Survival models and health sequences

Peter McCullagh*

July 3, 2014

Abstract

Survival studies often generate not only a survival time for each patient but also a sequence of health measurements at annual or semi-annual check-ups while the patient remains alive. Such a sequence of random length accompanied by a survival time is called a survival process. Ordinarily robust health is associated with longer survival, so the two parts of a survival process cannot be assumed independent. This paper is concerned with a general technique—reverse alignment—for constructing statistical models for survival processes. A revival model is a regression model in the sense that it incorporates covariate and treatment effects into both the distribution of survival times and the joint distribution of health outcomes. The revival model also determines a conditional survival distribution given the observed history, which describes how the subsequent survival distribution is determined by the observed progression of health outcomes.

Keywords: interference; preferential sampling; quality-of-life; revival process; semi-revival time; reverse alignment; stale values;

1 Survival studies

A survival study in one in which patients are recruited according to well-defined selection criteria and their health status monitored on a regular or intermittent schedule until the terminal event, here assumed to be fatal. Covariates such as sex and age are recorded at the time of recruitment, and, if there is more than one treatment level, the assignment is presumed to be randomized. In a simple survival study, the health status $Y(t)$ at time t is a bare-bones binary variable, dead or alive, and the entire process is then summarized by the time $T > 0$ spent in state 1, i.e. the survival time. In a survival study with health monitoring, $Y(t)$ is a more detailed description of the state of health or quality of life of the individual, containing whatever information—pulse rate, cholesterol level, cognitive score or CD4 cell count—is deemed relevant to the study. The goal may be to study the effect of treatment on survival time, or to study its effect on quality of life, or to predict the subsequent survival time of patients given their current health history.

*Department of Statistics, University of Chicago, 5734 University Ave, Chicago, Il 60637, U.S.A. E-mail: pmcc@galton.uchicago.edu

Survival studies with intermittent health monitoring are moderately common, and likely to become more so as health records become available electronically for research purposes. Within the past few years, several issues of the journal *Lifetime Data Analysis* have been devoted to problems connected with studies of exactly this type. For a good introduction, with examples and a discussion of scientific objectives, see Diggle, Sousa and Chetwynd (2006), Kurland, Johnson, Egleston and Diehr (2009) or Farewell and Henderson (2010). Section 8 of van Houwelingen and Putter (2012) is recommended reading.

In practice, the patient's health status is measured at recruitment ($t = 0$), and regularly or intermittently thereafter while the patient remains alive. To emphasize the distinction between the observation times and observation values, each time is called an appointment date, the set of dates is called the appointment schedule. Apart from covariate and treatment values, a complete uncensored observation on one patient $(T, \mathbf{t}, Y[\mathbf{t}])$ consists of the survival time $T > 0$, the appointment schedule $\mathbf{t} \subset [0, T)$, and the health status measurements $Y[\mathbf{t}]$ at these times. To accommodate patients whose record is incomplete, a censoring indicator variable is also included. In that case, the censoring time is usually, but not necessarily, equal to the date of the most recent appointment.

In the sense that the health status is measured over time on each patient, a survival study is a particular sort of longitudinal study. Certainly, temporal and other correlations are expected and must be accommodated. But the distinguishing feature, that each sequence is terminated by failure or censoring, gives survival-process models a very distinct character: as an absorbing state, death contradicts stationarity. For a good survey of the goals of such studies and the modeling strategies employed, see Kurland, Johnson, Egleston and Diehr (2009).

The goal of this paper is not so much to recommend a particular statistical model, as to explore a general mathematical framework for the construction of survival-process models, permitting easy computation of the likelihood function and parameter estimates, and straightforward derivation of predictive distributions for individual survival times. For example, the paper has nothing to say on the choice between proportional hazards and accelerated lifetimes for accommodating treatment effects. Apart from reservations concerning the use of time-evolving covariates, all standard survival models are acceptable within the framework. Nor has the paper anything to contribute to the choice between Bayesian and non-Bayesian methods of analysis; prior distributions are not discussed, so either approach can be used. Administrative complications of the sort that are inevitable in medical and epidemiological research will be ignored for the most part, so no attempt is made to provide a complete turnkey package. For example, the paper has little to say about how best to handle incomplete records other than to recognize that censoring and delayed reporting are issues that must be addressed—again using standard well-developed methods. Since most of the computations needed for model fitting and parameter estimation are relatively standard and need not involve specialized Markov chain or Monte Carlo algorithms, detailed discussion of computational techniques is omitted. The emphasis is on statistical ideas and principles, strategies for model formulation, sampling schemes, and the distinction between time-dependent variables and time-evolving variables in the definition of treatment effects.

2 Joint modelling

The medical and biostatistical literature contains numerous examples of studies involving both successive measurements on each patient, such as CD4 lymphocyte cell counts, together with survival time (Lagakos, 1976; DeGruttola and Tu, 1994; Faucett and Thomas 1996; Guo and Carlin, 2004; Fieuws, Verbeke, Maes and Vanrenterghem, 2008). Geriatric studies seldom focus exclusively on survival time, but tend to emphasize variables related to quality of life, such as overall physical and mental health, mobility, independence, memory loss, mental acuity, and so on. In the statistical literature, survival studies with health monitoring are called longitudinal studies with time-to-event data (Wulfsohn and Tsiatis, 1997; Henderson, Diggle and Dobson, 2000; Xu and Zeger, 2001; Tsiatis and Davidian, 2004; Rizopoulos, 2012). Although there are variations in model formulation and implementation, all authors are agreed on the need for a joint distribution covering both survival time and the progression of health outcomes.

The joint modelling approach begins with a pair (η, T) consisting of a latent recurrent temporal process $\eta(t)$ together with a positive random variable $T < \infty$. This joint distribution determines the distribution of the observable process

$$Y(t) = \begin{cases} \eta(t) & t < T \\ \flat & t \geq T \end{cases}$$

by restriction to $(-\infty, T)$, or censoring at T. The process constructed in this way has the same domain $t \in \Re$ as the unobservable recurrent process, but the state space includes an additional absorbing value, here labelled $\flat$. Equivalently, but slightly less conveniently, Y may be defined as a process on the random domain $(-\infty, T)$ or $(0, T)$. In either case the survival time T is a function of Y.

In practice, the recurrent process usually has a fairly simple form such as a stationary Gaussian process, and T is determined by a Cox model whose intensity is a latent process not independent of η. The values (η_i, T_i) are assumed independent for distinct individuals, and identically distributed for individuals sharing the same covariate values. Regardless of how the survival process is constructed, the joint distribution is determined by its finite-dimensional restrictions. In other words, we need to specify for each finite subset $\mathbf{s} \subset [0, \infty)$, the joint distribution of $Y[\mathbf{s}]$, or the joint distribution of $(T, Y[\mathbf{s}])$, or, equivalently, the joint distribution of $(T, Y[\mathbf{t}])$, where $\mathbf{t} = \mathbf{s} \cap [0, T)$. For a more detailed account of specific processes, see chapters 13–16 of Fitzmaurice, Davidian, Verbeke and Molenberghs (2009).

In order to avoid some of the technical difficulties associated with joint modelling the suggestion put forward in this paper is to approach the problem from a different angle—literally in reverse. Reverse alignment avoids the intermediate recurrent process by considering probability distributions for the non-recurrent observable process Y directly. It is discussed as one of several options in Table 2 of Kurland, Johnson, Egleston and Diehr (2009), and is mentioned in section 8.3 of van Houwelingen and Putter (2012), so the idea is not new. This paper explores the probabilistic and statistical implications of time-reversal, with a focus on exchangeability and distributional factorization for the entire survival process.

The implications for sampling, the consequences for survival prediction, and the interpretation and estimation of treatment effects are also considered.

3 Reverse alignment

3.1 The survival process

A survival process Y is a stochastic process defined for real t, in which $Y_i(t)$ is the state of health or quality of life of patient i at time t, usually measured from recruitment. In a simple survival process, the state space $\mathcal{R} = \{0, 1\}$ is sufficient to encode only the most basic of vital signs, dead or alive; more generally, the state space is any set large enough to encode the observable state of health or quality of life of the patient at one instant in time. Flatlining is the distinguishing characteristic of a survival process, i.e. $\flat \in \mathcal{R}$ is an absorbing state such that $Y(t) = \flat$ implies $Y(t') = \flat$ for all $t' \geq t$. The survival time is the time to failure:

$$T_i = \sup_{t \geq 0} \{t : Y_i(t) \neq \flat\};$$

it is presumed that $Y_i(0) \neq \flat$ at recruitment, so $T_i > 0$. This definition is quite general, and does not exclude immortality, i.e. $T = \infty$ with positive probability. In all of the models considered here, however, survival time is finite with probability one.

3.2 Administrative and other schedules

Since the appointment schedule is a random subset $\mathbf{t} \subset [0, T)$, it is obviously informative for survival: $T > \max(\mathbf{t})$. If better health is associated with longer survival, we should expect patients who are initially frail to have shorter health records than patients who are initially healthy. In other words, even if the trajectories for distinct individuals may be identically distributed, the first component of a short health-status record should not be expected to have the same distribution as the first component of a longer record. On the contrary, any model such that record length is independent of record values must be regarded as highly dubious for survival studies. It is necessary, therefore, to address the nature of the information contained in $\mathbf{t}$.

Consider a patient who has had appointments on k occasions $\mathbf{t}^{(k)} = (t_0 < \cdots < t_{k-1})$. The sequence $Y[\mathbf{t}^{(k)}]$ of recorded health values may affect the scheduled date t_k for the next appointment: for example, patients in poor health needing more careful monitoring may have short inter-appointment intervals. Whatever the scheduled date may be, the appointment is null unless $t_k < T$. The assumption used in this paper is sequential conditional independence, namely that

$$t_k \perp\!\!\!\perp Y \mid (T, \mathbf{t}^{(k)}, Y[\mathbf{t}^{(k)}]). \tag{1}$$

In other words, the conditional distribution of the random interval $t_k - t_{k-1}$ may depend on the observed history $Y[\mathbf{t}^{(k)}]$, but not on the subsequent health trajectory

except through T. Here, t_k may be infinite (or null) with positive probability, in which case the recorded sequence is terminated at t_{k-1}.

The schedule is said to be *administrative* if t_k is a deterministic function of the pair $(\mathbf{t}^{(k)}, Y[\mathbf{t}^{(k)}])$, implying that the conditional distribution (1) is degenerate. Eventually, for some finite k, the patient dies or is censored at time $T \in (t_{k-1}, t_k)$ while the next appointment is pending, so the recorded schedule is $\mathbf{t} = \mathbf{t}^{(k)} = \mathbf{t}^{(k+1)} \cap [0, T)$. Equivalently, the last recorded value is $(t_k, \flat)$.

While the sequential conditional independence assumption is mathematically clear-cut, the situation in practice may be considerably more muddy. Consider, for example, the CSL1 trial organized by the Copenhagen Study Group for Liver Diseases in the 1960s to study the effect of prednizone on the survival of patients diagnosed with liver cirrhosis. In this instance $Y(\cdot)$ is a composite blood coagulation index called the prothrombin level: details can be found in Andersen, Hansen and Keiding (1991). Beginning at death, the reverse-time mean intervals between appointments are 77, 210 and 251 days, while the medians are 21, 165 and 292 days. In other words, half of the patients who died had their final appointment within the last three weeks of life. It is evident that the appointment intensity increases as $s \to 0$ in reverse time, which is not, in itself, a violation of (1). However, one might surmise that the increased intensity is related to the patient's state of health or perception thereof. Condition (1) implies that the appointment intensity does not depend on the blood coagulation index other than at earlier appointments, and it is then unclear to what extent the condition may be violated by patient-initiated appointments. Liestøl and Andersen (2002, section 4.1) note that 71 off-schedule appointments occurred less than 10 days prior to death, the majority of which were patient-initiated. They also examine the effect on hazard estimates of excluding unscheduled prothrombin measurements.

Although we refer to $Y(\cdot)$ generically as the patient's *state of health,* this description is not to be taken literally. The actual meaning depends on what has in fact been measured: in general, $Y(\cdot)$ is only one component or one aspect of patient health.

3.3 The revival process

On the assumption that the survival time is finite, the time-reversed process

$$Z_i(s) = Y_i(T_i - s)$$

is called the revival process. Thus, $Z_i(s)$ is the state of health of patient i at time s prior to failure, and $Z_i(T_i) = Y_i(0)$ is the value at recruitment. By construction, $Z(s) = \flat$ for $s < 0$, and $Z(s) \neq \flat$ for $s > 0$. Although Z is defined in reverse time, the temporal evolution via the survival process occurs in real time: by definition, $Z(\cdot)$ is not observable component-wise until the patient dies.

The transformation $Y \mapsto (T, Z)$ is clearly invertible; it may appear trivial, and in a sense it is trivial. Its one key property is that the revival process Z and the random variable T are variation independent. In the statistical models considered here, variation independence may be exploited through the revival assumption, which

states that the revival process and the survival time are statistically independent. More generally, $Z \perp\!\!\!\perp T \mid X$ if covariates are present. The revival assumption provides a convenient starting point for model construction; it is not critical to any part of the theory presented here.

Table 1: Average prothrombin levels indexed by T and t.

Survival	Time t after recruitment (yrs)								
time (T)	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8+
0–1	58.0								
1–2	72.5	66.4							
2–3	72.6	73.2	66.0						
3–4	69.8	71.2	68.5	54.2					
4–5	68.5	75.7	72.5	74.6	57.7				
5–6	70.5	77.3	73.5	57.1	64.5	60.9			
6–7	81.8	73.6	81.1	80.6	79.4	75.5	75.8		
7–8	84.4	88.8	88.1	92.1	85.2	81.2	84.3	88.1	
8+	77.3	73.6	87.0	74.1	92.0	80.3	89.2	79.4	84.7

The chief motivation for time reversal has to do with the effective alignment of patient records for comparison and signal extraction. Are the temporal patterns likely to be more similar in two records aligned either by patient age or by recruitment date, or are they likely to be more similar in records aligned by reverse age (time remaining to failure)? Ultimately, the answer must depend on the context, but the context of survival studies suggests that the latter may be more effective than the former. Table 1 shows the averaged Y-values indexed by T and t for the prothrombin example discussed in more detail in section 6. It should be borne in mind that each cell is the average of 8–266 non-independent high-variability measurements, the larger counts occurring in the upper left cells. Alignment by reverse time is equivalent to counting leftwards from the main diagonal. Despite certain anomalies in the table of averages, e.g. row 6, column 4, it is clear that reverse-time is a more effective way of organizing the data to display the main trends in the mean response: the forward- and reverse-time sums of squares (equally weighted) are 543.0 and 1132.8 respectively, both on eight degrees of freedom.

Further confirmation is provided in Table 2, which shows the output from a standard equally-weighted analysis of variance applied to the table of averages, with three factors, row, column and diagonal (reverse time), denoted by R, C and D respectively. Compared with the residual mean square of 23.7, there is considerable excess variation associated with rows (117.0) and with the reverse-time factor (77.8), but not so much with columns (34.1). In other words, the means in Table 1 are expressible approximately as $\alpha_T + \beta_{T-t}$. Figures. 8.3 and 8.4 of van Houwelingen and Putter (2012), which are not substantially different from Fig. 2 of this paper, offer strong confirmation of this viewpoint in one further survival study involving white blood cell counts for patients suffering from chronic myeloid leukemia. For an application unrelated to survival, see example B of Cox and Snell (1981).

Model construction by reverse alignment may appear peculiar and unnatural in

Table 2: ANOVA decomposition for Table 1

$\mathcal{U}/\mathcal{V}$	$\|P_{\mathcal{U}}Y\|^2 - \|P_{\mathcal{V}}Y\|^2$	d.f.	M.S.
$(R+C+D)/(R+C)$	544.4	7	77.8
$(R+C+D)/(R+D)$	238.9	7	34.1
$(R+C+D)/(C+D)$	818.8	7	117.0
$RC/(R+C+D)$	498.3	21	23.7

biological work, where the accepted wisdom is that an effect such as death cannot precede its supposed causes, such as ill health. However, the contrarian viewpoint (Liestøl and Andersen (2002, section 4.0)), that proximity to death is the chief cause of ill health, seems neither less compelling nor more helpful. The author's attitude is that metaphysical discussion along such lines is seldom productive and best avoided.

3.4 Exchangeability

In the presence of covariates such as sex, age at recruitment or treatment status, exchangeability is understood in the sense of McCullagh (2008, section 2) i.e. it applies to each subset of patients having the same covariate value. For any such set of patients, it implies that the survival times $T_{i_1}, \ldots, T_{i_n}$ are identically distributed, the record lengths $\#\mathbf{t}_{i_1}, \ldots, \#\mathbf{t}_{i_n}$ are identically distributed, the health-status variables $Y_{i_1}(t), \ldots, Y_{i_n}(t)$ are identically distributed, and likewise for the revival values at any fixed revival time s. In this paper, therefore, baseline health status is the first component of Y, not a covariate. This is unavoidable for revival models: it is immaterial that $Y(0)$ is measured prior to randomization and treatment assignment. Exchangeability does not imply that $Y_i(0)$ is independent of the record length $\#\mathbf{t}_i$.

Assuming that the record is complete, the observation for one patient consists of a survival time T, a finite appointment schedule $\mathbf{t} \subset [0, T)$, and a sequence of length $\#\mathbf{t}$ taking values in the space of medical records, here denoted by $\mathcal{R}$. For simplicity of notation in what follows, it is assumed that the appointment date is included in $\mathcal{R}$. Then the sample space for the observation $(T, Y[\mathbf{t}])$ on one patient is

$$\mathcal{S} = (0, \infty) \times \bigcup_{k=0}^{\infty} \mathcal{R}^k$$

in which the second component is the space of finite-length $\mathcal{R}$-valued sequences.

For n patients $i_1, \ldots, i_n$, the observations are independent if the joint distribution on $\mathcal{S}^n$ factors in the usual way:

$$P_{i_1,\ldots,i_n}(A_1 \times \cdots \times A_n) = \prod_{j=1}^{n} P_{i_j}(A_j),$$

where each P_i is a probability distribution on $\mathcal{S}$, and $A_1, \ldots, A_n \subset \mathcal{S}$ are arbitrary events. In that circumstance, it is sufficient to describe the marginal distributions P_i on $\mathcal{S}$, which may depend on covariates x_i. The observations on patients are

infinitely exchangeable if $P_{i_1,\ldots,i_n}$ is the marginal distribution of $P_{i_1,\ldots,i_n,i_{n+1}}$, and all joint distributions are unaffected by permutation of patients, who are always assumed to be distinct individuals.

The implications of exchangeability are the same whether the record for each patient is expressed in terms of the survival process or the revival process. It implies that the revival processes are identically distributed. Together with the revival assumption, that Z and T are independent, it implies that $Z_{i_1}(s), \ldots, Z_{i_n}(s)$ are identically distributed independently of the survival times $T_{i_1}, \ldots, T_{i_n}$.

3.5 Covariates

In the absence of specific information to the contrary, responses for distinct units are presumed to be identically distributed. In the great majority of situations, specific information does exist in the form of covariates or classification variables or relationships. A covariate is a function $i \mapsto x_i$ on the units, in principle known for all units whether they occur in the sample or not. A covariate implies a specific form of inhomogeneity such that equality of covariates implies equality of response distributions: $x_i = x_j$ implies $Y_i \sim Y_j$. In practice, approximate equality of x-values also implies approximate equality of distributions. Likewise, a relationship is a function on pairs of units such that $R(i, i') = R(j, j')$ implies $(Y_i, Y_{i'}) \sim (Y_j, Y_{j'})$ for distinct pairs $i \neq i'$, $j \neq j'$, provided that the two pairs also have the same covariate values: $(x_i, x_{i'}) = (x_j, x_{j'})$. Geographic distance and genetic distance are two examples of symmetric relationships. The overarching principle is that differences in distribution, marginal or joint, must be associated with specific inhomogeneities in the experimental material.

The status of certain variables in specific survival studies may appear genuinely unclear. The conventional rationalization, in which certain variables used for prediction are notionally 'fixed' or non-random and treated as covariates, is not especially helpful for survival studies. Consider, for example, marital status as one variable in a geriatric study in which the goal is to study both quality of life and survival time. However it is defined, quality of life is a multi-dimensional response, a combination of mobility, independence, optimism, happiness, family support, and so forth. Marital status is a temporal variable known to be associated with survival and with quality of life; one goal may be to predict survival given marital status, or even to recommend a change of status in an effort to improve the quality of life. Another example of a similar type is air quality and its relation to the frequency and severity of asthmatic attacks (Laird, 1996). Should such a variable be regarded as a covariate or as one component of the response? For survival studies, and for longitudinal studies generally, the answer is very clear and very simple: *every time-evolving variable is necessarily part of the response process.*

By definition, a temporal variable x is a function defined for every $t \geq 0$. A temporal variable is a covariate if it is also a function on the units, meaning that the entire function is determined and recorded at baseline. Usually this means that x is constant in time, but there are exceptions such as patient age: see also section 3.6. Marital status and air quality, however, are not only temporal variables, but variables whose trajectories evolve over real time; neither is available as a

covariate at baseline.

With marital status as a component of the survival process, the joint distribution may be used to predict the survival time beyond t of an individual whose marital history and other health-status measurements at certain times prior to t are given. For that purpose, it is necessary to compute the conditional distribution of T, or more generally of Y, given the observed history $\mathcal{H}_t$ at the finite set of appointments prior to t. For such calculations to make mathematical sense, marital status must be a random variable, a function of the process Y. Thus, the statement 'marital status is a random process' is not to be construed as a sociological statement about the fragility of marriage or the nature of human relations; it is merely a mathematical assertion to the effect that probabilistic prediction is not possible without the requisite mathematical structure of σ-fields $\mathcal{H}_t \subset \mathcal{H}_{t'}$ for $t \leq t'$ and probability distributions.

3.6 Treatment

Treatment refers to a scheduled intervention or series of interventions in which, at certain pre-specified times following recruitment, the prescription for patient i is switched from one arm to another. Thus, $a_i(t)$ is the treatment arm scheduled for patient i at time $t \geq 0$. In general, but crucially for revival models, a null level is needed for $t \leq 0$, including the baseline $t = 0$. The entire temporal trajectory $a_i(t)$ for $t > 0$ is determined by randomization and recorded at baseline. It does not evolve over real time in response to the doctor's orders or the patient's perceived needs, so it is not a time-evolving variable. Ordinarily, the random variables $a_1(\cdot), \ldots, a_n(\cdot)$ are not independent. In the sense that it is recorded at baseline, $a_i(\cdot)$ is a covariate; in the sense that it is a temporal function, it is a time-dependent covariate.

Apart from crossover trials, the distribution of $a(\cdot)$ is such that a switch of treatment arms occurs only once, and then only immediately after recruitment. Nonetheless, more general formulation is retained to underline the fact that treatment is a scheduled intervention such that $a_i(t) \neq a_i(0)$, and thus not constant in time. Unlike the survival process, the treatment schedule does not evolve randomly in real time.

It should be understood that a treatment arm is a protocol, also called a dynamic treatment regime (Murphy, 2003), specifying the drug type, dose level, frequency, manner of ingestion, and even the next appointment date, as a function of the current medical circumstances and health history. A treatment arm is not necessarily associated with a specific drug or a specific medical therapy. Consider a hypertension study in which blood pressure $Y(\cdot)$ in conventional units is measured at regular appointments. Thus, *one blue pill to be taken three times daily while blood pressure exceeds 180, and one white pill twice daily if the pressure is between 160 and 180* is a treatment arm in which the actual therapy and dose level at time t' depend on the outcome $Y(t)$ at the most recent appointment $t \leq t'$. Another treatment arm might reverse the colours or adjust the doses or change the therapy or reduce the inter-appointment interval if preceding Y-values are not encouraging. In this setting, $a_i(t)$ denotes the assigned treatment arm, not the current drug or the dose, so $a_i(\cdot)$ is ordinarily constant for $t > 0$, even for dynamic chemotherapy strategies

such as those discussed by Rosthøj, Keiding and Schmiegelow, (2012). For a compliant patient, the drug type and ingested dose can, in principle, be determined from the treatment arm and previously recorded Y-values. The key point is that each treatment arm be fully specified in advance, and the assignment be randomized at recruitment.

Let $\bar{a}_i(s) = a_i(T_i - s)$ be the treatment arm expressed in revival time, so that, in the standard setting, $\bar{a}_i(s)$ is null for $s \geq T_i$. It is automatic that that $Z \perp\!\!\!\perp T \mid \bar{a}$, because T is a function of $\bar{a}$. In the case of treatment, however, the crucial assumption is lack of interference, i.e. the treatment assigned to one individual has no effect on the response distribution for other individuals, and the treatment protocol at one point in time has no effect on the response distribution at other times. For the latter, the statement is as follows. For each finite subset $\mathbf{s} \subset \Re^+$, the conditional distribution of $Z[\mathbf{s}]$ given the treatment schedule and survival time depends only on the treatment arms $\bar{a}[\mathbf{s}]$ prevailing at the scheduled times, i.e.,

$$Z[\mathbf{s}] \perp\!\!\!\perp \bar{a} \mid \bar{a}[\mathbf{s}].$$

For crossover trials in particular, this is a strong assumption denying carry-over effects from earlier treatments or later treatments. It implies in particular that $Z(s) \perp\!\!\!\perp T \mid \bar{a}(s)$, which is primarily a statement about the one-dimensional marginal distributions. Note, however, that the interference assumption is relatively benign if $a_i(t)$ is constant for $t > 0$, as is ordinarily the case.

It is common practice in epidemiological work for certain time-evolving variables to be handled as covariates, as if the entire trajectory were recorded at baseline. This approach is perfectly reasonable for an external variable such as air quality in an asthma study where lack of cross-temporal interference might be defensible. It has the advantage of leading to simple well-developed procedures for effect estimation using marginal moments (Zeger and Liang, 1986; Zeger, Liang and Albert, 1988; Laird, 1996; Diggle, Heagerty, Liang and Zeger, 2002). The same approach is less convincing for an evolving variable such as marital status in a survival study, because the entire trajectory—suitably coded for $t > T_i$—would often contain enough information to determine the survival time.

4 Survival prediction

4.1 Conditional distribution

Consider the simplest model in which observations for distinct patients are independent and identically distributed. To simplify matters further, problems related to parameter estimation are set aside. In other words, the survival time is distributed according to F, and the revival processes given $T = t$ is distributed as $G(\cdot \mid t)$. Given the joint distribution, we are free to compute whatever conditional or marginal distribution is needed to address the inferential target.

We consider here the question of how the partial trajectory of Y affects the subsequent survival prognosis. However the question is phrased, any suggestion of a causal mechanism is unwarranted: ultimately the answer is determined by

the appropriate conditional distribution. The problem is to predict the survival time of an individual given the survival process $Y[\mathbf{t}^{(k)}]$ at the first k appointments $\mathbf{t}^{(k)} = (t_0 < \cdots < t_{k-1})$.

For positive real numbers $\mathbf{s} = (s_1 > \cdots > s_k)$, let $g_k(z; \mathbf{s} \mid t)$ be the conditional joint density given $T = t$ of the health-status values

$$Z[\mathbf{s}] = (Z(s_1), \ldots, Z(s_k)) = (Y(T - s_1), \ldots, Y(T - s_k)).$$

Under the conditional independence assumption (1), which implies non-preferential appointment dates in the sense of Diggle, Menezes and Su (2010), the joint density of $(T, \mathbf{t}^{(k)}, Y[\mathbf{t}^{(k)}])$ at $(t, \mathbf{t}^{(k)}, y)$ is a product of three factors:

$$\begin{aligned} & f(t) \times \prod_{j<k} p(t_j, y_j \mid \mathcal{H}_j, T = t) \\ &= f(t) \times \prod_{j<k} p(y_j \mid \mathcal{H}_j, T = t) \times \prod_{j<k} p(t_j \mid \mathcal{H}_j, T = t) \\ &= f(t) \times g_k(y; t - \mathbf{t}^{(k)} \mid t) \times \prod_{j<k} p(t_j \mid \mathcal{H}_j, T = t), \end{aligned} \tag{2}$$

where $f = F'$ is the survival density, and $\mathcal{H}_j$ is the observed history $(\mathbf{t}^{(j)}, Y[\mathbf{t}^{(j)}])$ at time t_{j-1}. Without further assumptions, all three factors depend on t, meaning that all three components are informative for survival prediction.

In all subsequent discussion concerning prediction, it is assumed that the appointment schedule is uninformative for prediction in the sense that

$$p(t_k \mid \mathcal{H}_k, T = t) = p(t_k \mid \mathcal{H}_k, T = \infty) \tag{3}$$

for $t_{k-1} < t_k < t$. This means that the next appointment is scheduled as if $T = \infty$, but it is not recorded unless $t_k < T$. With this assumption, the third factor in (2) is constant in t and can be ignored. In other words, the distribution of the time to the next scheduled appointment may depend on the patient's medical history, but is independent of the patient's subsequent survival. Ordinarily, the scheduled appointment is included as a component of the patient's record only if it occurs in $[0, T)$ while the patient lives, implying that the partial appointment schedule $\mathbf{t}^{(k)}$ is uninformative for subsequent survival. In particular, an administrative schedule is uninformative.

A simple numerical example illustrates the idea. Suppose T is exponentially distributed with mean 10 years, and the revival process for $s > 0$ is a real-valued Gaussian process with mean $E(Z(s)) = \beta s/(1 + s)$ and covariance function $\delta_{ss'} + \exp(-|s - s'|)$ for $s, s' > 0$. The observed health-status values at $\mathbf{t} = (0, 1, 2, 3)$ are $y = (6.0, 4.5, 5.4, 4.0)$.

For $\beta = 0$, the conditional density is such that $T-3$ is exponential with mean 10; the conditional density is shown for various values $0 \le \beta \le 2$ in the left panel of Fig. 1, and for $4 \le \beta \le 8$ on the right. Evidently, the conditional distribution depends on both the observed outcomes and on the model parameters: the median

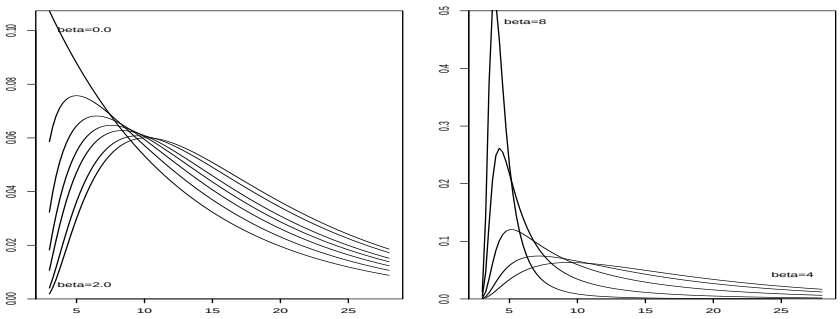

Figure 1: Conditional density of survival time for various values of β.

residual lifetime is not monotone in β. In applications where β is estimated with appreciable uncertainty, the predictive distribution is a weighted convex combination of the densities illustrated.

The conditional survival distribution given $Y[\mathbf{t}^{(k)}]$ depends not only on the current or most recent value, but linearly on the entire vector. In particular, the conditional distribution does not have the structure of a regression model in which the longitudinal variable enters as a time-dependent covariate without temporal interference. Thus, on the assumption that the joint model is adequate, issues related to covariate confounding do not arise.

4.2 A simple Gaussian revival process

Under assumptions (1) and (3), the ratio of the conditional survival density at t to the marginal density is proportional to the factor $g_k(y; t-\mathbf{t}^{(k)})$, in which $y, \mathbf{t}^{(k)}$ are fixed, and t the variable. This modification factor—the Radon-Nikodym derivative—depends only on the revival process, not on the distribution of survival times. On a purely mathematical level, it is precisely the likelihood function in the statistical model for the k-dimensional variable $Y[\mathbf{t}^{(k)}]$ whose conditional distribution given $T = t$ is $G_k(y; t - \mathbf{t}^{(k)} \mid t)$ for some value of the temporal offset parameter $t > t_{k-1}$.

Although not realistic for most applications, suppose that G is Gaussian with mean $\mu(s) = \alpha + \beta s$ independent of t and linear in reverse time, and covariance function $\mathrm{cov}(Z(s), Z(s') \mid t) = K(|s - s'|)$. Then the log density ratio factor

$$\log g_k(y; t - \mathbf{t}^{(k)} \mid t) = \text{const} - \tfrac{1}{2}(y - \mu[t - \mathbf{t}])' K^{-1} (y - \mu[t - \mathbf{t}]),$$

is quadratic in t for $t > t_{k-1}$. After substituting $\alpha + \beta(t - \mathbf{t})$ for the mean function, and expressing the log density ratio as a quadratic in t, it can be seen that the predictive density ratio at $t > t_{k-1}$ is the density at βt of the Gaussian distribution with mean

$$-\alpha + \mathbf{1}' K^{-1} (y + \beta \mathbf{t}^{(k)}) \big/ (\mathbf{1}' K^{-1} \mathbf{1}) = \bar{y} - \alpha + \beta \bar{\mathbf{t}}$$

and variance $1/(\mathbf{1}' K^{-1} \mathbf{1})$, where K has components $K(t_i - t_j)$. Ignoring the dependence on the data that comes from parameter estimation, the dependence of

the predictive density ratio on the data for one patient comes through the weighted averages

$$\bar{y} = \mathbf{1}'K^{-1}y/(\mathbf{1}'K^{-1}\mathbf{1}), \qquad \bar{\mathbf{t}} = \mathbf{1}'K^{-1}\mathbf{t}/(\mathbf{1}'K^{-1}\mathbf{1}) \tag{4}$$

for this particular individual.

4.3 Exchangeable Gaussian revival process

In a more natural Gaussian model, the revival processes for distinct patients are exchangeable but not necessarily independent. Revival models have much in common with growth-curve models (Lee, 1988, 1991) in which $Z_i(s) = \mu(s)+\eta_0(s)+\eta_i(s)$ is a sum of two independent zero-mean Gaussian processes, and the mean function $\mu(s)$ is constant across individuals. Usually the common deviation $\eta_0(\cdot)$ is moderately smooth but not stationary, perhaps fractional Brownian motion with $\eta_0(0) = 0$. The idiosyncratic deviations are independent and identically distributed and they incorporate measurement error, so $\eta_i(\cdot)$ is ordinarily the sum of a continuous process and white noise. Thus, the Gaussian process is defined by

$$\begin{aligned} E(Z_{is}) &= \mu(s) \\ \operatorname{cov}(Z_{is}, Z_{i's'}) &= K_0(s,s') + \delta_{ii'}K_1(s,s') + \sigma^2\delta_{ii'}\delta_{ss'} \end{aligned} \tag{5}$$

for some suitable covariance functions K_0, K_1, each of which can be expected to have a variance or volatility parameter and a range parameter. In the case of fractional Brownian motion, for example, $K(s,t) \propto s^\nu + t^\nu - |s-t|^\nu$ for some $0 < \nu < 2$, which governs the degree of smoothness of the random function.

For a new patient such that $Y[\mathbf{t}^{(k)}] = y$, the conditional survival density $\text{pr}(T \in dt \mid \text{data})$ given the data, including the outcomes for the new patient, is computed in the same way as above. The second factor in (2) is the density at the observed outcomes of the Gaussian joint distribution whose means and covariances are specified above. This involves all $n + 1$ patients.

4.4 Illustration by simulation

Figure 2 shows simulated data for 200 patients whose survival times are independent exponential with mean five years. While the patient lives, annual appointments are kept with probability $5/(5 + t)$, so appointment schedules in the simulation are not entirely regular. Health status is a real-valued Gaussian process with mean $E(Z(s)) = 10 + 10s/(10 + s)$ in reverse time, and covariances

$$\operatorname{cov}(Z(s), Z(s')) = (1 + \exp(-|s - s'|/5) + \delta_{ss'})/2$$

for $s, s' > 0$, so there is an additive patient-specific effect in addition to temporal correlation. Values for distinct patients are independent and identically distributed. This distribution is such that health-status plots in reverse time aligned by failure show a stronger temporal trend than plots drawn in the conventional way. The state of health is determined more by time remaining before failure than time since recruitment. These trends could be accentuated by connecting successive dots for

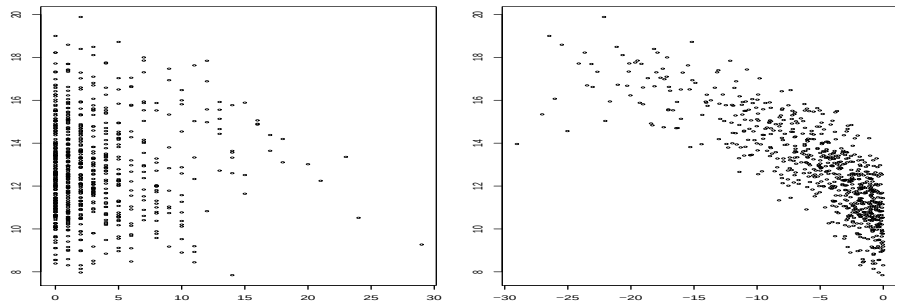

Figure 2: Simulated health status sequences aligned by recruitment time (left) and the same sequences aligned by failure time (right)

each individual, as in Fig. 2 of Sweeting and Thompson (2011), but this has not been done in Fig. 2.

Since the survival times are exponential with mean five, independent of covariates and treatment, the root mean squared prediction error using covariates only is five years. For fixed $k \geq 2$, and a patient having at least k appointments, the conditional survival distribution given the first k health-status values has a standard deviation depending on the observed configuration, but the average standard deviation is about 2.5 years, and the root mean squared prediction error is about 2.7 years. For this setting, the longitudinal variable is a reasonably effective predictor of survival, and the prediction error is almost independent of k in the range 2–5. This summary does not tell the full story because certain y-configurations lead to very precise predictions whereas others lead to predictive distributions whose standard deviation exceeds five years.

The parameter settings used in this simulation may not be entirely representative of the range of behaviours of the conditional survival distribution given $Y[\mathbf{t}]$. If the ratio of the between-patient to within-patient variance components is increased, the average variance of the conditional survival distribution decreases noticeably with k. For such settings, prediction using the entire health history is more effective than prediction using the most recent value.

4.5 Recurrent health-related events

In certain circumstances the health outcome Y is best regarded as a point process, recording the occurrences of a specific type of non-fatal event, such as epileptic or asthmatic attacks or emergency-room visits. In other words, $Y_i \subset \Re$ is the set of times at which patient i experiences the event. Then $\mathbf{t} = (0, t_k)$ is a bounded interval, and the observation $Y[\mathbf{t}] = Y \cap \mathbf{t}$ is the set of events that occur between recruitment and the most recent appointment. This observation records the actual date of each event, which is more informative than the counting process $\#Y[(0, t_1)], \ldots, \#Y[(0, t_k)]$ evaluated at the appointment dates. If there are recurrent events of several types, Y is a marked point process, and $Y[\mathbf{t}]$ is the set of all

events of all types that occur in the given temporal interval. The paper Schaubel and Zhang (2010) is one of several papers in the October 2010 issue of *Lifetime Data Analysis*, which is devoted to studies of this type.

In this situation, the frequency of the recurrent event may be constant over time, or it may vary in a systematic way. For example, the frequency may increase slowly but systematically as a function of either age or time since recruitment. Alternatively, the frequency may be unrelated to age at recruitment, but may increase in the last year of life as death approaches. In the former case, alignment of records by failure time is ineffective; in the latter case, the revival processes for different individuals have a common pattern, and alignment by failure tine is an effective device for exploiting this commonality.

We consider here only the simplest sort of recurrent-event process in which the revival process is Poisson, there is a single event type, and the subset $Y \cap \mathbf{t} = \mathbf{y}$ of observed event times is finite. The mean measure of the revival process is Λ, which is non-atomic with intensity λ on the positive real line. The density ratio at $t > \sup(\mathbf{t})$ is the probability density at the observed event configuration $t - \mathbf{y}$ as a subset of the reverse-time interval $t - \mathbf{t}$, i.e.,

$$g(\mathbf{y}; t - \mathbf{t}) = \exp(-\Lambda(t - \mathbf{t})) \prod_{y \in \mathbf{y}} \lambda(t - y).$$

In particular, if the intensity is constant for $s > 0$, the density ratio is constant, and the event times are uninformative for survival. In other words, it is the temporal variation of the intensity function that makes the observed configuration $\mathbf{y}$ informative for patient survival.

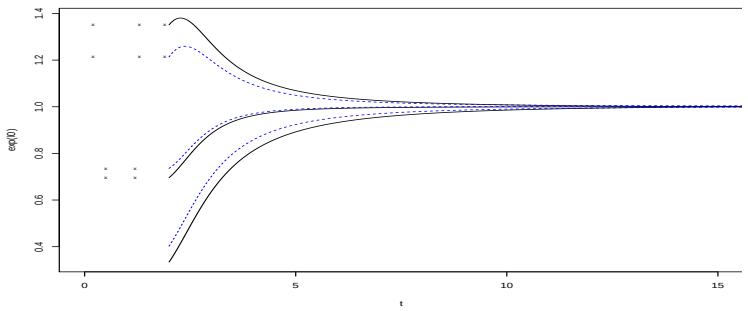

Figure 3: Likelihood functions (solid lines) for three point configurations, with predictive hazard ratios (dashed lines)

For a specific numerical example, let $\lambda(s) = (2 + s^2)/(1 + s^2)$ be the revival intensity, and let $\mathbf{t} = (0, 2)$ be the observation window. The revival intensity, monotone decreasing with an asymptote of one, implies that the recurrent events are moderately common at all ages, but their frequency increases as failure approaches. Figure 3 shows the likelihood as a function of $t \geq 2$ for three event configurations,

$\mathbf{y}_0 = \emptyset$, $\mathbf{y}_1 = \{0.5, 1.2\}$ and $\mathbf{y}_2 = \{0.2, 1.3, 1.9\}$. Since the likelihood function is defined only up to an arbitrary multiplicative constant, the curves have been adjusted so that they are equal at $t = 20$, or effectively at $t = \infty$. In place of the predictive survival distributions, we show instead the ratio of the predictive hazard functions to the marginal hazards as dashed lines on the assumption that the marginal failure distribution is exponential with mean 5. Because of the form of the revival intensity, which is essentially constant except near the origin, the predictive hazard functions are very similar in shape to the likelihood functions.

5 Parameter estimation

5.1 Likelihood factorization

The joint density for the observations in a revival model factors into two parts, one involving only survival times, the other involving only the revival process. More generally, if the revival assumption fails, the second factor is the conditional distribution of the revival process given $T = t$, so both factors depend on t. Although both factors may involve the same covariates and treatment indicators, the parameters in the two parts are assumed to be unrelated, i.e. variation independent. Thus the likelihood also factors, the first factor involving only survival parameters such as hazard modifiers associated with treatment and covariates, the second factor involving only health-status parameters such as temporal trends and temporal correlations. In other words, the two factors can be considered separately and independently, either for maximum likelihood estimation or for Bayesian operations.

The first stage in parameter estimation is to estimate the survival distribution F together with treatment and covariate effects if needed. Whether the model for survival times is finite-dimensional or infinite-dimensional, this step is particularly simple because the first factor involves only the survival times and survival distribution. The standard assumption of independent survival times for distinct patients simplifies the problem even further. Exponential, gamma and Weibull models are all feasible, with treatment effects included in the standard way.

For the Cox proportional-hazards model, the situation is a little more complicated. First, the survival time is finite with probability one if and only if the integrated hazard $\Lambda(\Re^+) = \int_0^\infty \lambda(t)\,dt$ is infinite, which is not satisfied at all parameter points in the model. Second, the proportional-hazards likelihood function depends only on baseline hazard values $\lambda(t)$ in the range $0 \le t \le T_{\max}$, where $T_{\max}$ is the maximum observed survival time, censored or uncensored. Thus, the likelihood does not have a unique maximum, but every maximum has the property that $\hat{\lambda}(t) = 0$ for all $0 \le t \le T_{\max}$ except for failure times, at which $\hat{\lambda}$ has a discrete atom. By common convention (Kaplan and Meier 1958; Cox 1972, §8) $\hat{\lambda}(t) = 0$ for $t > T_{\max}$, but this choice is not dictated by the likelihood function. Since the revival model requires survival times to be finite with probability one, it is essential to restrict the space of hazards to those having an infinite integral, which rules out the standard convention for $\hat{\lambda}$. Equivariance under monotone temporal transformations points to a mathematically natural choice $\hat{\lambda}(t) = \infty$ for $t > T_{\max}$; a less pessimistic

option is to use a finite non-zero constant such as

$$\hat{\lambda}(t) = \frac{\text{total number of failures}}{\text{total person time at risk}}$$

for $t > T_{\max}$. Both of these maximize the proportional-hazards likelihood function—restricted or unrestricted—and either one may be used in the revival model.

The second stage, which is to estimate the parameters in the revival process, is also straightforward, but only if all records are complete with no censoring. Serial dependence is inevitable in a temporal process, and there may also be independent persistent idiosyncratic effects associated with each patient, either additive or multiplicative. Gaussian revival models are particularly attractive for continuous health measurements because such effects are easily accommodated with block factors for patients and temporal covariance functions such as those included in the simulation in Fig. 2.

Thus the second stage involves mainly the estimation of variance components and range parameters in an additive Gaussian model. One slight complication is that the revival process is not expected to be stationary, which is a relevant consideration in the selection of covariance functions likely to be useful. Another complication is that the health status may be vector-valued, $Y(t) \in \Re^q$, so there are also covariance component matrices to be estimated. If the covariance function is separable, i.e.

$$\text{cov}(Z_{ir}(s), Z_{ir'}(s')) = \Sigma_{r,r'} K(s, s')$$

for some $q \times q$ matrix Σ, maximum-likelihood estimation is straightforward. But separability is a strong assumption implying that temporal correlations for all health variables have the same pattern, including the same decay rate, which may not be an adequate approximation. Nevertheless, this may be a reasonable starting point.

The second stage requires all health records to be aligned at their termini. Accordingly, a record that is right censored ($T_i > c_i$) cannot be properly aligned. If the complete records are sufficiently numerous, the simplest option is to ignore censored records in the second stage, on the grounds that the estimating equations based on complete records remain unbiased. This conclusion follows from the fact that the second factor is the conditional distribution given survival time. Thus, provided that the censoring mechanism is a selection based on patient survival time, the estimating equations derived from complete records are unbiased. The inclusion of censored records is thus more a matter of statistical efficiency than bias, and the information gained from incomplete records may be disappointing in view of the additional effort required.

5.2 Incomplete records

If we choose to include in the likelihood the record for a patient censored at $c > 0$, we need the joint probability of the event $T > c$, the density of the subset $\mathbf{t}_c = \mathbf{t} \cap [0, c]$, and the outcome $Y[\mathbf{t}_c]$ at y. On the assumption that censoring is uninformative, i.e. that the distribution of the subsequent survival time for a patient censored at

time c is the same as the conditional distribution given $T > c$ for an uncensored patient, the joint density is

$$\int_{t \geq c} f(t)\; p(\mathbf{t}_c \mid t)\; g(y; t - \mathbf{t}_c)\, dt$$

on the space of finite-length records. Assumption (3) implies that the second factor, the density of the appointment dates in $[0, c]$ for a patient surviving to time $t > c$, does not depend on the subsequent survival time $t - c$, in which case it may be extracted from the integral. It is also reasonable to assume that the distribution of appointment schedules is known, for example if appointments are scheduled administratively at regular intervals, in which case the second factor may also be discarded from the likelihood. Since the survival probability $1 - F(c)$ is included in the first-stage likelihood, the additional factor needed in the analysis of the revival model is

$$\frac{1}{1 - F(c)} \int_{t > c} f(t)\, g(y; t - \mathbf{t}_c)\, dt,$$

in which $\mathbf{t}_c$ may regarded as a fixed subset of $[0, c]$. Unfortunately, the integral involves both the survival density $f(t) = F'(t)$ and the density of the revival process, so the full likelihood no longer factors. For an approximate solution, f may be replaced with the estimate obtained from the first-stage analysis of survival times, and if $\hat{f}$ is purely atomic, the integral is converted to a finite sum.

The situation is considerably more complicated if, as in section 4.3, the revival processes for distinct patients are not independent,

5.3 Treatment effect: definition and estimation

We consider here only the simplest sort of revival model for the effect of treatment on patient health, ignoring entirely its effect on survival time. Health status in the revival process is assumed to be Gaussian, independent for distinct patients, and the treatment is assumed to have an effect only on the mean of the process, not on its variance or covariance. Consider two patients, one in each treatment arm,

$$a_i(t) = \bar{a}_i(T_i - t) = 1, \qquad a_j(t) = \bar{a}_j(T_j - t) = 0$$

such that $x_i = x_j$. The revival assumption asserts that the random variable $Z_i(s) - Z_j(s)$ is distributed independently of the pair T_i, T_j. By definition, the treatment effect as defined by the revival model is the difference of means

$$\tau_{10}(s) = E(Z_i(s)) - E(Z_j(s)) = E(Y_i(T_i - s)) - E(Y_j(T_j - s))$$

at revival time s. This is not directly comparable with either of the the conventional definitions

$$\gamma_{10}(t) = E(Y_i(t) - E(Y_j(t)) \quad \text{or} \quad \gamma'_{10}(t) = E(Y_i(t) - E(Y_j(t) \mid T_i, T_j > t)$$

in which the distributions are compared at a fixed time following recruitment. The expectation in a survival study—that healthy individuals tend to live longer than

the frail—implies that $E(Y(t) \mid T)$ must depend on the time remaining to failure. In that case, the conventional treatment definition $\gamma'_{10}(t)$ depends explicitly on the difference between the two survival times. In other words, it does not disentangle the effect of treatment on patient health from its effect on survival time.

If the revival assumption fails, i.e. if Z is not independent of T, then, in the simplest setting where the dependence on T is linear and additive, the difference of means

$$E(Z_i(s) \mid T) - E(Z_j(s) \mid T) = \tau_{10}(s) + \gamma(T_i - T_j).$$

contains both a treatment effect and an effect due to the difference in survival times. In other words, the failure of the revival assumption does not necessarily complicate the interpretation of treatment effects. By contrast with standard practice in the analysis of randomized trials with longitudinal responses, (Fitzmaurice, Laird and Ware 2011, section 5.6), it is most unnatural in this setting to work with the conditional distribution given the baseline outcomes $Y_i(0) \equiv Z_i(T_i)$. That is one reason for the recommendation in section 3.4 that the baseline response be regarded as an integral part of the outcome sequence, not as a covariate. Exchangeability implies distributional equality $Z_i(T_i) \sim Z_j(T_j)$ for individuals having the same covariate values, but it does not imply equality of conditional distributions given T. On the presumption that treatment assignment is independent of baseline response values, we also have $Z_i(T_i) \sim Z_j(T_j)$ conditionally on treatment, whether or not a_i, a_j are equal. Consequently, in order to satisfy the exchangeability assumption, it is necessary to introduce a null, pre-randomization, treatment level, $a_i(0) = a_j(0)$, common to all subjects.

5.4 Revival review

The easiest way to check the revival assumption is to formulate and fit a specific alternative model in which the revival process is not independent of the survival time. We consider here only the simplest design in which all records are complete, there are no covariates or treatment assignment, observations for distinct patients are independent, and the revival model is a family of Gaussian process. One way to do this is to replace (5) with

$$E(Z_i(s) \mid T) = \mu(s, T_i)$$

for some suitable family of functions $\mu(s, T)$, leaving the covariances unchanged. For example, if x denotes patient age at recruitment, the revival mean might be modeled as

$$E(Z_i(s) \mid T) = \mu(s) + \beta_1 x_i + \beta_2 T_i$$

depending additively on patient age and survival time. If $\beta_1 = \beta_2$, the dependence is on age at failure rather than age at recruitment. More general models involving multiplicative interactions between s and T_i may also be considered.

Consider, for instance, the non-linear Gaussian revival model with mean

$$\mu(s) = \alpha + \beta s/(\gamma + s),$$

which is such that $\mu(0) = \alpha$, $\mu(\infty) = \alpha + \beta$, and $\mu(\gamma) = \frac{1}{2}(\mu(0) + \mu(\infty))$, so that $\gamma > 0$ is the semi-revival time. Within this family, the revival trajectory for one patient could be different from that of another, depending on their survival times. In other words, α, β, γ could depend on T or $x+T$, either of which is a violation of the revival assumption. One of the simplest models of this type is the time-accelerated revival model in which the semi-revival time is inversely related to survival,

$$\mu(s, T) = \mu_0(sT) = \alpha + \beta sT/(\gamma + sT).$$

As a practical matter, it would be more effective to replace γ with $\gamma_0 + \gamma_1/T$ or $\exp(\gamma_0 + \gamma_1/T)$ to generate a test of the revival assumption. Likewise, we could replace α with $\alpha_0 + \alpha_1 T_i$, asserting that the outcome sequences for long-lived patients are elevated by a constant amount at all revival times. Similarly, if β is replaced with $\beta_0 + \beta_1 T_i$, the the asymptote is elevated in proportion to the additional lifetime.

Any modification of this sort is a violation of the revival assumption, so the survival time and the revival process are no longer independent. However, the factorization of the likelihood function remains intact, so the analysis remains relatively straightforward. For example, a likelihood ratio statistic to test the revival assumption can be constructed by fitting two nested models to the revival process, one satisfying the revival assumption, the other involving T.

6 A worked example: cirrhosis study

6.1 Prednizone and prothrombin levels

In the period 1962–1969, 532 patients in Copenhagen hospitals with histologically verified liver cirrhosis were randomly assigned to two treatment arms, control and prednizone. Only 488 patients for whom the initial biopsy could be reevaluated using more restrictive criteria were retained, yielding 251 and 237 patients in the prednizone and placebo groups respectively. Variables recorded at entry include sex, age, and several histological classifications of the liver biopsy. Clinical variables were also collected, including information on alcohol consumption, nutritional status, bleeding, and degree of ascites. However, these covariates were not included in the dataset used here, which was downloaded from the R library `http://cran.r-project.org/web/packages/joineR` maintained by Philipson Sousa, Diggle, Williamson, Kolamunnage-Dona and Henderson. At the end of the study period, the mortality rate was 292/488, or approximately 60%.

The focus here is on the prothrombin index, a composite blood coagulation index related to liver function, measured initially at three-month intervals and subsequently at roughly twelve-month intervals. The individual prothrombin trajectories are highly variable, both in forward and in reverse time, which tends to obscure patterns and trends. In Figure 4a the mean trajectory is plotted against time from recruitment for four patient groups placebo/prednizone and censored/not censored, with the solid lines denoting censored individuals. Naturally, only those patients who are still alive are included in the average for that time. Figure 4b shows the same plots in reverse alignment. While there are certain similarities in the two plots,

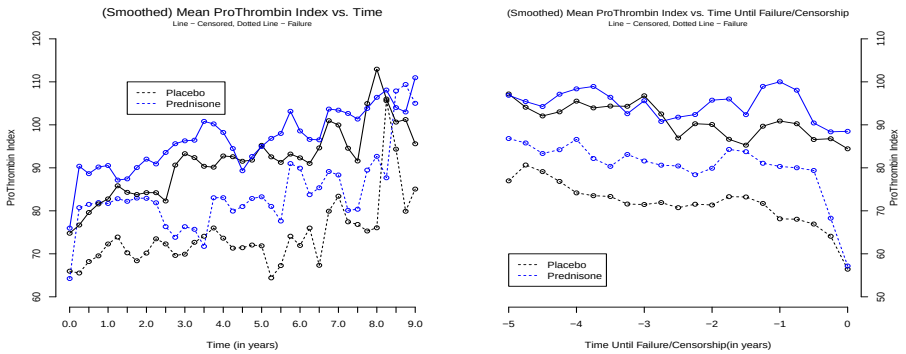

Figure 4: Prothrombin mean trajectories aligned by recruitment and by failure

the differences in temporal trends are rather striking. In particular, prothrombin levels in the six months prior to censoring are fairly stable, which is in marked contrast with levels in the six months prior to failure, as seen in the lower pair of curves.

Inspection of the graphs for uncensored patients in the right panel of Fig. 4, suggests beginning with the simplest revival model in which the sequences for distinct patients are independent Gaussian with moments

$$E(Z_i(s) \mid T) = \alpha + \tau_{\bar{a}_i(s)} + \beta_0 T_i + \beta_1 s + \beta_2 \log(s+\delta)$$
$$\mathrm{cov}(Z_i(s), Z_j(s') \mid T) = \sigma_1^2 \delta_{ij} K_1(s,s') + \sigma_2^2 \delta_{ij} + \sigma_3^2 \delta_{ij} \delta_{ss'}.$$

The non-linear dependence on s is accommodated by the inclusion of $\log(s+\delta)$ in the mean model with a temporal offset δ, which is equal to one day in all subsequent calculations. Inclusion of the survival time T_i is suggested by the increasing trend along the diagonals and sub-diagonals of Table 1. Since the value at recruitment is included as a response for each series, treatment necessarily has three levels, *null*, *control* and *prednizone*. The three covariance terms are associated with independent additive processes, the second for independent and identically distributed patient-specific constants, and the third for independent and identically distributed white noise or measurement error. The first covariance term governs the prothrombin sequences for individual patients, which are assumed to be continuous in time with covariance function $K_1(s,s') = \exp(-|s-s'|/\lambda)$ for $s, s' > 0$. The temporal range in all subsequent calculations is set at $\hat{\lambda} = 1.67$ years, implying an autocorrelation of 0.55 at a lag of one year. The implied one-year autocorrelation for the observed prothrombin sequences is considerably smaller, roughly 0.30, because of the white-noise measurement term.

For the likelihood calculations that follow, incomplete records are ignored; only the 1634 measurements for the 292 non-censored patients are used. The fitted variance components, estimated by maximizing the residual likelihood, are

$$(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2) = (211.4, 209.2, 179.7),$$

all significantly positive. Using these values to determine the covariance matrix, the weighted least-squares coefficients in the mean model are shown in Table 3. The standard error for the prednizone/control contrast is 1.77, somewhat larger than the standard error for the prednizone/null contrast because the former is a contrast between patients involving all three variance components, whereas the latter is a contrast within patients, which is unaffected by the second variance component.

Table 3: Regression coefficients in a revival model

Covariate	Coefficient	S.E.	Ratio
Null treatment	0.00	—	
Control	2.41	1.43	1.7
Prednizone	13.56	1.47	9.2
Survival (T)	1.75	0.47	3.7
Revival (s)	−2.12	0.47	−4.5
$\log(s+\delta)$	4.66	0.41	11.4

Various deviations from this initial model may now be investigated. In particular, it is possible to check whether there is an interaction between treatment and survival time, i.e., whether the treatment effect for long-term survivors is or is not the same as the treatment effect for short-term survivors. This comparison involves two variance-components models having different mean-value subspaces, so the residual likelihoods are not comparable. For likelihood comparisons, the kernel subspace must be fixed, and the natural choice is the mean-value subspace for the null model as described by Welham and Thompson (1997) or as implemented by Clifford and McCullagh (2006). The likelihood ratio statistic computed in this way is 0.83 on two degrees of freedom, showing no evidence of interaction. However, there is appreciable evidence in the data that the treatment effect (prednizone versus control) decreases as $t \to T$, i.e. as $s \to 0$, as is evident in Fig. 4 from the convergence of the two lines during last six months of life. The likelihood-ratio statistic for the *treat.s* interaction is 3.90 on two degrees of freedom, showing little evidence of a linear trend, but the value for the *treat.* $\log(s)$ interaction is 8.68, pointing to a non-linear trend, as is apparent from the convergence of the two lines in Fig. 4b.

We may also check the adequacy of the assumed form for the mean model by including an additional random deviation, continuous in reverse time, with generalized non-stationary covariance function such as $K_0(s,s') = -|\log(s+\delta)-\log(s'+\delta)|$. The fitted coefficient is 2.38, and the associated likelihood ratio statistic is 1.2 on one degree of freedom, showing no significant deviations that are continuous in reverse time. Finally, we check whether the sequences for different patients exhibit a characteristic pattern or trend associated with time measured from recruitment by including the generalized Brownian-motion covariance function $-|t-t'|$ in the covariance model. The fitted variance coefficient is 2.10, and the likelihood ratio statistic is 2.38 on one degree of freedom showing no significant characteristic patterns that are continuous in time measured from recruitment.

6.2 Effect of prothrombin on prognosis

Over a period of 5 years and one month following recruitment, patient u had eight appointments with prothrombin values as follows:

$\mathbf{t}_u$ (days)	0	126	226	392	770	1127	1631	1855
$Y_u[\mathbf{t}_u]$	49	93	122	120	110	100	72	59

This is in fact the record for patient 402 who was assigned to prednizone and was subsequently censored at 2661 days. As determined on day 1855, the survival prognosis for this patient depends on preceding sequence of measurements. Relative to the unconditional survival density for a patient on the prednizone arm, the conditional survival density at time $t > \max(\mathbf{t}_u)$ is modified multiplicatively by a factor proportional to the joint conditional density of the random variable $Z_u[t - \mathbf{t}_u]$ at the observed point y_u given $T_u = t$ and the data observed for all other patients.

For the model described in the preceding section—in which the records for distinct patients are independent—this factor is particularly simple. The conditional distribution of $Z_u[t - \mathbf{t}_u]$ given $T = t$ has a mean vector μ depending linearly on $t - \mathbf{t}_u$ and $\log(t - \mathbf{t}_u + \delta)$, and a covariance matrix Σ that is constant in t. The log density at y_u is a quadratic form

$$h(t, y_u) = \text{const} - (y_u - \mu)'\Sigma^{-1}(y_u - \mu)/2$$

depending on t only through μ. This estimated factor is shown in Fig 5a for three versions of the record in which the final prothrombin value is 59, 69 or 79.

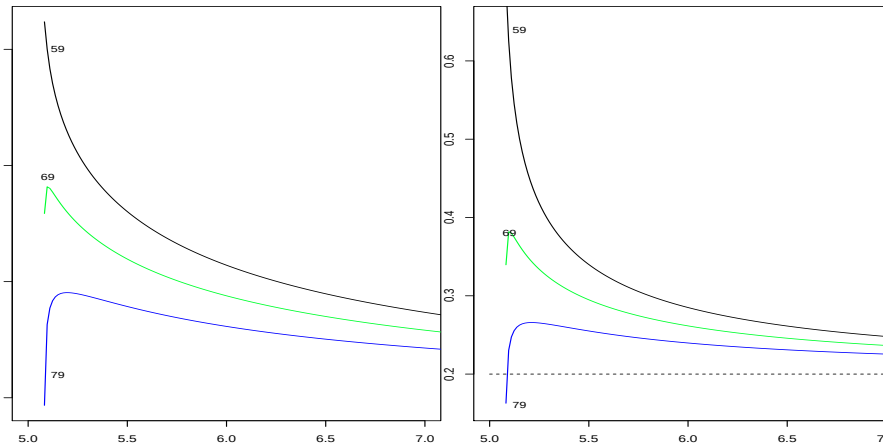

Figure 5: Three versions of the record for patient 402: log modification factors for the predictive survival density (left panel) and hazard functions (right panel).

It may be helpful to express the effect of the observed prednizone record on the conditional survival distribution through its effect on the hazard function at

times $t > \max(\mathbf{t}_u)$ rather than its effect on the conditional survival density. Suppose, therefore, that the unconditional survival time for a patient on the prednizone arm, is exponential with mean 5 years, so that the unconditional hazard function is constant. What is the conditional hazard at time $t > \max(\mathbf{t}_u)$ given the prothrombin sequence for patient u, with no further measurements made in the interval $(\max(\mathbf{t}_u), t)$ other than survival? The conditional hazard functions for the subsequent two-year interval $5 < t < 7$ are shown in Fig. 5b for the same three versions of the prothrombin record. It is evident from these plots that the conditional hazard for the real patient is substantially elevated following the last measurement, but the effect is transient and does not persist for the duration of a typical inter-appointment interval of one year. If the final value were 79 instead of 59, the hazard function is almost constant, initially increasing and subsequently reverting to the long-term value, which is slightly larger than the unconditional hazard.

The preceding analysis indicates that it may be misleading to treat the observed health sequence as a time-dependent covariate in the proportional-hazards model. At any one failure time t measured from recruitment, some of the health measurements are recent and fresh, while others are likely to be up to one year old. Figure 5b shows that stale measurements may have negligible prognostic value compared with fresh measurements, which contradicts the basic assumption in the proportional hazards model. The predictive revival model automatically takes into account the time that has passed since the last appointment, so that stale values are discounted appropriately.

6.3 Review of assumptions

The conditional independence assumption (1) does not require appointments to be scheduled administratively, nor does it forbid patient-initiated appointments. Consider two patients i, j at time s prior to failure, having similar prior appointment schedules and similar health values. Assumption (1) states that the conditional appointment-initiation intensity given the observed health record and subsequent survival time does not depend on subsequent health values. In other words, conditional independence implies that patients i, j are equally likely to initiate an appointment at time s; it is also assumed implicitly that they do so independently.

The evidence presented in section 3.2, and in Liestøl and Andersen (2002) shows clearly that the rate of patient-initiated appointments increases in the last few months of life. It is certainly possible that patient behaviour in this instance violates the conditional independence assumption, but the evidence presented does not directly address the matter. All in all, assumption (1) seems unavoidable and relatively benign.

The non-informative assumption (3) is much stronger than (1). It implies that appointments are scheduled as if the patient will live indefinitely, which is clearly contradicted by the evidence in this example. We now examine the consequences of failure of (3), retaining (1).

Assumption (1) implies that the sampling is non-preferential in the sense of Diggle, Menezes and Su (2010), which means that the second factor in (2) is the same as if the appointment dates had been fixed by design. Consequently, the

likelihood calculations in section 5 are unaffected by the failure of (3).

If the appointment for patient u on day 1855 were self-initiated in such a way that the last factor in (2) depends on subsequent survival, it would be technically incorrect to omit that factor in prognosis calculations. However, if it were known that all appointments for patient u were on schedule, the possibility of a dependence on subsequent survival is eliminated, and the prognosis calculations for this patient is technically correct even if the behaviour of other patients violates (3).

7 Summary

The paper examines the problem of model formulation for health sequences, whose defining characteristic is that the state space contains an absorbing value. Each health sequence is terminated ultimately by death, which is not equivalent to random restriction or censoring because subsequent values are known. Typically, sequence length and sequence values are not independent.

The principal suggestion is that it may be more natural in some circumstances to align health sequences by failure time than by age or by recruitment date. The following list describes various statistical implications of realignment.

1. The health sequence is regarded as a random process in its own right, not as a time-dependent covariate governing survival.
2. To a substantial extent, the model for survival time is decoupled from the revival model for the behaviour of the health sequence in reverse time.
3. Realignment implies that value $Y_i(0)$ at recruitment must not be treated as a covariate, but as an integral part of the response sequence. If they were available, values prior to recruitment could also be used.
4. The definition of a treatment effect is not the usual one because the natural way to compare the records for two individuals is not at a fixed time following recruitment, but at a fixed revival time. The treatment value need not be constant in revival time.
5. The predictive value of a partial health sequence for subsequent survival emerges naturally from the joint survival-revival distribution. In particular, the conditional hazard given the finite sequence of earlier values is typically not constant during the subsequent inter-appointment period.
6. Records cannot be aligned until the patient dies, which means that the revival process is not observable component-wise until T is known. As a result, the likelihood analysis for incomplete records is technically more complicated. This aspect needs further development.
7. The omission of incomplete records from the revival likelihood does not lead to bias in estimation, but it does lead to inefficiency, which could be substantial if the majority of records are incomplete.
8. The principal assumption, that appointment dates be uninformative for subsequent survival, does not affect likelihood calculations, but it does affect

prognosis calculations for individual patients. For that reason, it is advisable to label all appointments as scheduled or unscheduled.

Acknowledgements

Comments by D.R. Cox, D. Farewell, R. Gibbons, N. Keiding, S.M. Stigler and five referees on an earlier version of this paper are gratefully acknowledged. Walter Dempsey helped with the computation in section 6 and the preparation of Figure 4.

8 References

Andersen, P.K., Hansen, L.H. and Keiding, N. (1991) Assessing the influence of reversible disease indicators on survival. *Statistics in Medicine* **10**, 1061-1067.

Clifford, D. and McCullagh, P. (2006). The regress function. *R Newsletter* **6**, 6–10.

Cox, D.R. (1972) Regression models and life tables (with discussion). *J. Roy. Statist. Soc.* B **34**, 187–220.

Cox, D.R. and Snell, E.J. (1981) *Applied Statistics*. London: Chapman and Hall.

DeGruttola, V. and Tu, X.M. (1994) Modeling progression of CD-4 lymphocyte count and its relation to survival time. *Biometrics* **50**, 1003–1014.

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002) *Analysis of Longitudinal Data*. Oxford Science Publications: Clarendon Press.

Diggle, P.J., Farewell, D. and Henderson, R. (2007) Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal (with discussion). *Applied Statistics* **56**, 499–550.

Diggle, P.J., Sousa, I. and Chetwynd, A. (2008) Joint modeling of repeated measurements and tome-to-event outcomes: The fourth Armitage lecture. *Statistics in Medicine* **27**, 2981–2998.

Diggle, P., Menezes, R. and Su, T-L. (2010) Geostatistical inference under preferential sampling (with discussion). *Appl. Statist.* **59**, 191–232.

Farewell, D. and Henderson, R. (2010) Longitudinal perspectives on event history analysis. *Lifetime Data Analysis* **6**, 102–117.

Faucett, C.L. and Thomas, D.C. (1996) Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–1685.

Fieuws, S., Verbeke, G., Maes, B. and Vanrenterghem (2008) Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* **9**, 419–431.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2011) *Applied Longitudinal Data Analysis, 2nd edition*. New York: Wiley.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2009) *Longitudinal Data Analysis* Chapman & Hall.

Guo, X. and Carlin, B. (2004) Separate and joint modeling of longitudinal and event time data using standard computer packages. *American Statistician* **58**, 1–10.

Henderson, R., Diggle, P. and Dobson, A. (2000) Joint modeling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Kurland, B.F., Johnson, L.L., Egleston, B.L. and Diehr, P.H. (2009) Longitudinal data with follow-up truncated by death: match the analysis method to the research aims. *Statistical Science* 24, 211-222.

Lagakos, S.W. (1976) A stochastic model for censored-survival data in the presence of an auxiliary variable. *Biometrics* **32**, 551-559.

Lee, J.C. (1988) Prediction and estimation of growth curves with special covariance structures. *J. Amer. Statist. Assoc.* **83**, 432–440.

Lee, J.C. (1991) Tests and model selection for the general growth curve model. *Biometrics* **47**, 147–159.

Laird, N. (1996) Longitudinal panel data: an overview of current methodology. In *Time Series Models in Econometrics, Finance and Other Fields,* D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen, eds. Chapman & Hall Monographs on Statistics and Applied Probability 65.

Liestøl, K, and Andersen, P.K. (2002) Updating of covariates and choice of time origin in survival analysis: problems with vaguely defined disease states. *Statistics in Medicine* **21**, 3701–3714.

McCullagh, P. (2008). Sampling bias and logistic models (with discussion). *J. Roy. Statist. Soc.* B **70**, 643–677.

Murphy, S.A. (2003) Optimal dynamic treatment regimes (with discussion). *J. Roy. Statist. Soc.* B **25**, 331–366.

Rizopoulos, D. (2010) JM: An R package for the joint modeling of longitudinal and time-to-event data. *Journal of Statistical Software* **35**, 1–33.

Rizopoulos, D. (2012) *Joint Models for Longitudinal and Time-to-Event Data* Chapman and Hall.

Rosthøj, S., Keiding, N. and Schmiegelow, N. (2012) Estimation of dynamic treatment strategies for maintenance therapy of children with acute lymphoblastic leukaemia: an application of history-adjusted marginal structural models. *Statistics in Medicine* **31**, 470–488.

Schaubel, D.E. and Zhang, M. (2010) Estimating treatment effects on the marginal recurrent event mean in the presence of a terminating event. *Lifetime Data Analysis* **16**, 451–477.

Sweeting, M.J. and Thompson, S.G. (2011) Joint modeling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal* **53**, 750–763.

Tsiatis, A.A., DeGruttola, V. and Wulfsohn, M.S. (1995) Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. *J. Amer. Statist. Assoc.* **90**, 27–37.

Tsiatis, A.A, and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809–834.

van Houwelingen, H.C. and Putter, H. (2012) *Dynamic Prediction in Clinical Survival Analysis*. Monographs on Statistics and Applied Probability 123; CRC Press.

Welham, S.J. and Thompson, R. (1997) Likelihood ratio tests for fixed model terms using residual maximum likelihood. *J. Roy. Statist. Soc.* B **59**, 701–714.

Wulfsohn, M.S. and Tsiatis, A.A. (1997) A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Xu, J. and Zeger, S.L. (2001) Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics* **50**, 375–387.

Zeger, S.L. and Liang, K.-Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

Zeger, S.L., Liang, K.-Y. and Albert, P. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.