Statistical modelling of repeated and multivariate survival data
Wintrebert, C.M.A.

CHAPTER **7**

General Summary

In this thesis statistical models for complex survival data were studied; both when repeated failures and when different failure events were present in the same individuals. We used frailty models in which random effects per individual were introduced to explain the association between repeated, and multivariate failures, but in addition we considered weighted, and marginal survival models in which the association structure is left unspecified.

In chapter 2 we developed time-varying frailty models and applied these to survival data of patients included in the European Blood Marrow Transplant (EBMT) registry, where a decaying centre-effect was found. Centre-effects are usually investigated with constant shared frailty models (Andersen et al., 1999), and this presumes that this effect is constantly present during follow-up, even when the follow-up is very long. But with many applications this assumption is too restrictive. Therefore we studied ways to extend the constant frailty model to allow a time-dependence of the frailties. We found these models more realistic. We also compared our results with those obtained applying the time-dependent frailty model of Paik et al. (1994). Their model is very cumbersome to calculate and difficult to fit. We proposed two much easier models, and obtained satisfying results. Indeed, a clear centre-effect on survival after transplantation was found and this effect was mainly present in the first three to four months after transplantation, even when well-known risk factors were taken into account in the analysis.

In chapter 3, we analyzed survival data coming from wild animal ecology. We adopt survival approaches from clinical biostatistics, where models that take into account individual variation are extensively used, to assess the correlation between the survival of Kittiwake birds and repeated breeding attempts of the birds. In our analysis, we took into account that the data were both right and interval censored. An interesting point of the model we developed and applied is, that we used only one frailty in both models for the survival and the breeding, resulting in a more parsimonious model than the one used by Cam and Monnat (2000). Furthermore, our models were estimated using a classic frequentist method and also MCMC samplings in Winbugs (Spiegelhalter et al., 1996). This last statistical tool (MCMC samplings) to fit a model has been widely used in the statistical literature, as emphasized by Link et al. (2002), but there are still relatively few examples in wildlife-related applications. We used the tool to fit a model that has

never been used with data from animals populations. Evidence of a positive correlation between the survival and breeding in birds has been found, and this result is consistent with the one of Cam et al. (2002) who used the same data set but used other models from ecology.

In chapter 4, we developed models to analyze association between relatives with respect to age at onset of Huntington disease. In general, genetic models for survival data are hard to formulate and hard to fit. Several statistical models have been developed to investigate the correlation of failure type data induced by genetic or environmental models. These models, for instance the one of Li and Zhong (2002), are rather complicated, especially for large and complex pedigrees, and are not easy to fit either. Moreover, the association induced depends strongly upon assumptions made. In chapter 4, we used a different approach. We showed how martingale residuals from a (marginal) Cox model can be employed to estimate the presence of a genetic effect and to estimate genetic correlations depending on the genetic distance. From our results, it can be concluded that considerable association exists between age at onset of Huntington disease between individuals of the same family, even after correction for known risk factors and in particular for the mutation that is causing Huntington's disease. This association is much larger between siblings and between parents and children than for individuals at larger genetic distance. This observation suggests the existence of one or more genes, other than the Huntington disease gene, involved in the timing of onset of Huntington disease.Our new tools are simple, quick and applicable to pedigrees of any size, and can be performed with the user friendly computer package S-plus.

In chapter 5 we analyzed multivariate survival data from patients having suffered from atherosclerosis. Multivariate frailty models were used to estimate the correlation between 3 processes: cardiac, cerebral and peripheral atherosclerosis. Our starting point was the assumption that the 3 previous processes were different. The correlations between the 3 processes were estimated in patients who had second and more atherosclerotic events. To assess these correlations different frailty models were developed and applied. One of the complications encountered in this chapter is that all patients included in the data set experienced an index event but only a minority had a second event, or third event during follow-up. This meant that the failure time of the second or third event was censored for most patients. An other complication was that all patients were ascertained on the basis of having one event, excluding all individuals who were at risk for one or more events, but did not yet experience one. Our results indicated very strong relationships between the processes responsible for developing cardiac, cerebral and peripheral atherosclerosis. The association between different type of events was clearly as strong as the association between repeated events of the same type. Clinically, it seems important that our results suggest that the risk of a next event (any event) depends mainly on the age of patient at first event, and not on the type of the event.

Finally in chapter 6, we estimated marginal survivor curve in balanced and unbalanced longitudinal studies. Different approaches, parametric as well non-parametric have been developed to analyze possibly censored, repeated and unbalanced data. Two classes of models were proposed, namely weighted and frailty models. The two modelling approaches for estimating the marginal survivor curve were compared. The sensitivity of both methods to unbalanced data was assessed, and also the robustness against a misspecified model for the random effect was studied. The difficulty was that the choice of the frailty distribution determined also the dependence between the repeated measures: when the model does not fit well or when the specifications of the model are incorrect, the marginal survival curves can be badly estimated. The weighted model showed better or at least not worse characteristics than the estimate of the frailty model. It is unclear why the weighted model performs so well; perhaps the simplicity of the model and the lack of assumptions are responsible. The marginal survivor function proved to be difficult to estimate, and since there are few tools available to check its validity we therefore advice to evaluate many different models.