

A statistical model for parametric estimation of incidence and prevalence using databases with health care events without information on previous treatment history

Henrik Støvring

Research Unit of General Practice

University of Southern Denmark

hstovring@health.sdu.dk

April 1, 2004

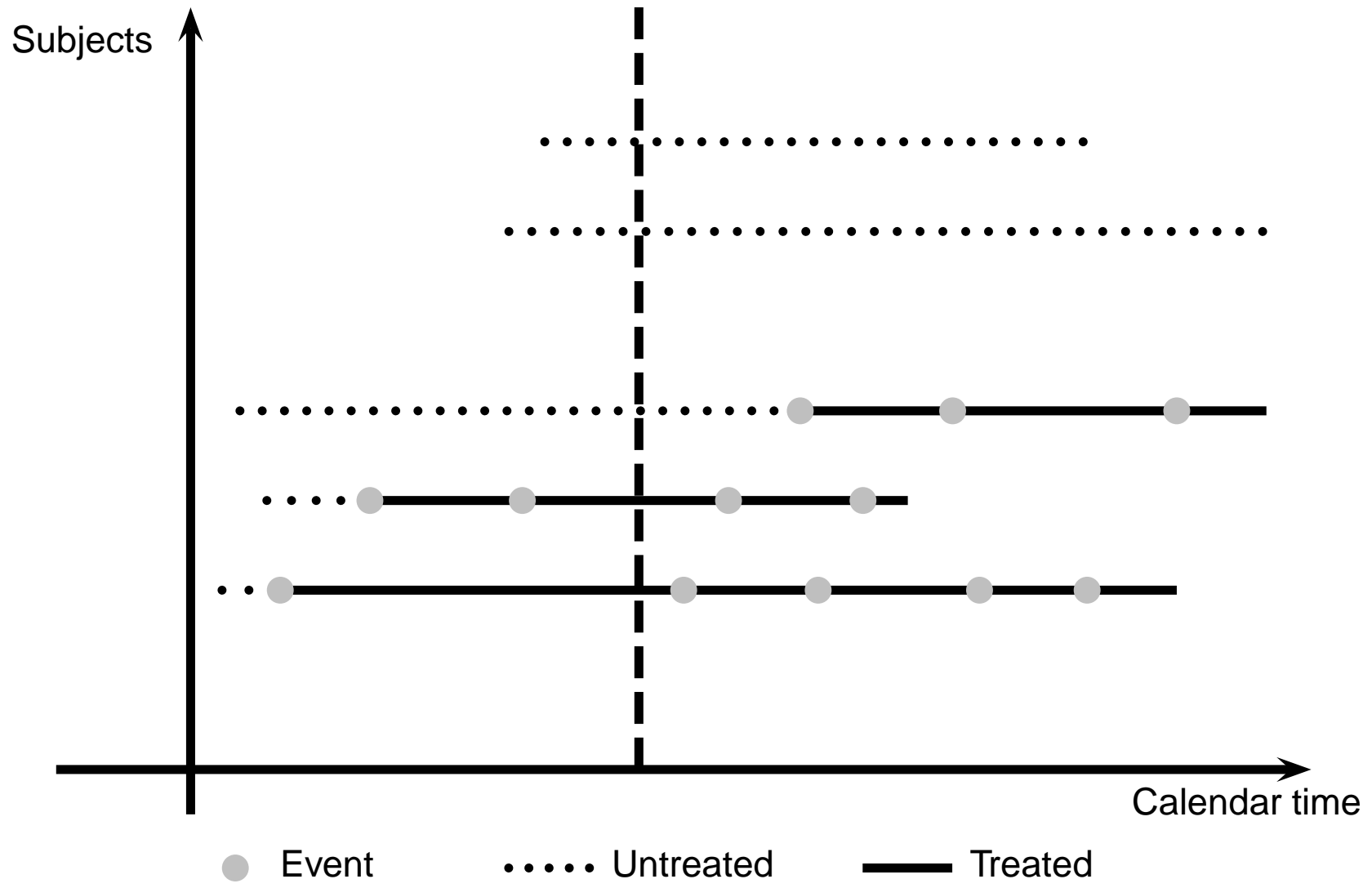
Joint work with Prof Werner Vach, Department of Statistics, University of Southern Denmark

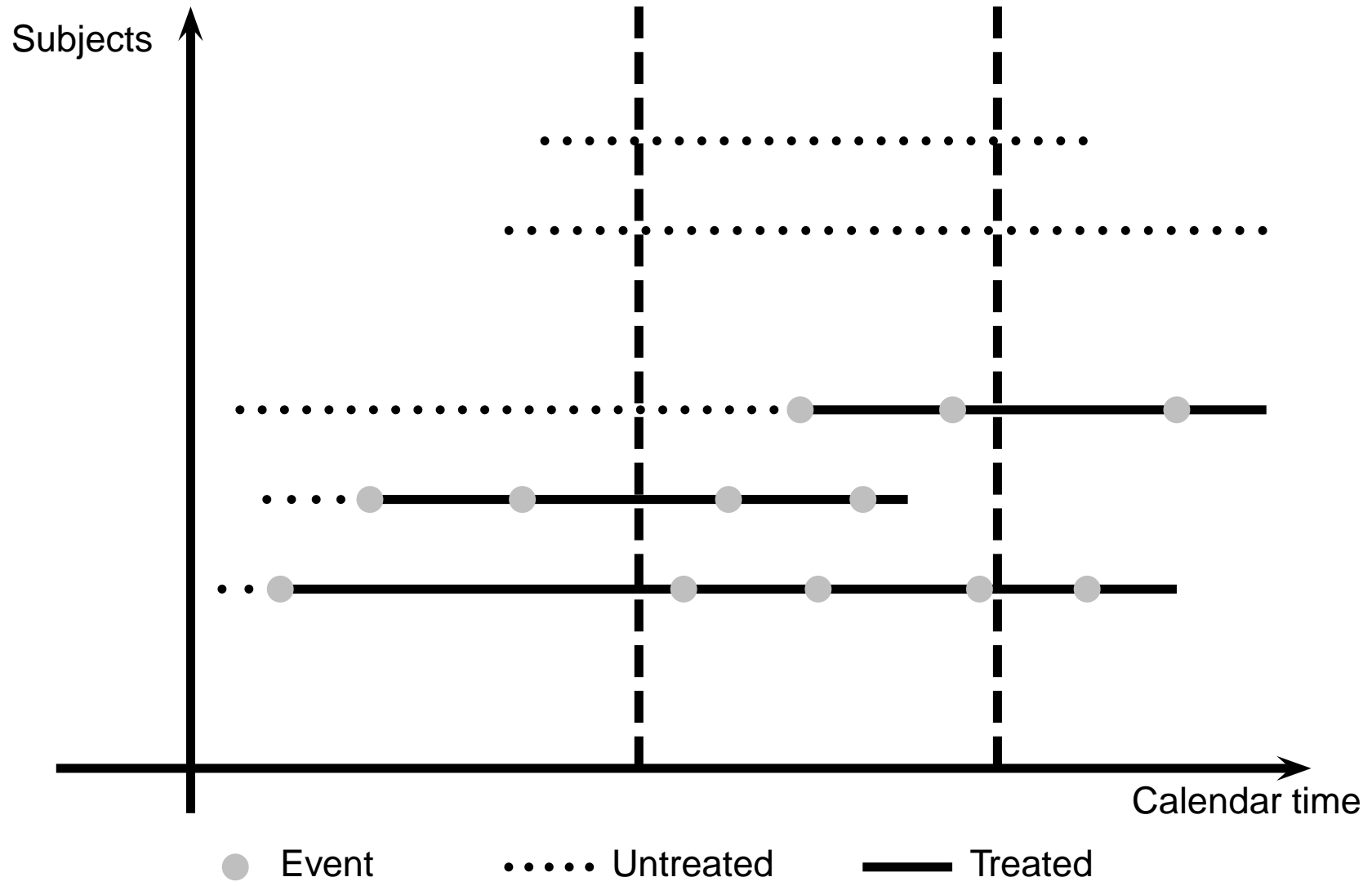
1 Outline

- Problem/Objective
- Model
- Application and diagnostics
- Outline of simulation results

2 Introduction

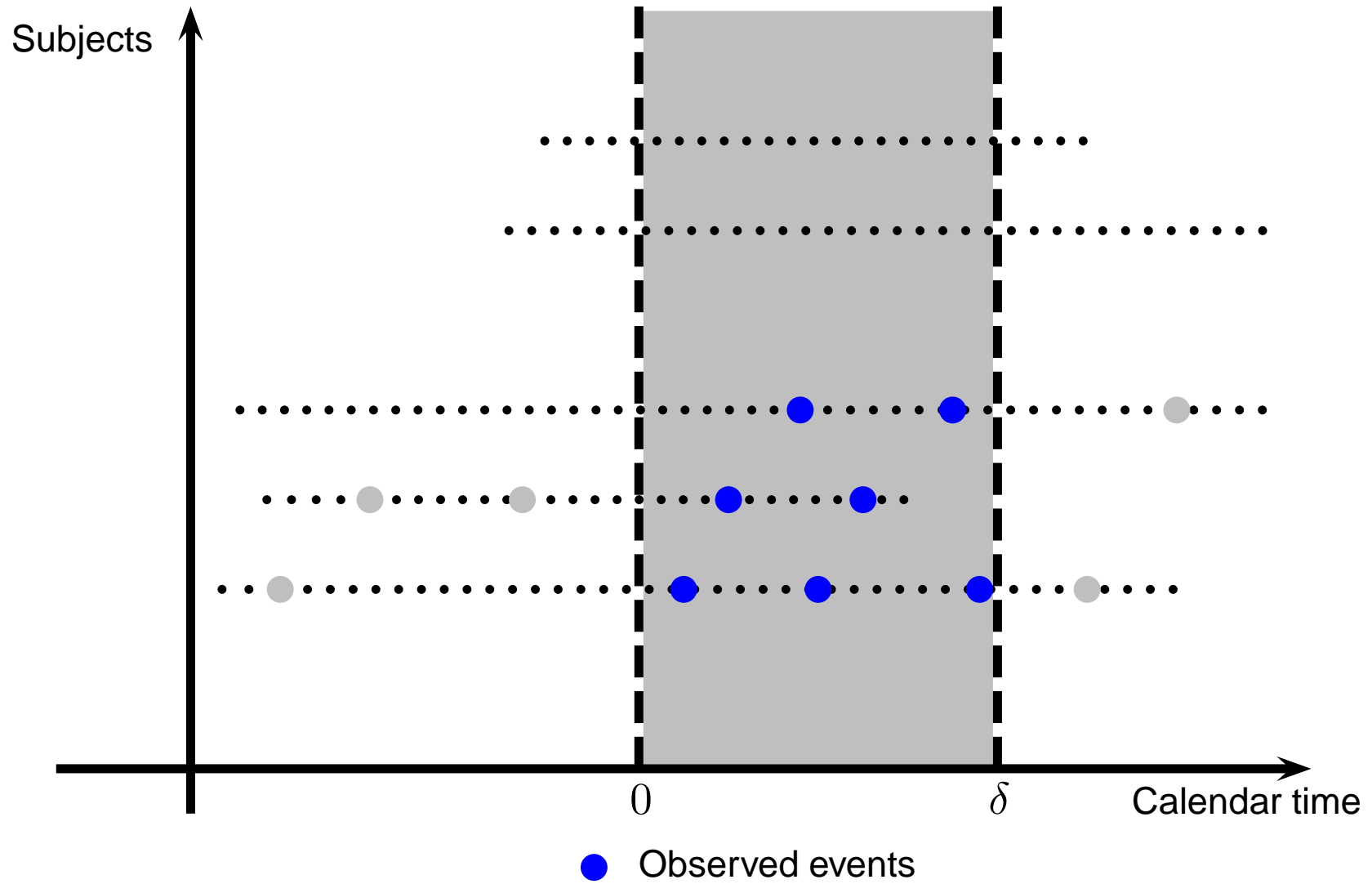
- Basic epidemiologic measures
 1. Prevalence proportion
 2. Incidence rate
- Traditional estimates are obtained by
 1. Cross sectional: Assessment of individual disease status at time t_0
 2. Follow up: Assessment of individual change from non-diseased to diseased
- **But:** Costly to obtain information on individual disease history



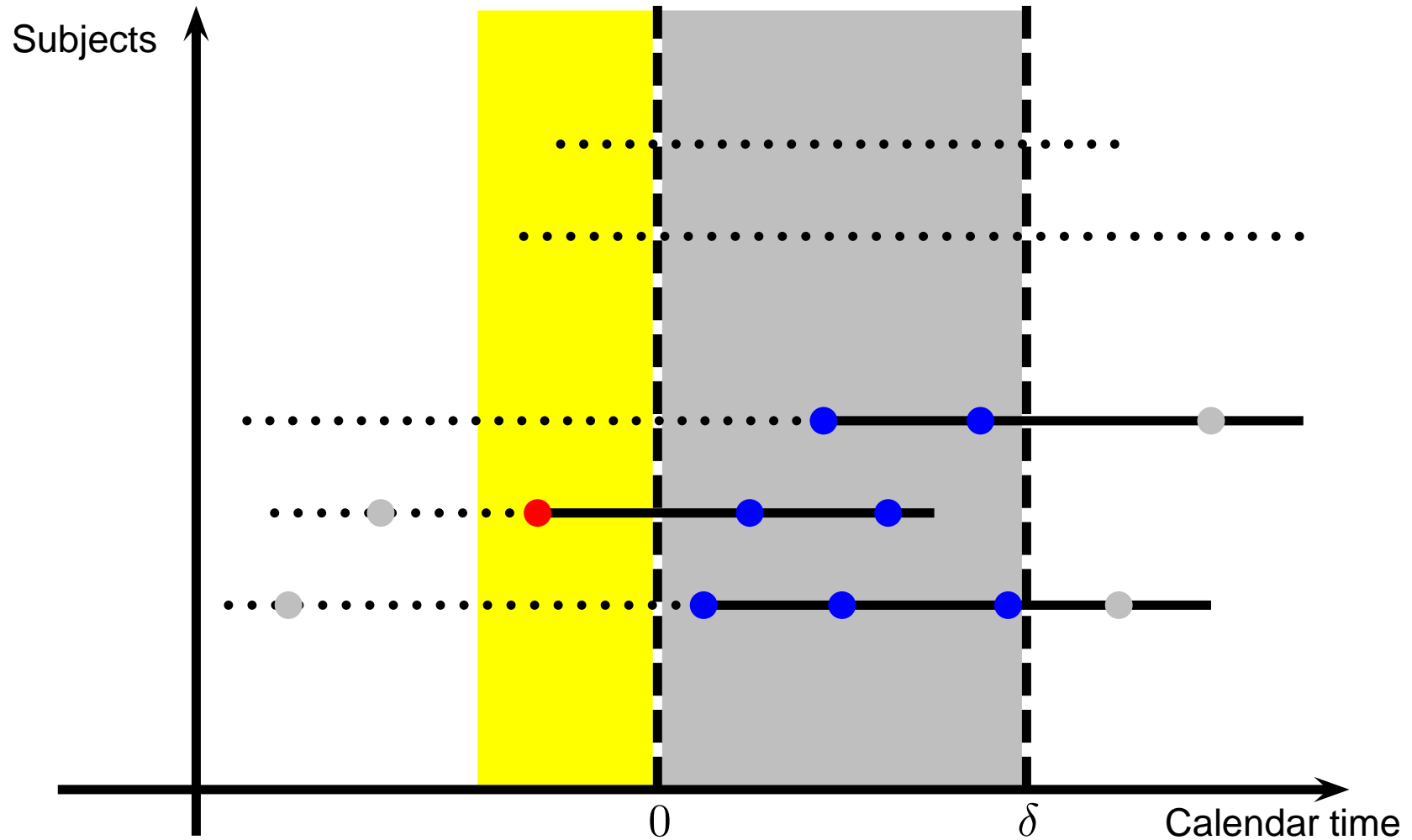


3 Characteristics of Health Service Events databases

- Records date and type of individual events
- Does not record
 - treatment status before events
 - treatment status between events
 - treatment status after events
- Corollary: No access to onset of treatment if it happens before observation window



3.1 Standard analysis: Use run-in period



Define treatment status at 0 based on run-in period (yellow area),
cf (Støvring, Andersen, Beck-Nielsen, Green, and Vach 2003)

4 Data: FLUKS

- FLUKS: *ForLøbsUndersøgelse og KvalitetsSikring*
Danish abbreviation for “investigation of sequences and quality control”,
cf (Schroll, Støvring, Houmand, and Kragstrup 2004)
- Registry on 2225 70+ year olds
- All were included on February 16, 1997, followed-up until December 24, 1997
→ Short observation period prohibits use of run-in period

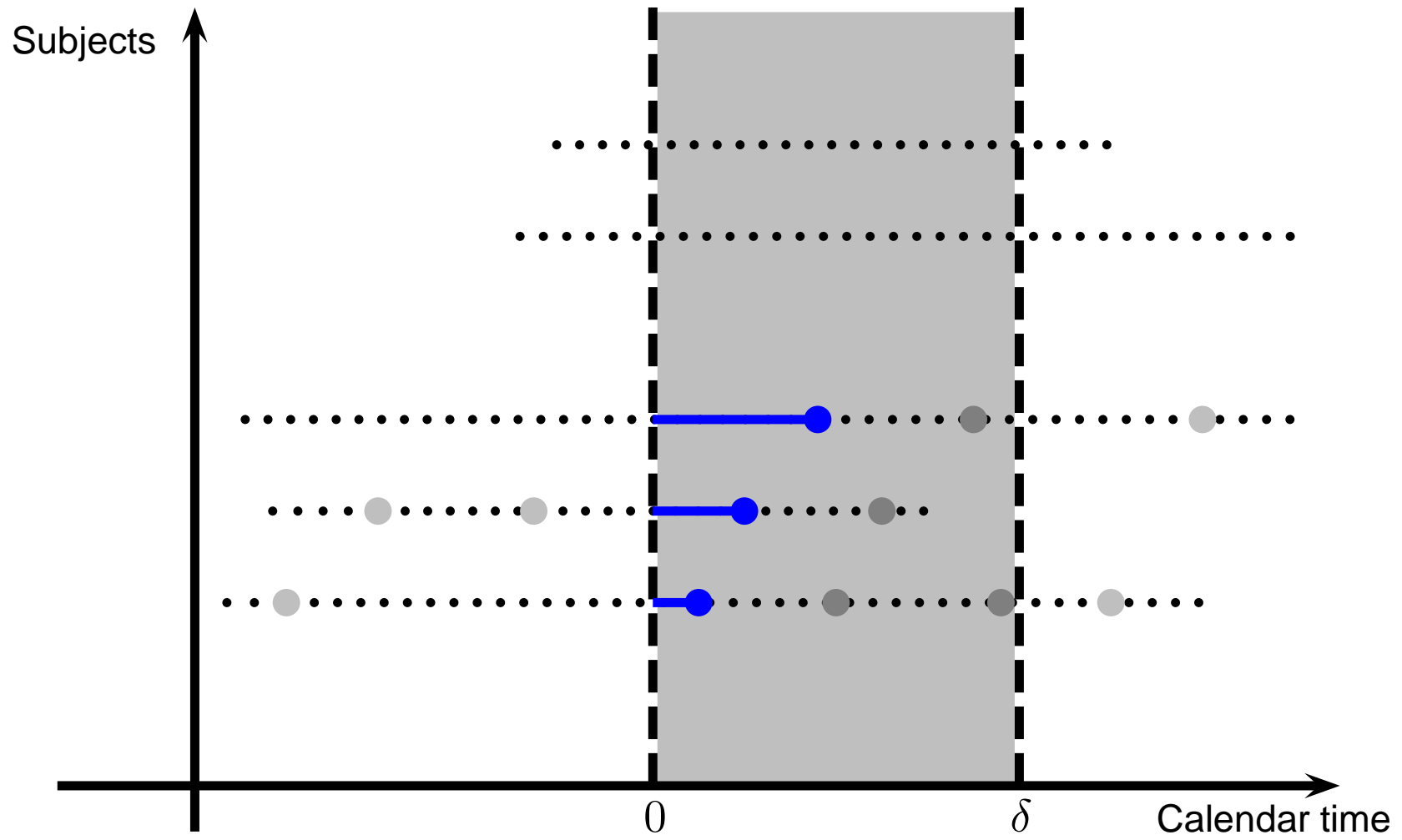
- Gives information on
 - Date of contact
 - Person ID, age, sex and moving/death
 - Contact diagnosis coded in terms of International Classification for Primary health Care (ICPC) (Lamberts, Wood, and Hofmans-Okkes 1993)
 - General Practitioner
- Focus exclusively on hypertension here

5 The challenge

- Databases with health-related events are available, for a sample or population wide
- Problem
 - The individual treatment status is *not* observed between events
- **Challenge:**
 - estimate prevalence and incidence based on these databases
 - should avoid run-in period
 - should be based on general principles
- Assumption: Disease is **chronic**

6 The idea of Hallas et al.

- Basic idea (Hallas et al. 1997):
 - Record time to first event for each individual
 - Make histogram for these waiting times
 - Choose cut-point such that histogram is constant afterwards
 - Constant level corresponds to incidence
 - Excess at beginning corresponds to prevalence



Only time to first observed events are considered

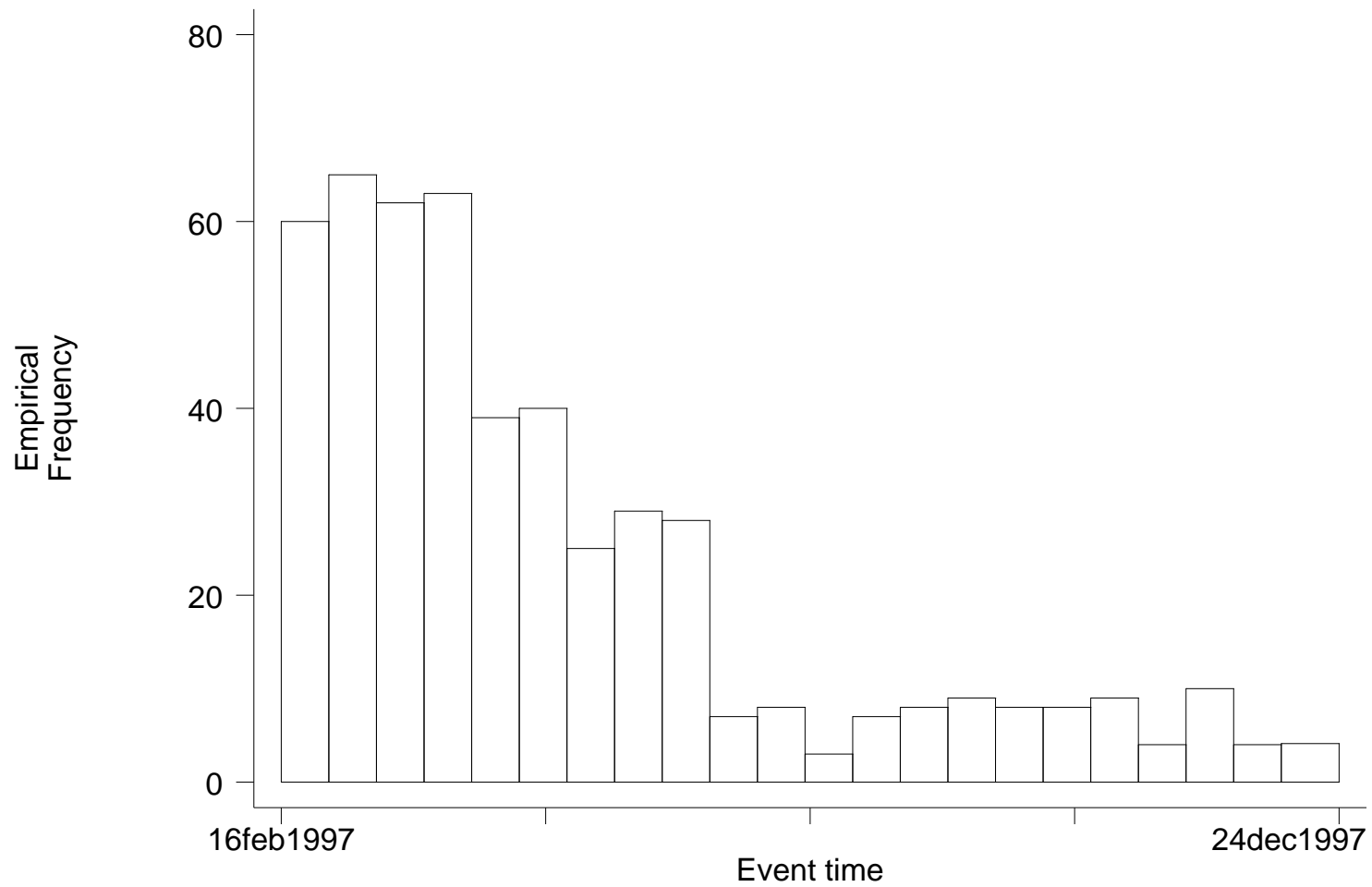


Figure 1: Empirical distribution of time until first hypertension event observed in FLUKS.

7 Basic model: a two-component mixture

- Consider the cohort present at 0 (follow-up on cross-section)

X : Initial disease or treatment status (unobserved/latent)

$$X = \begin{cases} 1 & \text{if treated at 0,} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

T : Time of first event after 0

- Interest parameters:

Prevalence: $p = P(X = 1)$ (2)

Incidence: $\lambda_t = \lim_{\Delta t \downarrow 0} \frac{P(X_{t+\Delta t} = 1 | X_t = 0)}{\Delta t} \equiv \lambda$

- Initial model

$$f^T = p \cdot f^{T|X=1} + (1 - p) f^{T|X=0} \quad (3)$$

8 Incorporating censoring

- Information on exit time is available

Z : First exit time after 0

- Note that:

- May observe both event and exit time for some (few) subjects
- Exit time is informative for treatment status
- Exit time is informative for event time

- Implications:

- censoring must be incorporated in model
- model for dependence structure of event and exit time should be simple

- For more details, see Appendix A

9 Likelihood construction

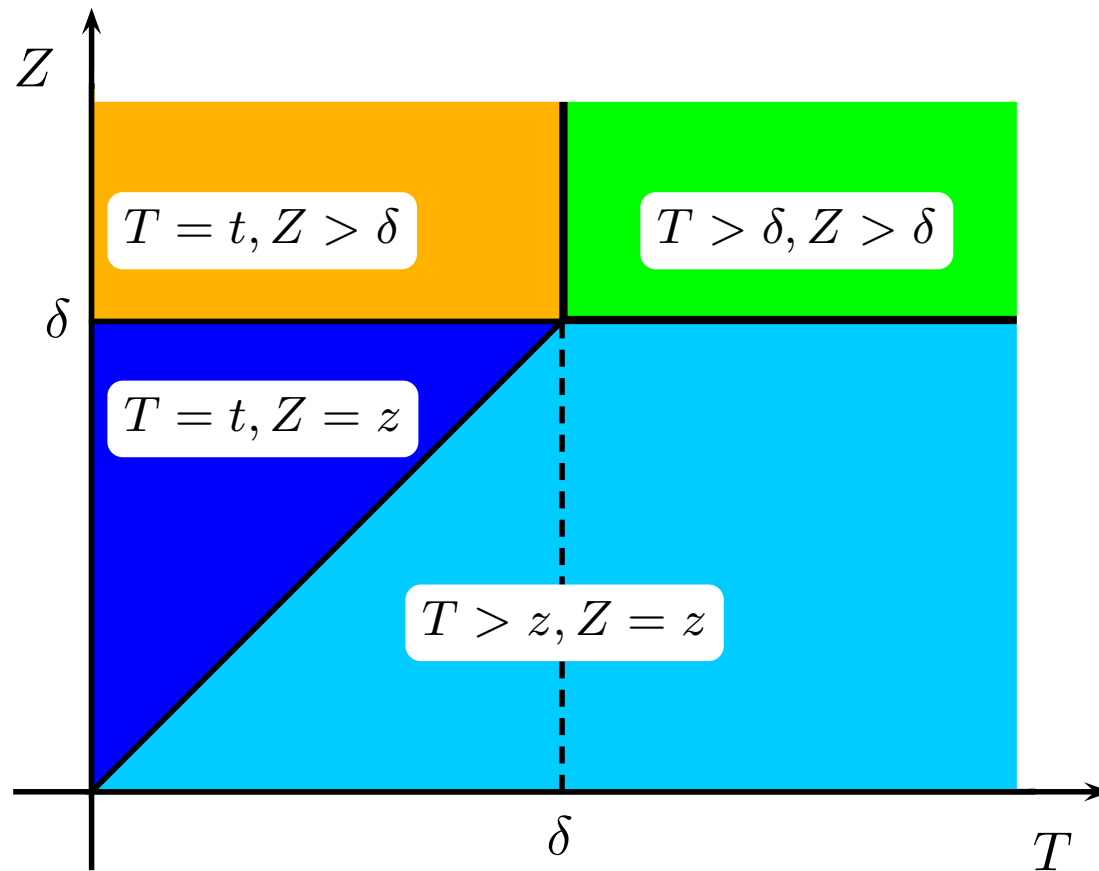


Figure 2: Observational areas of (T, Z)

9.1 Examples of likelihood contributions

- Consider the observation

$$T = t, Z = z$$

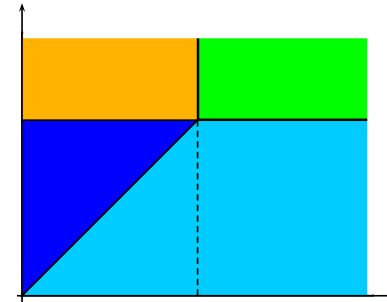
- The likelihood contribution has the structure

$$l^1(p, \theta) = p \cdot f^{(T,Z)|X=1}(t, z; \theta) + (1 - p) f^{(T,Z)|X=0}(t, z; \theta) \quad (4)$$

- In the remaining contributions we integrate over the missing data

- For example for $T = t, Z > \delta$

$$l^2(p, \theta) = p \int_{\delta}^{\infty} f^{(T,Z)|X=1}(t, s; \theta) ds + (1 - p) \int_{\delta}^{\infty} f^{(T,Z)|X=0}(t, s; \theta) ds \quad (5)$$



9.2 Choice of distributions: Forward recurrence density

- $g = f^{T|X=1}$ is a forward recurrence density, since:
 - Assume events of a single prevalent subject are a renewal process
 - We hit an interval at random, ie. length biased sampling
 - Hit time is uniformly distributed on the hit interval
 - Suppose the interarrival density is γ , then

$$g(t) = \frac{\int_t^\infty \gamma(s) ds}{\int_0^\infty \gamma(s) ds} \quad (6)$$

$$= \frac{S_\gamma(t)}{\mu} \quad (7)$$

where μ is the mean wrt. γ

- In reality we have heterogeneity across individuals and time
- Implications of heterogeneity:
 - The marginal forward recurrence density is a mixture of forward recurrence densities, ie. still a forward recurrence density
 - Non-trivial relationship between marginal forward recurrence density and marginal interarrival density
- We have worked with the Exponential, Weibull and Log-Normal as choices for γ

9.3 Choice of distributions: Incidence and exit

- All chosen as exponentials, since
 - Corresponds to a constant, homogeneous rate over the interval
 - Easy to handle numerically and analytically
 - Corresponds to crude rate estimate, had incidence been directly observed

10 Results for FLUKS: Hypertension

Model	$l(\hat{\theta})$	\hat{p}	$\hat{\lambda}$
Exponential	-1526.62	.191 (.160; .226)	.0606 (.0317; .116)
Weibull	-1518.76	.157 (.134; .185)	.1043 (.0825; .1095)
Log-Normal	-1520.14	.163 (.139; .190)	.0965 (.0751; .1242)

Table 1: Results for hypertension with 95%-confidence intervals based on robust variance estimates adjusting for clustering on GP. Model type refers to the parametric choice for the forward recurrence density g_1 .

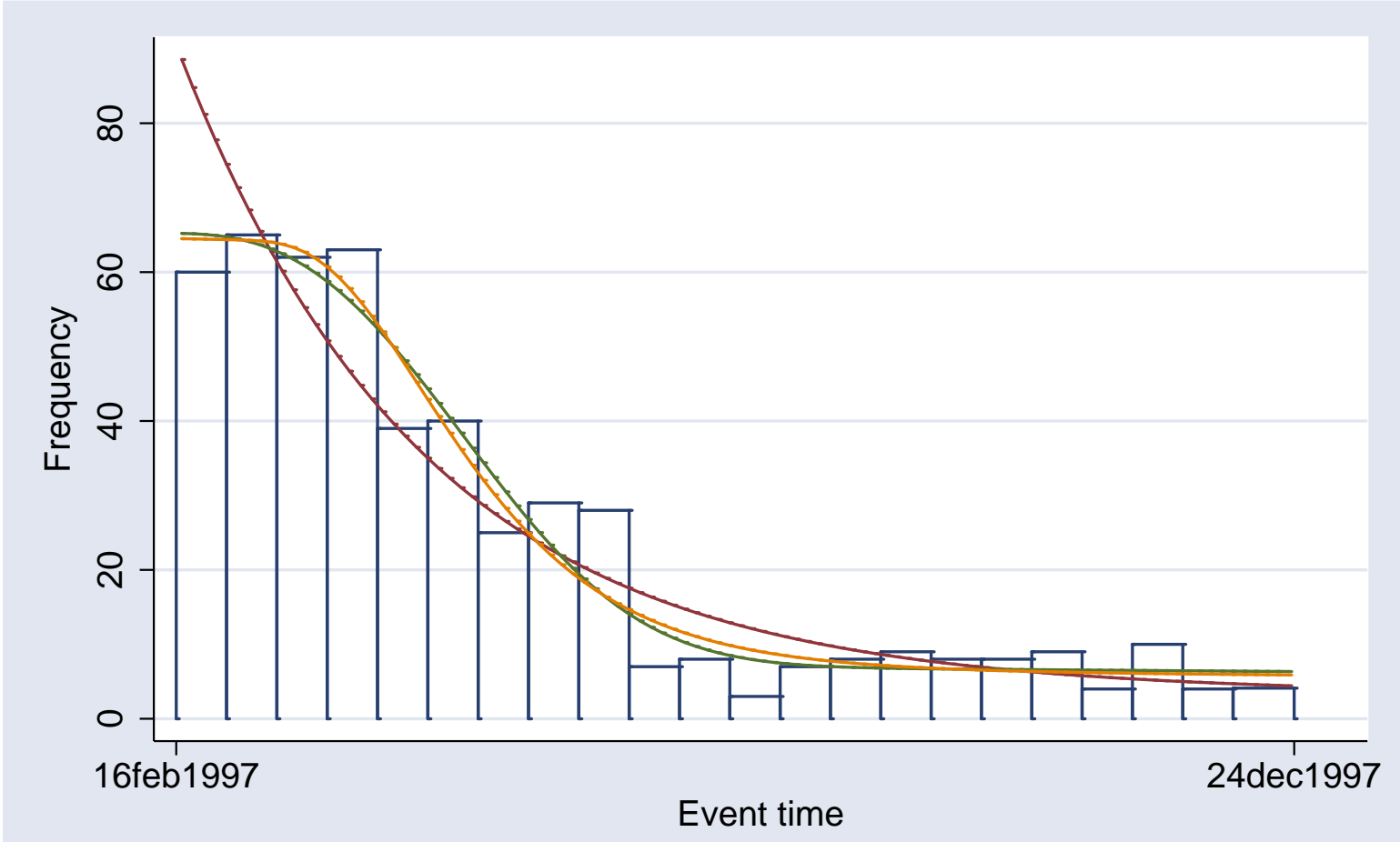


Figure 3: Diagnostic plot showing the fit of three different models,

Exponential , **Log-Normal** , and **Weibull** .

11 Finite sample properties

- Establishing a model in the likelihood framework yields asymptotic properties
- But how does the model perform in finite, realistic samples?
- Focused on three central questions
 - Sensitivity to misspecification
 - Validity of inference, ie. coverage probabilities
 - Loss of information due to true treatment status being unobserved
- Measured performance in terms of
 - relative bias
 - coverage probabilities of confidence intervals
 - variance inflation relative to studies with observed treatment status

- For correctly specified models
 - Bias is small
 - * Prevalence $< 1\%$
 - * Incidence $< 3.5\%$
 - In comparison to studies with observed disease/treatment history, relative sample sizes of two to five yields similar precision
 - Nominal levels of confidence intervals are maintained
 - Very small impact of censoring levels

- For misspecified models
 - Bias increases with level of misspecification
 - When both forward recurrence and incidence density is misspecified, bias increases to less than 9%
 - Precision decreases with level of misspecification
 - In comparison to studies with observed disease/treatment history, relative sample sizes of two to nine yields similar precision
 - Even in extreme situations, the coverage probability of 95%-CIs remains above 90%

12 Summary

- Using routine data which simply record event times, estimates of incidence and prevalence can be obtained with our method
- Formulated an intuitive idea using general principles
- Obtained model, which
 - incorporates censoring
 - allows ML estimation
 - allows construction of diagnostic procedures
- Thus, 'cheap' estimates of incidence and prevalence are available

- Perspectives

- Key elements in projecting the needed capacity of the health care system in future are incidence and prevalence of disease (along with mortality)
- Can now be obtained from within the system itself

References

- Hallas, J., D. Gaist, and L. Bjerrum (1997). The waiting time distribution as a graphical approach to epidemiologic measures of drug utilization. *Epidemiology* 8, 666–70.
- Lamberts, H., M. Wood, and I. Hofmans-Okkes (Eds.) (1993). *The International Classification of Primary Care in the European Community*. Oxford: Oxford University Press.
- Schroll, H., H. Støvring, A. Houmand, and J. Kragstrup (2004). Estimating incidence and prevalence of episodes of care in general practice. *Scandinavian Journal of Primary Health Care* 22, 60–64.
- Støvring, H., M. Andersen, H. Beck-Nielsen, A. Green, and W. Vach (2003). Rising prevalence of diabetes: evidence from a Danish pharmaco-epidemiological database. *The Lancet* 362, 537–38.

A Incorporating censoring: a detailed look

A.1 How to model correlation of event and exit times

- Consider first those **not** in treatment at 0
- Distinguish between two types of dependence
 1. Short term
 2. Long term
- Long term dependence leads to considering dependence of

$$I(T \leq \delta) \quad \text{and} \quad I(Z \leq \delta) \quad (8)$$

since we only have short observation period

- Examples of factors causing long term dependencies:
 - Genetic makeup

- Lifestyle
- Environmental factors
- Consequences of long term dependencies

$$P(T \leq \delta, Z \leq \delta) \neq P(T \leq \delta)P(Z \leq \delta) \quad (9)$$

- In particular we would expect that

$$P(T \leq \delta, Z \leq \delta) \geq P(T \leq \delta)P(Z \leq \delta) \quad (10)$$

since frail subjects are both at higher risk of initiating treatment **and** dying

- This suggests modeling $P(T \leq \delta, Z \leq \delta)$ in terms of

$$P(T \leq \delta), \quad P(Z \leq \delta), \quad \text{and} \quad \phi = \frac{P(T \leq \delta, Z \leq \delta)}{P(T \leq \delta)P(Z \leq \delta)} \quad (11)$$

- Examples of short term dependencies
 - Change in treatment status
 - Seasonal variation
 - Sudden changes in lifestyle, environment etc.
- We chose to ignore short term dependencies, since
 - No information on sudden changes
 - Data are too limited to allow explicit modeling of them
- Implies that event and exit times are independent **given** the indicators $I(T \leq \delta)$ and $I(Z \leq \delta)$

A.2 Dependence structure for prevalent

- Even less information available
- Most events for prevalent happen early in observation period
- Assume independence of event and exit time **given** treated at baseline