Dept. of Math. University of Oslo
Statistical Research Report No. 6
ISSN 0806–3842 June 2005

Dynamic path analysis – a new approach to analyzing time-dependent covariates

Johan Fosen
Department of Biostatistics
Institute of Basic Medical Sciences
University of Oslo

Egil Ferkingstad
Centre for Integrative Genetics, Aas, Norway
and
Department of Biostatistics
Institute of Basic Medical Sciences
University of Oslo

Ørnulf Borgan
Department of Mathematics University of Oslo

Odd O. Aalen
Department of Biostatistics
Institute of Basic Medical Sciences
University of Oslo

Abstract

In this article we introduce a general approach to dynamic path analysis. This is an extension of classical path analysis to the situation where variables may be time-dependent and where the outcome of main interest is a stochastic process. In particular we will focus on the survival and event history analysis setting where the main outcome is a counting process. Our approach will be especially fruitful for analyzing event history data with internal time-dependent covariates, where an ordinary regression analysis may fail. Our approach enables us to describe how the effect of a fixed covariate partly is working directly and partly indirectly through internal time-dependent covariates. For the sequence of times of event, we define a sequence of path analysis models. At each time of an event, ordinary linear regression is used to estimate the relation between the covariates, while the additive hazard model is used for the regression of the counting process on the covariates. The methodology is illustrated using data from a randomized trial on survival for patients with liver cirrhosis.

Keywords: event history analysis, graphical models, internal covariates, time-dependent covariates

1 Introduction

Graphical models (e.g. path analysis) have been widely used in areas such as computer science and social sciences. In survival and event history analysis as well as in many other branches of mathematical statistics, graphical models have been mostly absent. Hence there has not been much focus on the structure among variables beyond that between covariates and the response in a regression analysis. Graphical models such as path analysis have several appealing features. Assumptions about the relations between the variables in the model can be specified in a concise and unambiguous manner, which may be easily communicated to other researchers/analysts familiar with graphical modelling. The mathematical theory of graphical models such as Bayesian networks is well developed. However, one main disadvantage of traditional graphical modelling is that the role of time is not explicitly considered. In this paper we will introduce a new approach to graphical models, where we focus on how events and processes in the past influence the development in the future. We call this new approach dynamic path analysis.

We are going to study the situation where the main outcome is a stochastic process, and where we can model its compensator as a linear function of a

set of covariates. We will define a path analysis model, i.e. a set of hierarchical linear regression models where some covariates in one regression model will be the response in another regression model, thus forming a graph of vertices (variables) and edges showing how all variables are related to each other. One of these linear models will be the regression of the increment of the stochastic process. The path model will be fitted at each time we are collecting information about the process.

A special case is when the stochastic process is the counting process $N(t)$ counting the number of individual events until time t. For the regression of the event indicator $dN(t)$, we will use the additive hazard model (Aalen 1980, 1989). Since the model is linear at time t given the covariates $X_1(t), X_2(t), \ldots, X_p(t)$, the additive hazard model is well suited to be fitted into a path analysis framework. We have $dN(t) = dB_0(t) + dB_1(t)X_1(t) + \cdots + dB_p(t)X_p(t) + dM(t)$ where $dB_j(t)$ are the regression functions and $dM(t)$ is a martingale increment. The regression functions are estimated by the method of least-squares.

Because of the large variability in the estimates of $dB_j(t)$, it is common to consider the cumulative regression functions $B_j(t) = \int_0^t dB_j(s)$, and since each regression function in path analysis terminology is a direct effect of the covariate on the response, we will define $B_j(t)$ as the cumulative direct effect. Following the usual terminology of path analysis, the indirect effect along a path with regression functions $\psi_j(t)$ (linear regression), and $dB_j(t)$ is $\psi_j(t)dB_j(t)$. In a similar way as above we will then define the cumulative indirect effect as $\int_0^t \psi_j(s)dB_j(s)$. This method has previously been proposed by Fosen et al. (2004), but in this paper, in Section 2, we will define the dynamic path model and cumulative indirect and direct effects more formally and in larger generality.

An internal time-dependent covariate for a particular individual is usually the result of a process generated by that individual until time t, thus the internal covariate might carry information about time to event (Kalbfleisch & Prentice 2002). In the path analysis framework, an internal covariate is simply an intermediate variable, i.e. some (or all) of the effect of a fixed covariate on the response is working through the intermediate variable (also known as the mediator). The effect of interest of a fixed covariate, e.g. treatment, is then the sum of the direct effect of that covariate and the indirect effect through the intermediate variable. This sum is known as the total effect. Running a regression analysis without recognizing that one or more of the covariates are also intermediate variables for other covariates will hence lead to uncorrect estimation of effect since the indirect effect then is

discarded. This warning was also addressed, without referring to graphical models, in Section 6.4 of Kalbfleisch & Prentice (2002).

As described in e.g. Fosen et al. (2004), the total effect is simply the effect when excluding the intermediate variable(s). However, our main interest is to understand the nature of this total effect for survival data. We want to identify in which ways and to what extent the total effect is working directly and indirectly. In Section 3 we will use a liver cirrhosis data set as a case study.

Since we are trying to understand rather thoroughly how variables and processes are influencing each other, it is particularly important to remember that all conclusions rely on the assumption of no unmeasured confounders. All estimated effects are vulnerable to unobserved confounders, which could cause any direct effect in the model to be spurious. This is well-known in all regression analyses. However, indirect effects are more vulnerable than direct effects in the particular sense that the indirect effect depends on all the direct effects along the path, each of which may be a spurious effect due to an unmeasured confounder. A case where indirect effects constitute most of the total effect is more vulnerable to unmeasured confounders than a case where a direct effects is the main source of the total effect.

2 Dynamic graphical modelling

We will in this section first formally describe classical path analysis using graphical model notation. Then we will introduce a general version of Aalen's additive model, extending its use beyond the usual counting process framework. Finally we combine path analysis and Aalen's additive model into the dynamic path model.

2.1 Classical graphical models

The graphical models considered in this paper are based on the use of *directed acyclic graphs* (DAGs). A *directed graph* G is a pair (V, E) where V is a finite set and $E \subseteq V \times V$, i.e. E is a set of ordered pairs of elements from V. The elements of V are called *vertices*; the elements of E *edges*. If $(v_1, v_2) \in E$ we say that v_1 is a *parent* of v_2 and v_2 is a *child* of v_1. We denote the set of parents of a vertex v as $pa(v)$. A *directed path* from $v_{i_1} \in V$ to $v_{i_r} \in V$ is a sequence $(v_{i_1}, v_{i_2}, \ldots, v_{i_r}) \in V^r$ of vertices such that $(v_{i_j}, v_{i_{j+1}}) \in E$ for $j = 1, 2, \ldots, r-1$. A *directed cycle* is a directed path $(v_{i_1}, v_{i_2}, \ldots, v_{i_r})$ where $v_{i_1} = v_{i_r}$. Finally, a DAG is a directed graph with no directed cycles. For an overview on graphical models, see e.g. Lauritzen (1996) or Edwards (2000).

For statistical modelling, one possibility is letting $V = \{X_1, X_2, \ldots, X_p\}$, where the X_j are random variables. The basic idea is then to let $(X_h, X_j) \in E$ if X_h directly influences X_j with respect to the other variables. *Path analysis*, which was introduced by Sewall Wright in the 1920s (Wright 1921, 1934), was probably the first attempt of graphical modelling using this approach. For a general introduction to path analysis, see e.g. Loehlin (2004). In path analysis, all relationships between the variables are assumed to be linear with normally distributed errors. We are going to omit the normal distribution assumption.

Formally, we have p ordered variables $X_1, X_2, \ldots, X_p$ where $X_1 = \epsilon_1$ and

$$X_j = \sum_{h=1}^{j-1} \beta_{hj} X_h + \epsilon_j; \; j = 2, 3, \ldots, p; \tag{1}$$

where the ϵ_j are iid with expectation zero and variance σ^2. The β_{hj} are called *path coefficients*. We may interpret β_{hj} as the change that would occur in X_j if we could intervene and increase X_h by one unit without influencing the other variables in the model. In this respect we may say that β_{hj} is a measure of the causal effect of X_h on X_j. The associated *path diagram* is a DAG G with $V = \{X_1, X_2, \ldots, X_p\}$ where for $h < j$, $(X_h, X_j) \in E$ if and only if $\beta_{hj} \neq 0$.

A path model is constructed in two steps. Firstly, we order the variables such that X_j does not influence X_h for $h < j$. Secondly, we may make the further assumption that $\beta_{hj} = 0$ for one or more of the pairs (h, j). The causal ordering of the variables might be given by subject-matter knowledge or by logical considerations. For example, if the variables are ordered in time, then cause must precede effect. As another example, if a variable X_j is determined by randomization, then it seems reasonable to assume that $\beta_{hj} = 0$ for all $h \neq j$.

Because of the linearity and the acyclic structure, the path coefficients can be estimated by recursively regressing each variable onto all of its parents, using least squares: first we regress X_p onto all of the variables in $pa(X_p)$, then we regress X_{p-1} onto the variables in $pa(X_{p-1})$, and so on. Thus, path analysis may be seen as an extension of ordinary multiple regression.

2.2 The additive regression model

Before we introduce dynamic path analysis, we need to define the additive regression model. Previously, this has been done only for counting processes

(Aalen 1980, 1989), but we will give a more general definition. Let $Y(t)$ be a stochastic process defined on some probability space $(\Omega, \mathcal{F}, P)$ and adapted to the filtration $\{\mathcal{F}_t\}$, and assume that $Y(t)$ is the difference of two local submartingales. Then, by the Doob-Meyer decomposition theorem (e.g. Andersen et. al 1993, Section II.3.1), we may write $Y(t)$ as a sum of a finite variation predictable process $\Lambda(t)$ and a martingale $M(t)$,

$$Y(t) = \Lambda(t) + M(t).$$

The process $\Lambda(t)$ is called the compensator of $Y(t)$. Intuitively, the equation means that $Y(t)$ may be decomposed into a predictable part and a pure noise part. Writing informally $dY(t)$, $d\Lambda(t)$ and $dM(t)$ for the changes in the corresponding processes in the infinitesimal time interval $[t, t+dt)$, note that

$$\mathrm{E}(dY(t)|\mathcal{F}_{t-}) = \mathrm{E}(d\Lambda(t)|\mathcal{F}_{t-}) + \mathrm{E}(dM(t)|\mathcal{F}_{t-}) = d\Lambda(t),$$

since $\mathrm{E}(dM(t)|\mathcal{F}_{t-}) = 0$ by the definition of a martingale and $\mathrm{E}(d\Lambda(t)|\mathcal{F}_{t-}) = d\Lambda(t)$ by predictability of $\Lambda(t)$. Our aim is to model the influence of the covariates $X_1(t), X_2(t), \ldots, X_p(t)$ on the development of the process $Y(t)$.

The process $Y(t)$ might not be observable at all times. We use the "at risk" function $R(t)$ to indicate whether $Y(t)$ is observed; $R(t) = 1$ at any time t when $Y(t)$ is observed, $R(t) = 0$ otherwise. Further we let $R(t)$ be predictable. The additive regression model then is

$$d\Lambda(t) = R(t)(dB_0(t) + dB_1(t)X_1(t) + dB_2(t)X_2(t) + \cdots + dB_p(t)X_p(t)),$$

where $dB_j(t) = \beta_j(t)dt$ and the $\beta_j(t)$ are arbitrary regression functions. Estimation is based on the cumulative regression functions

$$B_j(t) = \int_0^t \beta_j(s)ds.$$

Assume that we have n independent copies $Y_1(t), Y_2(t), \ldots, Y_n(t)$ of the process $Y(t)$, corresponding to the observation of n individuals, and let $R_i(t) = 1$ if individual i is under observation just before time t, $R_i(t) = 0$ otherwise. Then, for each $i = 1, \ldots, n$, the Doob-Meyer decomposition gives

$$\begin{aligned} dY_i(t) &= d\Lambda_i(t) + dM_i(t) \\ &= R_i(t)(dB_0(t) + dB_1(t)X_{i1}(t) + \cdots + dB_p(t)X_{ip}(t)) + dM_i(t). \end{aligned}$$

To write this on matrix form, we introduce

$$\begin{aligned}
\mathbf{Y}(t) &= (Y_1(t), Y_2(t), \ldots, Y_n(t))';\\
\mathbf{M}(t) &= (M_1(t), M_2(t), \ldots, M_n(t))';\\
\mathbf{B}(t) &= (B_0(t), B_1(t), \ldots, B_p(t))';\\
\mathbf{W}_i(t) &= R_i(t)(1, X_{i1}(t), \ldots, X_{ip}(t)),\ i = 1, 2, \ldots, n;
\end{aligned}$$

and let $\mathbf{W}(t)$ be the $n \times (p+1)$ matrix with ith row equal to $\mathbf{W}_i(t)$. Then, we may write

$$d\mathbf{Y}(t) = \mathbf{W}(t)d\mathbf{B}(t) + d\mathbf{M}(t). \tag{2}$$

Note that this has the form of a linear model with $d\mathbf{Y}(t)$ as the response, $\mathbf{W}(t)d\mathbf{B}(t)$ as the systematic component, and $d\mathbf{M}(t)$ as the noise term. When $\mathbf{W}(t)$ has full rank, an estimator for $d\mathbf{B}(t)$ is given by

$$d\hat{\mathbf{B}}(t) = \mathbf{W}^-(t)d\mathbf{Y}(t),$$

where $\mathbf{W}^-(t) = (\mathbf{W}(t)'\mathbf{W}(t))^{-1}\mathbf{W}(t)'$. Letting $J(t)$ be the indicator of $\mathbf{W}(t)$ having full rank, an estimator for $\mathbf{B}(t)$ is then given by

$$\hat{\mathbf{B}}(t) = \int_0^t J(s)\mathbf{W}^-(s)d\mathbf{Y}(s). \tag{3}$$

Using (2) and (3), we see that

$$\hat{\mathbf{B}}(t) = \int_0^t J(s)d\mathbf{B}(s) + \int_0^t J(s)\mathbf{W}^-(s)d\mathbf{M}(s).$$

Here the latter term on the right-hand side is a stochastic integral. In particular the estimator is essentially unbiased, in the sense that

$$\mathrm{E}(\hat{\mathbf{B}}(t)) = \int_0^t \mathrm{E}(J(s))d\mathbf{B}(s). \tag{4}$$

2.3 Examples/Special cases

In the liver cirrhosis case study in Section 3, the outcome for each individual is a counting process indicating whether death has been observed to occur by time t. More generally we could be interested in how the covariates influence some event of interest, where the counting process $N(t)$ is the number of events that have occured up to and including time t. Then the compensator

is known as the integrated intensity process. With $Y(t) = N(t)$, Expression (3) may now be written as

$$\hat{\mathbf{B}}(t) = \sum_{T_k \leq t} J(T_k)\mathbf{W}^-(T_k)\Delta\mathbf{N}(T_k), \tag{5}$$

where $T_1 < T_2 < \cdots$ are the ordered event times of the n counting processes and

$$\Delta\mathbf{N}(T_k) = \mathbf{N}(T_k) - \mathbf{N}(T_{k-1}),\ k \geq 2;\ \Delta\mathbf{N}(T_1) = \mathbf{N}(T_1).$$

We will mainly focus on the counting process situation in the remainder of the present paper. Nevertheless, it is worthwhile to emphasize that the additive regression model may be useful in other situations as well, although the theory is most well-developed for counting processes. For example, consider the case where the outcome $Y(t)$ is a *diffusion process*: let $Y(t)$ be the solution of the following Itô stochastic differential equation

$$dY(t) = \alpha(Y(t), t)\, dt + \gamma(Y(t), t)\, dV(t),$$

where α and γ are deterministic functions, and $V(t)$ is a Wiener process (standard Brownian motion). Time-dependent covariates $X_1(t), \ldots, X_p(t)$ may influence the process. Since the Wiener process is a martingale, the integral $\int_0^t \alpha(Y(s), s)\, ds$ is the compensator of the process $Y(t)$, and the additive model then assumes that we may model $\alpha(Y(t), t)$ (known as the *drift* of the diffusion process) as

$$\alpha(Y(t), t) = \beta_0(t) + \beta_1(t)X_1(t) + \ldots + \beta_p(t)X_p(t).$$

Given several independent observations of the process $Y(t)$, we may then estimate the cumulative regression functions $B_j(t)$ based on (3). In practice, the diffusion process will only be discretely observed, so the stochastic integral in (3) cannot be evaluated directly, but must be approximated by a sum. Using the matrix notation of Section 2.2, if n iid copies of the process is observed at times $\tau_1, \tau_2, \ldots, \tau_m$, a natural estimate for the vector $\mathbf{B}(t)$ of cumulative regression functions is then

$$\hat{\mathbf{B}}(t) = \sum_{\tau_k \leq t} J(\tau_k)\mathbf{W}^-(\tau_k)\Delta\mathbf{Y}(\tau_k),$$

where $\Delta\mathbf{Y}(\tau_k) = \mathbf{Y}(\tau_k) - \mathbf{Y}(\tau_{k-1})$ for $k \geq 2$.

2.4 Dynamic path modelling

Having introduced path analysis and the additive hazard model in the previous sections, we are now ready to define 'dynamic path analysis'. The aim is to investigate the effects of the covariate processes on the infinitesimal changes of the outcome process $Y(t)$, and the relations between the covariates.

Dynamic path diagrams are defined in analogy with classical path diagrams (Section 2.1). A dynamic path diagram is a set of *time-indexed* DAGs $G(t) = (V(t), E(t)), t \in [0, \infty)$. At any time t, the corresponding vertex set $V(t)$ is partitioned into a *covariate set* $V_c(t) = \{X_1(t), X_2(t), \ldots, X_p(t)\}$ and an *outcome process* $Y(t)$: $V(t) = V_c(t) \cup \{Y(t)\}$ where $Y(t) \notin V_c(t)$. We will assume that $V(t)$ as well as the partition into covariates and outcome is time-invariant, so we may drop the subscript t from $V(t)$ and $V_c(t)$ and write $V = V_c \cup \{Y(t)\}$. However, the edge set $E(t)$ may vary with time. We will assume that, for all t, $E(t) \subseteq (V_c(t) \times V_c(t)) \cup (V_c(t) \times \{Y(t)\})$. This simply means that all edges are allowed, except edges pointing from the outcome to a covariate.

We have a sequence of dynamic path models, one for each time t when we collect information. The estimation of each dynamic path model is done by recursive least squares regression as usual in path analysis (Section 2.1), but now the first regression is of the increment of the outcome process $Y(t)$ onto $pa(Y)$, using the additive regression model. Informally, the dynamic path diagram may be seen as continuously evolving over time, with edges possibly appearing or disappearing at any time when we collect new information. In the case of a counting process outcome, this happens at each event time.

We focus on what happens locally in time, i.e. the effect of the covariates on the instantaneous change in the outcome process at each time point. This separates our approach from other dynamic graphical models such as Didelez' (2000) *local independence graphs*, where each node represents the entirety of a process (*globally* in time).

The additivity of classical path analysis is preserved in dynamic path analysis, hence total effects may be easily decomposed into direct and indirect effects, and this is possibly the most appealing feature of dynamic path analysis. A direct effect is an effect which is transmitted through a single edge in a graph, i.e. from a covariate to one of its children, while an indirect effect is an effect of a covariate working through a directed path of length greater than one. In other words, an indirect effect is an effect that is *mediated* through one or more other covariates. Note that there may be several indirect effects of a covariate on an outcome.

In nonlinear systems the notion of an indirect effect is quite problematic, and cannot be defined simply as the difference between the total effect and the direct effect, since such a "definition" would have no operational meaning (Pearl, 2001).

In our model, the indirect effect along each path is simply the product of the (ordinary linear or additive hazard) regression functions along the path, and there is no problem of interpretation. The direct effect and the indirect effects simply add up to the total effect. To formalize these notions, let $\psi_{hj}(t)$ denote the regression coefficient when $X_j(t)$ is regressed onto $X_h(t)$ (corresponding to an edge from $X_h(t)$ to $X_j(t)$), and let $dB_j(t)$ be the additive regression of $dY(t)$ onto $X_j(t)$. We will now define the direct and indirect effects of an arbitrary covariate $X_h(t)$ on the outcome $dY(t)$. Let $X_{c^h_{11}}(t), X_{c^h_{21}}(t), \ldots, X_{c^h_{m1}}(t)$ denote the set of children of an arbitrary $X_h(t)$. For each child $X_{c^h_{k1}}(t)$, there is a corresponding directed path

$$\{(X_h(t), X_{c^h_{k1}}(t)), (X_{c^h_{k1}}(t), X_{c^h_{k2}}(t)), \ldots, (X_{c^h_{kr_k}}(t), dY(t))\} \subseteq E$$

from $X_h(t)$ to $dY(t)$. The indirect effect of $X_h(t)$ on the infinitesimal change in $Y(t)$, $dY(t)$, is now given by

$$\text{ind}(X_h(t) \to dY(t)) = \sum_{k=1}^{m} \psi_{hc^h_{k1}}(t) \left(\prod_{l=1}^{r_k-1} \psi_{c^h_{kl}c^h_{k(l+1)}}(t) \right) dB_{c^h_{kr_k}}(t). \quad (6)$$

The direct effect $\text{dir}(X_h(t) \to dY(t))$ on $dY(t)$ is simply $dB_h(t)$, so the total effect is given by

$$\text{tot}(X_h(t) \to dY(t)) = dB_h(t) + \sum_{k=1}^{m} \psi_{hc^h_{k1}}(t) \left(\prod_{l=1}^{r_k-1} \psi_{c^h_{kl}c^h_{k(l+1)}}(t) \right) dB_{c^h_{kr_k}}(t). \quad (7)$$

The cumulative indirect, direct and total effects are then given by

$$\text{dir}(X_h(t) \to Y(t)) = B_h(t),$$

$$\text{ind}(X_h(t) \to Y(t)) = \int_0^t \sum_{k=1}^{m} \psi_{hc^h_{k1}}(s) \left(\prod_{l=1}^{r_k-1} \psi_{c^h_{kl}c^h_{k(l+1)}}(s) \right) dB_{c^h_{kr_k}}(s), \quad (8)$$

and

$$\text{tot}(X_h(t) \to Y(t)) = B_h(t) + \int_0^t \sum_{k=1}^{m} \psi_{hc^h_{k1}}(s) \cdot \left(\prod_{l=1}^{r_k-1} \psi_{c^h_{kl}c^h_{k(l+1)}}(s) \right) dB_{c^h_{kr_k}}(s). \quad (9)$$

The notions of direct and indirect effects are important when we want to understand the detailed workings of a system. A classical, simple example (Hesslow 1976) is the effect of birth control pills on thrombosis. It is suspected that the pills increase the risk of thrombosis (direct effect). Pregnancy also increases the risk of thrombosis, while, clearly, birth control pills lower the chance of getting pregnant. Thus, birth control pills have a negative indirect effect on thrombosis, through pregnancy. In this case, interest might focus on the direct effect. In other cases, both a direct effect and one or more indirect effects might be of interest, and we want to assess each of these separately.

2.5 Estimation in dynamic path models

From now on we are assuming the outcome $Y(t)$ to be a counting process, hence $Y(t) = N(t)$ in the expression (9) for the total effect of $X_h(t)$ on $Y(t)$. We write

$$\Delta\hat{\mathbf{B}}(T_k) \equiv J(T_k)\mathbf{W}^-(T_k)\Delta\mathbf{N}(T_k)$$

(where $T_1 < T_2 < \cdots$ are the ordered event times) and let $\Delta\hat{B}_h(T_k)$ be the hth element in this vector. Then the estimator $\hat{\mathbf{B}}(t)$ in (3) may be written as $\sum_{T_k \le t} \Delta\hat{\mathbf{B}}(T_k)$. From (9), using the notations $\widehat{\text{ind}}$, $\widehat{\text{dir}}$ and $\widehat{\text{tot}}$ for estimators of the indirect, direct and total effects, respectively, natural estimators are then given by

$$\begin{aligned}
&\widehat{\text{tot}}(X_h(t) \to N(t)) = \\
&= \widehat{\text{dir}}(X_h(t) \to N(t)) + \widehat{\text{ind}}(X_h(t) \to N(t)) \qquad (10)\\
&= \sum_{T_k \le t} \Delta\hat{B}_h(T_k) + \sum_{T_k \le t} \sum_{k=1}^{m} \psi_{hc^h_{k1}}(T_k) \left(\prod_{l=1}^{r_k - 1} \hat{\psi}_{c^h_{kl} c^h_{k(l+1)}}(T_k) \right) \Delta\hat{B}_{c^h_{kr_k}}(T_k),
\end{aligned}$$

where each $\hat{\psi}_{hj}(s)$ is an ordinary least squares regression estimate of the coefficient $\psi_{hj}(s)$ (for each time point s).

The distributions of the estimators for the cumulative indirect effects are complicated because they are sums of highly correlated terms, where each term is a product of a linear regression function and an additive hazard model regression function. Instead of trying to find the large sample distributions of the estimators, we will assess their variability using bootstrap confidence intervals. Since we then almost immediately also are provided with the confidence intervals for the direct effects, we will use bootstrapping here as well.

We will also use bootstrapping as method for selecting the path model, partly because of the non-normality of the ϵ_i of (1) which inhibits us from performing simultaneous testing of the complete path model (Bollen 1989, Bollen & Lang 1993). More precisely we will use simple tests of significance for each regression model contained in the path model to suggest which effects should be considered non-existing. Then, after arriving at a suggested model, we will use bootstrap confidence intervals for validating the model choice: if the confidence interval contains zero, then the covariate is unsignificant.

We will do a non-parametric bootstrap by sampling randomly with replacement from the set of all individuals (Efron 1981). Another approach would have been to rely more on the additive hazard model and thus resample the martingale residuals. The percentile method for bootstrap confidence intervals will be used, since this method is easily adapted to the bootstrap scheme above.

3 Case study: survival with liver cirrhosis

3.1 Data

We will apply the theory developed above to analyse data from a randomized trial of survival for patients with liver cirrhosis; see Andersen et. al (1993, pp. 19-20) for a detailed description of the data and further references to the study. The data consist of 488 patients diagnosed with liver cirrhosis between 1962 and 1969 at several Copenhagen hospitals. The patients were randomized into one group of 251 patients given treatment with prednisone and one group of 237 patients receiving placebo. Prednisone is a synthetic hormone of a class called glucocorticoids and is adminstered as a medicine due to its anti-inflammatory property, being e.g. significantly stronger than that of hydro-cortisone. The patients were followed until death or censoring at the closure of the study in October 1974.

Earlier analyses of the data suggest that there is interaction between ascites (whether or not presence of excess fluid in the peritoneal cavity) and treatment, hence we can't include ascites in the model without modifying the path analysis. We will restrict our attention to the larger group with no ascites, and then we have 386 patients, of whom 191 received treatment and 195 received placebo. The number of deaths observed were 211, of which 94 occurred in the treatment group and 117 in the placebo group.

We will concentrate on six covariates recorded at time zero (fixed covariates) and in addition the time-dependent covariate prothrombin recorded

at follow-up visits to a physician. The visits were scheduled at 3, 6, and 12 months after randomization, and then once every year. Also other time-dependent covariates were recorded at these visits, but we have chosen to concentrate on prothrombin, since this turned out to be the most interesting of them.

We then have the following list of covariates:

- whether receiving placebo or treatment with prednisone (0=placebo, 1=treatment).
- sex (0=female, 1=male).
- age, ranging mostly from 44 to 77 years, but with four persons at age 27 years.
- acetylcholinesterase: an enzyme that breaks the neurotransmitter acetylcholin down after the transmission of a nerve impulse. This is a necessary breakdown process in order to enable rapid neurotransmissions.
- inflammation in liver connective tissues (0=none, 1=present).
- baseline prothrombin (percent of normal value of a blood test of no. II, VII, and X of blood coagulation factors produced in the liver (Andersen et al. 1993, pp. 33)): prothrombin level measured at time zero.
- current prothrombin (time-dependent): prothrombin level measured most recently.

Based on a preliminary analysis, the prothrombin variables have been transformed such that if $V(t)$ is the untransformed prothrombin value (baseline prothrombin or current prothrombin) then we have

$$\text{Transformed Prothrombin} = \begin{cases} V(t) - 70 & \text{if} \quad V(t) < 70, \\ 0 & \text{if} \quad V(t) \geq 70. \end{cases}$$

Note that a prothrombin value above 70 is considered to be "normal". Thus the transformed prothrombin is zero for "normal" prothrombin levels and negative when the level is lower than "normal".

3.2 Illustration of the underestimation of treatment effect

Assume that we failed to recognize that current prothrombin might be an internal covariate, thus moved straight ahead and analysed the data as a traditional additive hazard model, giving the estimators (5) for the effect of the covariates. The lower panel of Figure 1 shows the effect of treatment in the model with all covariates. Using the concepts of direct and indirect effect of path analysis, we recognize this effect as the direct effect of treatment. However, if we remove current prothrombin from the analysis, we get the treatment effect of the upper panel of Figure 1 which is a significantly larger effect. This effect is the total effect (see e.g. Fosen et. al 2004), and is the effect of interest. Instead of simply doing this regression where the internal time-dependent covariate (current prothrombin) is removed from the analysis, our main focus in this article is to disentangle to what extent the total effect of treatment as well as the other fixed covariates is working directly or indirectly via current prothrombin, and to better understand the relation between the fixed covariates. Note that for treatment the indirect effect is the difference between the upper and lower panels of Figure 1.

3.3 Graphical representation of dynamic path models

We can use a time-indexed DAG to model the underlying data-generating process since all associations are directed, and Figure 2 is the DAG of one possible model. Here, $N(t)$ is the counting process recording the death of an individual, and $X_1, ..., X_6, Z(t)$ are the covariates treatment, sex, age, acetylcholinesterase, inflammation, baseline prothrombin and current prothrombin respectively.

In Figure 2, the variables $X_1, ..., X_6$ play identical rôles in the graph in the sense that there aren't any edges between them, and they potentially have an identical set of parents and children, although the analysis might reveal that some of the direct effects between parents and children may be zero. We say that these variables are on equal footing in this model, and for the sake of visual clarity, we will refer to variables on equal footing as a block. Figure 3 shows a block representation of the DAG of Figure 2. Note that a directed edge between two blocks means that there might be directed edges between all combinations of variables in these two blocks. Our purpose of a block representation is only to give a schematic representation of the relation between variables. We will not use any blocks in the dynamic path analysis, but only place variables from the same block close to each other. On the contrary, blocks are an integrated part of *chain graph models* (see

Cox & Wemuth 1996, Chapter 2). Blocks are then used throughout the analysis, and there are not edges between blocks, only between variables. Note also that the graphs in chain graph models are directed from right to left.

Since we are not going to use any blocks in the analysis, we will adopt the usual path analysis notation of representing variables by rectangles**.**

3.4 Estimation of direct, indirect, and total effects

We assume relation (1) to hold among the covariates, and since $N(t)$ is the main outcome, we can perform a dynamic path analysis as described in Sections 2.4 and 2.5.

When fitting the path analysis model for the model outlined in in Figure 3, we use $d\hat{B}_{x_h}(t)$ and $d\hat{B}_z(t)$ for the estimated direct effect of X_h and $Z(t)$ on $dN(t)$. For the estimated direct effect of X_h on $Z(t)$ we will use $\hat{\psi}_{x_h z}(t)$ instead of the more general $\hat{\psi}_{x_h c_{11}^h}$ from Section 2.4 since $Z(t)$ here replaces the general notation $X_{c_{11}^h}(t)$ for the first child along the first indirect path of the parent X_h.

The path between X_h and $dN(t)$ being mediated through $Z(t)$ is an indirect effect. By (6) this effect is $\psi_{x_h z}(t)dB_z(t)$. Since we only have one child along the path between X_h and $dN(t)$, thus only one mediating variable, we have no terms in the product $\prod_{l=1}^{r_k-1}$ in (6). Further, since this indirect path is the only indirect path between X_h and $dN(t)$, we have $m = 1$. The cumulative indirect effect then is

$$\text{ind}(X_h \to N(t)) = \int_0^t \psi_{x_h z}(s)dB_z(s),$$

and when inserting the estimators we get

$$\widehat{\text{ind}}(X_h \to N(t)) = \sum_{T_k \le t} \hat{\psi}_{x_h z}(T_k)\Delta\hat{B}_z(T_k).$$

The interpretation of the indirect effect is simply that X_h is working directly on $Z(t)$ and $Z(t)$ is working directly on $dN(t)$. If for instance X_h has a negative effect on $Z(t)$ and $Z(t)$ has a negative effect on $dN(t)$, then X_h has a positive indirect effect on $dN(t)$.

The total cumulative effect is then by (9)

$$\text{tot}(X_h \to N(t)) = B_{x_h}(t) + \int_0^t \psi_{x_h z}(t)dB_z(s), \tag{11}$$

and we estimate it with

$$\widehat{\text{tot}}(X_h \to N(t)) = \sum_{T_k \leq t} \Delta \hat{B}_{x_h}(T_k) + \sum_{T_k \leq t} \hat{\psi}_{x_h z}(T_k) \Delta \hat{B}_z(T_k)$$

The model of Figure 3 is a simple one. A model taking into account that age and sex is determined long before all other covariates, is given in Figure 4. We then have four blocks, that is four groups of variables being on equal footing in the graphical model. Being the exogenous variables, age and sex constitute block 1. The variables being only ascendants of block 1 variables, i.e. acetylcholinesterase, inflammation, baseline prothrombin, and treatment constitute block 2. Note, however, that we can tell a priori that there is no edge from block 1 to treatment since treatment is randomized. As a descendant of block 1 and block 2 variables, current prothrombin is block 3, and the outcome $dN(t)$ is block 4. Since we have only a small number of blocks, it is notationally convenient to let U, X, and Z denote the variables of block 1, block 2, and block 3 respectively instead of denoting all variables $X_1, ..., X_p$.

Now, Equation (7) for our extended model becomes

$$\text{tot}(X_h \to dN(t)) = \underbrace{dB_{x_h}(t)}_{\text{direct effect}} + \underbrace{\psi_{x_h}(t) dB_z(t)}_{\text{indirect effect}} \tag{12}$$

for the total effects of block 2 variables on the outcome block 4. The equation for the total effect of block 1 variables on the outcome is, with p_x being the number of covariates in block 2,

$$\begin{aligned} \text{tot}(U_h \to dN(t)) = & \underbrace{dB_{u_h}(t)}_{\text{direct effect}} + \sum_{i=1}^{p_x} \underbrace{\psi_{u_h x_i}(t) dB_{x_i}(t)}_{\text{indir. eff. through block 2}} \\ & + \underbrace{\psi_{u_h, z}(t) dB_z(t)}_{\text{indir. eff. through block 3}} + \sum_{i=1}^{p_x} \underbrace{\psi_{u_h x_i}(t) \psi_{x_i z}(t) dB_z(t)}_{\text{indir. eff. through block 2, and 3}} \end{aligned} \tag{13}$$

since we sum over all indirect effects only through block 2, all indirect effects only through block 3 (only one since only one $Z(t)$) and finally all indirect effects through both block 2 and 3. In a similar way,

$$\text{tot}(U_h \to Z(t)) = \underbrace{\psi_{u_h z}(t)}_{\text{direct effect}} + \sum_{i=1}^{p_x} \underbrace{\psi_{u_h x_i} \psi_{x_i z}(t)}_{\text{indirect effects}},$$

is the total effect of block 1 variables on block 3 variables.

Combining (10) and (12), the cumulative total effects of block 2 variables on the outcome are estimated by

$$\widehat{\text{tot}}(X_h \to N(t)) = \sum_{T_k \le t} \Delta \hat{B}_{x_h}(T_k) + \sum_{T_k \le t} \hat{\psi}_{x_h z}(T_k) \Delta \hat{B}_z(T_k), \qquad (14)$$

while the estimated effect of block 1 variables by (10) and (13) become

$$\begin{aligned} \widehat{\text{tot}}(U_h \to N(t)) = & \sum_{T_k \le t} \Delta \hat{B}_{u_h}(T_k) \\ & + \sum_{T_k \le t} \sum_{i=1}^{p_x} \hat{\psi}_{u_h x_i}(T_k) \Delta \hat{B}_{x_i}(T_k) \\ & + \sum_{T_k \le t} \hat{\psi}_{u_h z}(T_k) \Delta \hat{B}_z(T_k) \\ & + \sum_{T_k \le t} \sum_{i=1}^{p_x} \hat{\psi}_{u_h x_i}(T_k) \hat{\psi}_{x_i z}(T_k) \Delta \hat{B}_z(T_k). \end{aligned} \qquad (15)$$

Since we define a path analysis model at each time, this means that some edges might be non-zero at some times T_k and then zero at others. Typically we will have one path model for the interval $[0, \tau_1)$ and then another one for $[\tau_1, \tau_2)$ etc. To account for the change of model, for $T_k \in [\tau_1, \tau_2)$, some of the terms in (14) and (15) may become zero for $T_k > \tau_1$ that were not zero for $T_k \in [0, \tau_1)$ or vice versa.

3.5 The effects as functions of time

Based on the bootstrap inspection method described in Section 2.5, we choose the model with the path diagram given in Figure 5**.** We now want to study the estimated effects as functions of time. For convenience we will then refer to 'block i regression' as the regression of a block i covariate onto its parents. The results are given in Figure 6 for the block 2 regressions and in Figure 7 for the block 3 results. In order to get a better impression of the underlying structure, the estimates could be smoothed using a local scatter plot smoother (Cleveland 1979) or by local kernel regression (Hastie et al. 2001, section 6.1.2), but since the estimates are fairly stable we omit this. For the direct effects on $dN(t)$ we use the cumulative regression functions (5), and Figure 8 shows the results.

The direct effects of age and sex on the block 2 covariates acetylcholinesterase and inflammation, is varying with time even though all covariates

involved are constant in time. The time-dependence of the regression function is thus only due to changes of risk sets over time. The individuals in this study were followed to death or to the end of the study. Therefore, due to the randomization, censoring does not depend on covariates, and any systematic changes in the composition of the risk sets are due to death. Thus it seems from the right-hand panel of Figure 6 that sex might be more important for the presence of inflammation for those surviving beyond four years than for those surviving only a year: men are less susceptible to aqcuiring inflammation, and even more so for the group of survivors. This result is however uncertain. We also note that in the presence of a large proportion of lost-to-followup cases, the explanation could be selective censoring.

The estimated negative direct effect of age on acetylcholinesterase is less precise for the group of survivors than for all the individuals together but this can partly be explained by variation due to the decreasing risk set as time goes by.

Regarding the block 3 effects we see in Figure 7 that the direct effect of baseline prothrombin on current prothrombin is clearly positive until six years after randomization. Treatment is leading to an increase in current prothrombin, particularly for the first couple of years. Acetylcholinesterase has a positive effect on current prothrombin as well, but after about a year, the effect decreases and seems to vanish.

At block 4 (Figure 8) there is a clear positive direct effect of age on death and clear negative direct effect of inflammation as well as of current prothrombin.

The direct effect of treatment on $dN(t)$ is very small if present at all, as seen in the lower panel of Figure 1, while the total effect is larger (upper panel). By the decomposition of total effect, the difference between these two effects is the indirect effect working through current prothrombin. The indirect effect is given in the upper left panel of Figure 9 (and is also the product of the 1st panel of Figure 7 and increments of 3rd panel of Figure 8). We see that treatment leads to fewer deaths, and the way that this is happening is that treatment increases current prothrombin (upper left panel of Figure 7) and high level of current prothrombin decreases $dN(t)$, hence since the indirect effect is a product of these two effects, there is a negative indirect effect of treatment on $dN(t)$. Since there is no direct effect, all treatment effect is working through current prothrombin.

The indirect effect of baseline prothrombin on $dN(t)$ (upper right panel of Figure 9) is clearly negative. Note that the direction of this effect should be the same as that of current prothrombin since the two prothrombin variables are positively correlated.

There is an estimated negative indirect effect of acetylcholinesterase on $dN(t)$ but only until one year after randomization after which the effect seems to vanish (middle left panel of Figure 9).

Sex has an effect on $dN(t)$ indirectly through inflammation and it is positive (middle right panel of Figure 9). Men will have higher risk of death than women because men have less inflammation (Figure 6), and less inflammation leads to more deaths (Figure 8).

The indirect effect of age on $dN(t)$ is working through acetylcholinesterase and current prothrombin (Figure 5) and is positive because higher age leads to less acetylcholinesterase (Figure 6) and less acetylcholinesterase leads to less current prothrombin (Figure 7), and less current prothrombin leads to more deaths (Figure 8). The indirect effect of age seems to vanish after one year (lower left panel of Figure 9). The indirect path from age to $dN(t)$ contains the indirect path from acetylcholinesterase to $dN(t)$, the only difference is the edge from age to acetylcholinesterase. At each time of event, the indirect effect of age is thus equal to the indirect effect of acetylcholinesterase multiplied by the direct effect of age on acetylcholinesterase. Since the latter effect is negative and not changing much in time, the two indirect effects have a similar shape, only with different signs.

Age is working both directly and indirectly on $dN(t)$. Since the sign of the direct effect (Figure 8) and the indirect effect are identical, the mediation proportion (proportion of total effect being indirect) can be calculated. Figure 10 shows that the mediation proportion is small. It seems to be largest during the first year, but this might be caused by noise. As we noticed above, after one year the indirect effect might no longer be positive, and this also is the pattern for the mediation proportion.

4 Discussion

We have in this paper proposed a dynamic graphical model combining the additive hazard model and classical path analysis. At each time of event we fit a path analysis model by ordinary least-squares for all regression models but one: when $dN(t)$ is the dependent variable, we fit the additive hazard model.

Instead of estimating the effect of e.g. treatment by simply removing the internal covariates from the regression analysis, we want to understand how treatment effect is related to the intermediate variables. Dynamic path analysis is a way of doing this, since we decompose the total effect into direct and indirect effects. This decomposition has the advantage that we can see

to what extent the effect works directly and how much and in what ways the effect is working indirectly. This enables us to better understand the mechanisms involved in the processes.

Our analysis shows clearly how the treatment effect is mediated through a transient effect on prothrombine. Hence, a more detailed picture emerges than one would usually have in survival analysis. One gets a joint picture of the development of prothrombine and survival, including the effect of treatment. In the same manner we have seen that baseline prothrombin and sex is working indirectly through current prothrombin and inflammation respectively. Age is both working directly as well as indirectly along the path age-acetylcholinesterase-current prothrombin.

We have used an identical path analysis model throughout the entire observation period. As stated in Section 3.4, this is not a requirement for using dynamic path analysis. A thorough investigation of the bootstrap confidence intervals of different models, shows that acetylcholinesterase ceases to have an effect on current prothrombin one years after randomization (Figure 7). Then, after another two years, treatment also ceases to have an effect. A more accurate cumulative effect estimator would then be to use these three models at the three intervals in question. As described at the end of Section 3.4, this would just mean that some terms in e.g. (15) become zero for some times of event. Nevertheless, for simplicity we have kept the model of Figure 5 throughout the whole observation period. The choice of model at each time of event is an interesting topic for further research.

Acknowledgement

Johan Fosen was sponsored by Norwegian Cancer Society and Egil Ferkingstad was supported by The National Programme for Research in Functional Genomics in Norway (FUGE) from The Research Council of Norway.

References

Aalen, O. O. (1980), A model for nonparametric regression analysis of counting processes, *in* S. Fienberg, J. Gani, J. Keifer & K. Krickeberg, eds, 'Mathematical Statistics and Probability Theory', number 2 *in* 'Lecture Notes in Statistics', Springer, pp. 1–25. Proceedings of the Sixth International Conference, Wisla (Poland), 1978.

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

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993), *Statistical models based on counting processes*, Springer Series in Statistics, Springer.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, Wiley Series in Probability and Mathematical Statistics, Wiley.

Bollen, K. A. & Long, J. S. (1993), *Testing Structural Equation Models*, SAGE Publications, Newbury Park, California.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**(368), 829–836.

Cox, D. R. & Wermuth, N. (1996), *Multivariate Dependencies*, Monographs on Statistics and Applied Probability 67, Chapman & Hall.

Didelez, V. (2000), Graphical Models for Event History Data Based on Local Independence, PhD thesis, Universität Dortmund, Germany.

Edwards, D. (2000), *Introduction to Graphical Modelling*, 2nd edn, Springer-Verlag.

Efron, B. (1981), 'Censored data and the bootstrap', *Journal of the American Statistical Association* **76**(374), 312–319.

Fosen, J., Borgan, Ø., Weedon-Fekjær, H. & Aalen, O. O. (2004), Path analysis for survival data with recurrent events, Statistical research report 9/2004, University of Oslo, Department of Mathematics. http://www.math.uio.no/eprint/stat_report/2004/09−04.html.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning-Data Mining, Inference and Prediction*, Springer Series in Statistics, Springer.

Hesslow, G. (1976), 'Discussion: Two notes on the probabilistic approach to causality', *Philosophy of Science* **43**, 290–292.

Kalbfleisch, J. & Prentice, R. L. (2002), *The statistical analysis of failure time data*, Wiley series in probability and statistics, 2 edn, Wiley.

Lauritzen, S. (1996), *Graphical Models*, Oxford Statistical Science Series, Oxford University Press.

Loehlin, J. C. (2004), *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*, 4th edn, Lawrence Erlbaum, Hillsdale, N.J.

Pearl, J. (2000), *Causality : Models, Reasoning, and Inference*, Cambridge University Press.

Pearl, J. (2001), Direct and indirect effects, *in* J. Breese & D. Koller, eds, 'Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence', Morgan Kaufmann Publishers, San Francisco, pp. 411–420.

Wright, S. (1921), 'Correlation and causation', *Journal of Agricultural Research* **20**, 557–585.

Wright, S. (1934), 'The method of path coefficients', *Annals of Mathematical Statistics* **5**, 161–215.

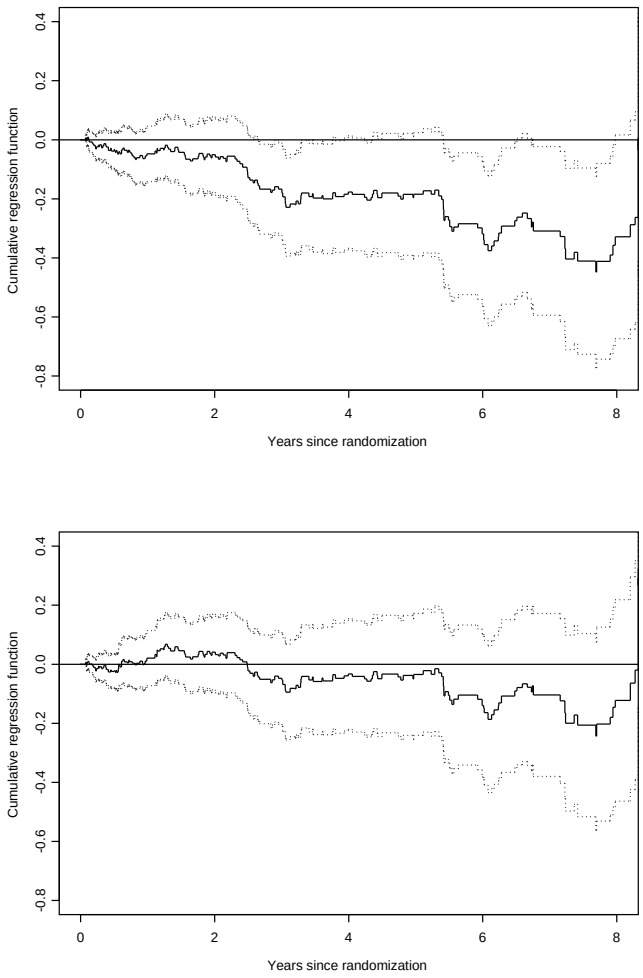

Figure 1: Cumulative regression functions estimates of treatment with 95% pointwise confidence intervals obtained by the standard martingale based method (splus-function aareg). The effect of treatment in the model with sex, age, treatment, acetylcholinesterase, inflammation, and baseline prothrombin (upper panel), and the effect of treatment when also current prothrombin is included in the model (lower panel).

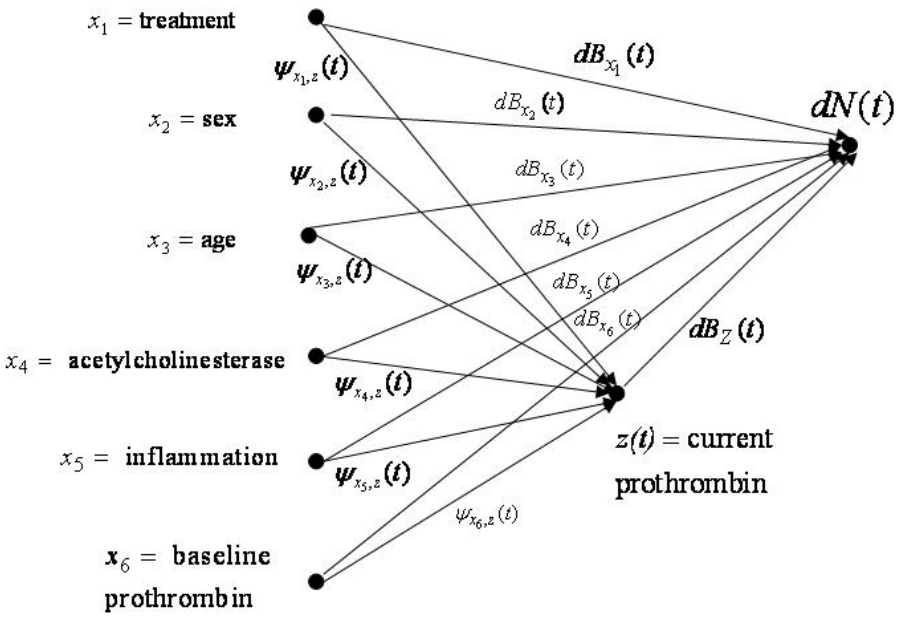

Figure 2: Directed acyclic graph (DAG) of one possible model of the liver cirrhosis data.

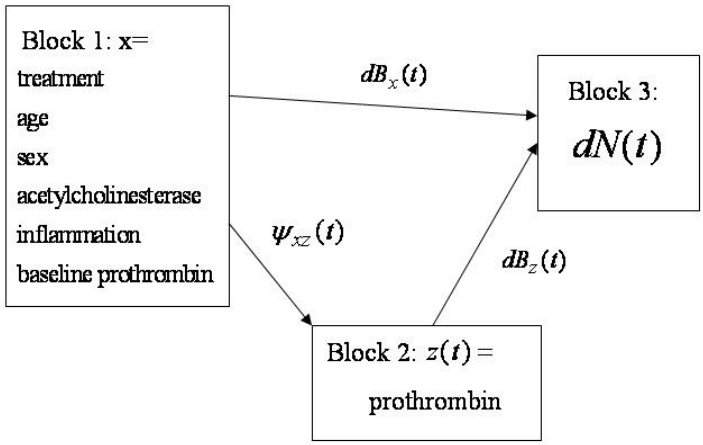

Figure 3: Block representation of path diagram of Figure 2.

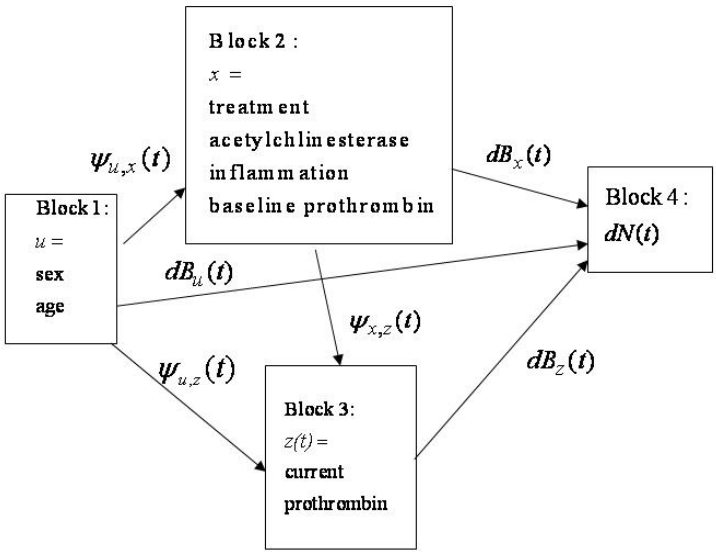

Figure 4: Block representation of a path model resembling that of Figure 3 but where the block of fixed covariates has been divided into a block of sex and age, and a block of treatment, acetylcholinesterase, inflammation, and baseline prothrombin. For treatment, the parameter $\psi_{u,x}$ is known to be zero due to randomization.

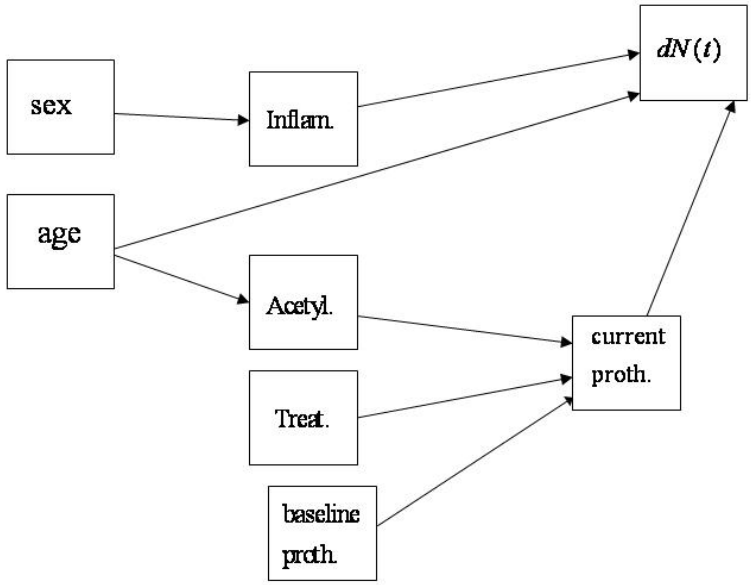

Figure 5: The path model best fitting the data.

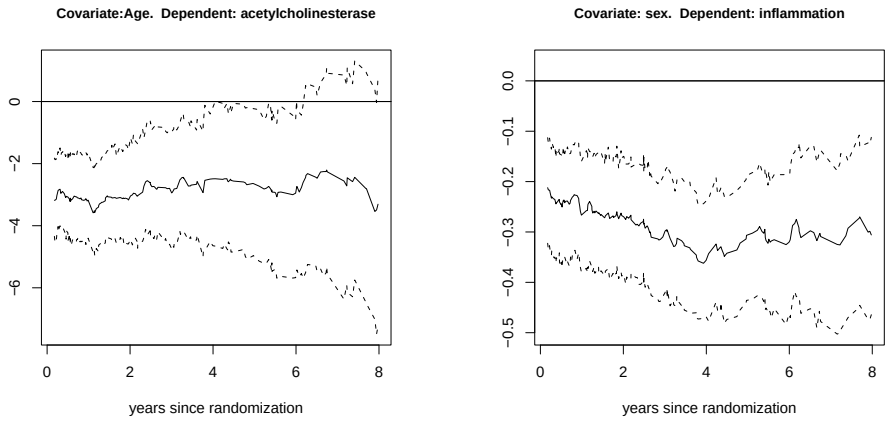

Figure 6: The estimated direct effects (regression functions) of the block 2 regressions: the effect of age on acetylcholinesterase (the left panel) and the effect of sex on inflammation (the right panel). Estimates with 95% bootstrap percentile pointwise confidence intervals, based on 500 bootstrap replications.

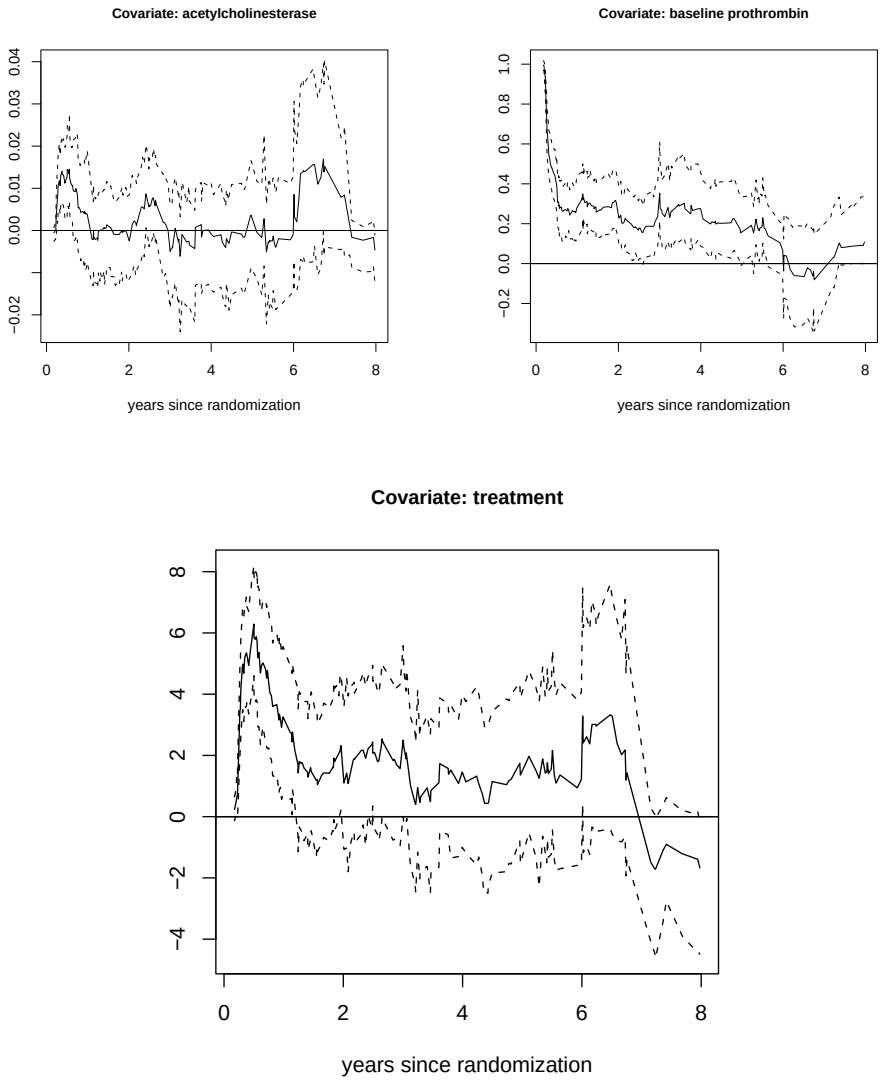

Figure 7: The estimated direct effects (regression functions) of the block 3 regressions: the effect of acetylcholinesterase, baseline prothrombin and treatment on current prothrombin. Estimates with 95% bootstrap percentile pointwise confidence intervals based on 500 bootstrap replications.

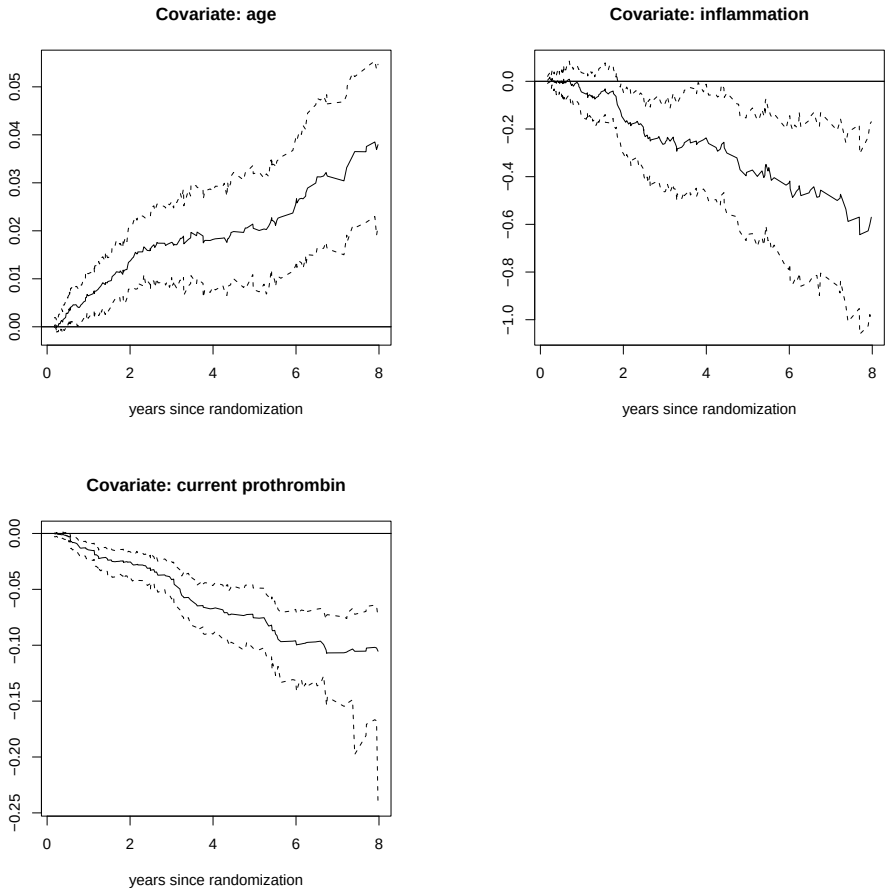

Figure 8: The estimated cumulative direct effects (cumulative regression functions) of the block 4 regressions: the effect of age, inflammation and prothrombin on $dN(t)$. Estimates with 95% bootstrap percentile pointwise confidence intervals, based on 500 bootstrap replications.

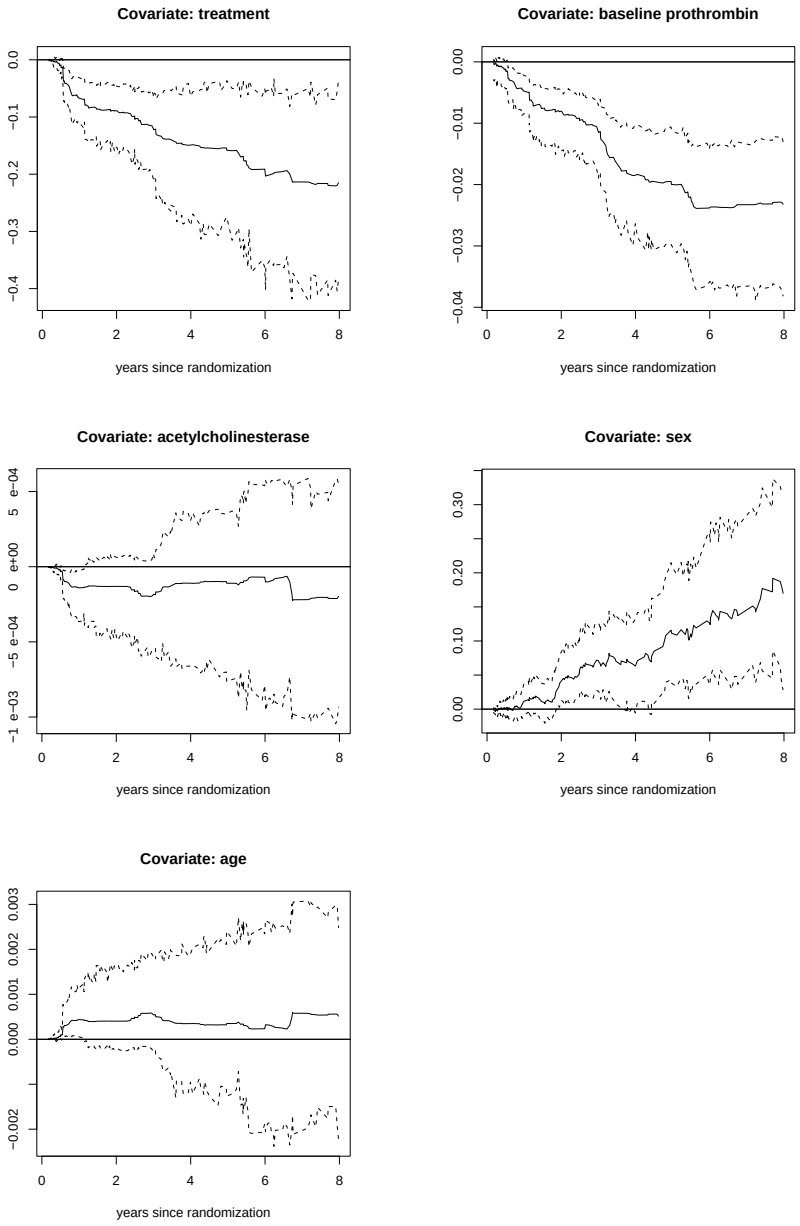

Figure 9: The estimated cumulative indirect effects on $dN(t)$ using the model of Figure 5: the effect of treatment through current prothrombin (the upper left panel), baseline prothrombin through current prothrombin (the upper right panel), acetylcholinesterase through current prothrombin (middle left panel), sex through inflammation (the middle right panel) and age through acetylcholinesterase and current prothrombin (the lower left panel). Estimates with 95% bootstrap percentile pointwise confidence intervals, based on 500 bootstrap replications.

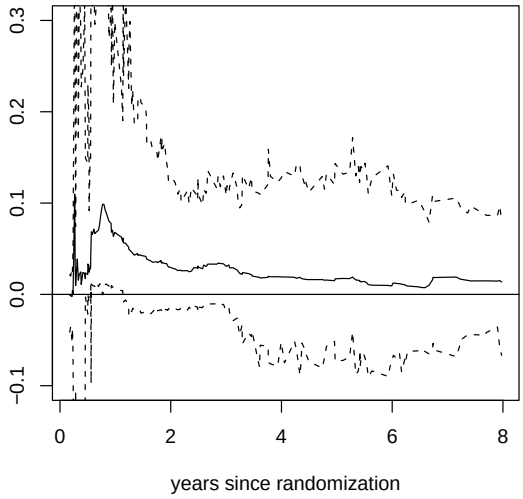

Figure 10: The mediation proportion (proportion of total effect working indirectly) of age on $dN(t)$. Estimates with 95% bootstrap percentile confidence interval, based on 500 bootstrap replications.