Dept. of Math. University of Oslo
Statistical Research Report No. 8
ISSN 0806–3842 August 2004

Dynamic analysis of multivariate failure time data

Odd O Aalen[1,*], *Johan Fosen*[1], *Harald Weedon-Fekjær*[1,2], *Ørnulf Borgan*[3] *and Einar Husebye*[4,5]

[1]Section of Medical Statistics, University of Oslo, P.O.Box 1122 Blindern, N-0317 Oslo, Norway

[2]The Cancer Registry of Norway, Montebello, N-0310 Oslo, Norway

[3]Department of Mathematics, University of Oslo, P.O.Box 1053 Blindern, N-0316 Oslo, Norway

[4]Dept.of Internal Medicine, Ullevål University Hospital of Oslo, N-0407 Oslo, Norway

[5]Hospital of Buskerud, N-3004 Drammen, Norway

[*]e-mail: o.o.aalen@basalmed.uio.no

Summary

We present an approach for analysing internal dependencies in counting processes. This covers the case with repeated events on each of a number of individuals, and, more generally, the situation where several processes are observed for each individual. We define dynamic covariates, i.e. covariates depending on the past of the processes. The statistical analysis is performed mainly by the nonparametric additive approach. This yields a method for analysing multivariate survival data, which is an alternative to the frailty approach. We present cumulative regression plots, statistical tests, residual plots and a hat matrix plot for studying outliers. A program in R and S-PLUS for analyzing survival data with the additive regression model is available on the web site `www.med.uio.no/imb/stat/addreg`. The program has been developed to fit the counting process framework.

Key Words: Aalen's additive model; Counting processes; Dynamic covariates; Event history analysis; Multivariate survival data; Repeated events; Survival analysis

1 Introduction

In event history analysis one is often confronted with series of events of a specific type observed for each of a number of individuals. Such data may also be termed multivariate failure time data. The following examples show different settings where such data arise.

1. Movements of the small bowel Aalen & Husebye (1991) analyzed the cyclic pattern of motility (spontaneous movements) of the small bowel in humans called the migrating motor complex (MMC). Phase III of MMC consists of a sequence of regular contractions migrating down the small bowel at irregular intervals; these intervals lasting from minutes up to several hours. In an ambulatory recording system, as used by Husebye et al (1990), several such intervals in a row are registered for each individual during prolonged recordings, with a censored one at the end, signifying the end of observation when the recording terminates at a predefined time.

One is interested in understanding various aspects of this process, e.g. wether there is a variation in frequency between individuals, how the frequency changes over time, and what governs the duration between MMC's.

2. Duration of amalgam fillings When making a study of the duration of amalgam fillings in teeth, one may include several patients who each have many fillings in their teeth. A study of this kind, including 32 patients, with from 4 to 38 fillings for each patient, was analyzed by means of frailty models by Aalen, Bjertness and Sønju (1995). One is interested in the duration of amalgam fillings, and how this depends on patient properties.

3. Analysis of sleep patterns Yassouridis et al (1999) study the occurrence of various sleep patterns, including REM (Rapid Eye Movements) sleep, by means of Cox type models with regression parameters varying over time. Sleep was monitored continuously for a number of individuals, and events where defined as falling asleep, waking up, going into REM sleep etc. Each individual may experience a number of such events during a night, and event history analysis may be used to analyze the pattern. One question asked is the relation of sleep to measurement of the stress hormone cortisol, which was monitored every twenty minutes throughout the night.

The datasets above fall into two different categories. In the case of the amalgam fillings, several dependent processes, namely one for each filling, is observed for each patient. In the other cases there is only one process, or a few processes defined for a set of specific event types, observed for each indvidual, but the event of interest repeats itself several times over each process. The sleep example is different from the others in that the effects of some covariates vary considerably over time. One aim of the additive model presented here is to handle just this kind of time-varying effects.

Data like those illustrated in these examples are typically analysed by frailty models, which is a kind of random effect model, see e.g. Hougaard (2000). It is well known that frailty models may alternatively be viewed in a dynamic fashion (see e.g. Aalen, 1988). By this we mean that instead of setting up a random effects, or frailty, model, one may alternatively condition with respect to past events to get a counting process model with suitable intensity processes. Frailty will induce dependence, such that, e.g., the rate of a new event is increased if many events have been observed previously for this individual, since this would indicate a high frailty. The aim of the present paper is to use a dynamic viewpoint for statistical

analysis of multivariate survival data. This turns out to be fruitful with the potential of giving more detailed information than a traditional frailty analysis. In fact, in the dynamic approach one does not have to postulate speculative latent, or frailty, variables, the actual existence of which is often uncertain. Furthermore, frailty models often appear as unrealistically simple, e.g. usually not including a varying frailty over time, and the present approach yields more flexibility.

Hence, we shall demonstrate the use of dynamic covariates, that is, covariates depending on the past of the process. The actual analysis may be carried out by different methods. Below we shall introduce the appropriate extension of an additive regression model, which shall be our main tool. Alternatively, one may apply a Cox model with dynamic covariates, as will be done briefly below. In fact, there is a paper by Cox (1972b) that extends his regression model to the case of observing stochastic processes, including dynamic covariates. This is a companion to his famous paper (Cox, 1972a), but has been largely ignored in the statistical literature. Still, dynamic covariates have been used by some authors, for instance Kalbfleisch and Prentice (2002, Chapter 9), Peña and Hollander (2003), Gandy and Jensen (2004), and Martinussen and Scheike (2000).

A program in R and S-PLUS for analyzing survival data with the additive regression model is available on the web site `www.med.uio.no/imb/stat/addreg`. The program has been developed to fit the counting process framework adopted in this paper. It can do analyses like those presented in this paper; with the limitation that dynamic covarates involving time since last occurrence have to be categorized.

2 A counting process formulation

We shall use the framework of counting processes, (Aalen, 1978). In the application of counting processes in survival analysis one usually only studies cases where there is at most one event for each individual, or maybe a few events of different types (see e.g. the monograph by Andersen et al (1993)). The multivariate event situation considered here is only rarely treated within the counting process framework. This is actually a highly unnatural limitation which does not fully exploit the value of counting processes, and we shall here show the value of considering processes with several events for each individual.

For individual i, let $N_i(t)$ be the process counting the number of occurrences of the event of interest up to time t, with $\underline{\mathbf{N}}(t)$ denoting the column vector of these counting processes. An individual counting process will be a step function starting at 0, and then jumping up one unit whenever an event happens. The intensity process, $\lambda_i(t)$, describes the risk that an event occurs at time t as a function of the past, and is therefore the natural mathematical concept to study dynamic covariates.

The existence of dynamic models follows from a general theorem for submartingales, namely the Doob-Meyer decomposition which states, essentially, that any submartingale can be decomposed into a martingale and a compensator. Note that this requires the notion of a history of past events, defined by an increasing family of σ-algebras. The history includes all previous occurrences in the relevant counting processes; in addition the history may include external information. A counting process is obviously a submartingale and, under certain regularity conditions, the compensator is just the integrated intensity process. What the Doob-Meyer decomposition tells us, is that there is essentially always an intensity process however the counting processes comes about. For instance, there might be an underlying random

effects, or frailty model, of a possibly complex and general nature, nevertheless the whole thing can be reformulated by intensity processes depending on the past.

One way of analysing counting processes, is to use a proportional hazards model. An appropriate extension of the Cox model to counting processes is given by Andersen and Gill (1982). Although, in such a formulation, the underlying hazard is arbitrary, and in fact nonparametric, the dependence on covariates is determined by strong parametric assumptions. This seems questionable when observing a process over some period of time, where one could very well imagine that conditions change. Therefore, the nonparametric additive model of Aalen (1989) represents an alternative. Since the Doob-Meyer decomposition guarantees the existence of an intensity process, it seems quite natural to attempt an additive model, in the absence of specific information as to other valid models.

2.1 The additive regression model

The additive model has been thoroughly studied for the usual survival situation where each individual experiences at most one event. We will here consider it for the case where there are several events for each individual, and focus on dynamic covariates.

We have to distinguish two situations: (i) Only one process is considered for each individual, like in examples 1 and 3 of the introduction. (ii) Several possibly dependent processes are observed for each individual, like the fate of individual amalgam fillings in example 2. We shall first give the details of situation (i), and then indicate necessary modications for situation (ii).

(i) *One process per individual.* Let $\underline{\boldsymbol{\lambda}}(t)$ be the column vector consisting of the

intensities $\lambda_i(t)$ for n individuals. We shall then apply the following model:

$$\underline{\boldsymbol{\lambda}}(t) = \underline{\mathbf{Y}}(t)\,\underline{\boldsymbol{\alpha}}(t) \tag{1}$$

where $\underline{\mathbf{Y}}(t)$ is a $n \times r$ matrix. If an individual i is in the risk set at time t, then the first element of row i equals 1, while the remaining elements are covariates. If the individual is for the moment not in the risk set, then the corresponding row is set equal to zero. Note that the elements of $\underline{\mathbf{Y}}(t)$ may be arbitrary (apart from regularity conditions) predictable stochastic processes. The $r \times 1$ vector $\underline{\boldsymbol{\alpha}}(t)$ are *abitrary regression* functions, hence the term *nonparametric*, indicating the effects of the various covariates. The first element of $\underline{\boldsymbol{\alpha}}(t)$ is called the baseline intensity since it corresponds to all covariates being equal to zero.

The integral of $\underline{\boldsymbol{\alpha}}(t)$, denoted $\underline{\mathbf{A}}(t)$, can easily be estimated in this case. The estimate of the integrated regression functions is given by (Aalen (1989), Andersen et al (1993, Section VIII.4)):

$$\widehat{\underline{\mathbf{A}}}(t) = \int_0^t \underline{\mathbf{Y}}(s)^-\, d\underline{\mathbf{N}}(s)$$

where $\underline{\mathbf{Y}}(t)^-$ denotes a generalized inverse. A common choice for the generalized inverse is the least square (or Moore-Penrose) inverse:

$$\underline{\mathbf{Y}}(t)^- = (\underline{\mathbf{Y}}(t)'\underline{\mathbf{Y}}(t))^{-1}\underline{\mathbf{Y}}(t)'$$

In some cases $\underline{\mathbf{Y}}(t)'\underline{\mathbf{Y}}(t)$ may be singular and the generalized inverse is not well defined. This may, for instance, occur at early times in processes with dynamic covariates, since the process has to run a while for the covariates to be well defined and reasonably stable. One solution is to understand $\underline{\mathbf{Y}}(t)^-$ to be identically equal to zero when $\underline{\mathbf{Y}}(t)'\underline{\mathbf{Y}}(t)$ is singular. Hence, estimation takes a pause, returning zero values, in the case of singularity. However, in singular or near-singular situations

one may alternatively use ridge regression to avoid stopping the estimation. The generalized inverse is then modified in the following way:

$$\underline{\mathbf{Y}}_R(t)^- = (\underline{\mathbf{Y}}(t)'\underline{\mathbf{Y}}(t) + k\underline{\mathbf{I}})^{-1}\underline{\mathbf{Y}}(t)'$$

where $\underline{\mathbf{I}}$ is the identity matrix and k is a suitable costant. Ridge regression is a standard tool in ordinary linear regression with well known properties in that context. We shall find it useful here.

Often, one may want to center the covariates. That is, for all columns in $\underline{\mathbf{Y}}(t)$, except the first, one subtracts the mean of those individuals at risk at a given time. Then column one of $\underline{\mathbf{Y}}(t)$ is orthogonal with respect to the remaining columns, implying that the first element of $\widehat{\underline{\mathbf{A}}}(t)$ is the same estimate as one would get if the covariates where not included in the analysis. This implies that first element of $\widehat{\underline{\mathbf{A}}}(t)$, that is the cumulative baseline intensity, is the Nelson-Aalen estimator. We shall center all covariates below, thus ensuring the validity of this interpretation.

Testing methods, estimates of variances, asymptotic results etc may be found in the mentioned references, particularly Andersen et al (1993) and Aalen (1993).

(ii) *Several processes for each individual.* The case with several units at risk for each individual is mentioned in example 2 above where the units are fillings at risk. One alternative is to aggregate over individuals, such that the counting process N_i for individual i, i.e. the i-th element of $\underline{\mathbf{N}}$, is the sum of all counting processes for the individual's units at risk. The intensity may be written:

$$\underline{\boldsymbol{\lambda}}(t) = \underline{\mathbf{K}}(t)\underline{\mathbf{Y}}_0(t)\,\underline{\boldsymbol{\alpha}}(t)$$

where $\underline{\mathbf{K}}(t)$ is a diagonal matrix with the diagonal elements containing the number of units at risk at time t for the individuals, and $\underline{\mathbf{Y}}_0(t)$ is defined like $\underline{\mathbf{Y}}(t)$ in case

(i). A reasonable estimator of $\underline{\mathbf{A}}(t)$ in this case is

$$\widehat{\underline{\mathbf{A}}}_0(t) = \int_0^t \underline{\mathbf{Y}}_0(s)^- \underline{\mathbf{K}}(s)^{-1} d\underline{\mathbf{N}}(s)$$

Simple modifications of known results for $\widehat{\underline{\mathbf{A}}}(t)$ yield corresponding results for $\widehat{\underline{\mathbf{A}}}_0(t)$.

It is not always natural to aggregate. Considering the amalgam fillings, it might be the case that individual fillings for a patient had tooth-specific covariates, and that the counting processes could thefore not be aggregated without loss of information. In that case there would be a counting process for each filling. The dynamic covariates would then have to contain information on the previous fates of all fillings belonging to a specific person. This creates a dependence between groups of counting processes. Such dependence might be relevant in many connections. In studies in social science it may be that events considered are dependent on peoples attitudes. The likelihood of a couple divorcing, e.g. may be dependent on the number of divorces having occurred among their family, friends and colleagues.

2.2 Residual processes. The hat matrix

We shall look at residual processes defined by taking the difference between the counting process and the estimated integrated intensity, i.e. a kind of "observed minus expected" difference. This can only be computed when the model is estimable, that is $\underline{\mathbf{Y}}(t)'\underline{\mathbf{Y}}(t)$ is non-singular (unless extended by ridge regression). More formally, the vector of martingale residual processes is defined by (Aalen, 1993):

$$\underline{\mathbf{M}}_{\text{res}}(t) = \int_0^t J(s)\, d\underline{\mathbf{N}}(s) - \int_0^t J(s)\underline{\mathbf{Y}}(s)\, d\widehat{\underline{\mathbf{A}}}(s).$$

where $J(t)$ is the indicator function of the event that $\underline{\mathbf{Y}}(t)'\underline{\mathbf{Y}}(t)$ is non-singular. The process $\underline{\mathbf{M}}_{\text{res}}(t)$ is, in fact, a martingale when the model is true (Aalen, 1993).

It is common to talk about martingale residuals also in connection with the Cox model. However, in that case there is no exact martingale structure, merely an approximation.

The quadratic variation process of $\underline{\mathbf{M}}_{\text{res}}(t)$ may be derived from the theory of stochastic integrals, to yield the matrix

$$\langle \underline{\mathbf{M}}_{\text{res}} \rangle (t) = \int_0^t J(s) \, (\underline{\mathbf{I}} - \underline{\mathbf{Y}}(s)\underline{\mathbf{Y}}(s)^{-}) \, \text{diag}(\underline{\boldsymbol{\lambda}}(s) \, ds)(\underline{\mathbf{I}} - \underline{\mathbf{Y}}(s)\underline{\mathbf{Y}}(s)^{-}).$$

Here $\text{diag}(\underline{\mathbf{V}})$ means the diagonal matrix with the vector $\underline{\mathbf{V}}$ as diagonal. In order to estimate this process, and hence the variance of the residual processes, one must substitute the intensity vector $\underline{\boldsymbol{\lambda}}(s) \, ds$ by the estimate derived from the regression model: $\underline{\mathbf{Y}}(s)\underline{\mathbf{Y}}(s)^{-} d\underline{\mathbf{N}}(s)$. Hence the following estimated covariance matrix of the martingale residual processes is suggested:

$$\underline{\mathbf{V}}(t) = \int_0^t J(s) \, (\underline{\mathbf{I}} - \underline{\mathbf{Y}}(s)\underline{\mathbf{Y}}(s)^{-}) \, \text{diag}(\underline{\mathbf{Y}}(s)\underline{\mathbf{Y}}(s)^{-} d\underline{\mathbf{N}}(s))(\underline{\mathbf{I}} - \underline{\mathbf{Y}}(s)\underline{\mathbf{Y}}(s)^{-}).$$

Standardized residual processes may simply be defined by dividing the residual process by its estimated standard deviation at any time t. By asymptotic theory for martingales, normal distributions will appear when a reasonable number of events occur, that is, when t is not too small. Hence, one would expect most standardized residuals to have values between -2 and +2 if the model is true; therefore, plotting the standardized residual processes will give information on model fit. A more formal test of the martingale property could also be constructed by the above theory, but so far we have not gone into this. Furthermore, by applying kernel estimation to $\underline{\mathbf{M}}_{\text{res}}(t)$ and $\underline{\mathbf{V}}(t)$, one may estimate standardized residuals with a local interpretation. Note that the residuals presented here are related to a robust variance estimator for the additive model suggested by Scheike (2002).

Also the hat matrix may be of use. This is an important quantity in ordinary

linear regression. It is here defined as $\underline{\mathbf{Y}}(s)\underline{\mathbf{Y}}(s)^-$, and the diagonal is of interest as a measure of influence. It may be quite informative to plot the diagnonal elements, $h_{jj,\mathrm{cum}}(t)$, of the cumulative hat matrix:

$$H_{\mathrm{cum}}(t) = \sum_{T_i \leq t} \underline{\mathbf{Y}}(T_i)\underline{\mathbf{Y}}(T_i)^- \tag{2}$$

where T_i are the successive times when jumps occur in any process. The idea is to look for processes with particularly high values.

From ordinary linear regression (see e.g. Wetherill, 1986) it is well known that the average value of the diagonal elements of a hat matrix is k/n where n is the number of observations and k is the rank of the hat matrix. In fact, k coincides with the sum of the diagonal elements of the hat matrix. Elements above k/n are said to have high leverage, and it is common to select points for investigation if a diagonal element of the hat matrix is greater than $2k/n$. In our case we can apply this theory at every jump time, and so the criterion for an outlying processes will be when

$$h_{jj,\mathrm{cum}}(t) > 2 \sum_{T_i \leq t} \frac{\mathrm{rank}(\underline{\mathbf{Y}}(T_i)\underline{\mathbf{Y}}(T_i)^-)}{n_i}$$

where n_i is the number of processes at risk at time T_i.

An application of the hat matrix is presented below under the analysis of sleep patterns.

2.3 Relationship to frailty theory

Although we shall not confine ourselves to models defined in frailty terms, it is of interest to see that there is some connection between frailty modelling and the additive regression model.

Assume that a simple frailty structure is valid for the individual counting pro-

cesses, i.e. the individual intensity given the frailty variable Z_i can be written:

$$\gamma_i(t) = Z_i\,\gamma(t)$$

where $\gamma(t)$ is a common baseline intensity and a fixed function, i.e. independent of the past, while the Z_i, $i = 1, ..., n$, are independent identically distributed variables giving the multiplicative factor that determines the risk of an individual. What is the observable intensity of an individual if the Z_i are unknown (which one would usually assume)? For the case that the Z_i are gamma distributed with scale parameter η and shape parameter ν, the answer may be found in Andersen et al (1993, p. 667):

$$\lambda_i(t) = (\nu + N_i(t-))\frac{\gamma(t)}{\eta + G(t)} \tag{3}$$

where $G(t) = \int_0^t \gamma(s)\,ds$. Note that the previous number of events, $N_i(t-)$, comes into the intensity in an *additive* fashion. Hence this gives some motivation for concentrating on additive models.

The right hand side of equation (3) may be viewed as having two covariates, namely the constant covariate 1, and the dynamic covariate $N_i(t-)$. The regression functions of these covariates are:

$$\alpha_1(t) = \nu\,\frac{\gamma(t)}{\eta + G(t)}, \qquad \alpha_2(t) = \frac{\gamma(t)}{\eta + G(t)}$$

Usually, one would want to include other covariates also. From equation (3) it is clear that if an additive covariate structure is put on the shape parameter ν of the frailty distribution, then the total model is still additive.

2.4 Why additivity?

Additivity produces exact martingale estimators and tests. Notice that additivity as such is not essential, but rather the linear structure of the model, i.e. interaction terms might be included. In the usual counting process approach to survival

analysis the exact martingale structure plays a fundamental role in two-sample (and k-sample) tests, but is then, surprisingly, dispensed with as soon as more than one covariate is considered. Then common models, like proportional hazards, preserve at most an approximate martingale structure. Testing for the effect of a covariate within such a framework produces tests where the effect is "smeared" out over an interval and where changes in the effect of a covariate is not rapidly picked up. In the present paper we argue for tests and estimators with a true martingale structure. The martingale structure is not only important for technical reasons, but is also intimately connected to measuring *local* effects of covariates. For instance, changes in effects of the covariates may be seen immediately. The weakness of additivity is that the hazard should be nonnegative, and this is not guaranteed by an additive model. However, in our opinion this is weighed up by the advantages of our model, especially for more complex datasets like the ones considered here. As we present it here the additive model is a pragmatic tool without deep theoretical justification, which in fact is the case for most models in statistics.

3 Dynamic covariates

An example of a dynamic covariate is the $N_i(t-)$ in Section 2.3, which is continuously updated as time goes by. Dynamic covariates should sum up important aspects of the previous development of the process that may contain prognostic information. Examples of such covariates can be

- *time since last event.* This could be a check of the Markov property. If the process is a Markov chain there should be no dependence on time since last event.

- *number of previous events in the process.* This could be seen as a check of frailty. In case of frailty effects one should expect that an excessive number of previous events would predict a greater intensity of events also in the future

- *estimated cumulative hazard.* If in the individual counting process there are several units at risk (like in the amalgam filling data referred to above), then the number of previous events is not appropriate. One has to consider also the number of units at risk at any time in a given counting process. A reasonable dynamic covariate in this case would be the estimated cumulative hazard of the event, estimated by a Nelson-Aalen estimator for each individual process. For instance, in the amalgam filling data, one could for each patient compute the Nelson-Aalen estimate of failure of amalgam fillings and use this as a dynamic covariate.

Of course, there are numerous other ways of constructing dynamic covariates, and it must be evaluated in any given case what are the most reasonable choices. One problem with dynamic covariates, is that they are not of any use before some occurrences have taken place, hence they cannot be used from the very beginning. Practically, the problem may be handled either by ridge regression, or by not starting estimations before a few events have occurred. There is nothing incorrect in this as long as one starts at an optional stopping time.

The martingale structure underlying the theory of the additive model implies, just as for the Cox model, that dynamic covariates do not have to be treated any differently than other covariates with regard to estimation procedures. The covariate functions can be arbitrary predictable processes (apart from regularity conditions, of course). The usefulness of the present approach can only be demonstrated by means of examples, as will be done in Section 4 to 7.

Dynamic covariates have been used by previous authors, see Kalbfleisch and Prentice (2002, Chapter 9) for references. Dynamic covariates are time-dependent covariates, and it is well known that one has to be careful with the handling of such covariates jointly with fixed covariates, see Kalbfleisch and Prentice (2002, p. 199). Dynamic covariates are "responsive" in the terminology of Kalbfleisch and Prentice. Their values may, for instance, be influenced by treatment assignment, and in the statistical analysis dynamic covariates may "steal" from the effect of treatment and other fixed covariates and weaken their effect. A solution for handling this treated in detail in Fosen et al (2003). The following solution appears to work in many cases: First one should carry out a marginal analysis with only the fixed effects included as covariates. The estimation goes as usual, but since important dynamic covariates are excluded, the martingale properties of the counting process theory are not valid. However, we have a rate model in the sense of Scheike (2002), and he gives a correct variance estimate for the marginal model. In the next step one should carry out a full analysis with all fixed and dynamic covariates. This will give correct estimates for the dynamic covariates and the martingale properties are valid if the residuals indicate a good fit as discussed above.

4 Simulation

4.1 Illustrating the connection to frailty models

We shall start by presenting a simulated example. This may serve the purpose of showing that the analysis gives sensible answers for a known model. We simulate a number, k, of independent Poisson processes where the rate in each process is simulated from an exponential distribution with expectation 2. Hence the rates

differ between processes, and so the rate serves as a frailty variable. For each of the k counting processes we define a covariate to be the number of previous events in the process divided by time elapsed, hence the covariate is dynamic, that is changing over time as a function of the past. The mean has been subtracted from the covariate at every time. This means that the cumulative baseline intensity is an estimate of the cumulative rate of the "average" process.

In the simulation we use $k = 40$. The results of the analysis is shown in Figure 1. The cumulative baseline intensity is close to a straight line, which would be expected since the underlying processes are homogeneous Poisson processes. The figure show a very clearly significant effect of the dynamic covariate, and hence demonstrates the frailty effect. Standardized residual processes are also shown in Figure 1. When the dynamic covariate is incorporated, the residuals are mostly confined between -2 and +2, which indicates a good model fit. Without the dynamic covariate the residuals are spread much more out.

4.2 Validity of asymptotic theory

In the cumulative regression plots we present pointwise confidence intervals for the curves. These intervals are based on asymptotic theory, and since the numbers of individual processes in the examples are not very large, one may wonder about the validity of the asymptotic results. A simulation has been carried out to illuminate this issue. The intensity is defined as follows:

$$\lambda_i(t) = \alpha_0 + \alpha_1 \frac{N_i(t-)}{t} \tag{4}$$

The integrated regression functions $\widehat{A}_0(t)$ and $\widehat{A}_1(t)$ have been estimated in each simulation, and it has been evaluated every time whether the 95% confidence intervals cover the true function at three specific time points. The percentage of intervals

covering is denoted the coverage. Ideally it should be 95%. Four different numbers of processes have been studied, i.e. 10, 25, 50 and 100 individual processes, and in each case 10 000 simulations have been carried out. The results are shown in Table 1. One sees that the coverage is good except for the early times when the number of events are small. Even with as small a number as 10 individual processes, the coverage becomes good when time is not too small.

5 Small bowel motility

Data The details, together with a statistical analysis differing from the present one, are given in Aalen and Husebye (1991). We here study a more extensive dataset, with data from 34 individuals. Pressures in the small bowel were recorded continuously from 5.45 pm to 7.25 am the following day. At 6 pm, a standardized mixed meal of 1700 kJ was given to each individual. This induced what is termed postprandial state, characterised by irregular contractions, lasting from 2 to 8 hours. The postprandial state is followed by a fasting state, during which a cyclic motility pattern occurs. Three phases of this may be defined (phases I-III); however, only the activity front (phase III), which is easy to distinguish, is needed for its recognition. Phase III therefore defines the fasting cycle, also called the migrating motor complex (MMC). The time interval between two phase IIIs is termed a MMC period. The start of fasting motility was defined by the first phase III occuring after the evening meal. Several MMC periods occurred in each individual (mean number 4.2) with a censored MMC period terminating the records. Censoring was due to the termination of measurement at 7.25 am. An example of data for an individual is given as follows: 112, 145, 39, 52, 21, 34, 33, 51, 54*. This means that the first

MMC period lasted 112 minutes, the second 145 minutes and so forth. The last period of 54 minutes is censored since observation is terminated as described above.

The great majority of MMC periods occurred after midnight, because the duration of the postprandial period was usually 4-6 hours. Therefore, we will here only consider the time period from midnight until 6.30.

Statistical analysis For each individual there is a counting process running in clock time starting with the first phase III and counting the later phase IIIs occuring. The intensity of the occurrence of a phase III shall be analysed, with covariates as follows.

- Covariate 1: This is a time-dependent covariate counting the number of previous phase IIIs for the individual. The intention is to decide whether there is dependence between the interevent times for an individual.

- Covariate 2: This is another time-dependent covariate measuring time since the last occurrence of a phase III. The object is to check whether the process is Markovian. The covariate is dichotomised to be smaller (or equal to) or larger than 50 minutes.

To estimate the influence of the covariates in a meaningful way, it is clear that some events must already have occurred since the covariates are defined relative to previous events. Here we decided to start estimation at midnight when already a few events had occurred.

The results of the analysis are shown in Figure 2. The cumulative baseline intensity appears to be approximately a straight line, indicating a constant intensity of new phase IIIs. The influence of the number of previous events is seen to be virtually nil (using the test of Aalen (1989) yields the normalized test statistic -0.33

based on observations from midnight to 6.30). This fits with a frailty analysis in Aalen and Husebye (1991) which indicates that there is very little variation between individuals as regards the occurrence of phase IIIs. In practice this means that the intraindividual variation dominates. The advantage of the present analysis is that one may get a picture of whether this is a constant phenomenon over time, as it appears to be. Finally, one sees from the figure that time since the last previous event for the individual in question has a strong effect on the intensity of a new event, the longer the time the more likely is a new event (normalized test statistic is 6.15, $p<0.001$, from midnight to 6.30). This shows that the process is non-Markovian, a conclusion that again fits well with results of Aalen and Husebye (1991) where the interevent time for each individual is estimated to have an increasing Weibull hazard. Again, an advantage of the present procedure is that one can see whether this changes with time, which does not appear to be the case here. Compared with the frailty analysis of Aalen and Husebye (1991), we find that this further analysis gives additional information, especially on whether effects change over time.

A Cox analysis with the same covariates has been carried out, giving for covariate 1 the coefficient 0.00 (s.e. 0.11) and for covariate 2 the coefficient 1.77 (s.e. 0.29, hazard ratio 5.88). The results are seen to be closely compatible with those of the additive analysis.

6 Analysis of sleep patterns

Yassouridis et. al (1999) describe an experiment where a number of people have been observed during one night. Every 30 seconds their sleep status have been registered. In addition the cortisol level was measured every 20 minutes. We have analysed a

set consisting of 27 individuals. Following Yassouridis et. al, we define the time for each individual as time since first time asleep.

An analysis of the data using a multiplicative hazards model is described by Yassouridis et. al (1999) and by Fahrmeir and Klinger (1998). Here, we do not perform an extensive analysis of the data, but just use them for illustrative purposes to indicate the potential of our approach.

In our analysis we have confined ourselves to the transitions from the state "asleep" to the state "awake". Our counting process of interest is thus the process which counts the cumulative number of transitions of this kind after the first time the individual falls asleep. The number at risk at each time point is the number of persons being asleep just prior to this time. We here estimate the regression functions by kernel smoothing techniques. For this purpose we have used the methods suggested in Aalen (1993) and Keiding & Andersen (1989). We have used ridge regression with parameter 0.001, and smoothing bandwidth of 1.67 hours.

The following time-varying covariates have been used:

- covariate 1: logarithm of cortisol level
- covariate 2: cumulative number of times awake, divided by elapsed time
- covariate 3: logarithm of time since last awake

The two latter covariates are dynamic covariates. The smoothed regression functions (not cumulative functions) are shown in Figure 3. The figure shows that cortisol has a positive effect (i.e. increasing the likelihood of waking up) during the later part of the night. The number of previous times awake also has a positive effect on the hazard of waking up during most of the night. The length of the current sleeping period has a negative effect, the longer it has lasted, the less likely is it for

the individual to wake up. This effect seems to be most pronounced early in the night.

To see whether any individual has a large influence on the results we plot the diagonal elements of the cumulative hat matrix in equation (2). In Figure 4, we see that one of the 27 individual cumulative hat processes exceeds the outlying process criterion, meaning that this observation is the one having the highest influence on the analysis. When looking more closely at the individual data, one sees that the person has unusually many awakenings early in the night.

7 Discussion: Interpretation of a dynamic model. Causality?

When finding dynamic effects in a data set, one may ask what is the natural interpretation of such effects. As pointed out above, the dynamic effects may simply reflect unobserved underlying variables, expressed for instance in a frailty model. Alternatively, they might represent real causal effects of past events.

This dilemma has been recognized for a long time. For instance, Feller (1971, p. 57-58), discusses the phenomenon of spurious contagion, pointing out that the Polya process may either be defined as a process of contagion, or as a mixture of non-contagious Poisson processes. He states: "We have thus the curious fact that a good fit of the same distribution may be interpreted in two ways diametrically opposite in their nature as well as in their practical implications."

Let us consider the possible interpretations of a dynamic intensity like that in equation (3). This may be a mechanistic model for individuals if it is actually the case that the number of previous occurrences through some mechanism influences

the likelihood of new occurrences. If, on the other hand, frailty is the true model, then (3) does not give a mechanistic explanation, but is nevertheless useful for individual prediction. Prediction may be carried out, as is often done in statistics, without necessarily knowing the underlying mechanisms.

It is of interest to note that in economics one has considered a similar problem. For instance, the year 2000 Nobel prize winner James J. Heckman, in his Nobel lecture (Heckman, 2000, p.287) discusses the problem of distuingishing "heterogeneity and state dependence", which is similar to what we have discussed above. Interestingly, Heckman asserts that it is possible in certain situations to make such a distinction statistically. However, this needs certain assumptions, otherwise one cannot know whether what one observes is due to individual heterogeneity or causal dynamic effects.

One should also note that in some situations one may have natural replications which could help the situation considerably. In the sleep example, one could make observations over several nights, and this would clearly give a better possibility of estimating the natural sleep pattern of each individual, and hence of distinguishing heterogeneity from state-dependent effects.

The type of causality which is discussed here is related to what is often termed "predictive causality". In fact, our approach constitutes one way of analyzing so-called "local dependence", which is again closely related to causality, see Pötter and Blossfeld (2001).

In spite of the mentioned difficulties in interpretation, a dynamic analysis along the lines suggested here may yield considerable insight into the nature of the data. The additive analysis is simple to carry out in practice. Since no likelihood is needed, the analysis does not require a particularly structured setting. Whenever a number

of events are observed over time one may introduce various information about the past. The method is a pragmatic way of using the information available at any given time. It is easier to carry out than alternatives, like frailty models, which may be difficult to fit.

In conclusion, we believe that the set of methods hitherto available to analyze complex event history data is far too limited, and that there is certainly a need of new methodology. The present approach is intended as a contribution to this.

Acknowledgement

We thank the Max-Planck-Institute for Psychiatry in Munich, Germany, for making available the data on sleep patterns. We also thank Nina Gunnes and Hege Leite Størvold for help with simulations.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics* **6**, 701-726.

Aalen, O. O. (1988). Dynamic description of a Markov chain with random time scale. *Mathematical Scientist* **13**, 90-103

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8**, 907-925.

Aalen, O. O. (1993). Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine* **12**, 1569-1588.

Aalen, O. O., Bjertness, E. and Sønju, T. (1995). Analysis of dependent survival data applied to lifetimes of amalgam fillings. *Statistics in Medicine* **14**, 1819-1829.

Aalen, O. O., Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* **10**, 1227-40.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 1100-1120.

Cox, D. R. (1972a). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Cox, D. R. (1972b). The statistical analysis of dependencies in point processes. In: P.A.W.Lewis, *Stochastic point processes: Statistical analysis, theory and applications*, Wiley, pp. 55-66.

Fahrmeir, L. and Klinger, A. (1998). A nonparametric multiplicative hazard model for event history analysis. *Biometrika* **85**, 581-592.

Feller, W. (1971). *An introduction to probability theory and its applications,* Vol. II, Wiley, New York.

Fosen, J., Aalen, O. O., Borgan, Ø. and Fekjaer, H. (2003). Separating fixed and dynamic covariates in survival analysis with repeated events. Unpublished manuscript.

Gandy, A. and Jensen, U. (2004). A Nonparametric Approach to Software Reliability, *Applied Stochastic Models in Business and Industry*, **20**, 3-15.

Heckman, J. J. (2000). Microdata, heterogeneity and the evaluation of public policy. Nobel prize lecture. http://www.nobel.se/economics/laureates/2000/

Hougaard, P. (2000). *Analysis of Multivariate Survival Data,* Springer-Verlag, New York

Husebye, E., Skar, V., Aalen, O. O., Osnes M. (1990) Digital ambulatory manom-

etry of the small intestine in healthy adults: Estimates of the variation within and between individuals and statistical management of incomplete MMC periods. *Digestive Diseases and Sciences* **35**, 1057-65.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* Wiley, New Jersey.

Keiding, N. and Andersen, P. K. (1989). Nonparametric estimation of transition intensities and transition probabilities: A case study of a two-state Markov process, *Applied Statistics* **38**, 319-329.

Martinussen, T. and Scheike, T. H. (2000). A nonparametric dynamic additive regression model for longitudinal data, *The Annals of Statistics* **28**, 1000-1025.

Peña, E. A. and Hollander, M. (2004). Models for Recurrent Phenomena in Survival Analysis and Reliability. In: *Mathematical Reliability: An Expository Perspective*, edited by T. Mazzuchi, N. Singpurwalla and R. Soyer, pp. 105-123, Kluwer.

Pötter, U. and Blossfeld, H.-P. (2001). Causal inference from series of events. *European Sociological Review* **17**, 21-32.

Scheike, T. H. (2002). The additive nonparametric and semiparametric Aalen model as the rate function of a counting process. *Lifetime Data Analysis* **8**, 247–262.

Wetherill, G. B. (1986). *Regression analysis with applications.* Chapman and Hall, London.

Yassouridis, A., Steiger, A., Klinger, A. and Fahrmeir, L. (1999). Modelling and exploring human sleep with event history analysis. *Journal of Sleep Research* **8**, 25-36.

Table caption

Table 1: Coverage of simulated confidence intervals. 10 000 simulations in each line.

Figure captions

Figure 1: Simulated example: Upper panels show cumulative baseline intensity (left) and cumulative regression function (right). Outer curves give pointwise 95% confidence intervals. Lower panels show cumulative residual processes for simulated example. Left panel: No covariate. Right panel: Dynamic covariate.

Figure 2: Occurrence of phase III events in small bowel motility: Cumulative baseline intensity (upper panel), cumulative regression function of covariate measuring previous number of phase III events (lower left panel) and cumulative regression function of covariate measuring time since last phase III event (lower right panel). Pointwise 95% confidence limits. The time axis goes from midnight until 6.30.

Figure 3: Sleep data: Smoothed baseline hazard function and regression functions together with pointwise 95% confidence limits. Bandwidth: 1.67 hour.

Figure 4: Sleep data: The 27 individual cumulative hat processes (diagonal hat matrices) as a function of the time intervals where the events happen, together with the expected cumulative hat matrix (the lower thick solid line) and the outlying process criterion (the upper thick solid line).

	Time (t)		
Number of individual processes	$t = 0.25$	$t = 0.50$	$t = 1.00$
n=10	0.85	0.94	0.96
n=25	0.92	0.95	0.96
n=50	0.93	0.95	0.95
n=100	0.94	0.95	0.95

Table 1: Coverage of simulated confidence intervals for $A_1(t)$; 10000 simulations in each line.

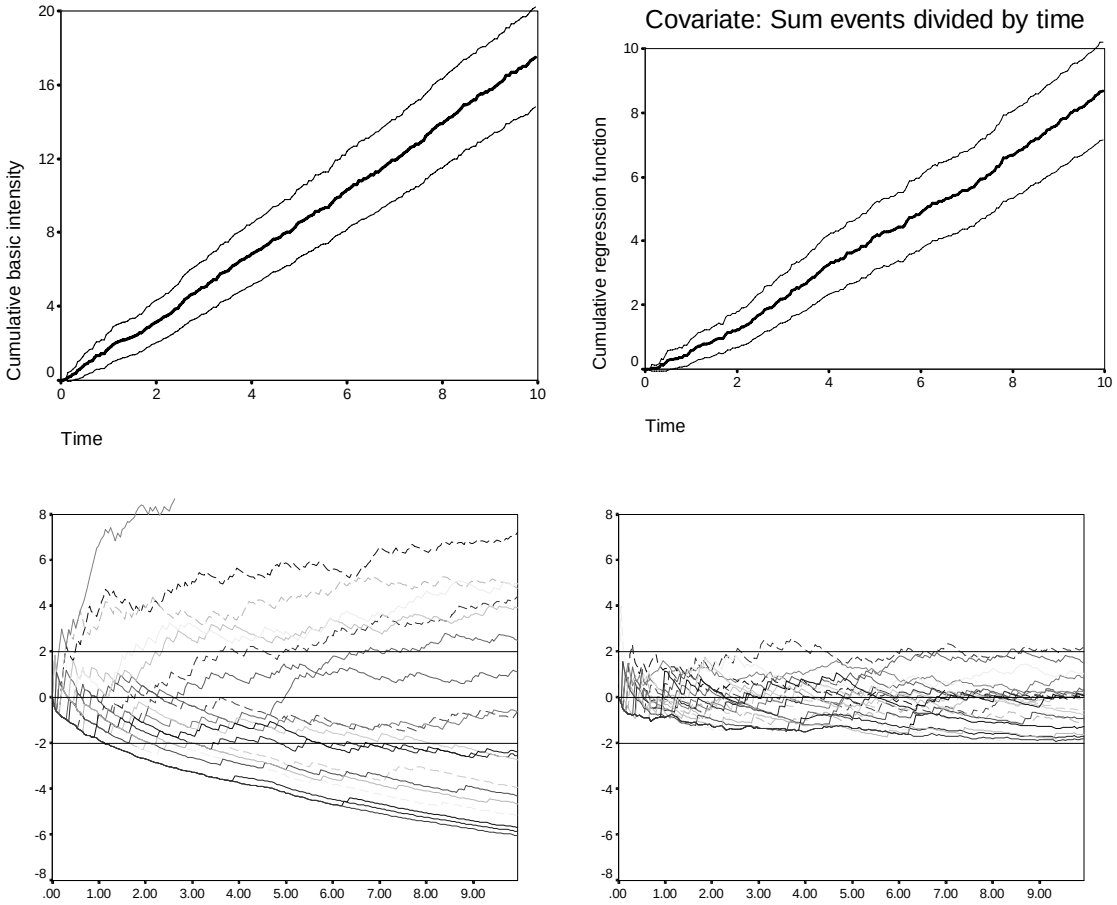

Figure 1: *Simulated example: Upper panels show cumulative baseline intensity (left) and cumulative regression function (right). Outer curves give pointwise 95% confidence intervals. Lower panels show cumulative residual processes for simulated example. Left panel: No covariate. Right panel: Dynamic covariate.*

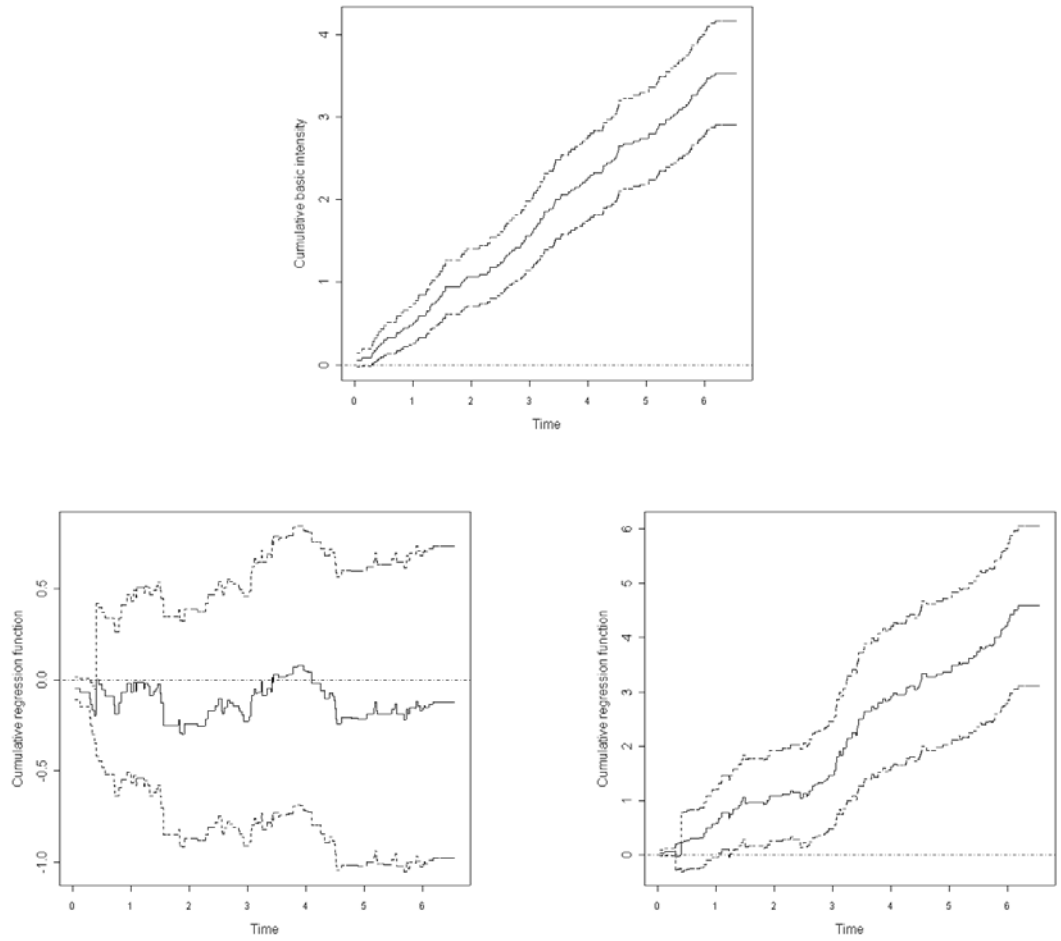

Figure 2: *Occurrence of phase III events in small bowel motility: Cumulative baseline intensity (upper panel), cumulative regression function of covariate measuring previous number of phase III events (lower left panel) and cumulative regression function of covariate measuring time since last phase III event (lower right panel). Pointwise 95% confidence limits. The time axis goes from midnight until 6.30.*

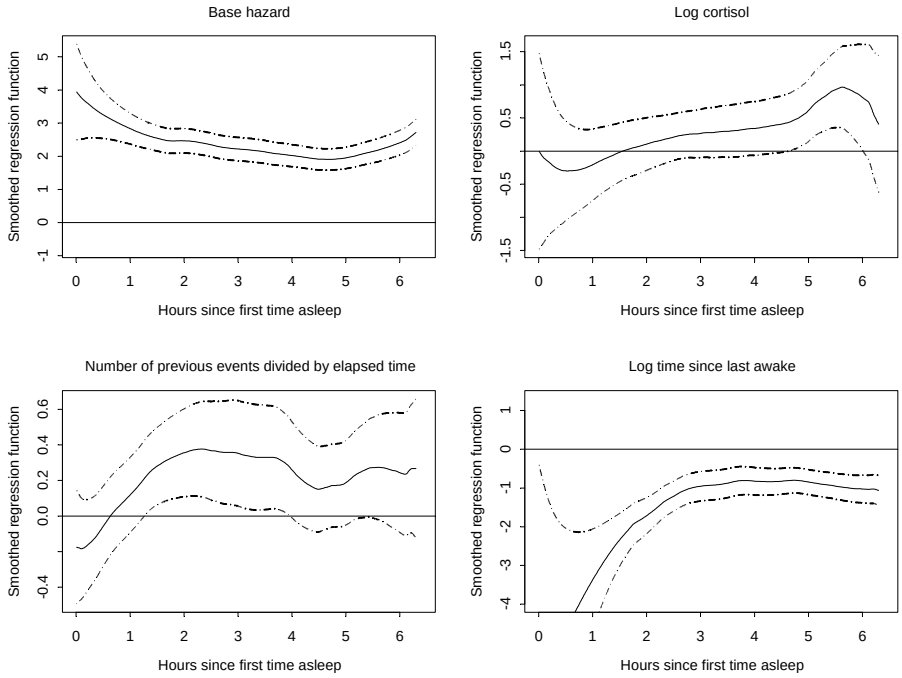

Figure 3: *Sleep data: Smoothed baseline hazard function and smoothed regression functions together with pointwise 95% confidence limits. Smoothing parameter (bandwidth) is 1.67 hour.*

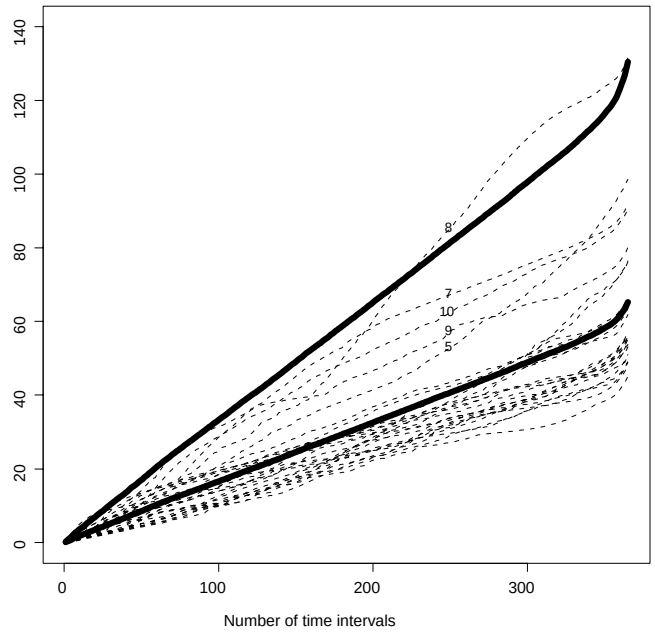

Figure 4: *Sleep data: The 27 individual cumulative hat processes (diagonal* elements of (2)) as a function of the number of half minute intervals with events, together with their average value (the lower thick solid line) and the outlying process criterion (the upper thick solid line).

.