HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Geilo Winter School Jan 2014

Causal inference in longitudinal studies: A Bayesian approach

Elja Arjas

UH, THL, UiO

Some background, for a start …

- After a very slow start, **causal inference** is now becoming a recognized field within statistics
- Its development has been much influenced by ideas coming from clinical trials (treatment assignment) and epidemiology (observational data, confounding)
- No deep philosophical (ontological, epistemological) issues relating to the existence / nature of causality in this talk …

Some background … (cont'd)

- Several statistical frameworks/approaches have been introduced for dealing with causality, e.g., using ideas based on
  - counterfactual random variables (Neyman, Rubin, Robins, …);
  - graphical models (Pearl, Lauritzen, Dawid, …)

- In spite of their giving probabilistic descriptions and statistical tools for considering similar problems, they have remarkably little in common.

- **'Time'** is an intrinsic part of all causal reasoning (in that: *'A cause must precede the effect in time*'), but this fact is rarely accounted for in the statistical modeling of causal problems.

Some background … (cont'd)

- In these more recent developments, also the Bayesian approach to inference has been almost completely ignored.
- This is remarkable in view of the fact that the approach of 'inverse probability' – a term used into the 1950'ies – was originally motivated as being a method that would provide causal explanations to observed facts (e.g., in astronomy).
- Hume vs. Bayes, Laplace, …

Simple “prototype setting” for causal inference

- Consider a time-ordered sequence of random variables (*U*, *X*, *A, Y),* providing them with the following interpretations:
- *U* and *X* are background variables, with *X* observed (‘covariate’) and *U* unobserved (‘potential confounder’);
- *A* is a contemplated causal variable (‘control’, or ‘treatment’), and
- *Y* is the considered ‘response’ or ‘outcome’ variable.

Simple “prototype setting” for causal inference

- We say that the unobserved variable *U is* a ‘potential confounder’ if the conditional distribution of the response *Y*, given (*U*, *X*, *A*), depends on *U*.

- If this is not the case, then (in this setting) we can simply forget about *U*.

Unconfounded causal inference in observational studies

- The Key Condition, in words:

  "The rule by which the value of the control variable *A* has been determined in the data is allowed to depend only on what has been observed in the past (= *X)*, but must not depend on the variables that have not been observed (= *U*)."

- This condition is automatically satisfied in experimental studies, where the choice of *A* ('treatment assignment') is under the control of the experimenter / statistician. RCT's are the prime example of this.

Unconfounded causal inference in observational studies

- But such a condition is needed in observational studies, to guarantee that **unconfounded** statistical inferences can be drawn in those situations.

- This leads to an idea where a direct comparison is made between these two types of designs: statistical inferences from data collected from an observational study should be ‘as if’ they would have come from an experiment with randomized treatment assignment.

Unconfounded causal inference in observational studies

- This motivates the use of two probability measures in our treatment of causality, and a corresponding notation:

  - P_{obs} is used as a description of the actual observational data;

  - P_{ex} is used as a description of a hypothetical experimental setting, where A would be randomized or 'exogenous'.

In a simple setting, the following Lemma (Lindley 2002, A & Parner 2004) formulates this idea in a precise manner:

- **Lemma.** Consider (U, X, A, Y) under P_{obs} and P_{ex} .

  Suppose that

  (i) The distribution of (U, X) is the same under P_{obs} and P_{ex};

  (ii) A is conditionally independent of U, given X, under P_{obs};

  (iii) The conditional distribution of Y, given (U, X, A), is the same under P_{obs} and P_{ex} .

  Then the posterior distribution of U, given (X, A), is the same under P_{obs} and P_{ex}, and it does not depend on A.

- Note: (ii) is automatically satisfied for P_{ex}.

Proof of the Lemma

- Proof: (with ‘α’ signifying proportionality in *u*)

$p_{obs}(u \mid x, a) \;\alpha\; p_{obs}(u, x)\, p_{obs}(a \mid u, x)$ by chain rule & Bayes

$= p_{obs}(u, x)\, p_{obs}(a \mid x)$ by (ii)

$\alpha\; p_{obs}(u, x)$

$= p_{ex}(u, x)$ by (i)

$= p_{ex}(u \mid x, a)$ by symmetry.

That this posterior does not depend on *a* is immediately seen from this proof. ■

Consequences for statistical modeling

- **Important observation:** If Condition (ii): $p_{obs}(a \mid u, x) = p_{obs}(a \mid x)$ is satisfied, the posterior $p_{obs}(u \mid x, a)$ does not depend on how $p_{obs}(a \mid x)$ would be specified (if it were!).

- In likelihood-based (including Bayesian) inference such models for treatment *a* would be viewed as proportionality constants, and can therefore be ignored.

- This forms a striking contrast to popular inferential methods based on ’inverse probability weighting’ (IPW), which make use of estimates of $p_{obs}(a \mid x)$, calling it ‘propensity score’.

Consequences for inference and prediction

- **Corollary.** Under the conditions of the Lemma,

  (i) $p_{obs}(u \mid x, a, y) = p_{ex}(u \mid x, a, y)$ 'posterior dsn, given the data'

  (ii) $p_{obs}(y \mid x, a) = p_{ex}(y \mid x, a)$ 'predictive' dsn for response.

- Proof:

  (i) $p_{obs}(u \mid x, a, y) \propto p_{obs}(u \mid x, a)\, p_{obs}(y \mid u, x, a)$ by Bayes

  $= p_{ex}(u \mid x, a)\, p_{ex}(y \mid u, x, a)$ by (iii) & Lemma

  $\propto p_{ex}(u \mid x, a, y)$ by Bayes

  (ii) $p_{obs}(y \mid x, a) = \int p_{obs}(u \mid x, a)\, p_{obs}(y \mid u, x, a)\, du$

  $= \int p_{ex}(u \mid x, a)\, p_{ex}(y \mid u, x, a)\, du$ by (iii) & Lemma

  $= p_{ex}(y \mid x, a).$ ■

Corollary explained

- "If in an observational study the likelihood of a contemplated cause (here *A*) does not depend on an unobserved (latent) variable / parameter (here *U*), then observation of *A* does not change the inferences concerning *U*".

- Therefore, when fixing *X* and *A* at their observed values and then predicting *Y* (but integrating *U* "away" from the conditional joint distribution of *Y* and *U*, given *X* and *A*), it makes no difference whether the value of *A* was a result of randomization, or merely "observed".

Corollary explained

- Put in a slightly different way: The response *Y* , given *U*, *X* and *A*, can be assumed to behave in the same way regardless of whether {*A* = *a*} was **done** or merely **seen**.

- Can write this as (Pearl, Lindley)

  $p(Y \mid X;\ \text{see}(A)) = p(Y \mid X;\ \text{do}(A))$.

- Note: The value of *A* influences the prediction *of Y* , but has no effect on the estimate of *U: A* may have a ”causal effect, forwards in time”, but no “inferential effect, backwards in time”.

Statistical inference: learning from data

- Suppose that our statistical inferences are based on observational data of the form $\{(X_i, A_i, Y_i); i = 1, 2, \ldots, n\}$, consisting of observations on n "exchangeable individuals drawn from P_{obs}" (cf. de Finetti).

Statistical inference (2)

- The corresponding unobserved variables U_i can contain characteristics of the considered individuals, and also (population) characteristics or structural parameters which are shared by all.

- This can be expressed by writing $U_i = (W_i , \theta)$.

- Although not observed, we assume that their values are similarly described by P_{obs}.

Statistical inference (3)

- Then **predictions of *Y*** (as in the Lemma) can be determined, *in the sense of (Bayesian)* **posterior predictive distribution given such data,**
  for a generic/hypothetical individual with chosen values of (baseline) covariate *x* and control *a.*

- This enables one to make a comparison of predictions of the form (as in the Lemma)

  $p(y \mid x, a, \text{data})$ vs. $p(y \mid x, a', \text{data})$

  (or corresponding expectations) with each other, where a and a' are two ‘causes’ whose effects are being considered.

Statistical inference (4)

- Here the subscripts have been dropped from p_{ex} or p_{obs} because they are now, under the assumptions of the Lemma, redundant.

- The ‘do’ notation (of J. Pearl) has been used to emphasize the idea of (in observational studies, only hypothetical) interventions whose effects are compared. Note also that the predictive distributions are calibrated to correspond to the same value of *x*. (However: the value of *x* is determined before treatment *a,* not after!)

How about connections to longitudinal and survival/duration analysis?

- This simple model structure can be extended in a fairly straightforward manner to **continuous time** and **event sequences** by using the framework of **marked point processes (MPP's).**

Data modeled as an MPP sample path

- Suppose that a random number N_τ events occur over the considered time interval $(0; \tau]$. At each event time T_k, covariates X_k are measured and an action, or treatment, A_k follows immediately upon this.

- Hence, the **recorded data** consist of $\{(T_k, (X_k, A_k)), k = 1, \ldots, N_\tau\}$, with $0 = T_0 < T_1 < T_2 < \ldots < T_{N_\tau}$, and finally, of a measured response Y.

Data modeled as an MPP sample path

- As a convention, and without restriction to generality, we can treat the considered response/outcome variable it as a marked point, identifying it with "the last observed covariate value" X_{N_T}.

- If we have data on n individuals indexed by $i = 1, 2, \ldots, n$, we can use a formulation in which the components of $Z_k = (X_k, A_k)$ are vectors with coordinates indexed according to the individuals.

Data modeled as an MPP sample path

- In some designs, e.g. randomized clinical trials, there may not be a covariate measurement X_k preceding a corresponding assignment A_k to a treatment, say a, in which case we could write ($Ø; a$) as the value of (X_k , A_k).

- On the other hand, in some sampling schemes a number of repeated covariate measurements are made before there is an actual assignment A_k to a treatment, and then we can similarly write ($x; Ø$).

Observed and unobserved (latent) processes

- The marked point process formalism is also able to accommodate **latent variables** and developments which are potentially relevant for describing the causal problem at hand but which were not observed.

- As before, we use the generic notation U for such variables, and then make the convention that they can be imbedded, as a sequence of additional random marks, into the marked point process.

- Having denoted by T_k the time of the k^{th} 'event ' in the considered MPP, we can extend the earlier notation of 'marks' to $Z_k = (U_k , X_k , A_k)$. Here (X_k , A_k) is **observed** in the data and U_k is **unobserved**.

Setting up a statistical model

- We now make the convention that the common ‘structural’ model parameters θ are imbedded into the latent mark U_0 as coordinates. Thus our inferences concerning the latent marks U_k , $k = 0, 1, 2, \ldots, N_T$, will cover also inferences on θ.

- In addition, U_0 could contain coordinates describing unobserved individual characteristics at the baseline.

Setting up a statistical model (2)

- Setting up a probability for the canonical sample paths of an MPP can be done by applying induction, always moving from a time point T_k to the next point at T_{k+1} and then considering it jointly with the corresponding mark $(U_{k+1}, X_{k+1}, A_{k+1})$.

- All this involves is sequential application of the chain multiplication rule!

Setting up a statistical model (3)

- In its k^{th} step we consider conditional probabilities of the form

  $p_{\text{obs}}(T_{k+1}, U_{k+1}, X_{k+1}, A_{k+1} \mid \mathcal{F}_k)$, where

  $$\mathcal{F}_k = \{(T_i, U_i, X_i, A_i ; i = 0, 1, \ldots, k\}$$

  is the (full) **history** of the marked point process up to time T_k.

- Denote similarly by

  $$\mathcal{H}_k = \{(T_i, X_i, A_i ; i = 0, 1, \ldots, k\}$$

  the **observed history** of the marked point process up to time T_k.

Unconfounded inference in the MPP setting

- **Definition**:

We say that a sequence of contemplated causal variables (A_k) in an observational study described by sample path $\mathcal{F}_{N_T} = \{(T_i, U_i, X_i, A_i); i = 0, 1, \ldots, N_T\}$ and probability p_{obs} is **unconfounded** relative to latent variables (U_k) if, for each k, A_k and $(U_i)_{0 \le i \le k}$ are conditionally independent given $(\mathcal{H}_{k-1}, T_k, X_k)$, that is,

$$p_{\text{obs}}(A_k \mid \mathcal{F}_{k-1}, T_k, U_k, X_k) = p_{\text{obs}}(A_k \mid \mathcal{H}_{k-1}, T_k, X_k);$$

$k = 1, 2, \ldots, N_T$.

Local independence

- The postulate of unconfounded inference could be stated as a **local independence** condition in continuous time *t,* (Schweder, 1970; Aalen 1987, Didelez, 2008).

- Technically, it says that the **local characteristics** in the statistical modeling of the sequence of treatment assignments are the same, regardless of whether they are considered relative to the observed histories $(\mathcal{H}_t)_{t>0}$ or to the 'full' histories $(\mathcal{F}_t)_{t>0}$ (the latter being generated, in addition, by the unobserved process $(U_t)_{t>0}$).

In words …

- This can be stated in words as follows:

Provided that the contemplated causal variable in an observational study is assumed to satisfy the **unconfounded inference / local independence** condition, it makes no difference when predicting a future response whether the value of the causal variable was “chosen” or merely “observed”.

Toy Example: Hospitals are hazardous places for many…

- Registry data provide ample evidence of the fact that hospitals are hazardous places, as most deaths (e.g., in Finland about 80 %) have happened in a hospital.

- So, why do we still want to have hospitals?

Hospitals … (2)

Could consider, here dropping the index *i* referring to an individual:

- $\mathcal{H}_k$ = observed (recorded in the data) individual baseline characteristics and pre-T_k history, including possible hospitalizations and death
- A_k = indicator of (possible) hospitalization at time T_k, contemplated causal variable
- Y = time of death, response
- $\mathcal{F}_k$ = 'full' individual pre-T_k history, including status of health (not recorded in the data)

Hospitals … (3)

- Unconfounded inference / Local independence in this case would mean that the probability of being taken to hospital would not depend on the individual's health status … which clearly does not make sense.

- Thus we cannot convert such information on hospital care and deaths the causal claim that "being taken to a hospital is more hazardous than not being taken".

P_{obs} and P_{ex} again …

- Let us now see how this postulate can be used in a context of a causal problem, viewing the observed variables (A_k) as “causes".

- Following the same idea as before, we connect the inferences, which can be drawn from the observational data and which are described in terms of a probability denoted by P_{obs}, to corresponding statements relative to another probability denoted P_{ex}.

P_{obs} and P_{ex} again … (2)

- To do so in the MPP framework, we link these two probabilities to each other by the following requirements:

$$p_{ex}(U_0, X_0) = p_{obs}(U_0, X_0)$$

and

$$p_{ex}(T_{k+1}, U_{k+1}, X_{k+1} \mid \mathcal{F}_k) = p_{obs}(T_{k+1}, U_{k+1}, X_{k+1} \mid \mathcal{F}_k),$$

$k = 0, 1, \ldots, N_T$.

Main result on inference

- **Theorem** Suppose that contemplated causal variables (A_k) are unconfounded in the above sense. Then, for each $k = 0, 1, \ldots, N_T$:

  The posterior distributions of the complete history $\mathcal{F}_k$, given the corresponding observed history $\mathcal{H}_k$, are the same in both schemes, that is,

  $$p_{\text{obs}}(\mathcal{F}_k \mid \mathcal{H}_k) = p_{\text{ex}}(\mathcal{F}_k \mid \mathcal{H}_k).$$

  Here neither of these posterior distributions depends on the latest treatment assignment A_k.

- The proof is by a repeated application of our first Lemma, which was concerned with (U, X, A, Y), to a sequence of 'marked points' in the MPP.

… and a Corollary concerning prediction

- The predictive distributions of the next marked point (T_{k+1}, U_{k+1}, X_{k+1}), given the corresponding observed history $\mathcal{H}_k$, are the same in both schemes, that is,

$$p_{\text{obs}}(T_{k+1}, U_{k+1}, X_{k+1}, \mid \mathcal{H}_k) = p_{\text{ex}}(T_{k+1}, U_{k+1}, X_{k+1}, \mid \mathcal{H}_k).$$

- This result extends our earlier Corollary concerning (U, X, A, Y).

Combining these two for treatment comparison

- The inferences that were drawn from the **data**

  $$\mathcal{H}_{N_T} = \{(T_i, X_i, A_i);\ i = 0, 1, \ldots, N_T\}$$

  on the common model parameter θ can then be utilized for predicting what will happen to “a generic individual” if he/she is to be given some specific sequence of treatments.

- Adding a star (*) to the notation to signify the considered generic individual, we would, in a simple case, be interested in **predicting the response *Y** under a given fixed sequence of “forced" treatment assignments**, say, $A_i^* = a_i^*$, $i = 0, 1, \ldots, k$.

Comparing dynamic treatment regimes

- More generally, there could be a **dynamic treatment regime**, say ***A***, such that each A_k could be allowed to be a function of the past observed history of that individual, consisting of past event times, covariate readings and possible earlier treatment assignments.

- More generally still, such a regime could be randomized, as long as the randomization mechanism does not depend on the past potential confounder variables U_k .

Predictive distributions for tratment comparison

- In order to make the role of the role of the regime ***A*** explicit in the notation, we write $p_{\mathrm{ex}(\mathbf{A})}$. The considered predictive distribution is be denoted by $p_{\mathrm{ex}(\mathbf{A})}(Y^* \mid \text{data})$.

- The exact specification of this probability will depend on the considered assignment mechanism ***A***.

- We can then consider any two such regimes of interest, say $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, and compare the corresponding predictive distributions $p_{\mathrm{ex}(\mathbf{A1})}(Y^* \mid \text{data})$ and $p_{\mathrm{ex}(\mathbf{A2})}(Y^* \mid \text{data})$ to each other.

Predictive distributions for tratment comparison

- Note: No conditioning on covariate values *X** after baseline is allowed here!

- If this were done, such conditioning on variables which are intermediate in time between a treatment (the contemplated cause) and a response (its effect), could potentially ’grab’ the entire statistical explanation to itself.

Predictive distributions for tratment comparison

- In practice, the necessary numerical integration can be carried out efficiently by Monte Carlo simulation, by applying data augmentation alongside the computations that are needed for statistical inference.

- Practical illustrations of this general method can be found, e.g., in Arjas and Liu (1995), Arjas and Liu (1996), Arjas and Haastrup (1996), Arjas and Andreev (2000), and Härkänen et al. (2000).

Example 2: “Home visits” (Didelez 2008; sometimes using direct quotations from the paper)

- “Programme to assist elderly by regular home visits, hoped to reduce unnecessary hospitalizations, improve quality of life, and increase survival time.”

- Simple causal question:

  “Can more frequent home visits increase the survival time?”

Home visits (cont'd)

- Suppose that timing of the home visits is determined externally, and in a way that is independent of all remaining processes;

- Therefore 'home visits', when considered as a treatment, clearly satisfy the 'unconfounded inference' condition.

Home visits (cont'd)

- However, their causal effect on survival, if it exists, is likely to be rather small.

- Could the power of the statistical data analysis, for the purpose of verifying such a causal claim, be increased by utilizing some appropriate covariate data?

Home visits (cont'd)

- "Home visits and hospitalizations are monitored. However, the underlying health status of an elderly person may be difficult to measure accurately in practice."

Question: Can we supplement our measured covariates by data on hospitalizations, including them as a part of the observed histories ($\mathcal{H}_k$), and then condition the response of interest (survival) on such covariate information?

Note: Technically, this would be very easy to do, e.g., in the context of the Cox the proportional hazards model. One might then try to answer a stated causal hypothesis by considering the estimated regression coefficients.

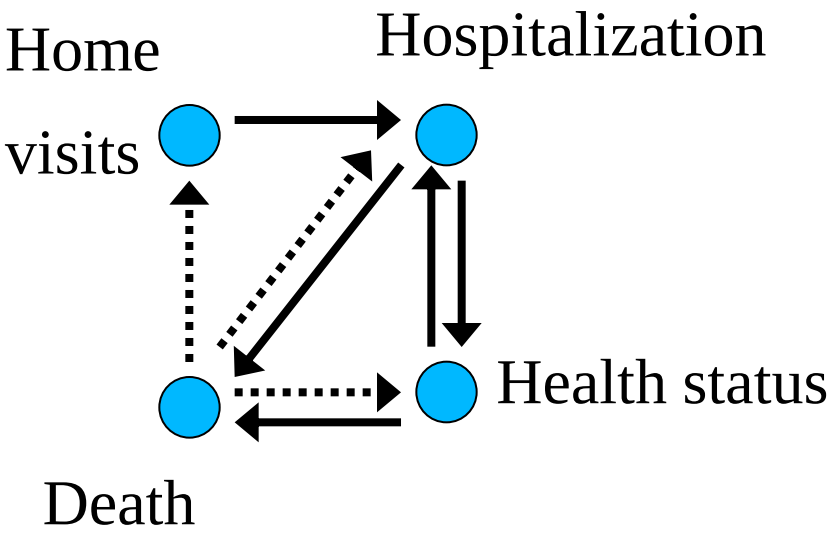
Home visits (cont'd …)

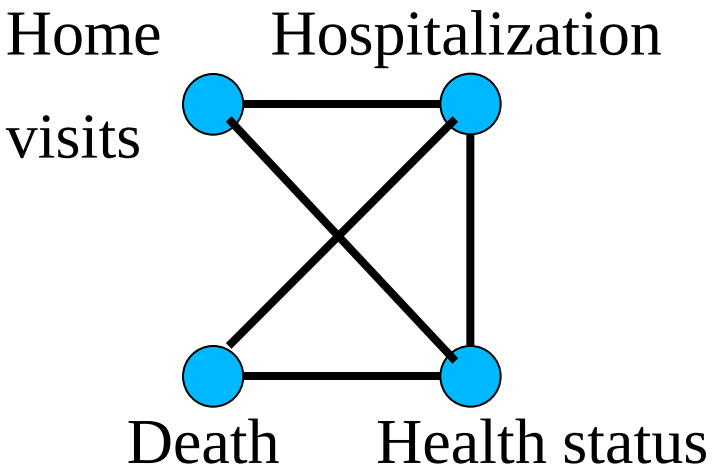
For considering the causal problem suppose that : For an individual still alive at time *t*,

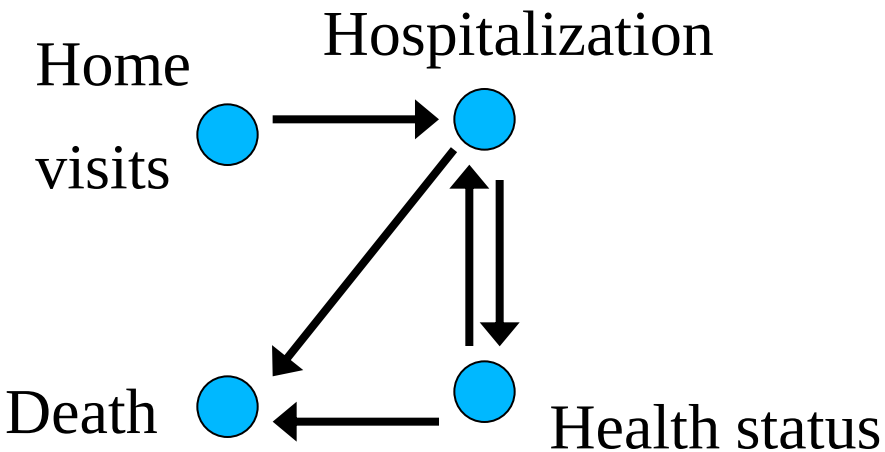
- initiation or termination of hospital care at time *t* can depend on the past (= up to time *t*) history of hospitalizations, home visits and health status;

- health status at time *t* can depend on past history of health status and hospitalizations, but is locally independent of home visits given these two;

Home visits (cont'd …)







Home visits (cont'd)

- There is nothing 'technically wrong' in estimating a hazard regression model (for survival) which is based on conditioning the hazards only on being still alive and on records on past home visits and possible hospitalizations as covariates.

- However, when considering the estimated hazards in a such a model, *observed differences in the estimates* arising from considering two different schedules for home visits, both combined with a shared history of hospitalizations, *cannot be viewed as representing causal effects* of those differences.

Home visits (cont'd)

- This is because joint consideration of home visits and hospitalizations carries (indirect) information also on unobserved health status.

- "Standard methods that just model the intensity for survival with time varying covariates for the previous home visits and hospitalizations will typically give misleading results due to the conditional association between 'Home visits' and 'Death' given 'Hospitalization'." (Didelez 2008)

Home visits (cont'd …)

- In a **longitudinal setting**, say, when comparing two protocols for home visits and studying whether their difference has an effect on survival, it is O.K. to collect records on past hospitalizations without adequate records on health status and use such information as a covariate for calibrating the survival prediction different individuals.

- However, for a *causal analysis*, such information can only be used at baseline, as a part of the observed past history.

Home visits (cont'd …)

- Hospitalization records should NOT be used as time-dependent covariates for conditioning the survival predictions when the protocols for home visits are already use!

- This conclusion would be obvious to most epidemiologists working with longitudinal data - even without more formal considerations, since such such hospitalizations are ”intermediate outcomes” between baseline and death.

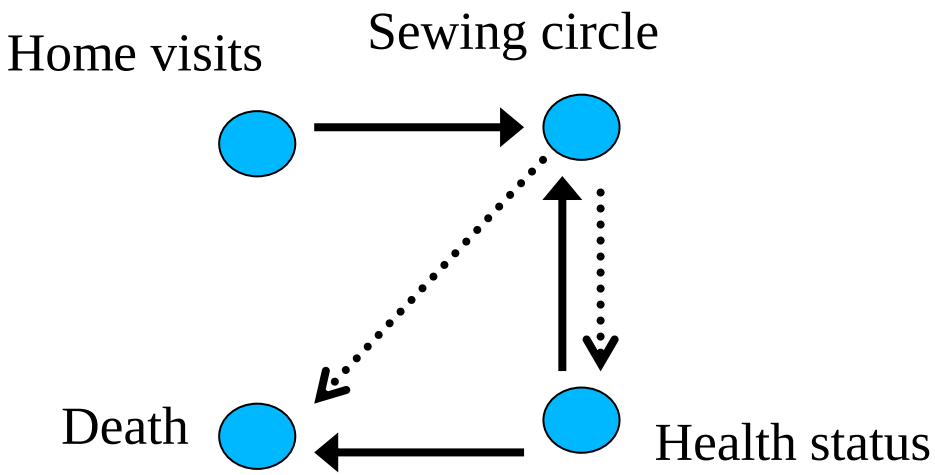
A modification: removing a causal link

- What happens if the causal link from ”hospitalization” to ”survival” is removed? For this to be realistic, suppose now that a recorded history of ”hospitalization” is replaced by recorded activity in a ”sewing circle”.

- Then it seems less likely that there would be a direct causal link from ”sewing circle” to either ”health status” or ”survival”, and thus there is no path from ”home visits” to ”survival”.

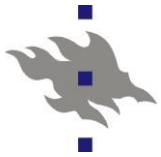
A modification (cont'd)

Dotted arrows indicate removed dependencies from previous example



A modification (cont'd)

- **Question:** Is it now OK, in an analysis based on survival/hazard regression and aimed at studying the potential causal influence of "home visits " to "survival", to condition on a recorded history of "sewing circle" participation as a time dependent covariate?



A modification (cont'd)

- The answer: NO, it's NOT!

A modification (cont'd)

- The reason is the same as before, with only 'hospitalization' being replaced by 'sewing circle':

  Joint consideration of 'home visits' and 'sewing circle' carries information about health status, and such updated information on health status influences the predictions of survival.

Take home points …

- *Unconfounded inference / Local independence* is *a* key condition for controlling potential confounding in causal inference from event history data.

- *Predictive distributions,* combined with MPP models and Bayesian/likelihood inference, provide a natural methodology for converting such data into intuitively understandable formulations of empirical evidence in support of causal claims.

- Introducing additional covariates into a hazard regression model is not always helpful in attempts to control confounding – sometimes it has the opposite effect!

References:

- Odd O. Aalen (1987) Dynamic modeling and causality. Scand Actuar J 12:177–190
- Elja Arjas and Jan Parner (2004) Causal Reasoning from Longitudinal Data. *Scandinavian Journal of Statistics* 31: 171–187.
- Elja Arjas 2012) 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.
- Vanessa Didelez (2008) Graphical models for marked point processes based on local independence. *JRSS B* 70: 245 – 264.
- Dennis V. Lindley (2002) Seeing and Doing: the Concept of Causation. *International Statistical Review* 70: 191–197.
- Judea Pearl (2000) Causality: Models, reasoning, and inference. New York: Cambridge University Press.
- Tore Schweder (1970) Composable Markov Processes. *Journal of Applied Probability,* 1970, 7: 400-410.