Geilo Winter School Jan 2014

Causal inference in longitudinal studies: An illustration

Elja Arjas

UH, THL, UiO

Motivation

- The purpose of this talk is to illustrate the important but often neglected role which time and, as a consequence, the use of modeling based on stochastic process formulations, have in the statistical treatment of causal problems.

- Another motivation is to provide a demonstration of how the tools offered by Bayesian modeling and inference, which likewise seem to have largely been ignored in the mainstream statistical literature on causality, can be usefully applied in this context.

Motivation (2)

- Causal concepts are best introduced and presented, not as abstract theoretical constructs, but as means that would facilitate an improved understanding of real world phenomena.

- In line with this, most space in this talk is allotted to considering an illustrative example, and introducing conceptual ideas and technical material only at the point at which they are needed for making progress.

Example: Type of daycare vs. AOM (Andreev & A 1998; A & Andreev 2000)

- **Problem description:** We are concerned with the aetiology of acute middle ear infection (acute otitis media, AOM) in small children.
- Earlier investigations have shown that the major risk factors for AOM are time-dependent: age, duration of breast-feeding, type of day care, and previous AOM history.
- For example, it is generally believed that breast-feeding has a protective effect, that the effect size depends on the age of the child, and that it will last for some time after breast-feeding has ended.

AOM: background

- **Causal question:** How does the type of daycare (home care, daycare in a family, daycare in a nursery / kindergarten) influence the risk of AOM in small children?

- **Data:** The study was based on a sample of 965 children from Oulu region (Finland), born between July 1 1985 and June 30 1986, who were followed-up for episodes of AOM maximally to the age of 33 months, with mean length of follow-up of 20.4 months.

AOM: data

For a child C_i indexed with i ($1 \le i \le 965$), the data contain values of the variables

- db_i = birth date of C_i
- de_i = date of end of follow-up of C_i
- dbf_i = date at which breastfeeding of C_i was stopped
- dca_i = date at which day care of C_i outside home began
- $tyca_i$ = type of day care of C_i after dca_i
  = 1 if in family care,
  = 2 if in nursing home/kindergarten
- t_{ij} = date at which the j^{th} episode of AOM in C_i was diagnosed.

AOM: data (2)

- In each case, the dates dbf_i, dca_i and t_{ij} can be right censored at de_i, and the value of $tyca_i$ remains undefined if $de_i < dca_i$.

- Such data can be interpreted, in an obvious manner, as a realization of a suitably defined Marked Point Process (MPP) or of a corresponding multivariate counting process.

- These processes can be considered either separately for each child, or also jointly for all children in the data.

AOM: ’*Real time approach’*

- In what follows we use calendar time as the basis of all considerations relating to effects of time. This is so for two reasons:

- Calendar time may itself represent an important causal factor due to the varying environmental infection pressure which may affect at the same time all children living in the considered area.

- All other time readings, including the child’s age, can be determined on the basis of calendar time (see below), but not conversely.

AOM: Model

***Model and inference**: a* non-parametric Bayesian intensity / hazard model of the form

(1) $\lambda_i(t) = Z_i\, f_0(t)\, f_1(a_i(t), bf_i(t))\, f_2(dc_i(t), I_i(sm), I_i(si))\, Y_i(t)$,

was assumed for initiating an episode of AOM, where

- i = index of a child, t = calendar time,
- $f_0(t)$ = baseline factor representing overall infection pressure at time t
- Z_i = latent 'frailty parameter' representing unobserved individual susceptibility to repeated occurrences of AOM
- $a_i(t) = (t - db_i)^+$ = age at time t,
- $bf_i(t) = (\min(t, dbf_i) - db_i)^+$ = duration of breast feeding until time t,
- $dc_i(t)$ = type of daycare at time t,
- $I_i(sm)$= indicator of parental smoking,
- $I_i(si)$ = indicator of siblings in the family,
- $Y_i(t)$ = 'at risk' indicator of being included in the follow-up at time t
- f_1, f_2 = non-parametrically specified relative risk functions.

AOM: Model (2)

- The functions f_0, f_1 and f_2 are treated as model parameters, assumed to be common to all children and are ultimately estimated from the data by applying some Bayesian non-parametric method. (Details are omitted here.)

- Including them in the definition of the AOM-intensity together with a multiplicative frailty parameter Z_i means that the resulting intensity / hazard is here specified relative to histories of the form

(2) $$\mathcal{F}_i(t) = \sigma\{f_0, f_1, f_2, Z_i\} \vee \mathcal{H}_i(t),$$

where

(3) $$\mathcal{H}_i(t) = \sigma\{Y_i(s), I_i(\mathrm{sm}), I_i(\mathrm{si}), a_i(s), bf_i(s), dc_i(s), I_i(t_{ij} \leq s), j \geq 1, s \leq t\},$$

is the internal history of child C_i up to time t, $t > db_i$, on the level of information that is available in the data.

AOM: Model (3)

- Since the values of f_0, f_1, f_2,Z_i , I_i(sm) and I_i(si) are all specified by the history $\mathcal{F}_i(t)$ at time $t = db_i$ at which child C_i was born, the AOM-intensity relative to $\mathcal{F}_i(t)$ is a deterministic function of t until the next recorded event time on that child in the data.

- At time $t = \min\{dbf_i, dca_i, de_i\}$ the intensity is then instantaneously updated to a new value in a way which depends on what particular event of these three possibilities happened first.

AOM: Model (4)

- If the first event was that breastfeeding was stopped, this information is coded into the definition (1) of $\lambda_i(t)$ by fixing, for $t > dbf_i$, the duration $bf_i(t)$ of breastfeeding at value $dbf_i – db_i$.

- If child C_i was then moved from home to some different type of day care during the follow-up, i.e., $dca_i < de_i$, this information is coded into (1) by changing, for times $t > dca_i$, the value of the variable $dc_i(t)$ from 0 corresponding to home care to either 1 or 2. After $t = de_i$ the intensity $\lambda_i(t)$ becomes zero.

AOM: Model (5)

- In this simple version, the value of AOM-intensity relative to the histories $\mathcal{F}_i(t)$ is not being updated at the times t_{ij} at which new AOM-infections are diagnosed. The reason is that this intensity is already conditional on the (latent) susceptibility parameter Z_i.

- But the intensity relative to the smaller internal histories $\mathcal{H}_i(t)$ would jump upwards at the times at which new AOM episodes are recorded. Applying Bayes' rule, ”Each new episode is an indication of a higher individual susceptibility to AOM”, and this corresponds to a (stochastically) larger value of parameter Z_i.

- In principle, the model for AOM intensity could be specified directly relative to the $(\mathcal{H}_i(t))$-histories. In practice, this would be very hard because of the widely different behavior of the individuals in the data.

AOM: Model (6)

- ‘Packaging’ of several covariates within the same non-parametrically specified relative risk function (here f_1 *and* f_2) allows for accounting for potentially complicated interactions of covariate effects.

- When considering the form of the ‘environmental risk’ component $f_2(dc_i(t), I_i(sm), I_i(si))$ we assumed that, for each value of $dc_i(t)$, the risk could only be increasing (= non-decreasing) in $I_i(sm)$ and $I_i(si)$. No corresponding *a priori* monotonicity property in $dc_i(t)$ was postulated.

- This is a concrete example of using existing (epidemiological) prior information for ’regularizing’ the estimates of non-parametrically specified parts in the model.

- For calibration, constraints $f_1(0, 0) = 1$ and $f_2(0, 0, 0) = 1$ were used.

AOM: Inference

- The likelihood contribution from the AOM events of child C_i is of the standard Poisson form

(4) $$\prod_j \lambda_i(t_{ij}) \exp\{- \int \lambda_i(s)\, ds\}.$$

- Note here that $\lambda_i(t) = 0$ outside the interval (db_i, de_i). If there were no AOM episodes for child i during the follow-up, we set $\prod_j \lambda_i(t_{ij}) = 1$.

- In the inferential problem based on a dynamic stochastic process formulation one needs to check whether also one or more of the events appearing at times de_i, dbf_i or dca_i in the data might result in non-trivial contributions to the overall likelihood.

AOM: Inference (2)

- Starting from de_i, it seems reasonable to assume that the times de_i of right-censoring are non-informative for such inference.

- This means that in an MPP model for all the data, considered as a realization in calendar time, the $\mathcal{F}_i(t)$-intensities for the right-censoring events would not depend on the model parameters f_0, f_1, f_2,Z_i of interest.

- Another way of saying the same thing would be: the $\mathcal{F}_i(t)$-intensities for the right-censoring events coincide with the corresponding $\mathcal{H}_i(t)$-intensities. This condition is called local independence.

AOM: Inference (3)

- Under this condition, the likelihood contribution due to the right-censoring events can be treated as a proportionality constant with respect to these parameters, and can therefore be ignored in likelihood-based (including Bayesian) inference. Thus there is no need to specify a model for the censoring events. (This is standard practice in ‘survival analysis’.)

AOM: Inference (4)

- Analogous conditions are now assumed to hold concerning the likelihood contributions coming from observing dates dbf_i at which breastfeeding was stopped, and dates dca_i at which a child was transferred to a new type of daycare.

- From the perspective of statistical inference, the events to stop breastfeeding or to transfer the child from home to some other type of daycare can be treated as being exogenous, or as if they were results of the well known ‘do’-conditioning operations of Pearl (1995).

AOM: Inference (5)

- These local independence / non-informativity conditions correspond to the kind of reasoning through which a statistician, or an epidemiologist, would normally have to go when contemplating about the possible presence of confounders in a planned causal analysis of observational data.

- They can be viewed as being dynamic versions of the well known and crucially important postulate of ’no unobserved confounders’ (e.g. Robins, 1986), or of ’strong ignorability’ (Rosenbaum and Rubin, 1983); these definitions are based on the concept of counterfactual or potential outcome.

AOM: Inference (6)

- In the large body of causality literature using (static) graphical models the corresponding condition is called the backdoor criterion, a term introduced by Pearl (1993).

- The concept of local independence was originally introduced in a somewhat different context by Schweder (1970), and has been considered later, e.g., by Aalen *et al.* (1980), Didelez (2008), and Arjas (2012). It is closely linked to the concept of Granger-causality used in time series analysis.

AOM: Inference (7)

- Making now explicit use of the concept of local independence, we summarize the assumptions in the above discussion into:

Assumption A1. For each $1 \le i \le N$,

(i) the AOM-intensities, when considered relative to the histories $\mathcal{F}_i(t)$, $t > 0$, are specified by formula (1);

(ii) the intensities of the events occurring at times db_i (birth), de_i (ending follow-up), dbf_i (stopping breastfeeding) and (dca_i, $tyca_i$) (transferring child C_i from home to a different type of day care), when considered relative to the histories $\mathcal{F}_i(t)$, $t > 0$, are locally independent from $\sigma\{f_0, f_1, f_2, Z_i\}$.

AOM: Inference (8)

- Interpretation of **A1** (ii): For example, deciding to transfer the child to a different type of day care (a ‘treatment assignment’) can depend on the past history of the child that is recorded in the data, including earlier episodes of AOM, but given that history, is conditionally independent of f_0, f_1, f_2 and Z_i.

- Due to the local independence postulate of **A1** (ii) we have actually been able to avoid the task of providing an explicit specification of models for the events at times db_i, de_i, dbf_i and dca_i.

- Up to a proportionality factor not depending on the latent variables f_0, f_1, $f_2$and Z_i, the likelihood arising from observing the complete data on child C_i retains the simple form (4), which was previously derived to correspond to the AOM episode data on that child.

AOM: Inference (9)

- Two more steps are still needed. First, we need to extend the above considerations from an individual child C$_i$ to all children in the data:

Assumption A2. The pairs $(\sigma\{Z_i\} \vee \mathcal{H}_i(\infty))$, $1 \leq i \leq N$, are conditionally independent given the common link functions f_0, f_1 and f_2, with the variables Z_i following the same distribution.

- When considered together, these assumptions **A1** and **A2** can be seen as postulating an exchangeability property of the children in the data, but only as far as it relates to specifying a model for AOM.

AOM: Inference (10)

- Under these assumptions, the overall likelihood becomes a product, over *i*, of expressions (4), then having the form

(5) $$\prod_{l} \prod_{j} \lambda_l(t_{lj}) \exp\{- \int \lambda_l(s)\, ds\}.$$

- If child C_l had no diagnosed AOM infections at all in the data, then we interpret the product $\prod_{j} \lambda_l(t_{lj})$ as being equal to 1.

AOM: Inference (11)

- For the proposed Bayesian approach for inference we still need to set up a prior for the latent variables in the model. In view of the independence assumption **A2**, it suffices to set up a prior for (f_0, f_1, f_2) and then a prior for all Z_i given these.

- The specification of such a joint prior is commonly done by postulating a product form for it:

$$p(f_0, f_1, f_2, Z_i;\ 1 \le i \le N) = p_0(f_0)\ p_1(f_1)\ p_2(f_2)\ p(\varphi)\ \Pi_i\ \ p_z(Z_i \mid \varphi) \tag{6}$$

for suitably chosen density functions p_0, p_1, p_2 , p and p_z.

- Here the common prior distribution p_z of the susceptibility parameters is assumed to be dependent on a parameter φ, whose distribution depends further on suitably chosen hyper-parameters. Often a convenient choice would be to postulate p_z to be either Gamma or log-normal. The assumed hierarchical model structure then allows its parameters to be updated from the data into a corresponding posterior.

AOM: Inference (12)

- **Remark.** Although assumption **A2**, according to which the observed individual histories are conditionally independent given the common link functions, seems appropriate in the present applied context, a comparable assumption may be too restrictive in some other applications.

- The concept of local independence can still be usefully employed, however, but then it needs to be considered in an extended form where all events in the data are first ordered according to the calendar time in which they occurred, thereby forming a ‘large’ multivariate marked point process as the superposition of all the individual processes.

- Details of such an extension are omitted in this presentation.

AOM: Using predictive probabilities in answering causal questions

- The causal question considered was as follows: How does the type day care (home care, day care in a family, or day care in a nursery/kindergarten) influence the risk of AOM in small children?

- More specifically, as an illustration, consider a situation in which the parents of a 14 month old daughter would contemplate between such alternative choices. For background, suppose the child has had ear infections at ages 9 and 13 months, breastfeeding lasted for 3 months, there are two siblings, and one parent is a smoker.

- Given this background, what can we say, on the basis of our statistical model and the information contained in the data, about future incidences of ear infection that this child might experience, provided that a particular day care option is chosen?

AOM: Using predictive probabilities in answering causal questions (2)

- A convenient answer can be formulated in terms of corresponding predictive distributions of future AOM episodes beyond the time at which type of daycare is chosen, by considering the three alternatives

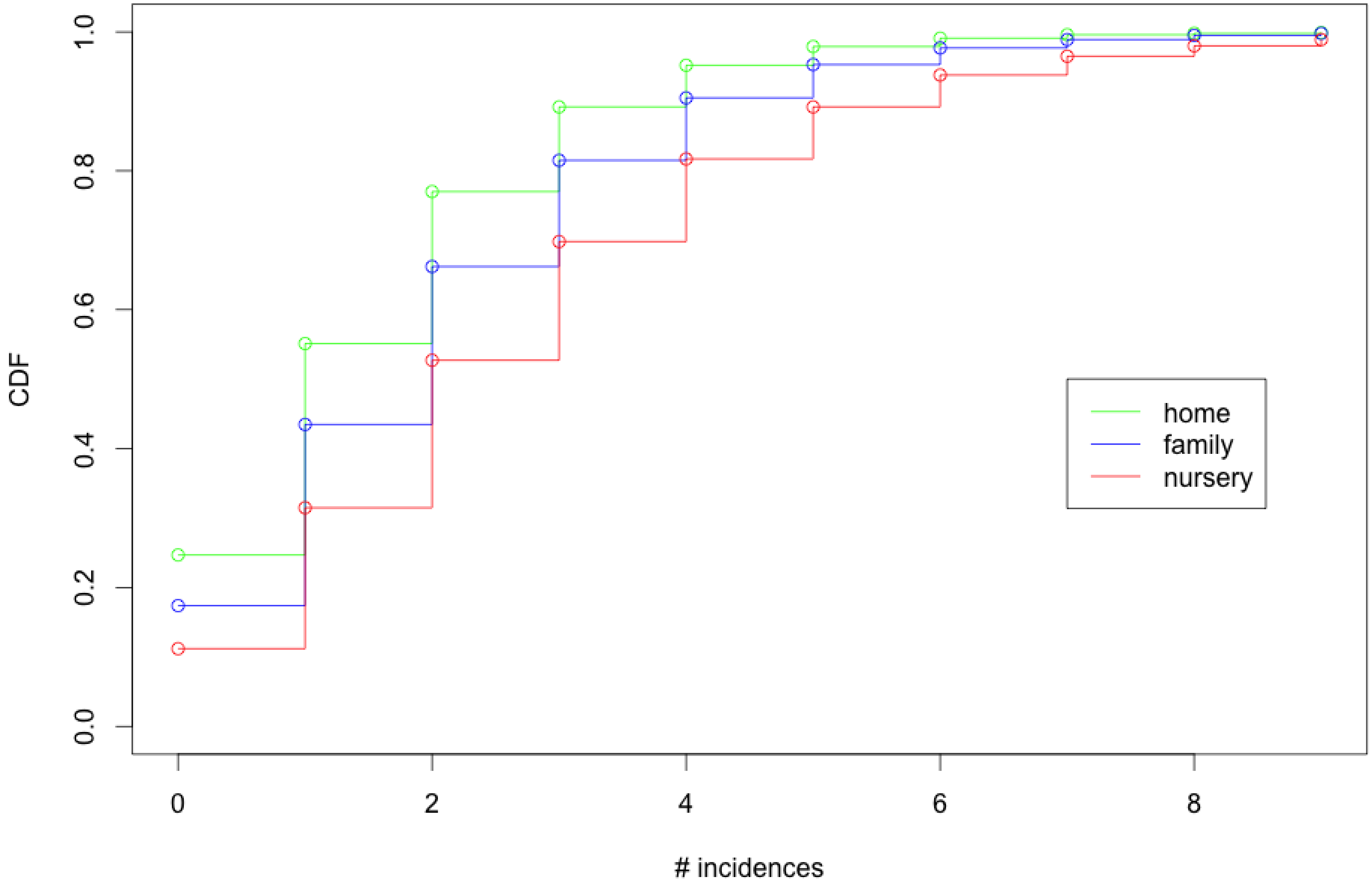
  home care / family day care / nursery day care.

- Numerical results can be obtained by making use of MCMC sampling and data augmentation, in a single Monte Carlo run.

AOM: Using predictive probabilities in answering causal questions (3)

- **Table.** Individual predictive probabilities of the number of future AOM incidences of a child during the age interval (14, 28], under three possible options of day care. (Based on the Oulu data, reproduced from Arjas and Andreev 2000).

		Type of day care		
	# incidences	home	family	nursery
•	0	0.247	0.174	0.112
•	1	0.304	0.261	0.203
•	2	0.219	0.227	0.212
•	3	0.122	0.153	0.171
•	4	0.060	0.090	0.119
•	5	0.027	0.048	0.075
•	6	0.012	0.024	0.046
•	7	0.005	0.012	0.027
•	8	0.002	0.006	0.015
•	9	0.001	0.003	0.009
• ·	≤9	0.999	0.997	0.989


1.0
0.8
0.6
0.4
0.2
0.0
CDF
0
2
4
6
8
# incidences
home
family
nursery

AOM: Using predictive probabilities in answering causal questions (4)

- More generally, it can be concluded that the number of AOM episodes is stochastically largest (w.r.t. the predictive distribution) in the case of kindergarten / nursery daycare, and smallest in the case of home care.
- Also expected numbers of such episodes, and their differences, can be computed with relative ease.

Final comments

- In causality modeling, essential aspects relating to time are often ignored (Notable exceptions: Odd Aalen, Vanessa Didelez, Daniel Commenges);
- When using graphical models, also explicit consideration of statistical inference is mostly by-passed, by assuming that there is a known joint distribution on the graph;
- In particular, Pearl wants to make a sharp distinction between causal postulates and statistics, by restricting the role of statistics (at its best) to identification of such a joint distribution.
- This is not how I see these things: In specifying the statistical model, one should use all relevant prior knowledge about the problem in question. This is actually demanded by the Bayesian approach! Moreover, probabilities are then viewed as quantitative expressions of the information that is available, and not as attributes of the considered physical objects or systems themselves.

Final comments (2)

- In reporting the findings from an empirical study, a majority of the statistical literature dealing explicitly with causality problems is concerned with hypothesis testing, relating to the parameters of simple (often simplistic!) statistical models.

- Often, however, the main problem is not in establishing that a causal effect exists, but rather in acquiring either

  - an improved understanding of the causal mechanism in question, or

  - a more refined understanding of different types of response.

- Performing such tasks may require elaborate statistical modelling, capable of dealing with different types of uncertainty and with problems arising from heterogeneity between individuals. (An Example: Individual responses to different types of HRT (Bhattacharjee & A (2005)).

Final comments (3)

- General formulations in continuous time, in terms of Marked Point Processes (MPP's) are readily available, leading to likelihood expressions of a standard (product) form.

- The product is over time, involving conditionings w.r.t. past histories (cf. product form in graphical models, where conditioning is on the parental nodes).

- Individual and treatment specific potential / counterfactual outcome random variables are not needed in the model specification. (They may have a role when explaining the meaning of the concept of predictive distribution.)

- In the general MPP framework, the 'no unmeasured confounders' postulate can be naturally expressed as a local independence condition, without referring to the concept of potential/counterfactual outcome.

Final comments (4)

- Technically the same local independence condition leads to simpler likelihood expressions in situations in which some aspects of the past history become redundant in the specification of the conditional intensities.

- Use of likelihood–based (or Bayesian) inference allows one to account for individual differences between subjects in the modelling, without a need to form ‘risk sets’ of exchangeable subjects and then consider corresponding estimators.

- Given that sufficient amounts of data are available, nonparametric Bayesian modelling (possibly of some suitably constrained form, e.g. assuming monotonicity), combined with algorithmic computational methods applying MCMC, offers an attractive and flexible alternative for statistical inference.

Final comments (5)

- The results from an empirical study should be preferably reported in the form of predictive distributions of the response of interest, with each such prediction corresponding to a specific (sequence of) intervention(s) or choice(s) of control variables.

- The required numerical computation of the predictive distribution(s) can mostly be carried out in a convenient manner by applying the technique of data augmentation, in a single run of the MCMC.

- Predictive distributions correspond to expressions obtained by applying the ‘g-computation algorithm’ of Robins, except that predictive distributions also account for the uncertainties in the estimated model parameters.

Final comments (6)

- In studies involving real data the computational challenge can become formidable - and even exceed what is feasible in practice.
- Nevertheless, I would view the flexibility, and the relative conceptual simplicity, of the present entirely probabilistic framework to be a valuable asset in the study of challenging causal problems.
- Stay within the domain of probability calculus as long as you can! This will lower your chances of getting stupid answers from your analysis …

Some references

- A. Andreev & E. Arjas: Acute middle ear infection in small children: a Bayesian analysis using multiple time scales. Lifetime Data Analysis 4 (1998): 121-137.
- E. Arjas: Causal inference from observational data: a Bayesian predictive approach. Book chapter in 'Causality: Statistical Perspectives & Applications' (C. Berzuini, A. P. Dawid, L. Bernardinelli, Eds.), Wiley (2012).
- E. Arjas & A. Andreev: Predictive inference, causal reasoning, and model assessment in nonparametric Bayesian analysis: a case study. Lifetime Data Analysis 6 (2000):187-205.
- E. Arjas & L. Liu: Assessing the losses caused by an industrial intervention: a hierarchical Bayesian approach. Applied Statistics **44** (1995): 357 - 368.
- E. Arjas: Time to consider time, and time to predict? Statistics in Biosciences, ‘Online First’ doi: 10.1007/s12561-013-9101-1

- Pdf-copies of my papers available on web page http://wiki.helsinki.fi/display/biometry/Elja+Arjas