 1 + (1 - \theta)^K - (1 - \theta S(t))^K \), with cure fraction \( (1 - \theta)^K \). Note that with \( K = 1 \), this reduces to \( S^*(t) = 1 - \theta F(t) \), which is equal to that in the BG model. This is intuitively clear; indeed with only a single latent factor, the first and last activations must be identical. For \( N \sim Geo(\theta) \), using the mgf of the geometric distribution we obtain, after some algebra, \( S^*(t) = 1 - (1 - \theta)\theta F(t)/(1 - \theta F(t)) \), with the cure fraction again being \( 1 - \theta \).

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.

Note that analogous to (4), (5) characterizes \( S^*(t) \) for last activation in terms of the moment-generating function of \( N \), and helps in deriving hazard functions and understanding regression (see Section 3).

### 2.3 Hierarchical activation schemes

Both the first- and last-activation schemes assume that \( r \), the number of promotions/activations needed for observed failure is either fixed or a completely specified function of \( N \). In spite of scientific motivations, these latent factors are never observed and \( r \) may be best modelled with a conditional probability distribution given \( N \). This suggests a “hierarchy” to our model, where we specify the joint distribution of \( r \) and \( N \) through a marginal specification for \( N \) and a conditional distribution for \( r \) given \( N \).

Recall that for \( N \geq 1 \) the support of \( r \) is the set of positive integers not exceeding \( N \). The conditional
distribution of $T$ given $N$ and $r$ remains as in (1), and $S^*(t)$ can be derived from (2) to yield

$S^*(t) = P(N = 0) + E_N \left[ E_{r|N} \left[ IB(S(t); N - r + 1, r) \right] \right]$

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

$= P(N = 0) + E_N \left[ 1(N \geq 1)N \int_0^{S(t)} E_{r|N} \left[ \left( \frac{N - 1}{r - 1} \right) u^{N-r}(1-u)^{r-1} \right] du \right]. \quad (6)$

A particularly simple but interesting result follows by specifying the conditional distribution of $r$ as $DiscreteUnif(1, N)$ (discrete uniform). Then (6) simplifies to (see item 1 in the appendix)

$S^*(t) = P(N = 0) + S(t)(1 - P(N = 0)), \quad (7)$

which is a classical BG-type model with cure fraction $P(N = 0)$ depending upon the distribution of $N$. This shows that even for a $N \sim Po(\theta)$ assumption, similar to the CIS model, using a noninformative $DiscreteUnif(1, N)$ for $r$ results in a BG-type model. Also note that if $N \sim Ber(\theta)$, then $N$ is binary and we recover precisely the BG model.

If $N \sim Po(\theta)$, we may specify $r | N \sim Bin(N, \pi)1(N \geq r \geq 1)$ which, though analytically intractable, can be computed using Markov chain Monte Carlo (MCMC). This scheme is more flexible in that the first and last activation schemes arise in this framework by assuming degenerate point-mass conditional distributions for $r$. Note that the hyperparameter $\pi$ can, at least conceptually, be assigned a suitable hyperprior (e.g. a uniform or beta distribution). However, in practice we do not expect the data to inform much about this hyper-parameter, so we can fix it (say at $\pi = 1/2$). In fact, this is a clear advantage of the Bayesian mechanism as a richer framework is obtained without impairing model identifiability.

3 Regression in cure models

In general, (3) shows $f^*(t)$ in a potentially complex relationship with the latent density function $f(t)$ and the other parameters. The corresponding hazard function, $h^*(t) = f^*(t)/S^*(t)$ is derived from the expressions in (2) and (3). Simplifications arise in the first- and last-activation schemes. Under the first-activation scheme, we set $r = 1$ in (2) and (3) to obtain

$h^*(t) = \frac{E_N[N(S(t))^N]}{E_N[(S(t))^N]} h(t), \quad 9$
where \( h(t) = f(t)/S(t) \) is the latent hazard function. Analogously, for the last-activation scheme we set \( r = N \) in (2) and (3) to obtain the hazard as

\[
h^∗(t) = \frac{E_N[N(F(t))^{N-1}1(N \geq 1)]}{1 + P(N = 0) - E_N[(F(t))^N]}f(t),
\]

while for hierarchical-activation with \( r | N \sim \text{DiscreteUnif}(1, N) \), we have

\[
h^∗(t) = \frac{P(N \neq 0)}{P(N = 0) + S(t)P(N \neq 0)}f(t).
\]

In all the above schemes \( h^∗(t) \) depends upon the distributions of \( N \) and \( r \), which can sometimes lead to closed form expressions. Also, as for \( S^∗(t) \), we can characterize the hazard functions for the first and last activation schemes using the mgf of \( N \) as

\[
h^∗(t) = \frac{\frac{d}{dt}[m(\log\{S(t)\})]}{m(\log\{S(t)\})} \quad \text{and} \quad h^∗(t) = -\frac{\frac{d}{dt}[m(\log F(t))]}{1 + m(\log F(t))},
\]

respectively.

In the implementations that follow, 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 \( x \) in the Weibull scale parameter \( \eta = \eta(x) \), which ensures 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 \( S^∗(t|x) \). In particular, whenever the regressor \( x \) is modelled via the activation time using an accelerated life model structure as \( S(t|x) = S(t\phi(x)) \), the corresponding cure rate model under the first activation scheme is also going to be an accelerated life model with \( S^∗(t|x) = m(\log\{S(t\phi(x))\}) = S^∗_0(t\phi(x)) \). This is the case irrespective of the form of the mgf \( m(u) \) as long as it is free of \( x \). Under the condition that \( m(u) \) exists, we can show, using the uniqueness of the mgf, when \( S(t) \) does not depend on \( x \) the cure model cannot have an accelerated failure time structure. Similarly, we can show that when \( S(t|x) = S(t\phi(x)) \) and the \( S^∗(t|x) \) is of accelerated failure time form then the mgf is free of \( x \). 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 CIS model, when \( N \) is Poisson with mean \( \theta = \theta(x) \) and \( S(t) \) is free of \( x \), we obtain a proportional hazards structure for \( h^∗(t|x) = \theta(x)f(t) \). On the other hand with \( N \sim Bin(K, \theta) \), we obtain \( h^∗(t) = K\theta f(t)/(1 - \theta F(t)) \) which does not render a proportional hazards structure, irrespective of whether we regress through \( \theta(x) \) or \( \eta(x) \). Similarly, with \( N \sim Geo(\theta) \), we find \( h^∗(t) = \theta f(t)/(1 - \theta S(t)) \),
again precluding proportional hazards for $\theta(x)$ or $\eta(x)$. In fact, using the uniqueness property of the Poisson mgf, we can show that when $S(t)$ is free of $x$ and $h^*(t|x)$ has a proportional hazards structure then $N|x$ has a Poisson distribution with mean $\theta = \theta(x)$, a function of $x$. These characterization results show that the available observed data can inform about the form of $S^*(t|x)$, and thus help us 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. With the first we obtain

$$h^*(t) = \frac{\theta e^{-\theta S(t)} f(t)}{1 + e^{-\theta (1 - e^{\theta F(t)})}},$$

while for the binomial we have

$$h^*(t) = \frac{K \theta (1 - \theta S(t))^{K-1} f(t)}{1 + (1 - \theta)^K - (1 - \theta S(t))^K},$$

which reduces to $\theta f(t)/(1 - \theta F(t))$ for $K = 1$, as in the first activation scheme. Finally for the geometric we obtain

$$h^*(t) = \frac{\theta (1 - \theta)}{(1 + \theta (\theta - 2) F(t))(1 - \theta F(t))} f(t).$$

Clearly, the hazards structure for the last activation scheme is more complex, and perhaps less intuitive, than for the first activation setting. Nevertheless, these provide valid regression structures and may provide better fits in situations where proportional hazards or other simpler models are inappropriate.

With regard to a proportional hazards structure for $S^*(t)$, therefore, the advantage of regressing on $\theta$ is seen only in the Poisson mgf under the first activation scheme (CIS model). For a Bernoulli or Binomial mgf, regressing on $\theta$ is precluded with improper priors on regression coefficients as they produce improper posteriors, while with Poisson mgf’s we recover proper posteriors under fairly general conditions (Chen et al., 2003). In the appendix (item 2) we show that under similar conditions, this propriety carries over to the last-activation setting. On the other hand, regressing through $\eta$ in the latent Weibull distribution is always valid (i.e. yields proper posteriors, even with improper priors) with the appropriate link, mimicking generalized linear modelling.

To see how the foregoing assumptions lead to flexible classes of models, consider the setting with $I$ subjects, where the $i^{th}$ subject has a vector of risk factors (covariates) $x_i$. Letting $\eta_i$ and $\theta_i$ be the respective analogues of the foregoing $\eta$ and $\theta$ for subject $i$, we can build upon the cure models in the first or last-
activation settings by letting either $\eta_i = x_i^T \beta$ or $\theta_i = g(x_i^T \beta)$, where $g$ is a suitable link mapping onto the positive real line (recall that $\theta_i$ has positive support).

We thus classify cure rate models based upon the activation scheme, the distribution of $N$, and the regression. In particular, we will consider each of the following models under different activation schemes:

Model 1(a): $\eta_i = x_i^T \beta; \theta_i = \theta; N \sim Po(\theta)$

Model 1(b): $\eta_i = x_i^T \beta; \theta_i = \theta; N \sim Ber(\theta)$

Model 1(c): $\eta_i = x_i^T \beta; \theta_i = \theta; N \sim Bin(K, \theta)$ ($K$ known)

Model 1(d): $\eta_i = x_i^T \beta; \theta_i = \theta; N \sim Geo(\theta)$

Model 2: $\log(\theta_i) = x_i^T \beta; \eta_i = \eta; N \sim Po(\theta)$

Models 1(a)–1(d) regress on the latent Weibull mean, allowing proper models irrespective of the distribution of $N$, but keep a constant cure parameter $\theta$. Model 2 regresses on the cure-parameter $\theta$ (which 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). The first and last activation schemes can be implemented in each of the models, but the hierarchical activation scheme with $r|N \sim DiscreteUnif(1, N)$ is precluded in Model 2. Recall from (7) that this results in a BG-type model that does not permit regression in the cure fraction.

Note that the interpretation of the parameters will be different for the two sets of models. In models 1(a)–(d), higher values of $\eta_{ij}$ indicate longer latent event times and hence prolonged survival, so covariates influence the latent factors that cause failure. On the other hand, in model 2, covariates influence the cure fraction directly. Almost certainly most covariates that affect the latent event-times will also affect cure. For these covariates we have a consistency in behavior under the two groups of models. To see this, note that the Weibull mean is a decreasing function of $\eta_i$ (given by $\exp(-\eta_i/\rho)\Gamma(1 + 1/\rho)$) and the cure fraction is a decreasing function of $\theta_i$. This is sensible because covariates that adversely affect survival time will likely affect the cure fraction adversely as well (and vice-versa). Therefore, it is expected that their coefficients have the same sign under both sets of models.

A natural question that arises here is that of feasibility and identification of regressions in both $\eta$ and $\theta$. Over and above what has been said about regression in the above models, note that such a setting will
not necessarily yield proportional hazards or accelerated failure structures in $h^*(t)$, which would lead to unidentifiable models. In a different context, we foresee this as a more relevant issue in developing spatial models for geographically referenced cure data (see, e.g., Banerjee and Carlin, 2004). Here one encounters competing models concerning geographically varying cure fractions or geographically varying link functions and merits separate investigation.

4 Bayesian estimation and model comparisons

For the $i$th individual, our observed data consists of $D_i = (t_i, \nu_i, x_i)$, where $t_i$ is the observed failure time, $\nu_i$ is the failure indicator, and $x_i$ is a set of covariates. We collect the model-specific parameters (and hyperparameters) into $\Omega_i$. Generally we write $\Omega_i = (\eta_i(\beta), \theta_i(\beta), \rho, \psi)$ 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 $\rho$, the contribution of subject $i$ to the data likelihood (in a right-censored setting) is

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

where $P(T \geq t_i | N_i, r_i)$ is as in (1), and $-\frac{d}{dt} P(T \geq t_i | N_i, r_i) = N(N, N-1)[S(t_i)]^{N-r_i}[F(t_i)]^{r_i-1} f(t_i)$. We seek the posterior distribution of $\{\Omega_i\}$ (marginalized over $N_i$ and $r_i$), which can be written down in generic notation as

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

where $P(\Omega_i, N_i, r_i)$ are the joint prior probabilities. Note that for Models 1(a)–(c) we have $P(\theta_i(\beta)) \equiv P(\theta)$ and $P(\eta_i(\beta)) \equiv P(\beta)$, 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 considering a marginalized likelihood. In fact, marginalizing the $(N_i, r_i)$’s out of (8) reduces the estimation space considerably, and amounts to evaluating

$$E_{N_i, r_i|\theta_i} \left[\left(-\frac{d}{dt} P(T \geq t_i | N_i)\right)^{\nu_i} \right] (P(T \geq t_i))^{1-\nu_i}.$$
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 $\left(h^*(t_i)\right)^{\nu_i}S^*(t_i)$, yielding the data likelihood as

$$L(\{D_i\}; \{\Omega_i\}) = \prod_{i=1}^{I} \left(h^*(t_i)\right)^{\nu_i}S^*(t_i).$$

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

With the broad range of models that the above formulation encompasses, model evaluation and selection become important issues. In order to assess the goodness of fit for the different models, we perform posterior predictive comparisons. In general, letting $\Omega$ be the set of parameters for which inference is desired, we generate (future) data replicates $t^* = \{t^*_i\}_{i=1}^{I}$ from the posterior predictive distribution,

$$P( t^* | \{D_i\}) = \int P( t^* | \Omega, \{\nu_i, x_i\}) P( \Omega | \{D_i\}) d\Omega, \quad (9)$$

where $P( t^* | \Omega, \{\nu_i, x_i\})$ is the underlying probability distribution in the data likelihood. We do this by composition: given post-convergence posterior samples of size $M$ from MCMC output, say $(\Omega_1, ..., \Omega_M)$ from $P( \Omega | \{D_i\})$, for each $\Omega_j$ we generate a predictive data replicate $t^*_j$ (an $I \times 1$ vector) by drawing from $P( t^*|\Omega_j, \{\nu_i, x_i\})$. This yields an $I \times M$ posterior predictive ensemble matrix, say $T^* = [t^*_1 : \cdots : t^*_M]$, with the $i$-th row being 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 $P( t^* | \Omega, \{\nu_i, x_i\})$. 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 $\Omega_j$, $j = 1, \ldots, M$ for the required parameter estimates.

The algorithm proceeds by simulating the latent events leading to the observed failure/censoring time for each individual, using the posterior sample values for the parameters in $\Omega$. For instance, if $\nu_i = 1$, then subject $i$ has failed, so we first generate $N_i$ from the truncated $Po(\theta_i(\beta))1(N_i \geq 1)$ 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 $Y_1, \ldots, Y_N$ from the Weibull family and set $T_i = Y_{(r_i)}$, its $r$-th order statistic. 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 $Po(\theta_i(\beta))$ 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 that is very large relative to 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 any number of models may provide adequate fit to the data, each model represents a hypothesis (or a set of hypotheses), and so it is beneficial to have a framework for choosing among them. Here we adopt the posterior predictive L-measure (Laud and Ibrahim, 1995), computed as the sum of a goodness-of-fit measure and a penalty term. This approach has decision-theoretic justifications (see, e.g., Gelfand and Ghosh, 1998), treating model choice as a minimizing decision rule for a squared-error loss function. Here, the model assessment measure \( L \) may be written as

\[
L = E_{t^* | \{D_i\}} \left[ (t^* - \mu^*)^T (t^* - \mu^*) \right] + \frac{\delta}{\delta + 1} (t - \mu^*)^T (t - \mu^*),
\]

where \( \mu^* = E_{t^* | \{D_i\}} [t^*] \) is the posterior predictive mean of the replicated data, and \( t \) is a vector of the observed time-points \( \{t_i\} \). Denote the first term in (10) (which equals \( tr(Var(t^*|t)) \)) by \( P \) and let \( G = (t - \mu^*)^T (t - \mu^*) \), since the former acts as a penalty term while the latter acts as a goodness-of-fit measure (lower values of \( G \) indicate better fit). Note that \( L \) tends to \( P + G \) as \( \delta \to \infty \). Computation of \( G \), \( P \) and \( L \) proceeds by first computing \( \mu^* = \frac{1}{M} \sum_{i=1}^{M} t_i^* \) and then plugging in to compute \( P = \frac{1}{M-1} \sum_{i=1}^{M} (t_i^* - \mu^*)^T (t_i^* - \mu^*) \) and \( G = (t - \mu^*)^T (t - \mu^*) \). 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 values of the L-measure suggest better models in terms of predictive fit, and simulation draws from different likelihoods can be fairly compared. We note that alternative model comparison measures such as the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), while computationally simpler, are not suitable here because of the lack of concavity in our log-likelihoods, which can seriously damage the quality of DIC estimates. 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. The use of DIC for comparing such non-nested likelihoods entails computation of normalizing constants that greatly detracts from its computational simplicity.
We illustrate our models by implementing the algorithms discussed in the previous section by first designing a simulation study for assessing performances of different activation schemes, and then for data sets of two different cancers (melanoma and breast), for comparing both activation schemes as well as modelling assumptions. For the melanoma data, failure is defined to be the relapse of melanoma. These data have been analyzed using CIS modelling (Chen et al., 1999), who showed a pronounced “plateau” effect in its survival curve. The breast cancer data, where the failure is defined to be death from breast cancer, does not reveal such a dominant plateau effect, at least in the time frame of the study. Yet, cure modelling for breast cancer survival is particularly pertinent, since prognosis has shown marked improvement over recent years, with data tending to reveal substantial cure fractions. Indeed, investigators seek cure fraction estimates, rather than pure survival probabilities, for assessing treatment efficacy. We remark that often deaths from many aggressive cancers are caused due to a relapse of the disease, and the time of relapse is practically indistinguishable from the time of death.

Under all three activation schemes in our analysis of the cancer data, we assumed that the regression coefficients $\beta$ had a non-informative $N(0, 10^3)$ prior (flat priors are admissible here as well), while a relatively weak Gamma$(0.001, 0.001)$ prior was used for $\rho$. For Model 2, a $N(0, 100)$ prior for $\eta$ was used, while a Gamma$(0.001, 0.001)$ prior was used for $\theta$ in Model 1a, and $U(0, 1)$ priors were used for $\theta$ in Models 1b and 1c. These priors are vague enough to allow the data to drive the posterior inference, while still leading to acceptable MCMC convergence. We experimented with other hyperparameter values and did not observe much sensitivity to these choices in the posterior distributions. For the hierarchical activation scheme we selected $r \mid N \sim Bin(N, \pi)$ with $\pi = 1/2$. A beta or uniform prior on $\pi$ would also be valid, but experimentation did not reveal much learning about this parameter from the data, so we present our results only with this fixed value.

For each analysis, we ran two initially dispersed parallel MCMC chains for 30,000 iterations each, where convergence was monitored using sample autocorrelations within the chains, cross-correlations between the parameters, and plots of the sample traces. These tools suggested discarding the first 5,000 iterations from each chain as pre-convergence burn-in. Retaining every 10th of the remaining $2 \times 25,000 = 50,000$ iterations yielded a final sample of size 5,000 for posterior analysis. The first- and last-activation models were imple-
mented in WinBUGS (see www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml) using R (www.r-project.org) for final posterior summarization. Computations for the hierarchical-activation model had to be carried out entirely in R, particularly for accessibility to the incomplete beta functions.

5.1 Simulation study

In order to assess the performance of different activation schemes, we designed a study that generated data from known model specifications and activation schemes that were subsequently analyzed by fitting the various models. For each activation scheme, first (FA), last (LA) and hierarchical (HA), we generate data sets from Model 1c, Model 1d and Model 2. In order to keep the design simple, we worked with an intercept as the sole regressor. This renders a single $\theta$ parameter and a single $\eta$ parameter, whence Model (1a) becomes a simple reparametrization of Model 2. Also, Model 1b, the classical BG-type model, does not distinguish between the different activation schemes, so we do not consider it here.

For Model 1c and Model 2, each data set was generated with a cure fraction of 20% and a random censoring rate of around 25%. This cure fraction implies $\theta = 0.15$ for Model 1c and $\theta = 1.61$ for Model 2. For Model 1d, because of the heavier tail of the geometric distribution, we generated data using a cure fraction of 40% (correspondingly $\theta = 0.6$) to maintain comparable values of the simulated $N$ and ensure numerical stability. Also, we set the Weibull link parameter $\eta = -3$ and the scale parameter $\rho = 1.5$ for Model 1c, Model 1d, and Model 2. For each generated data set, the respective model was fitted under all three activation schemes. We employed priors similar to those described above, except we used a $N(0, 10^3)$ vague prior for $\log(\theta)$ in Model 2, since we do not have $\beta$ here. Subsequent MCMC iterations were carried out as already described and consistent estimates of the model parameters were found. Figure 1 shows the Kaplan-Meier plots of the generated data along with the posterior median plots of the estimated $S^*(t)$ under the respective model and activation assumptions. Almost all the plots seem to indicate adequately good fits to the data, capturing the “plateau” effect in all cases.

Table 1 shows the model comparison scores using the L-measure ($L = G + P$) for each of the models. Each row corresponds to a different data set, so comparisons are valid only along a row. The first two columns represent the model and activation scheme used to generate the data set (the “true” model). Subsequent columns indicate the performance of a particular model and activation scheme for these data.
Entries corresponding to the best-fitting (smallest L-measure) models are shown in boldface. The excellent performance of the hierarchical model, even when the true mechanism is first or last, features prominently in the table. Except for the first-activation scheme under Model 1d and Model 2, the hierarchical activation schemes always perform the best – even better than the true underlying scheme. This is not totally unexpected, given that the hierarchical encompasses the first and last activation schemes, without over-parametrization. We also observe that, while the goodness-of-fit measure (G) does not vary much between these schemes, the penalty (P) from the misspecified models do, and are usually better for the hierarchical models. Between just first and last activation, the true model specification always outperforms the other.

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 3) 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 all three activation schemes coincide; hence, the L-measures are same across this row. Model 1d (regressing on the Weibull mean) for the first-activation scheme has the lowest L-measure score among all the different models under any activation scheme with a marginal improvement over the CIS model and more pronounced improvements over the BG-type model. Model 1b seems to provide the best model under the last-activation implementation, while Model 1c wins under the hierarchical activation scheme.

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 probability of cure 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 are consistent with the results reported by Chen et al. (2001) (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.

The first column of Figure 2 plots the individual-specific cure fraction estimates (posterior medians) for the three data activation schemes as provided by Model 2 (regressing on the cure-fraction) 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 3 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 posterior predictive L-measures for the different models for the breast cancer data set. Clearly, the first-activation scheme (CIS model) does not perform well here. Indeed, the hierarchical- and the last-activation frameworks with regression in the cure fraction (Model 2) seem to be more desirable, with the hierarchical activation Model 2 having the lowest L-measure score.

In Table 5, we present parameter inference results from last activation Model 1d along with those from Model 2 for the first and hierarchical schemes (the best models under each scheme). For all the three models, we find a significant positive impact for age at diagnosis (with later diagnosis corresponding to greater hazard). The number of primary nodes has a positive estimate, significantly so for last and hierarchical activation. Also, as expected, regional or distant stage lead to significant hazard increases relative to local. Finally, the Weibull scale parameter $\rho$ is robustly estimated across the models.

Turning our attention now to the median cure fraction estimates for the breast cancer data in the second column of Figure 2, we note that the variation in the cure fraction estimates is not as tightly centered about the fixed cure fraction estimate as for the melanoma data. Unlike the melanoma data, the differences in performance for the breast cancer data 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 4, where the Kaplan-Meier plot for the survival data runs only up to 84 months, but we plot our fitted curves up to 200 months to see the plateau. Generally, we find the fit of Models 1(a)–1(d) (including BG-type models) to be less satisfactory than that of Model 2, showing the inadequacy of a constant cure fraction. In Model 2 the last and hierarchical activation schemes offer better fit than first activation (CIS model), and nicely capture the tendency of the survival curve to flatten for the observed data after 50 months. These plots also bring out the differences in the plateaus much more distinctly than the simulated and the melanoma data. Generally, we find that last-activation reaches the plateau fastest, first activation the slowest, and (not surprisingly) hierarchical is between. This is expected because of the more stringent conditions for failure in last activation than first. Finally, we point out 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 CIS models and the improvements available using more flexible hierarchical methods.

6 Summary

In this article we have proposed a general class of cure models motivated from latent activation mechanisms. We have outlined a Bayesian hierarchical framework allowing for a rather general class of models, which can be fitted using Markov chain Monte Carlo (MCMC) algorithms – some of them using standard software. Formal statistical inference and model comparisons can be carried out using posterior simulations.

While mechanistic modelling based upon underlying scientific phenomena for the specific disease has become an important modelling tool, we believe that our general framework, along with the implementation algorithms, will provide important benefits to clinical researchers and biostatisticians. We often come across survival data for diseases whose 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 well conceal such a plateau. Our illustrations with melanoma data reveal our methods to be performing consistently with existing models, while our breast cancer example clearly demonstrated better performances by the hierarchical methods. In situations where mechanistic assumptions are in doubt, our templates should be useful for the data analysts and modelers for fitting and comparing different mechanistic hypotheses.

We conclude by noting that our framework can be extended in a number of directions. For instance, the parametric Weibull distribution for the latent event times may appear restrictive. Any of the non-parametric alternatives common in traditional survival analysis can be used. We believe that this will lead to more adaptive model fits than provided by the Weibull scheme. In particular, there appears more sensitivity to the modelling for data that may not be as smooth, such as our breast cancer data, and nonparametric modelling might be desirable there. Along a different route, it is now common to find spatially referenced survival data sets. In fact, Banerjee and Carlin (2004) analyzed data from a spatially referenced interval-censored smoking-cessation study in southern Minnesota using the simple BG-
type models. Spatial associations are incorporated into such models using random effects or frailties that are correlated. Our framework can certainly accommodate such modelling, but interesting issues arise regarding capturing the associations in the latent link or in the cure fraction, or perhaps in both. Finally, one might explore temporal and spatiotemporal extensions by modelling cure fractions as functions of time and space-time.

**Appendix**

1. Here we derive (7), which shows how the hierarchical framework leads to a Berkson-Gage type model.

Taking the conditional distribution with respect to $r \mid N \sim \text{DiscreteUnif}(1, N)$ in (6), we obtain

\[
S^*(t) = P(N = 0) + E_N \left[ \sum_{r=1}^{N-1} \frac{(N-r)(1-u)^{r-1}}{r} \right] 
= P(N = 0) + E_N \left[ \sum_{r=0}^{N-1} \frac{(N-1-r^*)(1-u)^{r^*}}{r^*} \right], \text{ where } r^* = r - 1,
\]

\[
= P(N = 0) + E_N [1(N \geq 1)S(t)] \quad (\text{the above binomial expansion equals 1})
\]

\[
= P(N = 0) + S(t)E_N [1(N \geq 1)] = P(N = 0) + (1 - P(N = 0))S(t).
\]

2. We now prove the propriety of parameter estimates (under mild conditions) for the regression in the cure fraction for the last-activation scheme. Consider the last-activation scheme with $N \sim \text{Po}(\theta)$. Without loss of generality assume that the model matrix $X$ is of full column rank. Without loss of generality assume that there are $k$ individuals who experienced the events and that $\nu_1 = \ldots = \nu_k = 1$. Let $u_i = x_i^T \beta$, $i = 1, \ldots, k$, where $x_1^T, \ldots, x_k^T$ are the regressor vectors for those $k$ individuals. This is summarized as $u = X^* \beta$, where $X^*$ is the $k \times k$ full-rank matrix, resulting in a one-one onto relationship between $u$ and $\beta$. So, for $i = k + 1, \ldots, n$ there exists unique $v_i = (v_{1i}, \ldots, v_{ki})^T$ such that $x_i^T \beta = \sum_{j=1}^{k} u_j v_{ji}$.

Note that the likelihood is given by

\[
L(\beta, \psi \mid \{t_i\}, X, \{\nu_i\}) \propto \prod_{i=1}^{n} \left( S^*(t_i) \right)^{1-\nu_i} \left( f^*(t_i \mid \psi) \right)^{\nu_i} 
\]

\[
\propto \prod_{i=1}^{n} \left( 1 + e^{-\theta_i} (1 - e^{\theta_i} F(t_i \mid \psi)) \right)^{1-\nu_i} \left[ f(t_i) \theta_i e^{-\theta_i S(t_i \mid \psi)} \right]^{\nu_i},
\]

22
where $\theta_i = e^{(x_i^T \beta)} = e^{u_i}$. Also note that

$$
\prod_{i=1}^n \theta_i e^{-\theta_i \nu_i S(t_i)} = \left[ \prod_{i=1}^k e^{u_i - \nu_i S(t_i)} \right] \left[ \prod_{i=k+1}^n e^{\nu_i \sum_{j=1}^k u_j v_{ji} - \nu_i \sum_{j=k+1}^n u_j v_{ji} S(t_i)} \right]
$$

$$
\leq \prod_{i=1}^k e^{A_i u_i - \nu_i S(t_i)} \forall u_1, \ldots, u_k,
$$

where $A_i = 1 + \sum_{m=i+1}^n \nu_m v_{mi}$.

Using the facts that

$$
[1 + e^{-\theta_i} (1 - e^{\theta_i F(t_i | \psi)})]^{1-\nu_i} \leq 2 \forall i,
$$

and

$$
\int \prod_{i=1}^k e^{A_i u_i - \nu_i S(t_i)} du_1 \cdots du_k < \infty,
$$

the propriety of the posterior distribution follows as long as we use proper priors for the parameters in $\psi$.

REFERENCES


### Simulated data

<table>
<thead>
<tr>
<th>Model</th>
<th>Activation scheme</th>
<th>First</th>
<th>Last</th>
<th>Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(c)</td>
<td>First</td>
<td>402.65</td>
<td>439.87</td>
<td>842.52</td>
</tr>
<tr>
<td>1(c)</td>
<td>Last</td>
<td>241.75</td>
<td>236.81</td>
<td>478.56</td>
</tr>
<tr>
<td>1(c)</td>
<td>Hierarchical</td>
<td>342.86</td>
<td>438.77</td>
<td>871.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Activation scheme</th>
<th>First</th>
<th>Last</th>
<th>Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(d)</td>
<td>First</td>
<td><strong>495.59</strong></td>
<td><strong>577.65</strong></td>
<td><strong>1073.24</strong></td>
</tr>
<tr>
<td>1(d)</td>
<td>Last</td>
<td>226.15</td>
<td>387.29</td>
<td>613.44</td>
</tr>
<tr>
<td>1(d)</td>
<td>Hierarchical</td>
<td>258.61</td>
<td>360.94</td>
<td>619.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression structure</th>
<th>First</th>
<th>Last</th>
<th>Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>First</td>
<td><strong>430.55</strong></td>
<td><strong>419.39</strong></td>
<td><strong>849.94</strong></td>
</tr>
<tr>
<td>2</td>
<td>Last</td>
<td>220.85</td>
<td>269.79</td>
<td>490.63</td>
</tr>
<tr>
<td>2</td>
<td>Hierarchical</td>
<td>407.15</td>
<td>411.97</td>
<td>819.12</td>
</tr>
</tbody>
</table>

**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 provide the goodness-of-fit (G), the penalty (P) and the L-measure values under first-, last- and hierarchical-activation schemes. The model with the best L score in each row is highlighted.

### Melanoma data

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression structure</th>
<th>First</th>
<th>Last</th>
<th>Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(a)</td>
<td>$\eta_i = x_i^T \beta; \theta_i = \theta; N \sim Po(\theta)$</td>
<td>206.98</td>
<td>421.93</td>
<td>628.91</td>
</tr>
<tr>
<td>1(b)</td>
<td>$\eta_i = x_i^T \beta; \theta_i = \theta; N \sim Ber(\theta)$</td>
<td>206.29</td>
<td>438.30</td>
<td>644.59</td>
</tr>
<tr>
<td>1(c)</td>
<td>$\eta_i = x_i^T \beta; \theta_i = \theta; N \sim Bin(K = 10, \theta)$</td>
<td>207.93</td>
<td>423.87</td>
<td>631.80</td>
</tr>
<tr>
<td>1(d)</td>
<td>$\eta_i = x_i^T \beta; \theta_i = \theta; N \sim Geo(\theta)$</td>
<td><strong>209.35</strong></td>
<td><strong>409.10</strong></td>
<td><strong>618.45</strong></td>
</tr>
<tr>
<td>2</td>
<td>$\log(\theta_i) = x_i^T \beta; \eta_i = \eta; N \sim Po(\theta_i)$</td>
<td>206.59</td>
<td>414.22</td>
<td>620.80</td>
</tr>
</tbody>
</table>

**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.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>First (Model 1d)</th>
<th>Last (Model 1b)</th>
<th>Hierarchical (Model 1c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median (2.5%, 97.5%)</td>
<td>median (2.5%, 97.5%)</td>
<td>median (2.5%, 97.5%)</td>
</tr>
<tr>
<td>Intercept</td>
<td>–1.835 (–2.242,–1.479)</td>
<td>–0.924 (–1.196,–0.675)</td>
<td>–1.063 (–1.316,–0.817)</td>
</tr>
<tr>
<td>Age</td>
<td>–0.019 (–0.244, 0.215)</td>
<td>–0.023 (–0.186, 0.144)</td>
<td>–0.017 (–0.185, 0.159)</td>
</tr>
<tr>
<td>Gender (male=0)</td>
<td>–0.041 (–0.568, 0.441)</td>
<td>0.025 (–0.385, 0.396)</td>
<td>–0.030 (–0.330, 0.348)</td>
</tr>
<tr>
<td>Performance Status</td>
<td>–0.283 (–1.151, 0.472)</td>
<td>–0.246 (–1.023, 0.372)</td>
<td>–0.196 (–0.981, 0.175)</td>
</tr>
<tr>
<td>Cure fraction (exp(–θ))</td>
<td>0.342 (0.270, 0.408)</td>
<td>0.359 (0.297, 0.419)</td>
<td>0.356 (0.295, 0.423)</td>
</tr>
<tr>
<td>ρ</td>
<td>1.426 (1.232, 1.622)</td>
<td>1.193 (1.038, 1.353)</td>
<td>1.230 (1.077, 1.391)</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression structure</th>
<th>First</th>
<th>Last</th>
<th>Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(a)</td>
<td>ηᵢ = xᵢᵀ β; θᵢ = θ; N ~ Po(θ)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>G  P  L</td>
<td>G  P  L</td>
<td>G  P  L</td>
</tr>
<tr>
<td>1(b)</td>
<td>ηᵢ = xᵢᵀ β; θᵢ = θ; N ~ Ber(θ)</td>
<td>149.44 242.15 391.59</td>
<td>149.44 242.15 391.59</td>
<td>149.44 242.15 391.59</td>
</tr>
<tr>
<td>1(c)</td>
<td>ηᵢ = xᵢᵀ β; θᵢ = θ; N ~ Bin(K = 5, θ)</td>
<td>180.26 283.63 463.89</td>
<td>137.89 235.04 372.93</td>
<td>138.81 224.57 363.38</td>
</tr>
<tr>
<td>1(d)</td>
<td>ηᵢ = xᵢᵀ β; θᵢ = θ; N ~ Geo(θ)</td>
<td>188.19 306.76 494.95</td>
<td>117.49 204.24 321.72</td>
<td>151.11 263.89 355.01</td>
</tr>
<tr>
<td>2</td>
<td>log(θᵢ) = xᵢᵀ β; ηᵢ = η; N ~ Po(θᵢ)</td>
<td>98.54 261.03 359.57</td>
<td>98.71 231.43 330.14</td>
<td>95.85 200.87 296.72</td>
</tr>
</tbody>
</table>

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.
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: The posterior estimates (median) of the survival plots for the simulated data sets under the different models and activation schemes. Adequate model fits are revealed by all the activation schemes.
Figure 2: 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 3: 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 4: 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.