Statistical modelling of repeated and multivariate survival data

Wintrebert, C.M.A.

CHAPTER 6

Marginal Survival Curve Estimates in Unbalanced Grouped Failure Time Data

Abstract

In biomedical studies more and more examples of grouped failure times data are encountered. The analysis of this kind of data is complicated by the censoring and by the dependence of the related failure times, for instance, within one patient. Moreover, often the number of measures per patient is different for each patient. This is called unbalanced data. In this chapter different approaches, parametric as well non-parametric to analyze these censored, grouped and unbalanced data have been developed. We proposed two classes of models, weighted models and frailty models. Both models yield consistent estimates for (unbalanced) clustered data with independent censoring. Focussing on assessing the marginal survival curve we compared the two modelling approaches for estimating this curve. The specific aim was to assess the sensitivity of both methods to unbalanced data with dependent censoring and the robustness against a misspecified model for the random effect. This was studied in simulated data.

This chapter has been submitted for publication as: C. M. A. Wintrebert, A. H. Zwinderman and J. C. van Houwelingen . Marginal Survival Curve Estimates in Unbalanced Grouped Failure Time Data.

6.1 Introduction

More and more examples of repeated failure times data are encountered in biomedical research (Hougaard, 2000; Vaida and Xu, 2000; Wintrebert et al., 2004). The analysis of such data is complicated by the censoring and by the dependence of the related failure times within one patient. Indeed for one patient, one failure time influences the risk of the others. In addition to the censorship and to the correlation between repeated failure times within patients, an other difficulty of many of repeated failure time data is that they are often unbalanced (Wintrebert et al., 2004); this is the case when patients have different numbers of repeated measures.

To analyze this kind of data, different approaches, parametric as well as non-parametric, have been developed in the setting of clustered data. In this framework, two classes of models have been proposed; in *weighted models* observations in a specific patient are weighted according to the number of observations in that patient (Binder, 1992), and in *mixed-effects models*, also called frailty models, a non-observed random effect is introduced in the model to account for the correlation between the repeated failure times (Clayton, 1978; Hougaard, 2000; Klein and Moeschberger, 1997b). In the first approach, the unbalance is taken into consideration by specifying different weights for the various observations in the same patient. The dependence structure between repeated measures is not specified but is adjusted for it in the inference by the way of weights. On the other hand, the frailty approach formulates an explicit model for the dependency between the repeated measurements in a patient; the unbalance problem is not specifically addressed in frailty models.

In this chapter, we focussed on estimating the marginal survival curve, and we compared the two modelling approaches for estimating this curve. The marginal survival curve gives the cumulative survival probability for a random period in a random patient. The specific aim was to assess the sensitivity of both methods to unbalanced data, and the robustness against a misspecified model for the random-effect. This was studied in simulated data. The rest of this chapter is organized as follows. In section 6.2, the two models that we used, are described in detail; first the weighted model, and afterwards the mixed-effects model. In section 6.3, the simulations and their results are given. Finally, some concluding remarks are given in section 6.4.

6.2 Methods

Standard methods to estimate a marginal survival curve were developed by Kaplan and Meier (1958), and by Aalen (1978); Nelson (1969), amongst others. These approaches, however, do not take the structure of repeated measures, nor unbalance into account. For such data weighted models, and mixed-effects models have been suggested, which will be described below. We therefore considered the situation of n patients ($i = 1, ..., n$) in which k_i ($j = 1, ..., k_i$) repeated measures of a failure time T_{ij} have been observed. The

number of repeated measures (k_i) may vary considerably between patients, for instance due to differences in mortality. Also the number of available observations may depend on T_{ij}: when patients are followed for a fixed time-period, the number of observations will be larger when T_{ij} were small than when they were large.

Further, we will also consider a parametric model (Weibull) for the survival curve.

6.2.1 Weighted Model

Before applying a weighted model (non-parametric or parametric) to the data we first apply the corresponding non-weighted model. In the case of the non-parametric approach, we use the Nelson-Aalen model, and in the parametric approach the Weibull model.

First, we ignore heterogeneity between patients. In this case, the simplest non-parametric estimates of the marginal survival function are the Kaplan-Meier and Nelson-Aalen estimates. When using the Kaplan-Meier or the Nelson-Aalen function, we implicitly assume the T_{ij} to be all mutually independent.

With balanced data the estimate is approximately correct. When the number of failure times in a patient is fully random, then the estimate is not too bad either; this is similar to the Missing-Completely-At-Random situation in the missing data literature (Little and Rubin, 1987). But when the Missing-Completely-At-Random situation does not hold, the estimate may be seriously biased. A possible correction for the unbalance is to weight each observation such that each patient contributes equally to the marginal survival function. *Weighted estimates can be understood in the setting of clustered data where the survival function and the number of observations vary from patient to patient and the number of observations may be related to the survival function of the patients (but does not depend on the actual observations).* This is not quite the practical situation we have in mind, but it gives a rationale for the weighted approach. Whether this weighting works in practice will be studied by simulation.

Let the survival function in a patient i be equal to $S_i(t) = Pr(T_i > t)$, $h_i(t)$ its corresponding hazard function and k_i the number of independent (within the patient) observations of patient i. Let also $C(t)$ be the survival function of the censoring time, which does not vary between patients. Moreover, we suppose that each observation of each patient is censored independently in the same way. The marginal survival is then equal to $S(t) = \frac{1}{n}\sum_{i=1}^{n} S_i(t)$ and the marginal hazard is equal to $h(t) = \frac{f(t)}{S(t)} = \frac{\sum_{i=1}^{n} h_i(t)S_i(t)}{\sum_{i=1}^{n} S_i(t)}$.

Our aim is to estimate these marginal survival and hazard functions. Let T_{ij} be the j^{th} failure time of patient i and D_{ij} the corresponding censoring indicator ($D_{ij} = 1$ if a failure occurred, $D_{ij} = 0$ if not). Since the observations are independent (within patients) we can act as if all observations start at the same time $t = 0$. For each patient, we define $Y_i(t) = \sum_{j=1}^{k_i} Y_{ij}(t) = \sharp\{j : T_{ij} \geq t\}$, which

represents the number of "observational units" within a patient at risk at time t and $N_i(t) = \sum_{j=1}^{k_i} N_{ij}(t) = \sharp\{j : T_{ij} \leq t, D_{ij} = 1\}$ which represents the total number of events at before or at time t for patient i. Following martingale theory (Fleming and Harrington, 1991), the expectation of dN_i at time t given the observed history in patient i is written by $E[dN_i(t)|history] = h_i(t)Y_i(t)dt$ and the expectation of Y_i at time t by $E[Y_i(t)] = k_i S_i(t)C(t)$. Since $S(t) = \exp(-H(t))$ where $H(t)$ is the marginal cumulative hazard function, we would first compute several Nelson-Aalen estimates (Andersen et al., 1993) of the cumulative hazard. A naive estimate of $H(t)$ is

$$\hat{H}_{naive}(t) = \int_0^t \frac{d(\sum_{i=1}^n N_i(u))}{\sum_{i=1}^n Y_i(u)},$$

which estimates

$$\int_0^t \frac{\sum_{i=1}^n h_i(u)k_i S_i(u)C(u)}{\sum_{i=1}^n k_i S_i(u)C(u)}du = \int_0^t \frac{\sum_{i=1}^n h_i(u)k_i S_i(u)}{\sum_{i=1}^n k_i S_i(u)}du.$$

Observe that the number of periods in patient i (k_i) does not cancel from the preceding equation and when k_i varies between patients, then $H_{naive}(t)$ may be inconsistent, especially when k_i depends on $h_i(u)$ or $S_i(u)$. To avoid this we choose to use a weighted estimate equal to:

$$\hat{H}_{weighted}(t) = \int_0^t \frac{d(\sum_{i=1}^n \frac{1}{k_i} N_i(u))}{\sum_{i=1}^n \frac{1}{k_i} Y_i(u)},$$

which estimates, by the same reasoning, to the marginal cumulative hazard function:

$$\int_0^t \frac{\sum_{i=1}^n h_i(u)S_i(u)}{\sum_{i=1}^n S_i(u)}du = \int_0^t h(u)du = H(t),$$

and should therefore be a consistent estimate of $H(t)$, and be robust against unbalanced data. Notice that the explicit object of this approach is to estimate the marginal survival or hazard functions, and that the dependence between the different repeated survival times within a patient is not specifically addressed.

When a parametric model is used for $H(t)$ it can be easily estimated by maximizing the weighted log likelihood

$$l = \sum_{i=1}^n \frac{1}{k_i} \sum_{j=1}^{k_i} l_{ij}(\Theta)$$

where $l_{ij}(\Theta)$ is the log likelihood contribution of the j^{th} failure time of patient i:

$$l_{ij}(\Theta) = -H_\Theta(t_{ij}) + D_{ij} ln(h_\Theta(t_{ij})),$$

and Θ is a vector of model parameters. Again, if the number of observations within a patient varies and is independent of the observations themselves, but might be related

to the patient's survival function, this weighted log-likelihood leads to a consistent estimate of the marginal $H(t)$, provided it belongs to the parametric class. Notice that the weighted Nelson-Aalen is also obtained by maximizing this log likelihood in a non-parametric way.

6.2.2 Mixed-effects model

In the former model no attempt was made to model the heterogeneity. When we choose to consider the patients to be a random sample from an undefined population this means that we suppose a distribution function for the S_i's. The use of frailty models is one way to do this. More generally, one speaks of mixed-effects models.

The conditional hazard function of patient i given his/her frailty, Z_i, is specified as

$$\lambda(T_{ij} \mid Z_i) = Z_i \lambda_0(T_{ij})$$

where $\lambda_0(T_{ij})$ is the unspecified baseline hazard function. The frailties are assumed to be sampled from some distribution $g(Z)$. For sake of convenience, $g(Z)$ is usually chosen to be the gamma density, $g(\frac{1}{\gamma}, \frac{1}{\gamma})$ ($\gamma > 0$), with expectation $EZ_i = 1$ and variance $VZ_i = \gamma$. In case $g(Z)$ is taken to be the gamma distribution, the marginal survivor curve is given by

$$S(t) = \big(1 + \gamma \Lambda(t)\big)^{-\frac{1}{\gamma}} \tag{6.1}$$

where $\Lambda(t)$ is the cumulative hazard function.

The difficulty is that the choice of the frailty distribution determines also the dependence, so when the model doesn't fit well or when the specifications of the model are not correct, the marginal survival curves could be badly estimated.

This approach does not have problems with varying number of observations. Indeed, given the frailty Z_i for patient i the T_{ij} are independent meaning that there is no difficulty with different number of repeated measures per patient apart from the accuracy of the (conditional) survivor curve estimate. However, it might still run into trouble if the censoring depends on the preceding observations. In that sense, it is not guaranteed to work properly for the unbalanced data we want to consider.

Extensions of the model of Huang and Wolfe (2002) might help to remedy this. However, that is beyond the scope of this chapter.

In this approach we also fit several models. First, the semi-parametric gamma frailty model where the baseline hazard is estimated in a non-parametric way. In order to compute the marginal survival curve of this gamma frailty model, we use functions from the Splus package, developed by Therneau and Grambsch (2000). These functions enable us to estimate γ and $\Lambda(t)$ from formula (6.1) above. Then, we also fit the gamma-weibull frailty model, where in that case the baseline hazard is supposed to follow a Weibull distribution.

6.2.3 Simulations

Balanced data

We first studied a situation with balanced data, i.e. where all patients had the same number of repeated measurements. We considered 4 situations with relatively few patients (80) or relatively many patients (200), each with relatively few repeated failure times (3), or relatively many failure times (30). Failure times (T_{ij}) were simulated using a

TABLE 6.1: *Overview of simulations*

data	number of patients	number of measures per patient
balanced	80	3
	80	30
	200	3
	200	30
unbalanced	80	max=12
	80	max=30
	200	max=12
	200	max=30

Weibull-gamma frailty model; failure times (T^*_{ij}) were sampled from a Weibull distribution ($\alpha = 1; \beta = 0.05$), and for each patient a frailty Z_i was sampled from a gamma distribution with mean$= 1$ and a variance of 4. Failure time T_{ij} was then calculated as $T_{ij} = T^*_{ij}/Z_i$.

For each T_{ij} a censoring time C_{ij} was sampled from a log normal distribution with parameters (μ, σ^2) and T_{ij} was observed if $T_{ij} \leq C_{ij}$ and was censored at C_{ij} if $T_{ij} > C_{ij}$. The parameters μ and σ^2 were chosen such that approximately 50% of observations was censored.

To construct a slightly more complex frailty distribution example, we also considered a situation where $ln(Z_i)$ was sampled from a mixture of two normal distributions: $p = 0.116$ percent of $ln(Z_i)$ was sampled from the normal distribution $N(0.911, 0.88)$, and $(1-p)$ $(p = 0.116)$, percent from $N(-0.911, 0.88)$. We have made this choice since in that case we again have a frailty distribution with mean 1 and a variance of 4.

Note that in this situation, the longitudinal aspect does not come into play. The repeated data can be considered to be clustered data and the methods described above can be expected to work properly.

Unbalanced data

To simulate unbalanced data we removed a random number of (censored and uncensored) observations per patient from a balanced data set. In order to construct a complicated situation that mimics the longitudinal data that we meet in practice, we also removed those observations in a patient coming after the one for which the cumulative sum of (possibly censored) failure times is bigger than 4 times the median of the survival times. In this way the number of available repeated observations depends on the size of the first few failure times; patients in whom the first failure time is large will have a few repeated observations while patients of whom the first failure time is small will have many repeated observations. The fact that each measure of a patient can be censored introduces an extra censoring process and also influences the independence of the censoring process. The situation gets really complicated, we do not intend to analyze it theoretically, but hope that the simulations can give some insight about the validity of the different approaches.

For every simulated data set we subsequently estimated the weighted and unweighted Nelson-Aalen estimate of the marginal survival curve, and the marginal survival curve of the gamma-frailty model. In addition, we calculated weighted and unweighted marginal survival curves assuming that the baseline hazard function was the hazard function of a Weibull distribution.

Comparison of the models

For each data set we calculated the marginal survival curve using the Nelson-Aalen estimator ($exp(-\hat{H}_{naive})$), using the weighted Nelson-Aalen estimator ($exp(-\hat{H}_{weighted})$), and using the gamma-frailty Cox model. We also estimated parametrically the marginal survival curves according to the Weibull, the weighted Weibull, and the gamma-frailty Weibull models.

Since the true survival function was known in the simulations, we evaluated the different estimates of the marginal survival curve by calculating the mean Bias (MB), and the mean integrated squared errors ($MISE$) for each of the different models and for each data set. Let $[0, t]$ be the interval on which we want to evaluate the performance of the estimated survival distribution. The mean Bias is defined as

$$MB = \frac{1}{t}\int_0^t (\hat{S}(u) - S(u))du,$$

where $\hat{S}$ is the estimated marginal survival curve and S the simulated survival curve. The mean integrated squared error is defined as

$$MISE = \frac{1}{t}\int_0^t (\hat{S}(u) - S(u))^2 du.$$

These quantities are only well defined if t is smaller than the largest observation. In our simulation study we take $t = 160$ when the frailty is sampled from a gamma distribution and we take $t = 90$ when the frailty is sampled from a mixture of two log normal distributions. In all our simulations the largest observation is larger than these values of t.

6.3 Results

Mean Bias(MB) and mean integrated squared errors ($MISE$) of the models with non parametric baseline hazard estimates (gamma Weibull simulated data) are given in Table 6.2.

TABLE 6.2: *Mean Bias(MB) and mean integrated squared errors(MISE) of the models with non parametric baseline hazard estimates.*

	MB			MISE		
data	NA[1]	WNA[2]	GF[3]	NA[1]	WNA[2]	GF[3]
80∝3	0.0102	0.0102	- 0.0018	0.0011	0.0011	0.0005
80∝30	0.0212	0.0212	0.0309	0.0006	0.0006	0.0011
200∝3	- 0.0083	- 0.0083	- 0.0041	0.0001	0.0001	0.0001
200∝30	0.0234	0.0234	0.0268	0.0006	0.0006	0.0007
80∝max12	- 0.2819	- 0.1466	- 0.2615	0.0890	0.0364	0.0785
80∝max30	- 0.2946	- 0.0380	- 0.2256	0.0914	0.0029	0.0568
200∝max12	- 0.2043	- 0.0223	- 0.1749	0.0465	0.0019	0.0356
200∝max30	- 0.2375	0.0242	- 0.1370	0.0594	0.0022	0.0219

[1] Nelson-Aalen
[2] Weighted Nelson-Aalen
[3] Gamma frailty

In the case of balanced data, MB and $MISE$ of the Nelson-Aalen and weighted Nelson-Aalen estimators are exactly the same, as they should be since in that case the two estimates are exactly the same. The Nelson-Aalen and gamma-frailty models work out equally well, both models giving approximately the same MB and $MISE$. This is illustrated in Figure 6.1. In Figure 6.1, several survival curves are plotted: on the one hand the marginal survival curve of a gamma Weibull frailty model used to simulate the data set; on the other hand the fitted survival curves of the three non-parametric models (Nelson-Aalen, weighted Nelson-Aalen and gamma frailty) that are applied to a simulated balanced data set with 80 patients and 3 measurements per patient.

In the case of unbalanced data, the weighted Nelson-Aalen model clearly fits better than the Nelson-Aalen and gamma frailty models. The latter two models give approximately the same results. See Figure 6.2.

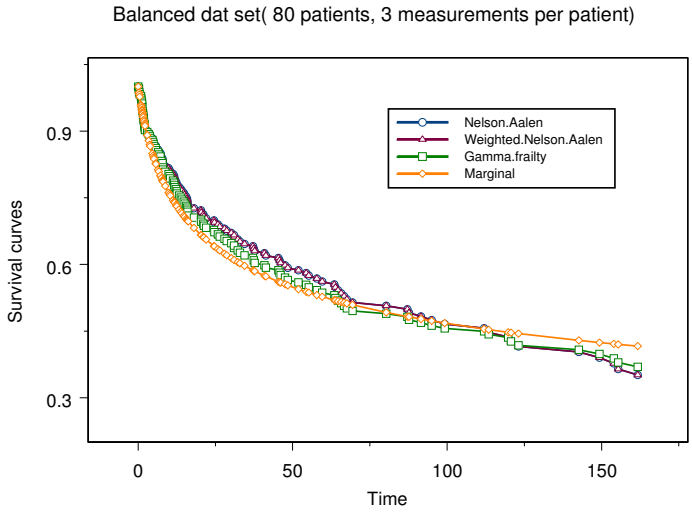

FIGURE 6.1: *Survival curves(marginal and fitted) concerning a simulated data set (*80 *patients,* 3 *measurements per patient) from a gamma Weibull frailty model.*

In Figure 6.2, the same curves are plotted as in Figure 6.1. The difference is that here the models are applied to an unbalanced simulated data set with 80 patients and with a maximum of 12 measurements per patient, varying from patient to patient.

Mean Bias(MB) and mean integrated squared errors(MISE) for the models assuming a baseline Weibull hazard function (gamma Weibull simulated data) are given in Table 6.3. The Weibull model does not give satisfactory results, and this is as expected since the model assumptions are wrong(in each patient the hazard function has Weibull form, but the average hazard function over all patients does not). The gamma Weibull frailty model gives good results because the model assumptions are right. Applied to unbalanced data sets the weighted Weibull model fits better than the gamma Weibull frailty model. The gamma Weibull frailty model fits better than the Weibull model. As an illustration Figures 6.3 and 6.4 are given. These figures are the corresponding Figures to Figures 6.1 and 6.2 in the parametric situation.

In the case of balanced data, parametric models do not fit better than non-parametric models.

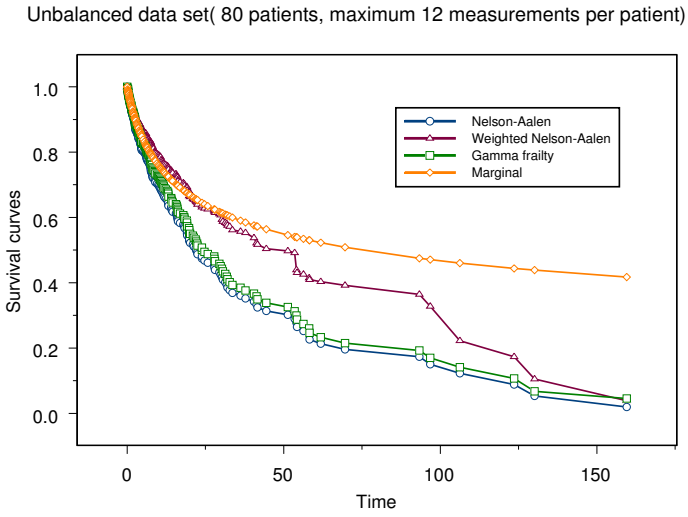

FIGURE 6.2: *Survival curves(marginal and fitted) concerning an unbalanced simulated data set (*80 *patients, between* 1 *and* 12 *measurements per patient) from a gamma Weibull frailty model.*

Mean Bias (MB) and mean integrated squared errors (MISE) for the models assuming a non-parametric hazard function (mixture log normal Weibull simulated data) are given in Table 6.4. When frailty parameters are sampled from a mixture of two log normal distributions the three models applied to balanced simulated data sets give almost similar results. The gamma frailty model that does not fit here, has slightly lower MB and MISE. Applied to unbalanced simulated data sets the unweighted Nelson-Aalen model fits less well than the two other models. The results of MB and MISE of the weighted Nelson-Aalen and the gamma frailty models differ slightly from each other. The survival curves are illustrated in Figure 6.5 with a balanced simulated data set and in Figure 6.6 with an unbalanced simulated data set.

6.4 Concluding Remarks

In this chapter we compared statistical characteristics of the estimates of the marginal survivor curve of weighted and gamma-frailty models. We found comparable results

TABLE 6.3: *Mean Bias (MB) and mean integrated squared errors (MISE) of the models with Weibull baseline hazard estimates.*

	MB			MISE		
data	WB[1]	WWB[2]	GWB[3]	WB[1]	WWB[2]	GWB[3]
80∝3	0.0192	0.0192	- 0.0034	0.0020	0.0020	0.0003
80∝30	0.0400	0.0400	0.0334	0.0020	0.0020	0.0012
200∝3	0.0073	0.0073	0.0025	0.0007	0.0007	0.00001
200∝30	0.0371	0.0371	0.0217	0.0021	0.0021	0.0005
80∝max12	- 0.2843	- 0.1428	- 0.2544	0.0913	0.0301	0.0743
80∝max30	- 0.3021	- 0.0273	- 0.2099	0.0985	0.0027	0.0488
200∝max12	- 0.2115	- 0.0354	- 0.1765	0.0526	0.0055	0.0370
200∝max30	- 0.2545	- 0.0033	- 0.1451	0.0713	0.0041	0.0244

[1] Weibull
[2] Weighted Weibull
[3] Gamma Weibull Frailty

TABLE 6.4: *Mean Bias (MB) and mean integrated squared errors (MISE) of the models assuming a non-parametric hazard function (mixture log normal Weibull simulated data).*

	MB			MISE		
data	NA[1]	WNA[2]	GF[3]	NA[1]	WNA[2]	GF[3]
80∝3	0.0086	0.0086	- 0.0070	0.0002	0.0002	0.0003
80∝30	0.0016	0.0016	- 0.0027	0.0003	0.0003	$2.73\ 10^{-5}$
200∝3	- 0.0251	- 0.0251	- 0.0227	0.0009	0.0009	0.0008
200∝30	0.0209	0.0209	0.0097	0.0005	0.0005	0.0001
80∝max12	- 0.0770	0.0548	- 0.0519	0.0074	0.0070	0.0045
80∝max30	- 0.0931	0.0186	- 0.0425	0.0130	0.0030	0.0035
200∝max12	- 0.1472	- 0.0209	- 0.1068	0.0229	0.0069	0.0131
200∝max30	- 0.1357	0.0096	- 0.0628	0.0213	0.0011	0.0049

[1] Nelson-Aalen
[2] Weighted Nelson-Aalen
[3] Gamma frailty

while using a parametric or a nonparametric form for the baseline survivor curve. With balanced data we found little difference between the weighted approach and the gamma-frailty model with -perhaps- slightly smaller mean bias, and mean integrated squared errors for the gamma-frailty model, even if the true frailty distribution was not like a

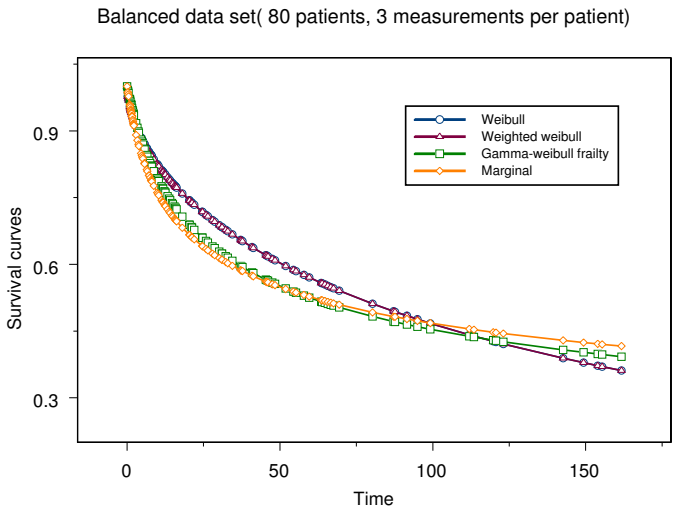

FIGURE 6.3: *Survival curves(marginal and parametric fitted) concerning a simulated data set (*80 *patients,* 3 *measurements per patient) from a gamma Weibull frailty model.*

gamma-distribution. Bias and squared errors clearly decreased with increasing sample size, and increasing number of observations per patient/cluster.

With unbalanced data both estimates showed considerable larger bias and squared errors, but the weighted approach showed clearly less bias and error than the gamma-frailty model. The gamma-frailty estimate was always lower than the true survivor curve, and the weighted estimate was lower than the true survivor curve when we sampled the frailties from a gamma-distribution, but it was higher when we sampled the log- frailties from a mixture of two normal distributions. This is probably due to the fact that with sampling from the gamma-distribution, most frailties were close to one with only few high values, while with sampling from the mixture there were both high but also a considerable (even larger) proportion with very low values. Given the mechanism that we used to generate unbalanced data, low frailties correspond to relatively few and large observed T-values in a patient, while high frailties correspond to relatively many and short T-values. We conclude that the gamma-frailty model will likely lead to an underestimate of the true marginal survivor curve, and that the weighted approach

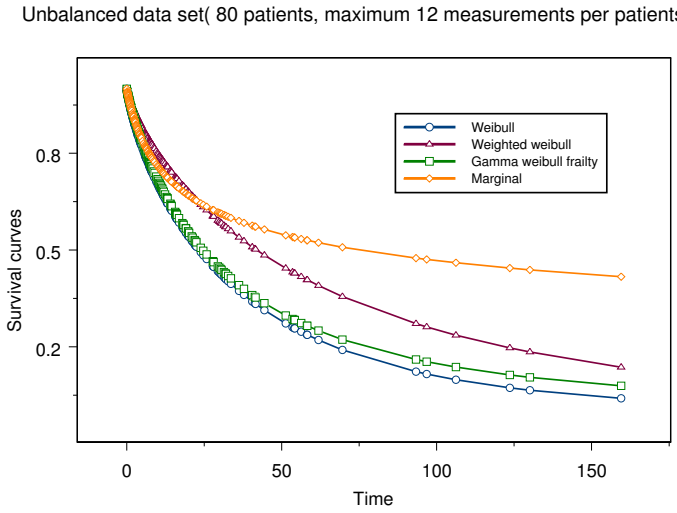

FIGURE 6.4: *Survival curves(marginal and parametric fitted) concerning a simulated data set (*80 *patients, between* 1 *and* 12 *measurements per patient) from a gamma Weibull frailty model.*

can yield both under- or an overestimation, depending on the mechanism responsible for the unbalance in the data.

The present simulation is much too simple and too small to derive any conclusions on when which approach should be preferred. But even with the small simulation study that we performed it is clear that there exist situations where both estimates show large differences. There are few tools available to check the validity of both estimates, and we therefore suggest to use both approaches in practice.

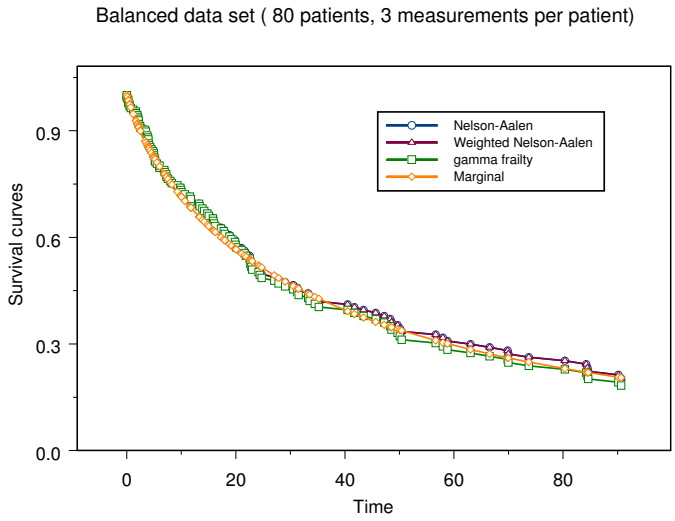

FIGURE 6.5: *Survival curves(marginal and parametric fitted) concerning a simulated data set (*80 *patients,* 3 *measurements per patient) from a log normal Weibull frailty model.*

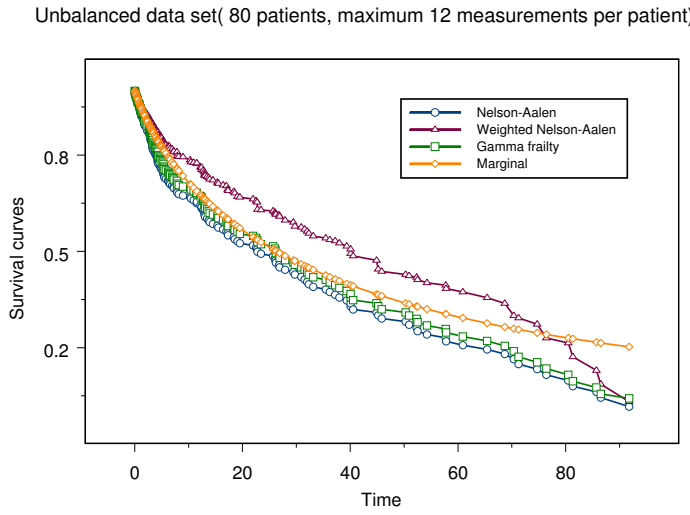

FIGURE 6.6: *Survival curves(marginal and parametric fitted) concerning a simulated data set (*80 *patients, between* 1 *and* 12 *measurements per patient) from a log normal Weibull frailty model.*