Chapter 3

Score test for association between recurrent events and a terminal event

Abstract

The statistical analysis of recurrent events relies on the assumption of independent censoring. When random effects are used, this means, in addition, that the censoring cannot depend on the random effect. Whenever the recurrent event process is terminated by death, this assumption might not be satisfied. Joint models for recurrent and terminal events are often difficult to fit. Thus, clinicians rarely check whether they are preferred to separate models. In this chapter, we propose and compare simple, yet efficient ways of testing whether the terminal event and the recurrent events are associated or not. The proposed methods are evaluated in a simulation study and are illustrated through a data set consisting of repeated observations of skin tumors on T-cell lymphoma patients.

3.1 Introduction

Recurrent event data have become increasingly common in clinical studies, in reliability theory, and in other fields (Cook and Lawless, 2007). The shared frailty model (Nielsen et al., 1992) is a popular method for analyzing this type of data, because it retains a

This chapter has been published as: T.A. Balan, S.E. Boonk, M.H. Vermeer, H. Putter (2016). Score test for association between recurrent events and a terminal event. *Statistics in Medicine* 35(18), 3037-3048.

similar semiparametric specification with the well known Cox model, it is supported by asymptotic results (Murphy, 1995a; Parner, 1998) and is available in standard statistical software (Therneau and Grambsch, 2000). The *frailty* (Vaupel, Manton, and Stallard, 1979) is a random effect which accounts for heterogeneity that can not be explained by observable covariates. In other words, it describes whether a subject or a cluster of subjects is at a higher risk (large frailty) than others (small frailty). In the recurrent events framework, the frailty accounts for the dependence between the observations on the same individual. Conditional on the frailty, one hopes that the stochastic processes underlying the individuals are independent. Thus, frailty models allow an elegant and parsimonious explanation of the mechanism which generates the data.

In a clinical context, recurrent events are often a symptom of a medical condition which might lead to the end of follow-up in the form of dependent censoring by terminal event, such as death. In particular, a more frail subject might not only be associated with a higher recurrence rate, but also an increased or decreased risk of experiencing the terminal event, to a greater or lesser extent. If this is the case, the recurrences and the terminal event should be jointly modeled, allowing for the frailty to describe both the unaccounted differences in the risk for both recurrences and death. Such a model was introduced in Liu, Wolfe, and Huang (2004), who adapted a model for clustered failures with informative censoring (Huang and Wolfe, 2002). For estimation of a semiparametric joint frailty model, the Expectation-Maximization (EM) algorithm can be used, the method being very similar to the estimation of the shared frailty model (Nielsen et al., 1992; Klein, 1992).

There are however disadvantages of the joint model. It is notably easier to consider separate models for the recurrences and death, both in terms of difficulty of fitting and interpretation; a comparison between the estimation methods of the shared frailty model (Nielsen et al., 1992) and the joint model (Liu, Wolfe, and Huang, 2004) can attest to this. Furthermore, expressions for marginal features of the recurrent events or terminal event processes are not readily obtained, and the interpretation of features of interest, such as treatment effects, is not as straightforward as for the separate models. Although software for parametric models for recurrent and terminal events exists (Rondeau and Gonzalez, 2005), there is no method to check a priori whether separate models are similarly appropriate or not. This may lead to situations when clinical practitioners will ignore the dependence between the two event types.

In this chapter, we aim to develop a simple statistical test for association between the recurrent events and the terminal events, which does not require the estimation of a joint model. This provides an answer to a clinically relevant problem and it also indicates whether the joint modeling of the processes is more suitable. The idea that we follow is similar to a test for informative censoring (Huang, Wolfe, and Hu, 2004) and heterogeneity (Commenges and Andersen, 1995) in the context of clustered failures.

The outline of the article is as follows. In Section 3.2, we review a joint model closely related to that of Liu, Wolfe, and Huang (2004). In Section 3.3, we review possible tests for association and introduce the robust score test, and in Section 3.4 we discuss the

efficiency and validity of our approach in a simulation study. Finally, in Section 3.5 we illustrate the proposed methods on a data set of successive hospital readmissions.

3.2 Models

Let D_i and C_i denote the time of the terminal event and right censoring time respectively, both of which correspond to the end of followup. Also define $T_i = \min(D_i, C_i)$, and $Y_i(t) = 1(t \leq T_i)$ the "at risk" indicator. While $Y_i(t) = 1$, we observe two counting processes, $N_i^D(t) = 1(D_i \leq t)$ corresponding to the terminal event and $N_i^R(t)$ which is equal to the number of recurrences in $(0, t]$, or equivalently their increments $\Delta N_i^R(t)$ and $\Delta N_i^D(t)$, equal to the number of respective events in the small interval $(t, t + \Delta t]$. We can consider a $p \times 1$ vector of possibly time-dependent covariates $\{x_i(t) : t \geq 0\}$ and denote their path up to time t as $x_i^{(t)} = \{x_i(s) : 0 \leq s \leq t\}$. We require the time-dependent covariates to be external, in the sense of Kalbfleisch and Prentice (2002). The history up to time t is then

$$H_i(t) = \left\{ \left(N_i^R(s),\, N_i^D(s)\right) : 0 \leq s \leq t;\, x_i^{(t)} \right\}. \tag{3.1}$$

The intensities of N_i^R and N_i^D can be associated, meaning that the rate of recurrences and that of the terminal event can depend on elements of (3.1). It is, for example, plausible that a high rate of recurrent events is associated with a reduced survival. Often, this can be an indication of a "hidden" factor, such as a severe disease, which influences both intensities of N_i^R and N_i^D.

As in the model of Liu, Wolfe, and Huang (2004), we consider a frailty variable $\mathbf{Z} = (Z_1, ..., Z_n)$ with Z_i's i.i.d. with a distribution function $G(z; \theta)$, with mean 1 and variance θ. Conditional on $\mathbf{Z} = (z_1, ..., z_n)$, the intensities of N_i^R and N_i^D are:

$$\begin{aligned} r_i(t|z_i) &= \lim_{\Delta t \to 0} \frac{Pr\left\{\Delta N_i^R(t) = 1 | z_i,\, H_i(t-)\right\}}{\Delta t}, \\ \lambda_i(t|z_i) &= \lim_{\Delta t \to 0} \frac{Pr\left\{\Delta N_i^D(t) = 1 | z_i,\, H_i(t-)\right\}}{\Delta t}. \end{aligned}$$

Further, we assume that both N_i^R and N_i^D can not increase after D_i. Although a natural assumption for the terminal event, for the recurrent events death is an instance of potentially informative censoring. In particular, a violation of the assumptions of the classical shared frailty model (Nielsen et al., 1992) occurs if z_i can not be dropped from the expression of λ_i. Finally, we follow Liu, Wolfe, and Huang (2004) in choosing a multiplicative model for the intensities, so that r_i and λ_i can be expressed as

$$\begin{cases} r_i(t|z_i) &= z_i 1(D_i > t)\, e^{\beta' x_i^R(t)} r_0(t) \\ \lambda_i(t|z_i) &= z_i^\gamma 1(D_i > t)\, e^{\alpha' x_i^D(t)} \lambda_0(t) \end{cases}. \tag{3.2}$$

The baseline intensities r_0 and λ_0 are assumed for now to be continuous positive functions. The regression coefficients α and β have the dimensions of the corresponding covariates x_i^D and x_i^R.

The question of association between N_i^R and N_i^D is closely related to the parameter γ in (3.2), which describes the direction and magnitude at which the frailty influences the hazard λ_i. Thus, the interest lies in testing the hypothesis $H_0 : \gamma = 0$ against $H_A : \gamma \neq 0$. Under H_0, the expressions of λ_i and r_i do not share any parameters, and then both processes can be analyzed separately; in particular, the censoring of the recurrent event process by the terminal event is non-informative, in the sense of Nielsen et al. (1992).

Assume that the baseline intensities are fully described by some parameters ϕ_r and ϕ_d, i.e. $r_0(t) \equiv r_0(t;\, \phi_r)$ and $\lambda_0(t) \equiv \lambda_0(t;\, \phi_d)$. If ϕ_r and ϕ_d are finite dimensional, then the model is parametric; otherwise, the model is semi-parametric, as originally proposed by Liu, Wolfe, and Huang (2004). Nevertheless, we denote the nuisance parameter vector by $\eta = \{\beta,\, \alpha,\, \theta,\, \phi_r,\, \phi_d)\}$.

For subject i, we denote the observed data O_i as the event "n_i observed recurrent events at t_{i1}, ..., t_{in_i} over $[0, t_i]$ and $\delta_i = 1(D_i < C_i)$". Under the regularity conditions of Liu, Wolfe, and Huang (2004), the "conditional likelihood" based on $(H_i(\infty);\ i = 1...n;\ \mathbf{Z})$ is formed from the conditional probabilities

$$Pr(O_i|z_i) = \prod_j \left\{ r_i(t_{ij}|z_i) \right\} \exp\left\{ -\int_0^\tau Y_i(s) r_i(s|z_i) ds \right\} \lambda_i(t_i|z_i)^{\delta_i} \times \\ \times \exp\left\{ -\int_0^\tau Y_i(s) \lambda_i(s|z_i) ds \right\}.$$

Similarly, the "marginal likelihood" based on $H_i(\infty)$ alone is obtained from the marginal contributions to the likelihood $Pr(O_i) = \int_0^\infty Pr(O_i|z) dG(z;\, \theta)$. The marginal log-likelihood is then

$$\ell\,(\gamma, \eta) = \sum_i \left[\sum_{j=1}^{n_i} \left\{ \beta' x_i(t_{ij}) + \log r_0(t_{ij}) \right\} + \delta_i \left\{ \alpha' x_i(t_i) + \log \lambda_0(t_i) \right\} + \right. \\ \left. + \log \int_0^\infty K_i(z, t_i) f_\theta(z) dz \right] \quad (3.3)$$

where $K_i(z,t) f_\theta(z)$, is the kernel of the "posterior" distribution $Z_i|H_i(t)$ computed with the data available until time t. We denote the cumulative given $z_i = 1$ as $R_i(t) = \int_0^t Y_i(s) e^{\beta' x_i^R(s)} r_0(s) ds$ and $\Lambda_i(t) = \int Y_i(s) e^{\alpha' x_i^D(s)} \lambda_0(s) ds$, and then

$$K_i(z,t) = z^{N_i^R(t-) + \gamma N_i^D(t-)} \exp\left\{ -z R_i(t) - z^\gamma \Lambda_i(t) \right\}. \quad (3.4)$$

Under $\gamma = 0$, K_i is the kernel of a Gamma distribution, so a convenient choice for G is the Gamma distribution as well (Nielsen et al., 1992); also see Duchateau and Janssen (2007).

If $\gamma \neq 0$ the Expectation-Maximization algorithm must be employed to maximize the log-likelihood, using numerical methods to approximate integrals at every iteration (Liu, Wolfe, and Huang, 2004). The numerical approximations and the slow convergence of the EM algorithm result in an overall slow and complicated method.

One way out is to consider a parametric version of the joint model. At the expense of introducing assumptions about the functional form of r_0 and λ_0, one can obtain a numerically tractable form of the log-likelihood (3.3), which can be maximized with standard maximum likelihood methods (Rondeau, Mathoulin-Pelissier, et al., 2007). This approach is implemented in the R package **frailtypack** (Rondeau and Gonzalez, 2005; Rondeau, Mazroui, and Gonzalez, 2012), which also offers the option to choose flexible parametric specifications for r_0 and λ_0, such as piecewise constant or spline-approximated.

There are however reason not to employ the joint model. First, clinicians prefer more familiar models such as a frailty model for the recurrent events (available in e.g. R, SAS, Stata) or a Cox model for the terminal event (also available in SPSS), if there is no need of doing something more complicated. The parametric assumptions have their price as well. Splines, for example, require the specification of two "smoothing parameters", which may or may not be easy to obtain. We will return to considerations about computation in section 3.4. Thus, it would be useful to be able to see if there is evidence against H_0 even before the joint model is used. While the Likelihood Ratio Test (LRT) or the Wald test require the maximization of (3.3), the score test does not. If the null hypothesis is rejected, the shared frailty model is not appropriate and the terminal event should be jointly modeled (Liu, Wolfe, and Huang, 2004; Ye, Kalbfleisch, and Schaubel, 2007).

In the following section, we describe tests for H_0 based on (3.3), with a focus on those that do not require the maximization of (3.3).

3.3 Tests for independence

Our goal is to test $H_0 : \gamma = 0$, in the presence of the nuisance parameters η; a complete specification of the null hypothesis is $H_0 : (\gamma, \eta) = (0, \eta)$. Abiding by our purpose of developing a simple test for this hypothesis, we first focus on how this can be achieved while avoiding the direct maximization of (3.3). This can be done by considering the maximum likelihood estimate $\hat{\eta}_0$ under $\gamma = 0$ and measuring the variation of (3.3) around $\gamma = 0$. This forms the basis of the score test in section 3.3.1. Other approaches, for which estimation of the joint model is needed, are detailed in Section 3.3.2.

3.3.1 Score Test

The starting point for this is the score function for γ under H_0, defined as the derivative with respect to γ in (3.3):

$$U_\gamma(0, \eta) = \frac{\partial}{\partial \gamma} \ell(\gamma, \eta)\Big|_{\gamma=0}.$$

If we denote $\hat{\eta}_0$ the estimate of η under H_0, then

$$\frac{\left\{ U_\gamma \left(0, \hat{\eta}_0\right)\right\}^2}{\text{Var}\left\{ U_\gamma \left(0, \hat{\eta}_0\right)\right\}} \tag{3.5}$$

follows asymptotically a χ^2 distribution with 1 degree of freedom. The variance of the score is

$$\text{Var}\left\{ U_\gamma \left(0, \hat{\eta}_0\right)\right\} = \left(I_{\gamma\gamma} - I_{\gamma\eta} I_{\eta\eta}^{-1} I_{\eta\gamma}\right)\Big|_{\gamma=0, \eta=\hat{\eta}_0}, \tag{3.6}$$

where the Is are obtained from the Fisher information matrix

$$I(\gamma, \eta) = \begin{pmatrix} I_{\gamma\gamma} & I_{\gamma\eta} \\ I_{\eta\gamma} & I_{\eta\eta} \end{pmatrix}.$$

If the model is semi-parametric, the score function and information matrix of η are replaced by a score and an information operator (Rabinowitz, 2000; Kosorok, 2008). Although this does not lead to a closed form of (3.6), any "good" estimate of the variance of the score can be used. The first choice is to replace the denominator of (3.5) with $I_{\gamma\gamma}\big|_{\gamma=0}$, which is the variance of the score if η were *known* to be equal to $\hat{\eta}_0$. By this, the variance will be underestimated, thus leading to a conservative test statistic. We refer to this as the **naive score test** (NST).

Further insight can be obtained by calculating U_γ:

$$U_\gamma(\gamma, \eta) = \sum_i \frac{\int N_i^D(t_i) \log z - \Lambda_i(t_i|z) z^\gamma \log z K_i(z) f_\theta(z) dz}{\int K_i(z) f_\theta(z) dz}.$$

Setting $\gamma = 0$ and replacing η with $\hat{\eta}_0$, we obtain

$$\begin{aligned} U_\gamma(0, \hat{\eta}_0) &= \sum_i \frac{\int \left\{ N_i^D(t_i) - \widehat{\Lambda}_i(t_i|x_i,\ z)\right\} \log z \widehat{K_i}(z) f_\theta(z) dz}{\int K_i(z) f_\theta(z) dz} \\ &= \sum_i \widehat{M_i^D} \cdot \widehat{\log z_i}, \end{aligned} \tag{3.7}$$

where $\widehat{M_i^D}$ and $\widehat{\log z_i}$ are the estimates of $M_i^D = N_i^D(t_i) - \int_0^{t_i} Y_i(s)\lambda_i(s)ds$, the martingale residual of the terminal event, and of $E\left[\log Z_i | O_i(t_i)\right]$, where the expectation is taken with respect to the "posterior" distribution $\widehat{K_i}(z) f_\theta(z)$ of (3.4), with R_i and Λ_i replaced by their estimates under H_0.

A similar expression involving a correlation between martingale residuals and aspects of the posterior distribution of random effects was obtained in Jacqmin-Gadda et al. (2010) in the context of joint latent classes and survival models.

Both estimates in (3.7) are only asymptotically independent samples; in practice, there is a dependency between the estimates (Therneau and Grambsch, 2000). In particular, the martingale residuals $\widehat{M_i^D}$ are constrained to have mean 0 , therefore (3.7)

is proportional to the sample covariance of the martingale residuals and expected log-frailties, which is a measure of linear dependence. In fact, if an ordinary linear regression model is considered:

$$\widehat{M_i^D} = a + b\,\widehat{\log z_i} + \varepsilon_i,$$

then the departure of (3.7) from 0 is equivalent to the departure of the regression coefficient b from 0. Thus, for testing H_0, the regular t statistic can be used:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{3.8}$$

where $r = \text{Corr}(\widehat{M_i^D}, \widehat{\log z_i})$ and t follows asymptotically a t distribution with $n-2$ degrees of freedom under the null hypothesis under H_0. We refer to the test based on (3.8) as the **robust score test** (RST).

Heuristically, a justification for the RST can be derived by interpreting the quantities which appear in (3.7). The martingale residuals $\widehat{M_i^D}$ can be informally interpreted as an "observed - expected" quantity for the terminal event. For example, if $\widehat{M_i^D} > 0$, then the rate of the terminal event is *larger than expected, taking only the x_i into account*, and how much *larger* is determined by how large $\widehat{M_i^D}$ is. A large (log-)frailty estimate corresponds to a subject who is at high risk for recurrences. Hence, the larger the value of (3.7), the stronger the evidence for the association between recurrent and terminal events is. More frail subjects are more likely to experience the terminal event earlier if $r > 0$, or later if $r < 0$, so the sign of the RST statistic also indicated the direction of the association.

3.3.2 Alternative tests

The **likelihood ratio test** (LRT) can be computed by maximizing the likelihood (3.3) via the expectation-maximization algorithm, as described in Liu, Wolfe, and Huang (2004), and comparing it to the likelihood under H_0. If (3.3) is maximized in $(\widehat{\gamma}, \widehat{\eta})$, then the LRT statistic is

$$D = -2\log\left\{\frac{l(0, \hat{\eta}_0)}{l(\widehat{\gamma}, \widehat{\eta})}\right\}$$

and it asymptotically follows a χ^2 distribution with one degree of freedom under H_0.

The **efficient score test** (EST) is described by (3.5) and the efficient information (3.6), and as previously mentioned it can be computed numerically. As shown in Murphy and Vaart (2000), the efficient information can be obtained as minus the second derivative of the profile likelihood

$$\ell_{\text{prof}}(\gamma) = \sup_{\eta} \ell(\gamma, \eta). \tag{3.9}$$

In practice, we can approximate $\tilde{I}_\gamma\Big|_{\gamma=0} = -E\left(\frac{d^2}{d\gamma}\ell_{\text{prof}}(\gamma)\Big|_{\gamma=0}\right)$ with the numeric Hessian of (3.9) in $\gamma = 0$. This can be obtained from general purpose optimization software, such

Table 31: Average number of recurrent event in simulated data sets.

	θ		
γ	0.5	1	1.5
-0.5	2.36	2.44	2.52
-0.25	2.32	2.36	2.41
0	2.27	2.27	2.27
0.25	2.21	2.16	2.12
0.5	2.13	2.05	1.96

as the function `optim` in `R` or `S-Plus` or the package `numDeriv` in `R`. We comment in the Appendix on computational considerations regarding the EST and how it is related to the NST in this light.

Alternatively, the $\hat{\gamma}$ can be obtained from maximizing (3.9) with respect to γ. The variance of the estimate $\text{Var}(\hat{\gamma})$ can be obtained from the numeric Hessian, and then the **Wald test** statistic is

$$W = \frac{\hat{\gamma}}{\sqrt{\text{Var}(\hat{\gamma})}}$$

and it asymptotically follows a standard normal distribution under H_0. The sign of W also corresponds to the direction of the tested association.

3.4 Simulation

A simulation study has been conducted to assess the validity of the Robust Score Test (RST) and compare the small sample properties to those of the other tests described in Section 3.3. The simulations have been carried out in the following setting: data sets consist of $n \in 100, 200, 500$ subjects; for each subject the data is generated according to model (3.2), for scenarios pertaining to $\gamma \in \{-0.5, -0.25, 0, 0.25, 0.5\}$. The frailty is generated from a Gamma distribution with mean equal to 1 and variance $\theta \in \{0.5, 1, 1.5\}$. One binary covariate is generated from a Binom(n, 1/2) distribution with fixed regression coefficients $\beta = \alpha = 1$. An exponential baseline hazard is used, with $\lambda_0(t) = 1/2$ and $r_0(t) = 2$. The follow-up is ended by either the terminal event, or by an administrative censoring time $C_i = 1$, whichever occurs first. Note that the recurrent event rate and the terminal event rate are independent only under H_0. Every simulation cycle consist of 1000 replications under the same conditions.

In Table 31 we show an indication on the size of the simulated data sets. It can be seen that the number of recurrent events decreases with γ and with θ. The asymmetry is explained by the fact that when $\gamma < 0$ the recurrent events have a "protective" effect and subjects with many events exit the data set later. The degree to which there is more variance in the frailty amplifies this effect.

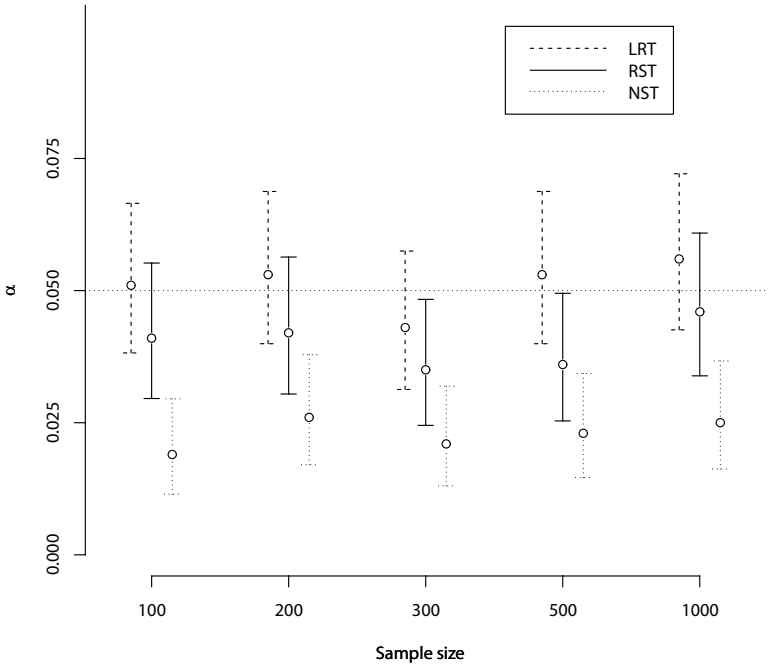

Figure 31: Estimated α levels with simulated data under H_0. Wald and EST (not shown here) are close to the LRT estimates. Binomial confidence intervals are also shown, where a "success" is a p-value smaller than 0.05

We use the abbreviations of the tests as described in Sections 3.3.1 and 3.3.2. Furthermore, we also consider the Wald test from a parametric model where the baseline intensities are considered piecewise constant with 3 intervals, from the R package `frailtypack`; this approach is described in Section 3.2, and we see it as an approximation to the semiparametric joint model.

Figure 1 compares the type 1 error (false rejections) of the LRT, RST and NST, as a function of n, under H_0, in the case $\theta = 1$. Although the estimated α level seems consistently lower for RST than for LRT, bionomial confidence intervals for the proportion of rejections have a notable overlap, and both seem to approach the desired 0.05 with a sufficiently large sample. In this comparison, it can also be seen that the naive score test (NST) is indeed over-conservative, as it is argued also in Appendix 3.6: even as the sample size becomes larger, the proportion of rejections is significantly lower then the nominal α level of 0.05. Finally, we note that the results for $\theta \in \{0.5, 1.5\}$ (not shown) are very similar.

To better illustrate the relation between the different tests, we plotted the p-values obtained in the case $\gamma = 0$, $\theta = 1$, $n = 500$ in Figure 32. Under the null hypothesis, one

would expect the p-values of a valid test to be approximately uniformly distributed on $[0, 1]$. The Wald test and EST are virtually indistinguishable from the LRT in this case. The figure indicates that RST approximates the LRT for small deviations as well. The parametric Wald (WaldPar) test is also shown in the plot; it can be seen that the p values can differ wildly from those of the semi-parametric Wald; this can be seen as a trade off for the parametric assumption. For other values of n or θ very similar figures were obtained.

Finally, we analyze the power of the aforementioned tests against the alternatives $\gamma \in \{-0.5, -0.25, 0.25, 0.5\}$, for $\theta \in \{0.5, 1, 1.5\}$. The results are summarized in Table 32. Two trends are visible regardless of sample size. First, the power of the tests grows with the frailty variance, meaning that it is more likely to reject the null hypothesis of no association in more heterogeneous data sets, if this association exists. Second, in particular for LRT, Wald and EST, the tests fare slightly better for alternatives with $\gamma < 0$, which can be explained by the asymmetric size of the simulated data sets showed in table 31.

As expected, the tests are more powerful when there is a higher number of individuals in the data set. The RST performs better than Wald for small sample sizes ($n = 100$), however there is no clear difference for others. Generally, the power of the RST is slightly lower but reasonably comparable with the other tests. In Figure 33 we compare the power of the tests for $\theta = 1$. It can be seen that, except for NST which is over-conservative, the LRT, Wald, EST and RST are quite similar. It looks like for small samples there is a slight advantage in power of LRT and EST, while the RST is closer to the Wald test.

Finally, we note that the computation time is much smaller for the RST, as compared to the other tests, including the parametric Wald test, WaldPar. Average computation times from the simulations are shown in Table 33.

3.5 Application

We illustrate our methods using data from a study on Mycosis Fungoides (MF). MF is the most common type of cutaneous T-cell lymphoma that generally presents with patches and plagues Doorn, Scheffer, and Willemze (2002). Over time a number of patients progress to tumor stage disease (stage IIB) and a minority develop extracutaneous localization of the disease. It is well known that there is considerable variability in the number of recurrent skin tumors and is believed that an increased number of recurrent skin tumors is associated with disease progression and survival. In addition, it has been reported that folliculotropism of neoplastic cells is associated with an adverse prognosis. In Boonk et al. (2014), 46 patients with stage IIB MF were selected from the cutaneous lymphoma database of the Dutch Cutaneous Lymphoma Group. During follow-up, data on recurrences of skin tumor and disease progression and survival were collected. We consider overall survival as the terminal event. Median follow-up was 88 months. Covariates considered in this application are age (median 69, range 39–90), gender (33

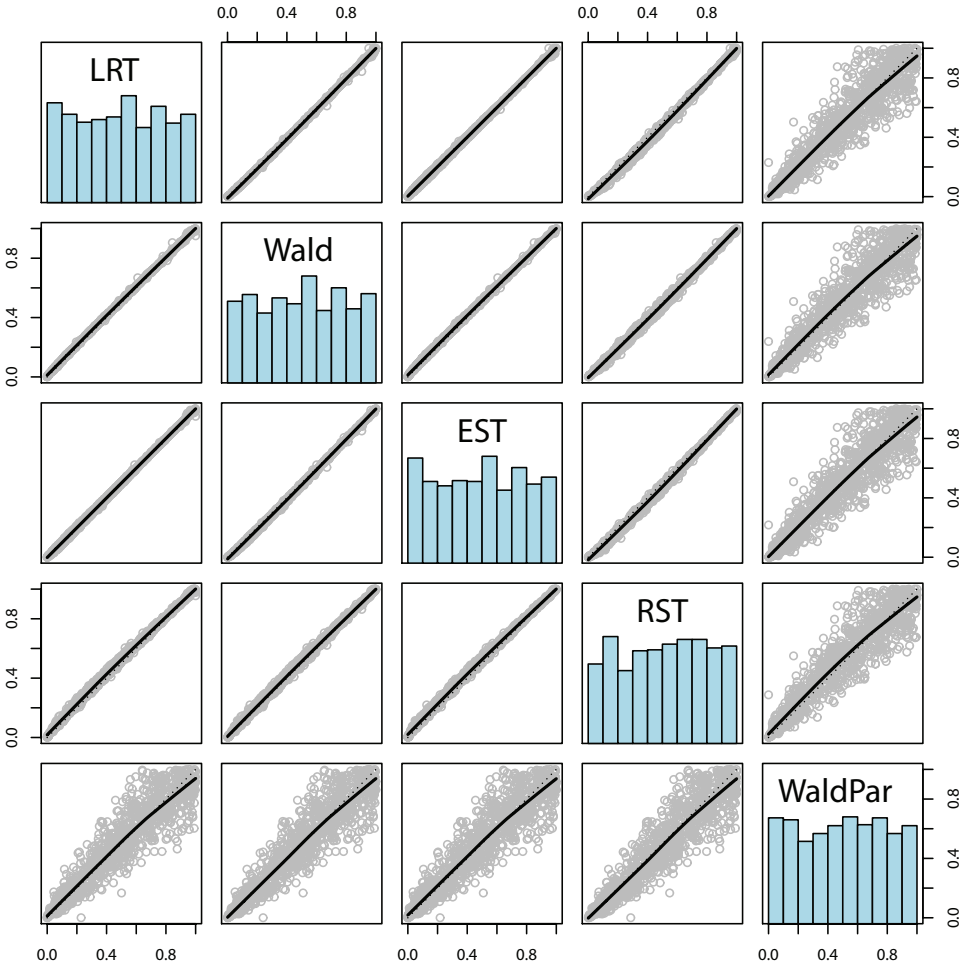

Figure 32: Histograms and scatterplots of p-values corresponding to 1000 datasets simulated under $H_0 : \gamma = 0$. Within the scatter plots, a straight line with equation $y = x$ has been added, as well as a dotted nonparametric smoother. The data sets follow the simulation scenarios of Section 3.5 with $n = 500$.

Table 32: Power against alternative hypotheses with varying sample size $n \in \{100, 200, 500\}$ and frailty variance $\theta \in \{0.5, 1, 1.5\}$

		LRT			Wald			EST			RST		
θ	γ	100	200	500	100	200	500	100	200	500	100	200	500
0.5													
	-0.5	0.37	0.60	0.94	0.25	0.56	0.94	0.42	0.63	0.95	0.32	0.54	0.92
	-0.25	0.13	0.19	0.49	0.08	0.19	0.47	0.18	0.28	0.52	0.11	0.18	0.42
	0.25	0.14	0.17	0.51	0.08	0.17	0.47	0.14	0.22	0.50	0.10	0.18	0.44
	0.5	0.33	0.54	0.96	0.21	0.54	0.93	0.33	0.6	0.94	0.27	0.54	0.92
1													
	-0.5	0.70	0.94	1.00	0.64	0.93	1.00	0.72	0.94	1.00	0.65	0.92	1.00
	-0.25	0.29	0.55	0.91	0.22	0.51	0.90	0.37	0.60	0.93	0.26	0.50	0.89
	0.25	0.30	0.49	0.89	0.20	0.42	0.87	0.28	0.46	0.88	0.25	0.44	0.88
	0.5	0.72	0.95	1.00	0.61	0.92	1.00	0.69	0.94	1.00	0.65	0.92	1.00
1.5													
	-0.5	0.85	0.99	1.00	0.82	0.99	1.00	0.75	0.92	0.99	0.82	0.98	1.00
	-0.25	0.47	0.80	0.99	0.39	0.76	0.98	0.56	0.83	0.99	0.44	0.77	0.98
	0.25	0.46	0.77	0.99	0.35	0.71	0.98	0.44	0.75	0.98	0.4	0.71	0.98
	0.5	0.88	0.99	1.00	0.82	0.99	1.00	0.87	0.99	1.00	0.86	0.98	1.00

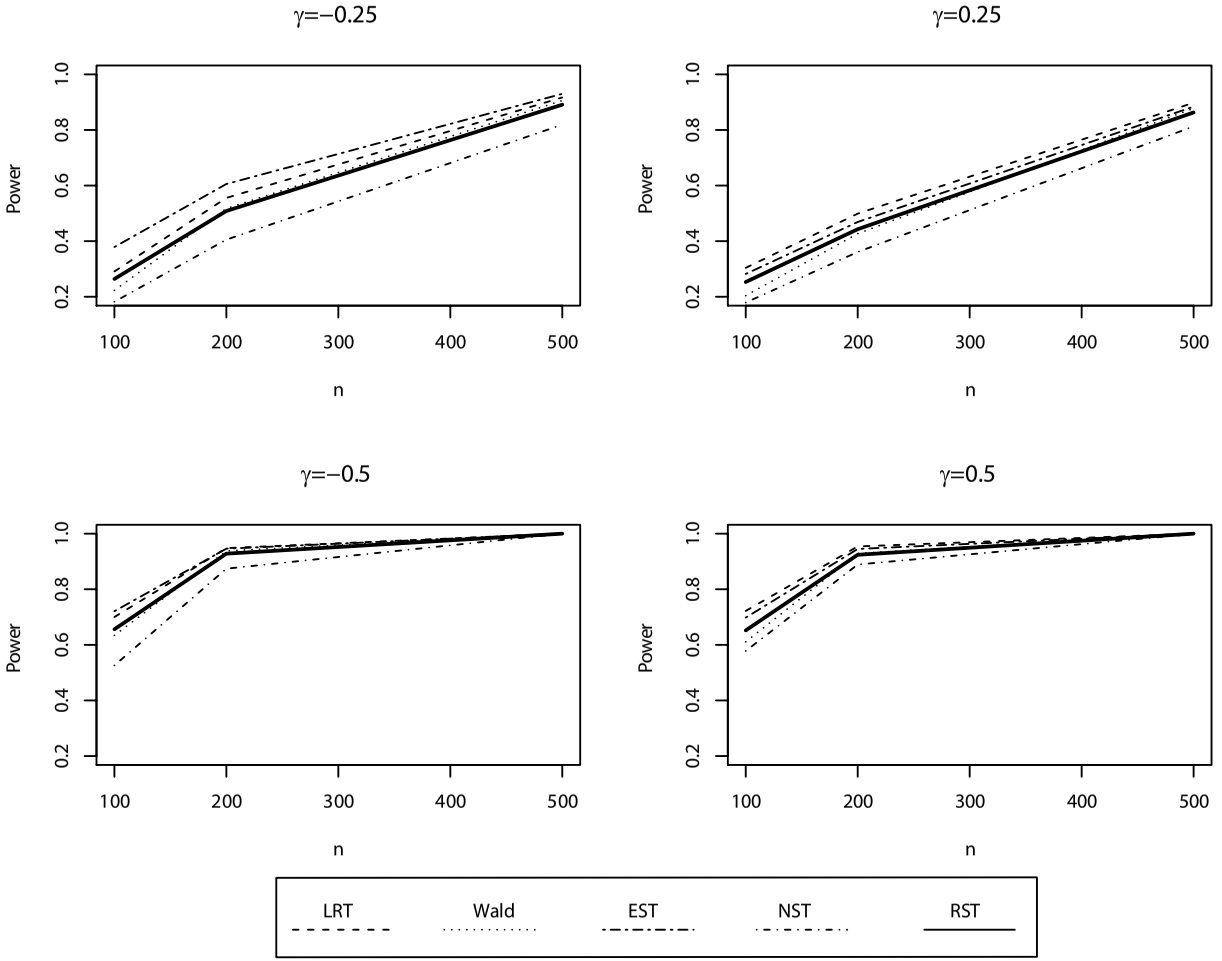

Figure 33: Power of LRT, Wald, EST, NST, and RST compared for $\theta = 1$

Table 33: Average computation time for different tests. For RST the standard `survival` package was used, for WaldPar the `frailtypack` package, and for EST and LRT or Wald a self-written algorithm was used, similar to that described in Liu, Wolfe, and Huang (2004).

	Computation time (s)		
	100	200	500
RST	0.04	0.09	0.31
EST	16.18	48.05	138.03
WaldPar	1.04	1.64	2.68
LRT/Wald	44.57	128.25	331.61

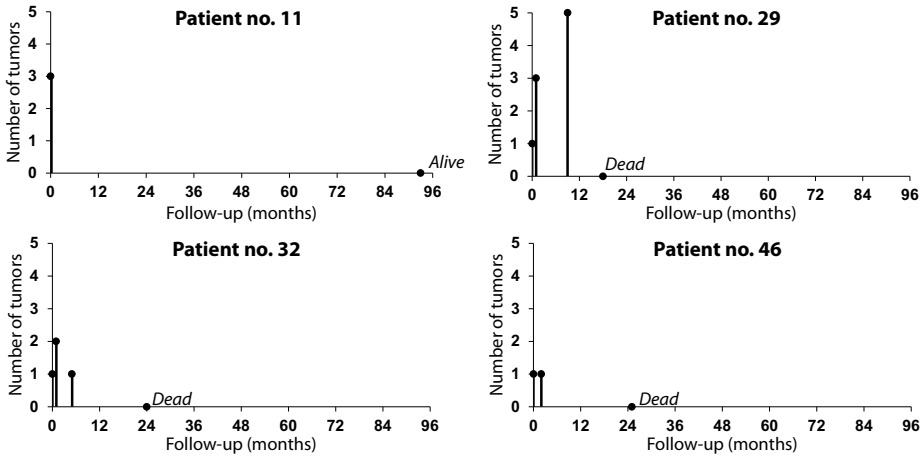

Figure 34: Recurrent event history and survival outcome of 4 patients

males, 13 females), and the presence of folliculotropic MF (26 absent, 20 present). Figure 34 shows examples of the variability in the number of tumors and time intervals between tumor recurrences. It can be seen that some patients experienced multiple recurrences at a single follow-up visit; the ties caused by these simultaneous recurrences were randomly broken. 11 patients (23.9%) experienced 0 recurrences, 5 (10.8%) 1 recurrence, 6 (13.0%) 2 recurrences, and 24 (52.1%) more than 2 recurrences. The maximum number of recurrences was 21. The original publication (Boonk et al., 2014) used the number of recurrent skin tumors in the first year as explanatory variable in a landmark Cox model at 1 year for overall survival, and showed that the number of recurrent skin tumors was highly prognostic for subsequent survival.

A gamma frailty model ignoring possible informative censoring due to the terminal event death, yielded the results shown in Table 34, under "Separate models". The frailty variance was estimated to be 1.574. The estimates of the Cox model for the terminal

Table 34: Estimated regression coefficients for recurrent events and terminal event, using separate models and the joint model.

	Separate models			Joint model		
	Beta	SE	p-value	Beta	SE	p-value
Recurrent events						
Male gender	0.230	0.687	0.74	0.286	0.476	0.54
Age	0.039	0.020	0.058	0.039	0.018	0.035
Folliculotropic MF	0.019	0.595	0.97	0.039	0.276	0.88
Frailty variance	1.574		< 0.0001	1.358	0.323	< 0.0001
Association parameter (γ)				0.778	0.276	0.004
Terminal event						
Male gender	0.616	0.486	0.20	0.747	0.648	0.24
Age	0.048	0.019	0.012	0.067	0.023	0.004
Folliculotropic MF	0.378	0.402	0.35	0.127	0.486	0.79

event, ignoring the recurrent events is also shown under "Separate models". Figure 35 shows a scatterplot of the posterior log frailties from the gamma frailty models against the martingale residuals of the Cox model for the terminal event. The correlation was estimated to be 0.488, and the p-value of the robust score test was 0.0006. The result of this quick test indicates that a joint model is really needed to reliably model the association between the recurrent skin tumors and death. The result of this joint model, using a self-written EM-algorithm, is shown in Table 34, under the "Joint model". The regression coefficients in the joint model are generally comparable with the ones from the separate models. The association parameter γ was estimated to be positive and highly significant, indicating an increased death rate for the subjects with a high propensity of recurrent events, in agreement with the findings in Boonk et al. (2014).

3.6 Discussion

We have shown that the estimated correlation between the martingale residual and the estimated log-frailties can be used as the basis for a test of association between recurrent events and a terminal event. The advantage of the robust score test is that it is easy to compute and does not require fitting the joint model. Thus, it can serve as a simple preliminary check whether models for the recurrent events and for the terminal events can be fitted separately or whether more complex joint models are needed to obtain reliable estimates.

We note that heterogeneity with respect to the recurrent events is required not only for the joint model to be estimated, but also for the implementation of the RST. This can be assessed via a likelihood ratio test (Nielsen et al., 1992; Therneau and Grambsch, 2000). In addition, we note that the model described in Section 3.2 leads to the interpretation

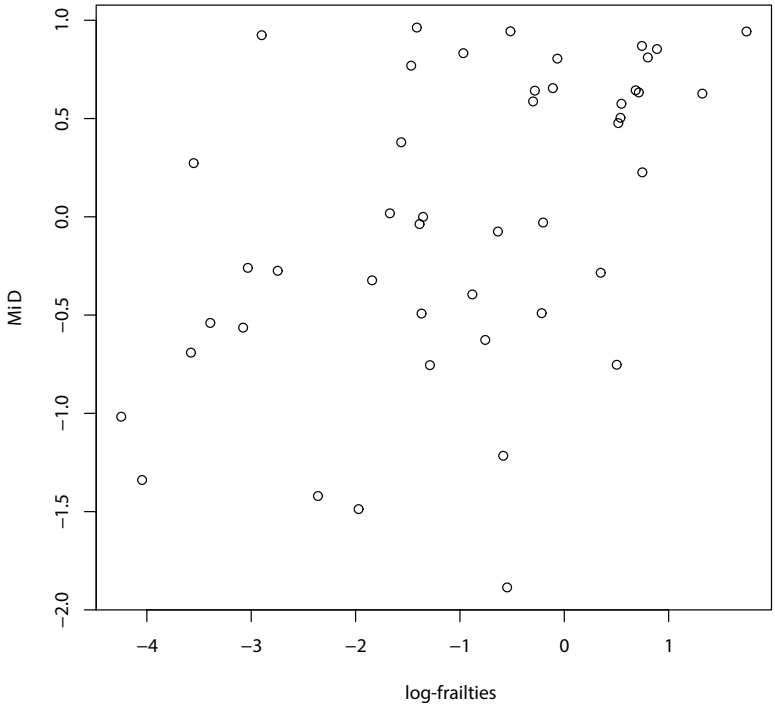

Figure 35: Martingale residuals of terminal event versus the posterior log-frailties estimated from the recurrent events

of a common hidden factor influencing both risks of experiencing recurrent events and the terminal event. The plausibility of this should be assessed separately, because more models can describe the type of data encountered in this chapter (Cook and Lawless, 2007, ch. 6.6) and the effects of internal time dependent covariates are often difficult to separate from that of the frailty (Aalen, Borgan, and Gjessing, 2008, ch. 8.5).

The fact that the martingale residuals and the estimates of the log-frailty are not samples coming from a bivariate normal distribution should also lead to a cautious interpretation of correlation coefficients and of the test statistic (3.8). In the simulations of Section 3.4 we did not notice any increase in the estimated α levels of the RST, but this might depend on the data set on which the method is employed. Finally, note that there is no closed form connection between the parameter which describes the association between recurrent events and terminal event γ and the correlation ρ used to calculate the RST statistic (3.8).

Although we have not explicitly stated that the frailty should follow a gamma distribution throughout Section 3.2, we still employed this assumption in Sections 3.4 and 3.5. The RST can accommodate any distribution for the frailty, including, for example,

a two-point mixture or a compound Poisson distribution, as long as the shared frailty model for recurrences can be estimated. It can be seen from (3.8) that the choice of the frailty distribution will affect only the estimation of $\widehat{\log z_i}$. We expect the RST to have the largest power if the true frailty distribution is used, however this was not checked in the simulation study.

The idea of a simple test, here in the form of RST, could be extended to more models which inherit the issues which would prevent practicians to use a more complicated joint model. Because a recurrent event data in the presence of a terminal event is a particular case of a multistate model with competing risks (Cook and Lawless, 2007, ch. 6.6), similar methods could be found by generalizing RST to multistate models with frailty (Putter and Houwelingen, 2015).

Appendix: Estimation via profile likelihood

In Sections 3.3.2 and 3.4 we used the profiling out of the nuisance parameters from the log-likelihood (3.3), in the sense shown by the definition (3.9). First, note that, if $(\hat{\gamma}, \hat{\eta})$ maximizes (3.3), then $\hat{\gamma}$ maximizes (3.9), and $\hat{\eta}$ is the estimate of η obtained by maximizing $\ell(\hat{\gamma}, \eta)$. It is clear that

$$\ell_{\text{prof}}(0) = \ell(0, \hat{\eta}_0)$$

and $\ell_{\text{prof}}(\gamma) \geq \ell(\gamma, \hat{\eta}_0)$ with equality only when $\gamma = 0$. It follows that $\ell_{\text{prof}}(\gamma) - \ell(\gamma, \hat{\eta}_0) \geq 0$. Thus,

$$\frac{d}{d\gamma} \left\{ \ell_{\text{prof}}(\gamma) - \ell(\gamma, \hat{\eta}_0) \right\} \Big|_{\gamma=0} = 0,$$

which shows that $U_\gamma(0, \hat{\eta}_0)$ from (3.7) is equal to the efficient score function, $U_\gamma(0) = \frac{d}{d\gamma} \ell_{prof}(\gamma) \Big|_{\gamma=0}$. This justifies why (3.7) is the correct score function for testing H_0. Further, because $\ell_{prof}(\gamma) - \ell(\gamma, \hat{\eta}_0)$ is always positive and it has a minimum, it follows that

$$\frac{d^2}{d\gamma^2} \left\{ \ell_{\text{prof}}(\gamma) - \ell(\gamma, \hat{\eta}_0) \right\} \geq 0$$

for any value of γ. This implies that $\frac{d^2}{d\gamma^2} \ell_{prof}(\gamma) > \frac{d^2}{d\gamma^2} \ell(\gamma, \hat{\eta}_0)$ for all values of γ, which is equivalent to

$$-\frac{d^2}{d\gamma^2} \ell_{\text{prof}}(\gamma) = I_\gamma \leq I_{\gamma\gamma} = -\frac{d^2}{d\gamma^2} \ell(\gamma, \hat{\eta}_0)$$

for all γ. We conclude that $\ell_{\text{prof}}(\gamma)$ and $\ell(\gamma, \hat{\eta}_0)$ have the same value and the first derivative in $\gamma = 0$, but the curvature of $\ell(\gamma, \hat{\eta}_0)$ is more pronounced. This is the intuition behind the reason why the likelihood $\ell(\gamma, \hat{\eta}_0)$ with fixed nuisance parameters can be used to obtain the correct score, but not the correct information.