

**New Statistical Methods for Estimation of
Prevalence, Incidence and Mortality based on
Pharmacoepidemiological
and
other Health-related Databases**

Henrik Støvring

December 18, 2002

Supervisors:

Prof. **Werner Vach**, Department of Statistics and Demography, SDU

Prof. **Jakob Kragstrup**, Research Unit of General Practice, SDU

Prof. **Xiao-Li Meng**, Department of Statistics, Harvard University

1 Motivating example

Henrik Støvring, Morten Andersen, Henning Beck-Nielsen, Anders Green, and Werner Vach. A new look at the epidemic of diabetes: Evidence from a Danish pharmaco-epidemiological database. *Submitted, 2002*

1.1 Background

- International Diabetes Federation

“The alarming increase of diabetes prevalence is projected to occur because of:

- Population ageing
- Unhealthy diet
- Obesity
- A sedentary lifestyle”

1.2 Data

- Data from OPED (1992-1999):
 - Date of redemption
 - ATC-code of dispensed drug (drug type)
 - Civil Registration Number of the subject to whom the drug was prescribed, contains information on date of birth and gender
- From the central person register (Statistics Denmark):
 - Civil Registration Number
 - Dates of subjects moving in and out of the county
 - Date of death

for all subjects present at some point in the County of Fyn during 1992-1999

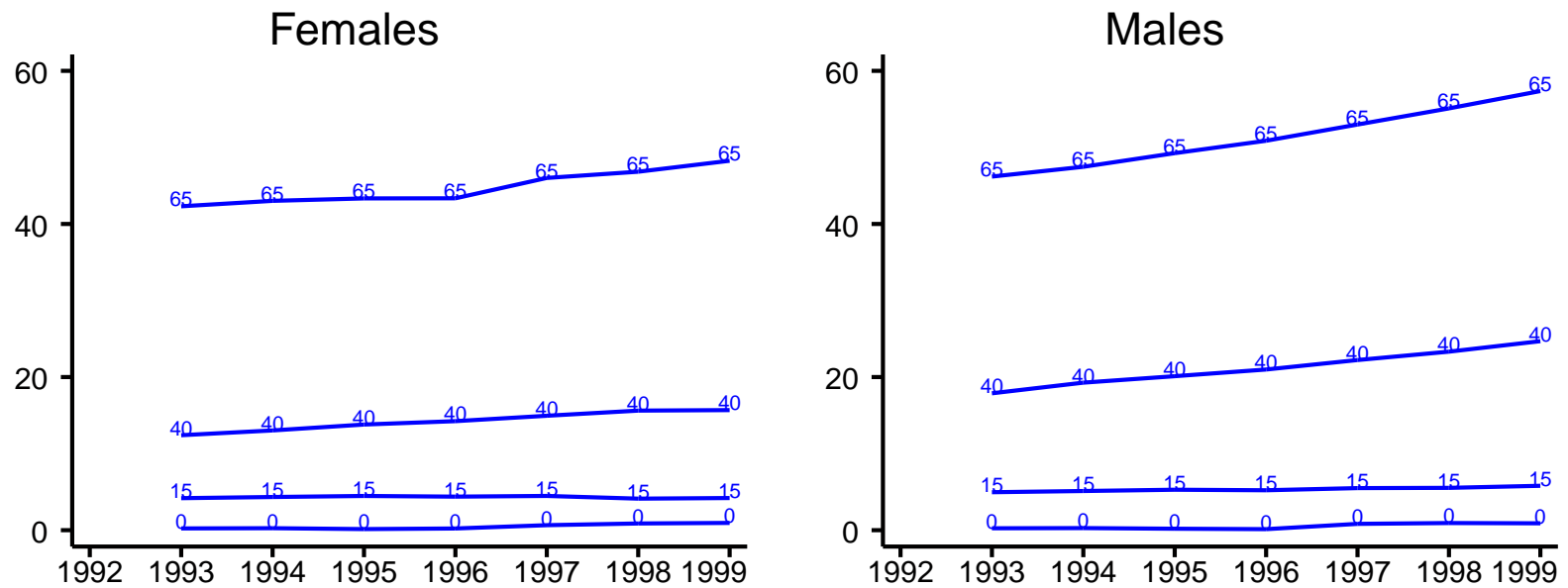
1.3 Objective

- Estimate and analyze trends of pharmacologically treated diabetes with respect to:
 - prevalence
 - incidence
 - mortality

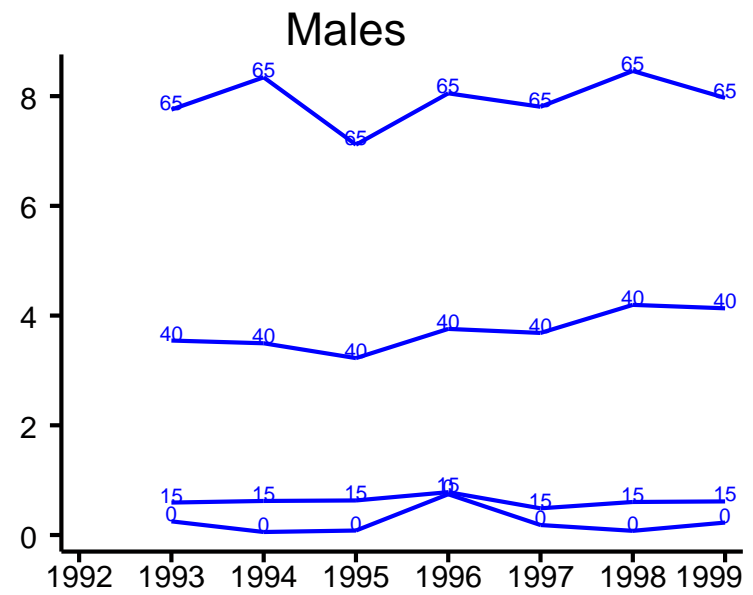
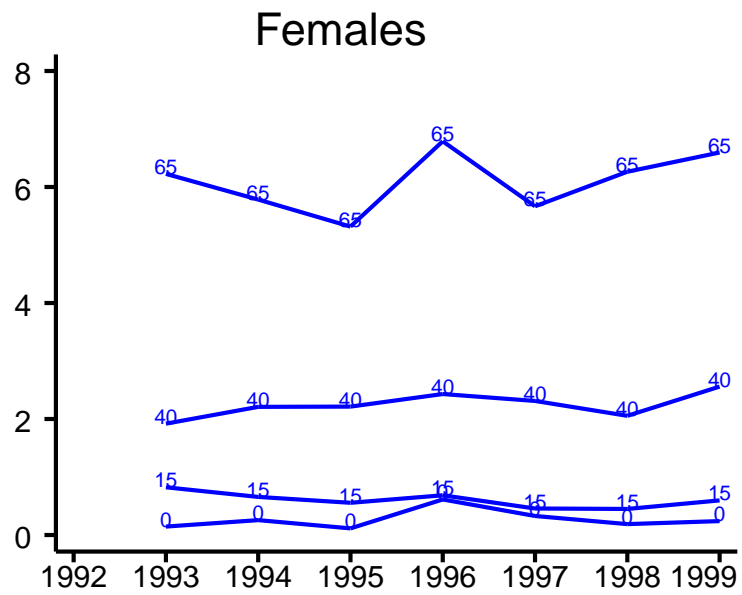
1.4 Methods

- Identified all subjects present on Jan 1 of each calendar year
- Determined treatment status based on previous years redemptions:
 - No redemptions → untreated
 - At least one redemption → treated
- Note: All immigrating during previous year were discarded
- We could then compute
 - Prevalence on Jan 1
 - Incidence in the calendar year
 - Mortality in the calendar year among those treated on Jan 1

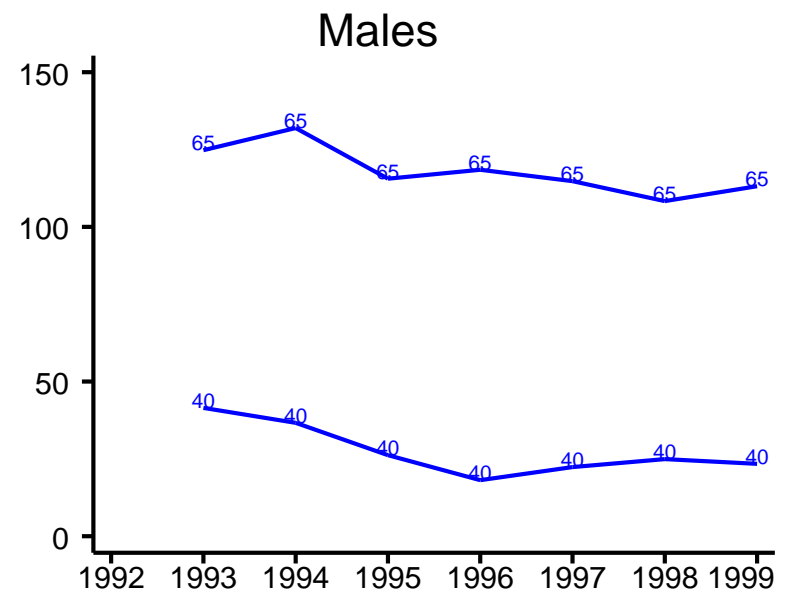
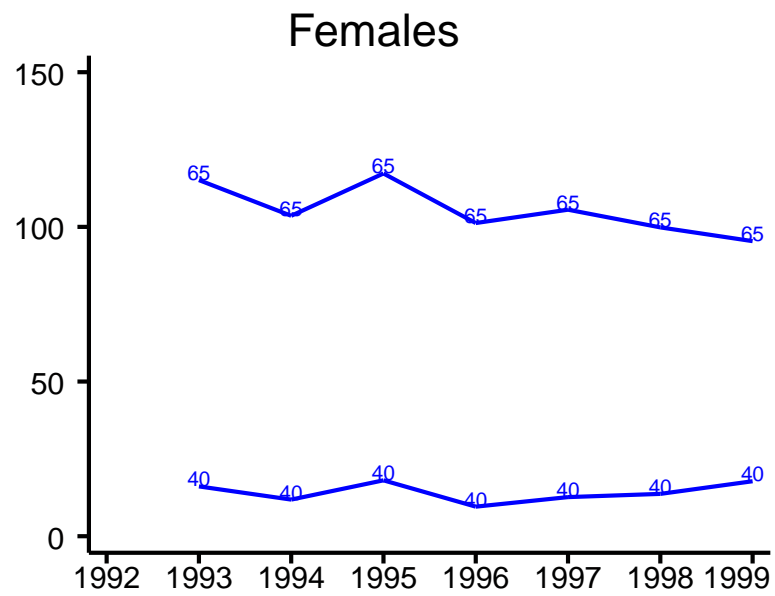
1.5 Results



(a) Prevalence per 1000



(b) Incidence per 1000 person-years



(c) Mortality among treated per 1000 person-years

2 Motivating example II: FLUKS

- FLUKS: *ForLøbsUndersøgelse og KvalitetsSikring*
Danish abbreviation for “investigation of sequences and quality control”
- Registry on 2225 70+ year olds
- All were included on February 16, 1997
- Followed-up until December 24, 1997
- 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)
 - General Practitioner

- Motivated by earlier studies in Copenhagen^a
- Key finding of this early study: Very difficult to make use of the theoretical concept of *episodes of care*

^aIn 1989, Hollnagel, Pedersen, Gannik, Heldrup and Frimodt-Møller published six papers in *Ugeskrift for Læger* 151(3), 142-172 & 230-235.

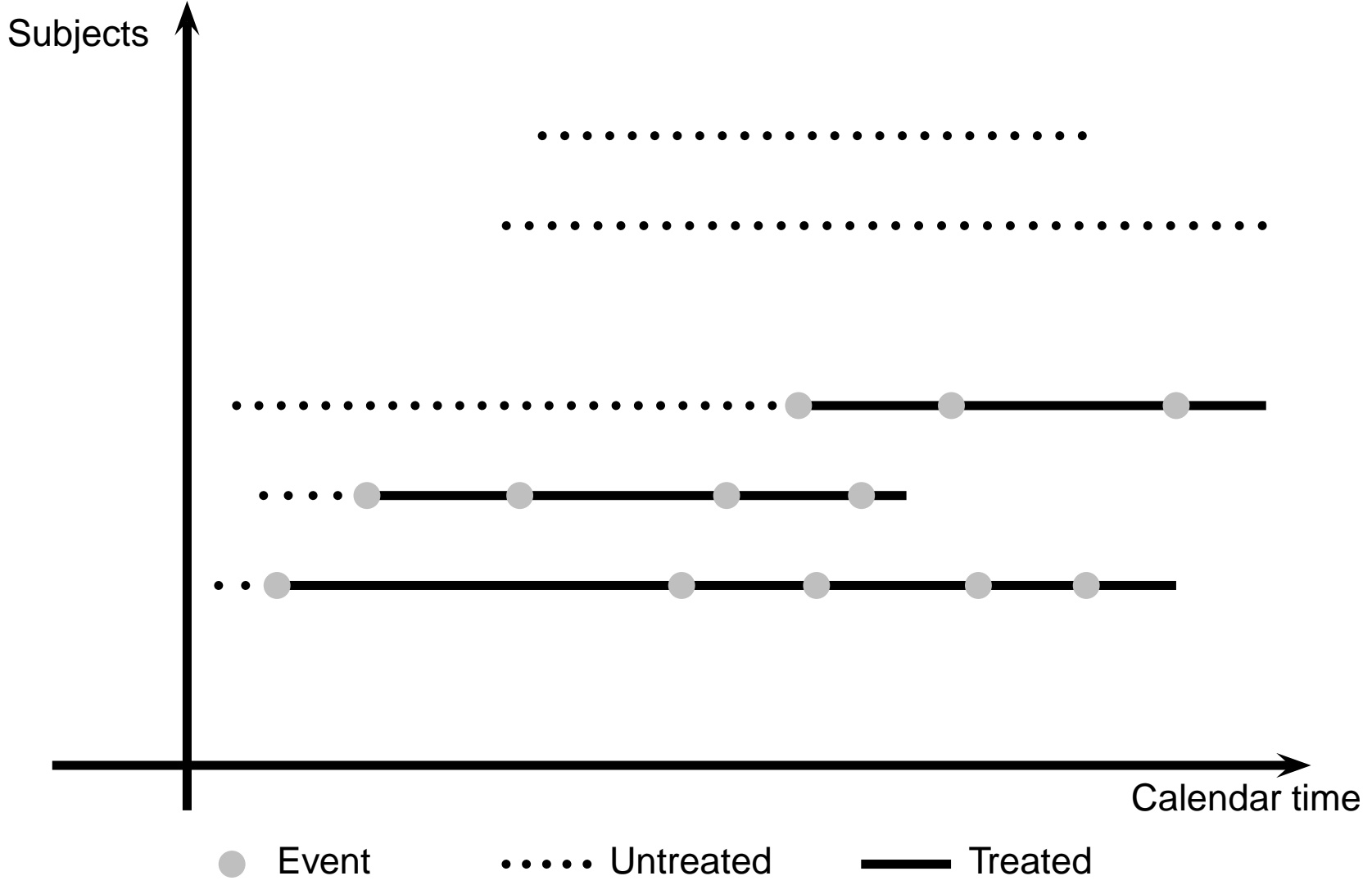
2.1 Initial observations on FLUKS

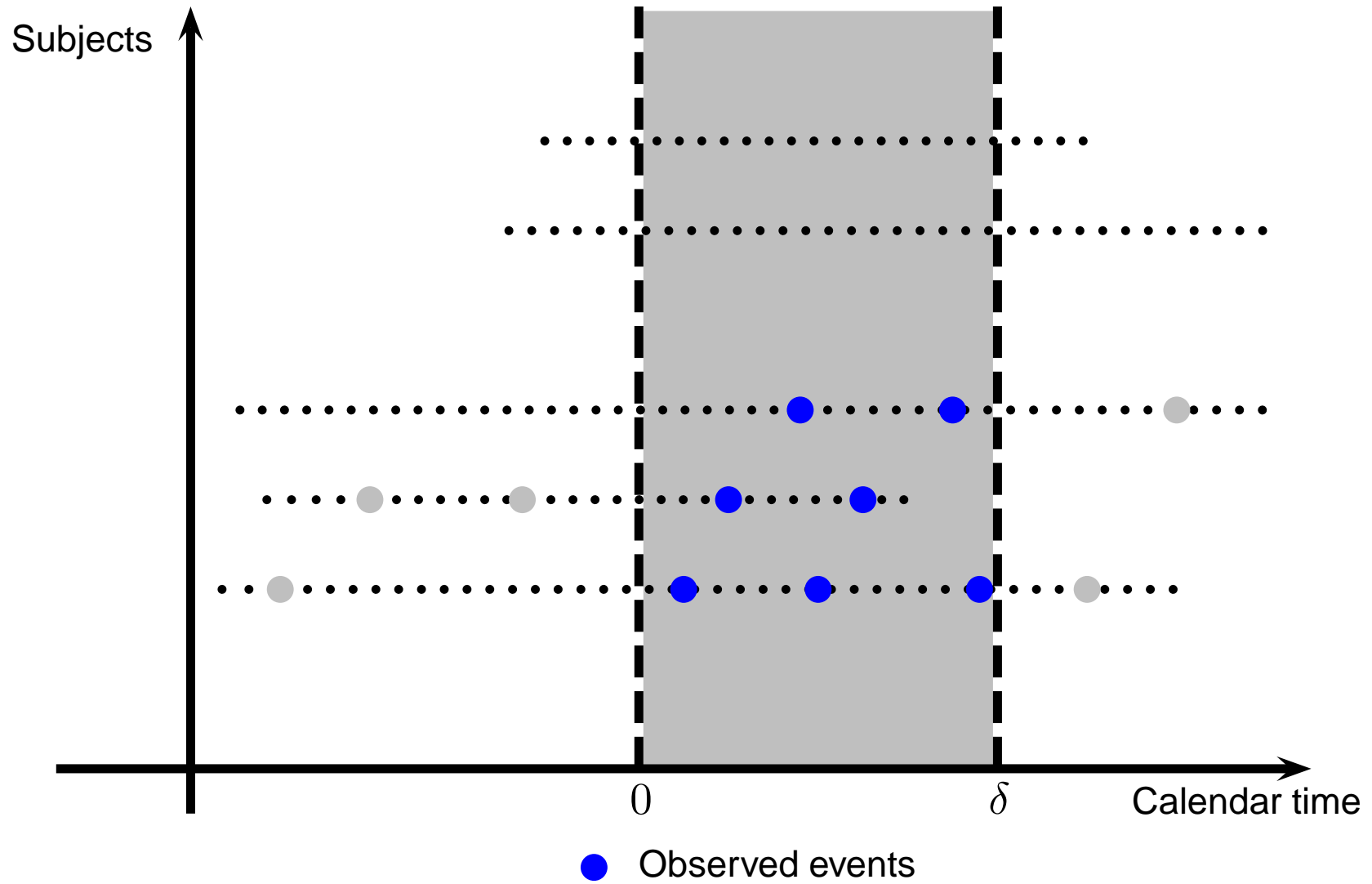
- Follow-up on cross-sectional sample
- Short observation period prohibits use of run-in period

3 The challenge

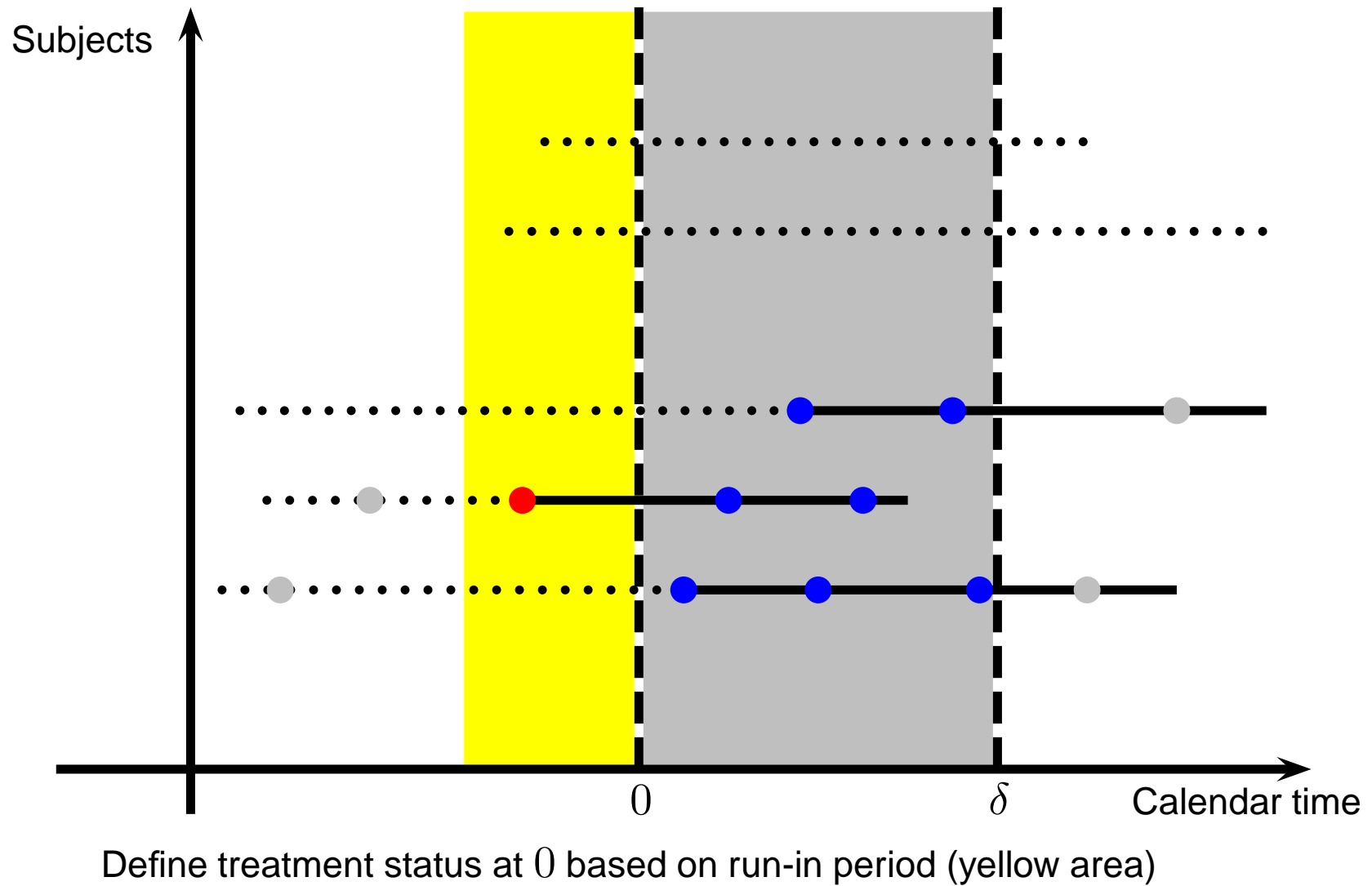
- Databases with health-related events are available, for a sample or population wide
- Problem
 - The individual treatment status is *not* observed between events
- Challenge:
 - estimate prevalence, incidence and mortality based on these databases
 - should avoid run-in period
 - should be applicable to both OPED and FLUKS type databases

4 Characteristics of Health Service databases



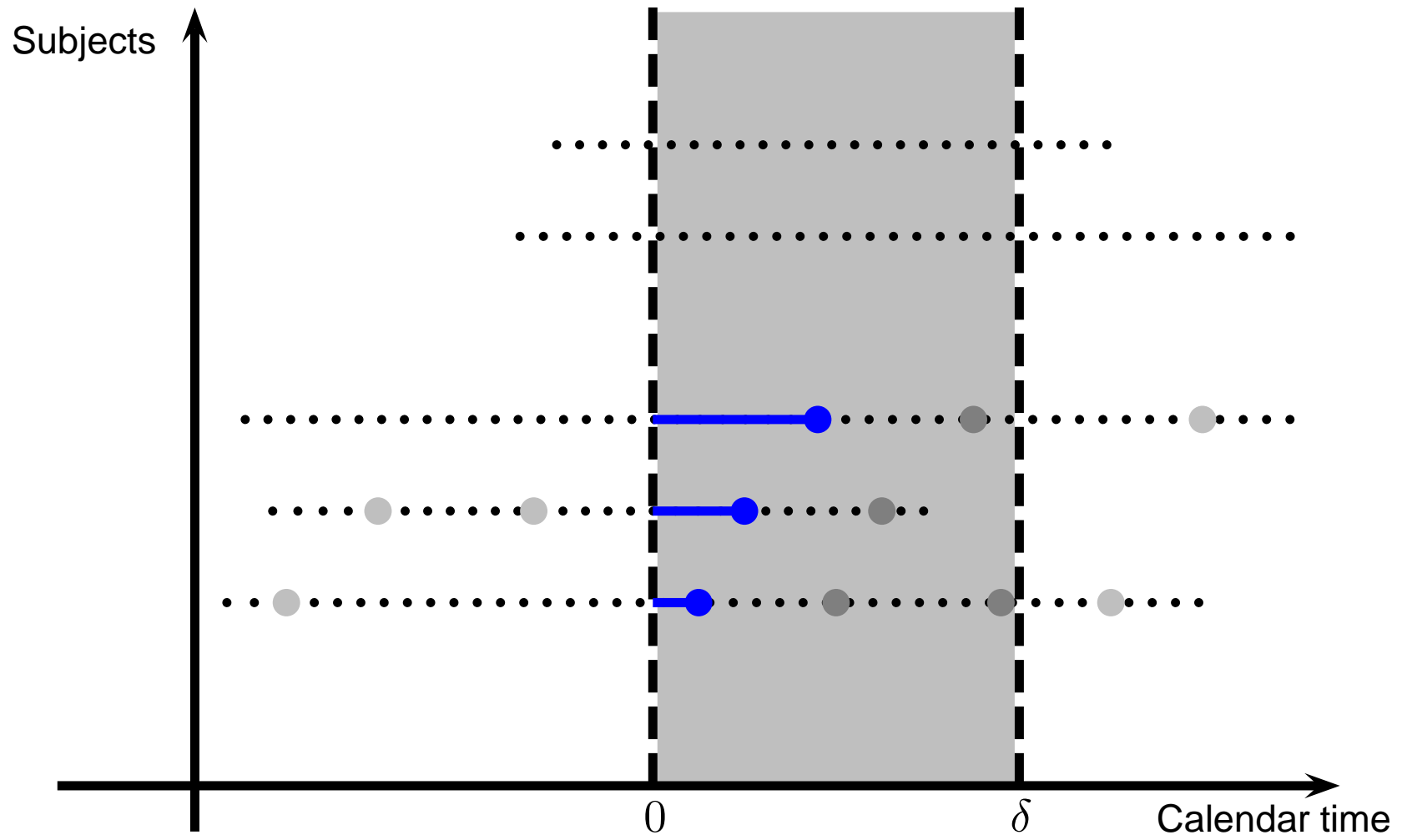


4.1 Standard analysis: Use run-in period



5 The idea of Hallas et al.

- Jesper Hallas, David Gaist, and Lars Bjerrum. The waiting time distribution as a graphical approach to epidemiologic measures of drug utilization.
Epidemiology, 8:666–70, 1997
- Basic idea:
 - 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

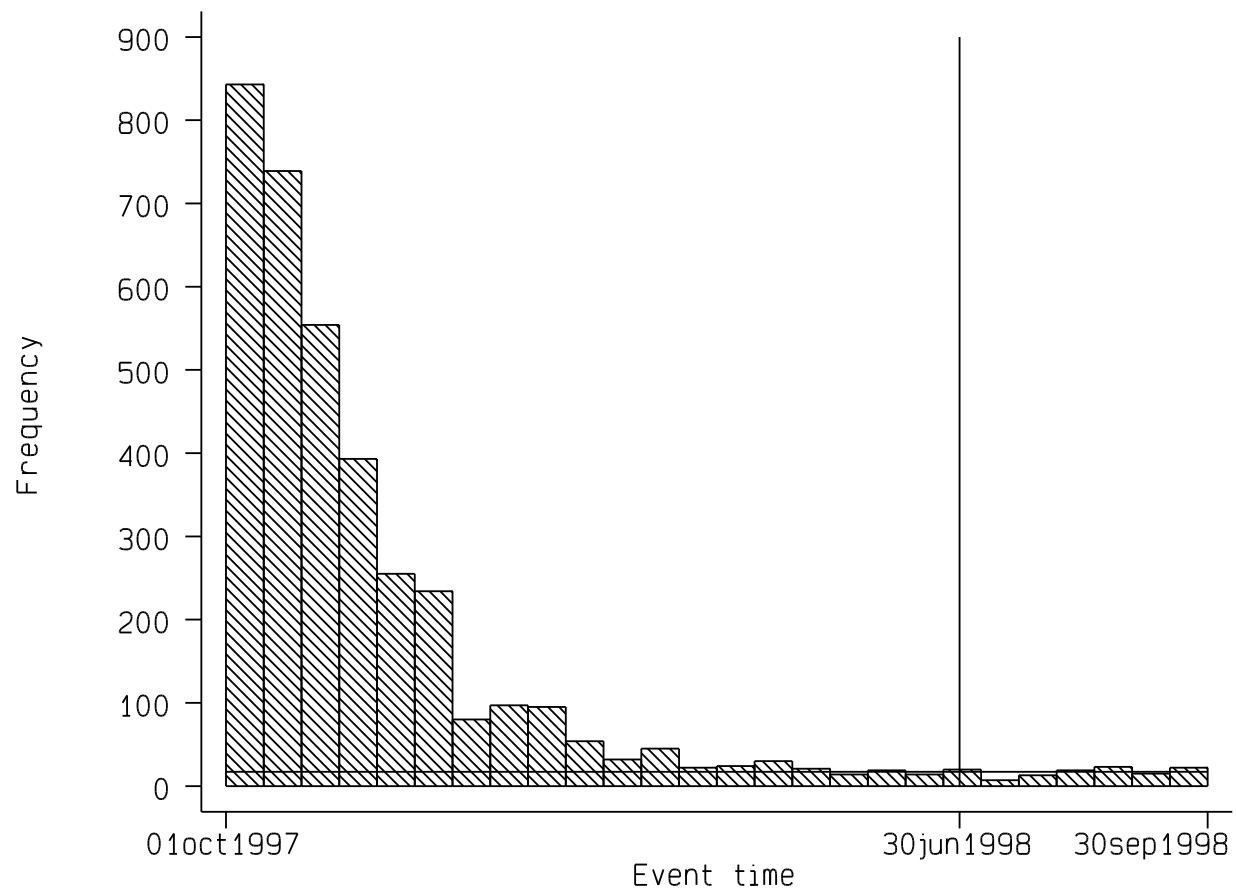
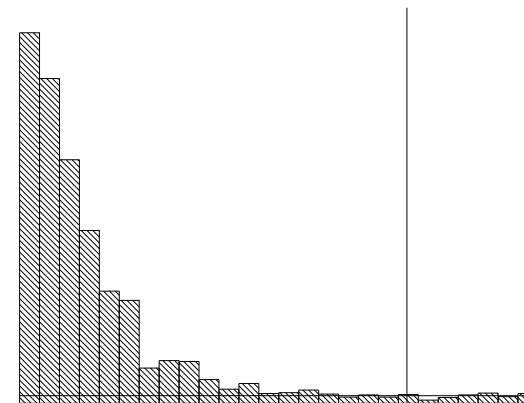


Figure 1: Insulin

5.1 Evaluation of the approach of Hallas et al.

- Three types of assumptions
 1. All subjects prevalent at 0 have first event before cut-point
 2. Population is considered stable, ie. on average:
 - (a) incidence is constant
 - (b) emigrants are replaced by immigrants
 3. first event of immigrants arriving after cut-point is an incident event

- Is applicable at start of database
- Assumption 3 is generally questionable
- Assumption 2 → **not applicable to FLUKS**



6 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: $\lambda(0) = \lim_{\Delta t \downarrow 0} \frac{P(X_{0+\Delta t} = 1 | X_0 = 0)}{\Delta t} = f^{T|X=0}(0)$

- Initial model

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

7 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

7.1 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

$$I(T \leq \delta) \quad \text{and} \quad I(Z \leq \delta) \quad (4)$$

since we only have short observation period

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

- Consequences of long term dependencies

$$P(T \leq \delta, Z \leq \delta) \neq P(T \leq \delta)P(Z \leq \delta) \quad (5)$$

- In particular we would expect that

$$P(T \leq \delta, Z \leq \delta) \geq P(T \leq \delta)P(Z \leq \delta) \quad (6)$$

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

$$P(T \leq \delta), \quad P(Z \leq \delta), \quad \text{and} \quad \phi = \frac{P(T \leq \delta, Z \leq \delta)}{P(T \leq \delta)P(Z \leq \delta)} \quad (7)$$

- 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 $I(Z \leq \delta)$

7.2 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

8 Likelihood construction

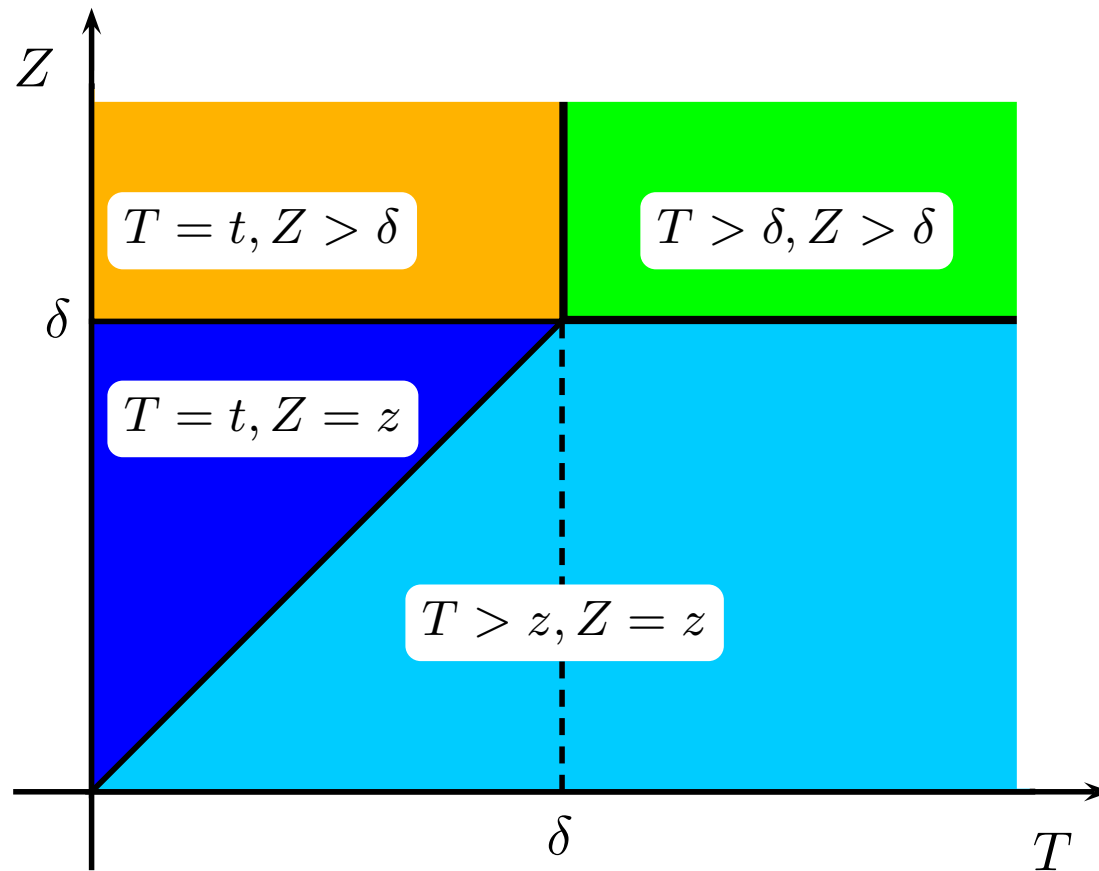


Figure 2: Observational areas of (T, Z)

8.1 Examples of likelihood contributions

- Consider the observation

$$T = t, Z = z$$

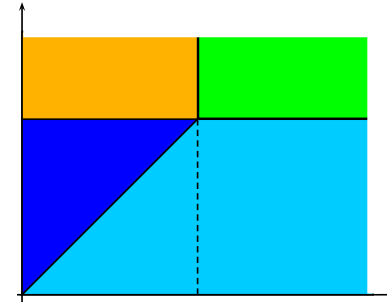
- The likelihood contribution has the structure

$$l^1(p, \theta) = p \cdot f^{(T,Z)|X=1}(t, z; \theta) + (1 - p) f^{(T,Z)|X=0}(t, z; \theta) \quad (8)$$

- In the remaining contributions we integrate over the missing data

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

$$l^2(p, \theta) = p \int_{\delta}^{\infty} f^{(T,Z)|X=1}(t, s; \theta) ds + (1 - p) \int_{\delta}^{\infty} f^{(T,Z)|X=0}(t, s; \theta) ds \quad (9)$$



8.2 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

$$g(t) = \frac{\int_t^\infty \gamma(s) ds}{\int_0^\infty \gamma(s) ds} \quad (10)$$

$$= \frac{S_\gamma(t)}{\mu} \quad (11)$$

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

8.3 Choice of distributions: Incidence and exit

- All conditional densities are chosen as $U(0; \delta)$, since
 - Ensures identifiability for incidence
 - Very easy to handle numerically and analytically
 - For small rates very similar to exponential on short intervals
- Note:
 - Not the same uniform as in Hallas et al's approach
 - Ignores seasonal variation

9 Results for FLUKS: Hypertension

Model type	Likelihood	Prevalence	Incidence
Exponential	-1529.49	.192 (.162; .227)	.0412 (.0214; .0778)
Weibull	-1520.72	.159 (.135; .186)	.0712 (.0570; .0885)
Log-Normal	-1522.04	.164 (.140; .191)	.0664 (.0523; .0838)

Table 1: Results for hypertension with 95%-confidence intervals based on robust variance estimates adjusting for clustering on GP.

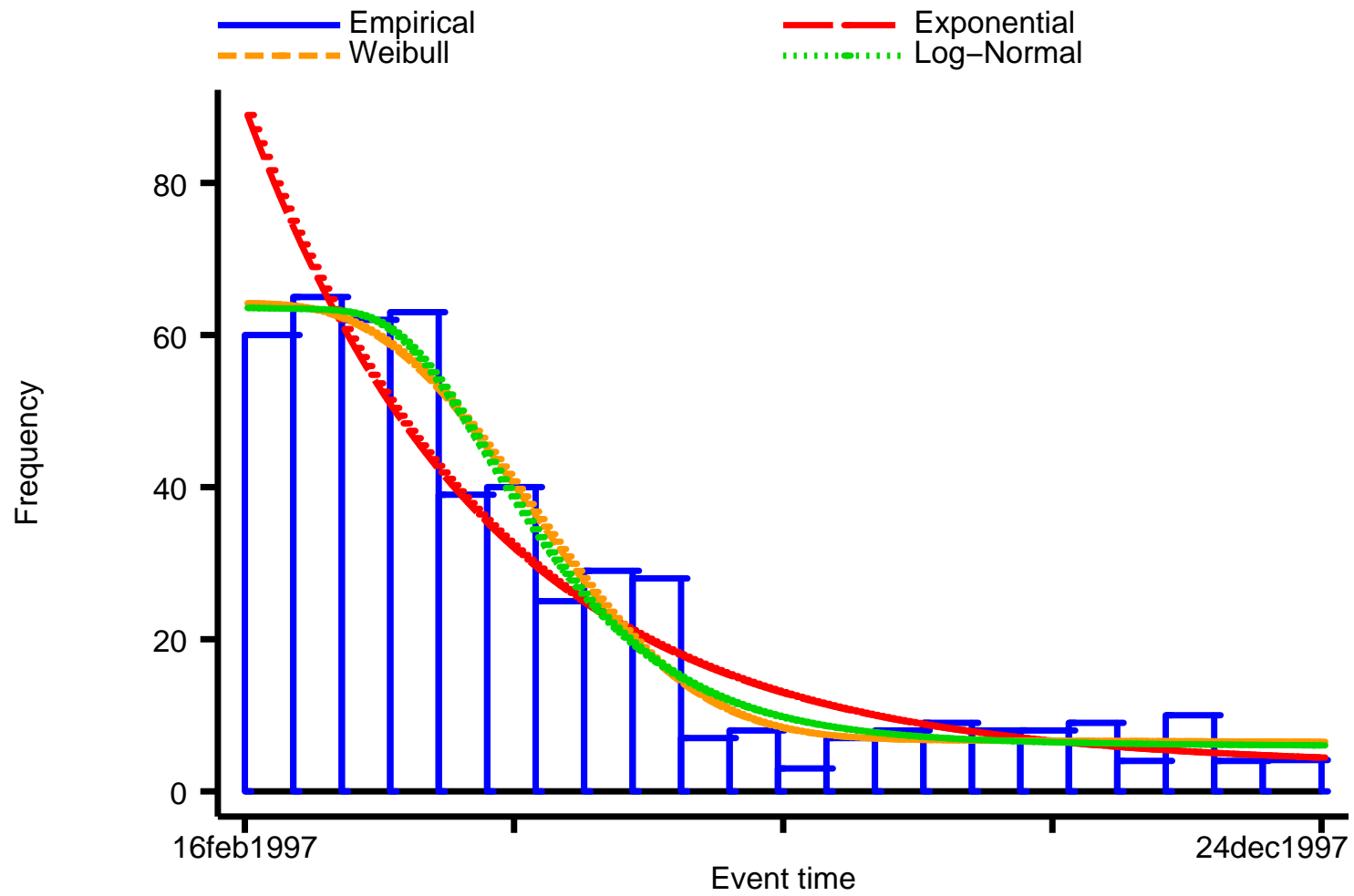


Figure 3: Diagnostic plot showing the fit of three different models.

10 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 two central questions
 - Sensitivity to misspecification
 - Validity of inference, ie. coverage probabilities
- 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 negligible
 - In comparison to studies with observed disease/treatment history, relative sample sizes of two to four yields similar precision (prevalence 1.5, incidence 4)
 - Nominal levels of confidence intervals are maintained
 - Very small impact of censoring levels
- For misspecified models
 - Bias increases with level of misspecification
 - Precision decreases with level of misspecification
 - Even in extreme situations, the coverage probability of 95%-CIs remains approximately 90%

11 Estimation of Mortality

- How can we estimate mortality among treated?
- We can not include duration of treatment in model – it is unobserved
- We can however consider estimating the following:

$$\lambda^P = \lim_{\Delta t \downarrow 0} \frac{P(D_{0+\Delta t} = 1 | D_0 = 0, X_0 = 1)}{\Delta t} \quad (12)$$

where D_t is an indicator of survival status at time t

- Interpretation: The mortality rate at 0 among those in treatment at time 0
- Note: Does not account for time of disease/treatment onset

11.1 Fundamental setup

- Now four stochastic variables

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

$$X = \begin{cases} 1 & \text{if treated at } 0, \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

T : Time of first event after 0

U : Time of death after 0

Z : Emigration time after 0

- Note: We don't get to observe all three, since U and Z censors each other^a

^aActually we do get info on death after emigration, since CPR-registry is nationwide. Ignored here.

11.2 Likelihood contributions

- The fundamental model is again a two component mixture

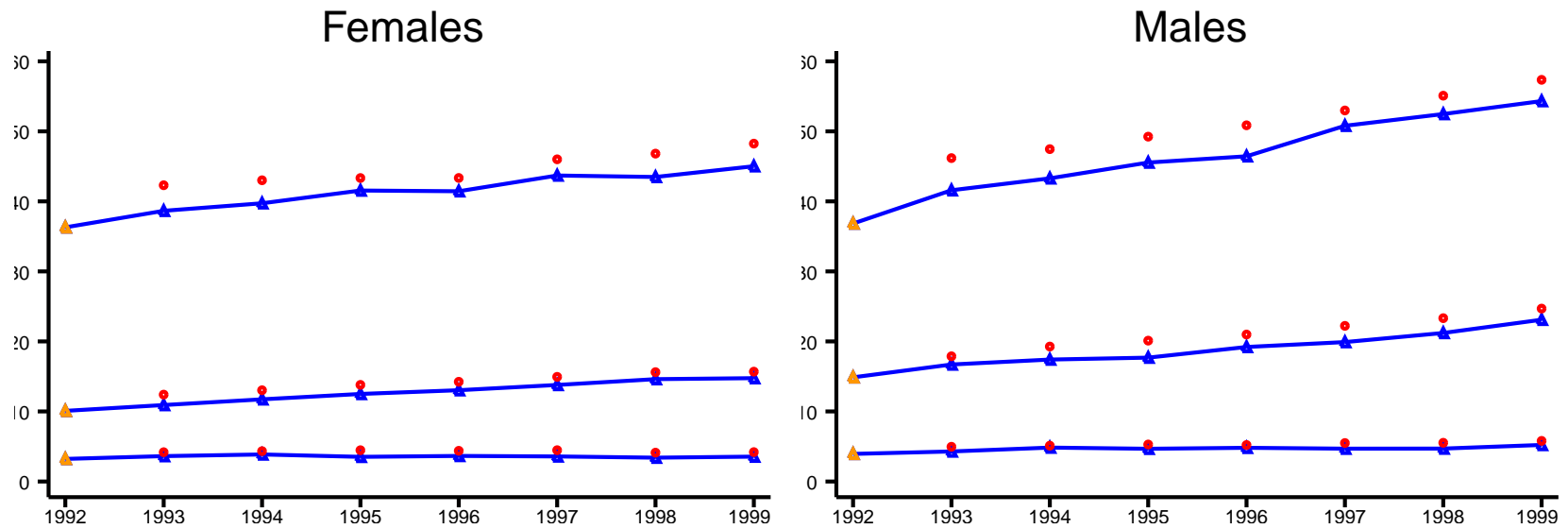
$$\begin{aligned} f^{(T,U,Z)}(t, u, z; \theta) &= p \cdot f^{(T,U,Z)|X=1}(t, u, z; \theta) \\ &\quad + (1 - p) f^{(T,U,Z)|X=0}(t, u, z; \theta) \end{aligned} \quad (14)$$

- Now we have six different types of contributions
- For example we might observe $T = t, U = u$ and $Z > u$:

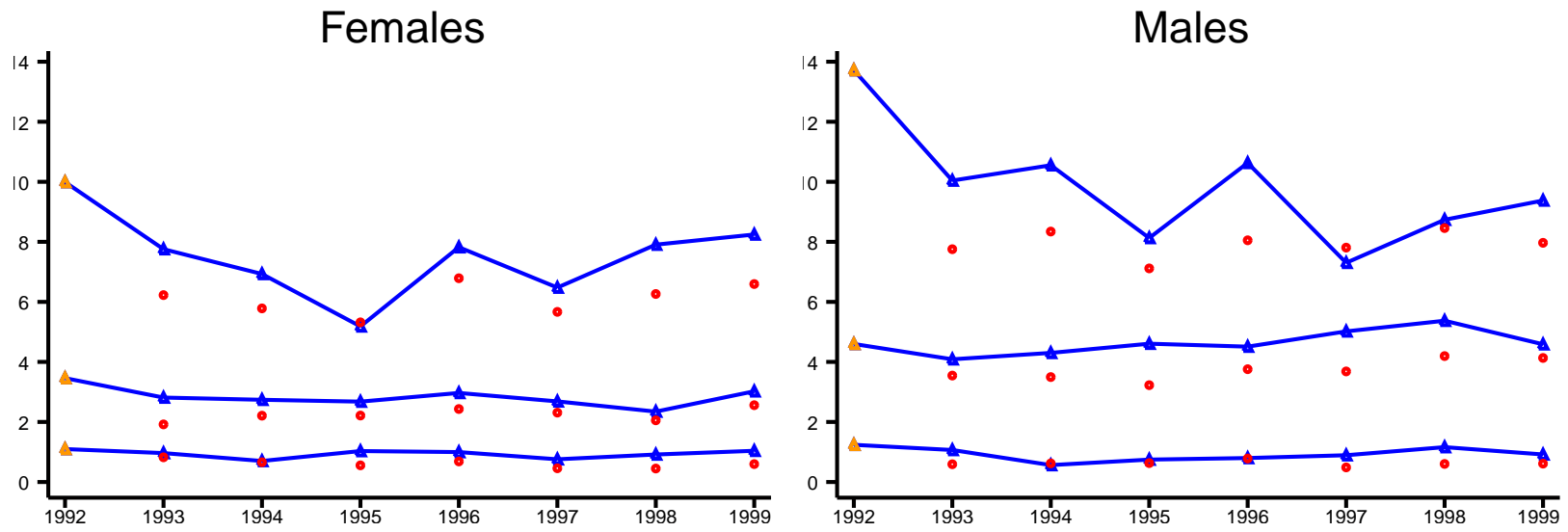
$$\begin{aligned} l^1(p, \theta) &= p \int_u^\infty f^{(T,U,Z)|X=1}(t, u, z) dz \\ &\quad + (1 - p) \int_u^\infty f^{(T,U,Z)|X=0}(t, u, z) dz \end{aligned} \quad (15)$$

- As before we allow for dependence of indicators for non-prevalent subjects
- For treated subjects we assume independence of event, death and exit

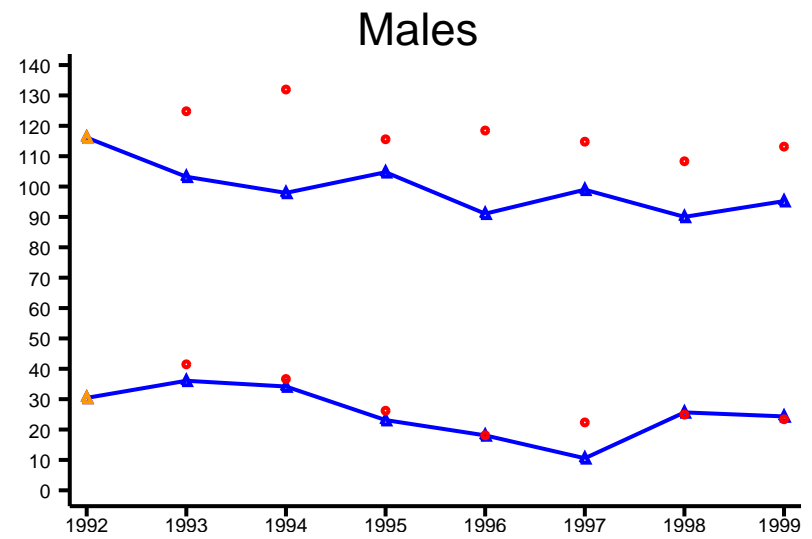
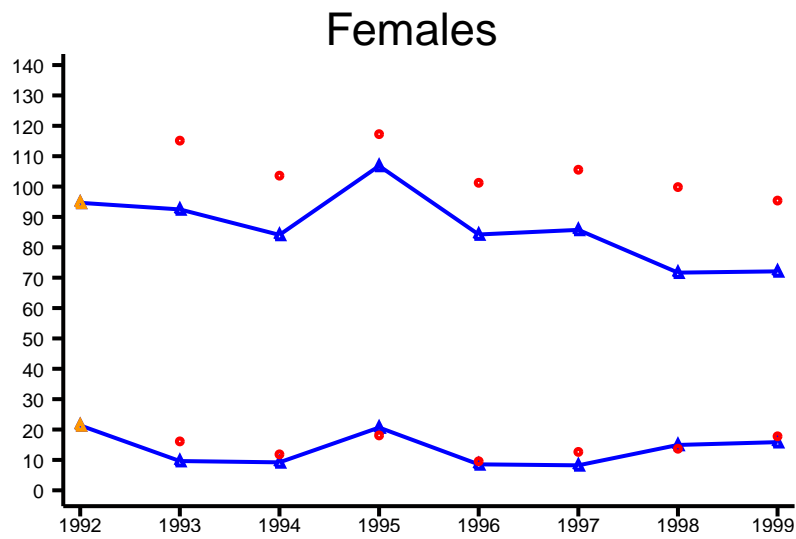
12 Diabetes epidemic: Motivating example re-analyzed



(a) Prevalence per 1000



(b) Incidence per 1000 person-years

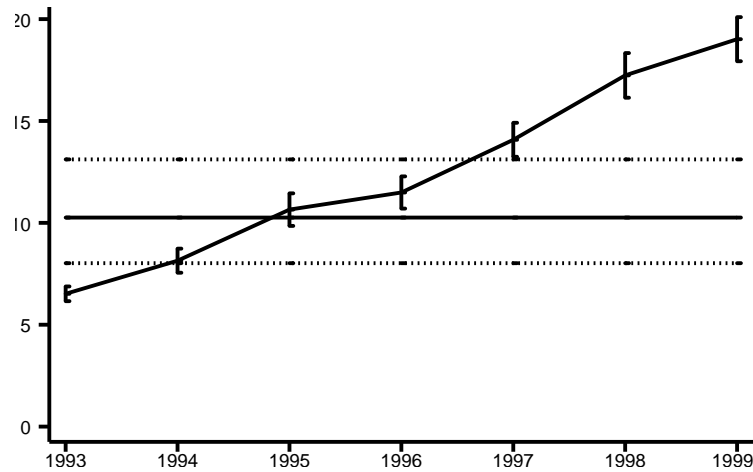


(c) Mortality among treated per 1000 person-years

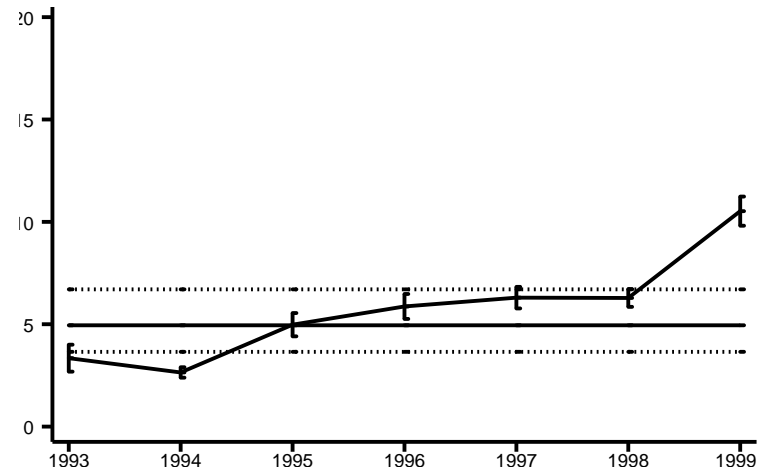
13 Treatment status among immigrants

- What is the prevalence among immigrants?
- What is the incidence among immigrants subsequent to entry?
- May be studied if we modify the initial model as follows
 - Let X be treatment status at entry
 - Let T be time until first event after entry
 - Let Z be time until first exit after entry
- Use a period of length δ as follow-up for all immigrants
- Again, we have a two-component mixture model for (T, Z)

13.1 Example: Use of antidepressants among immigrants

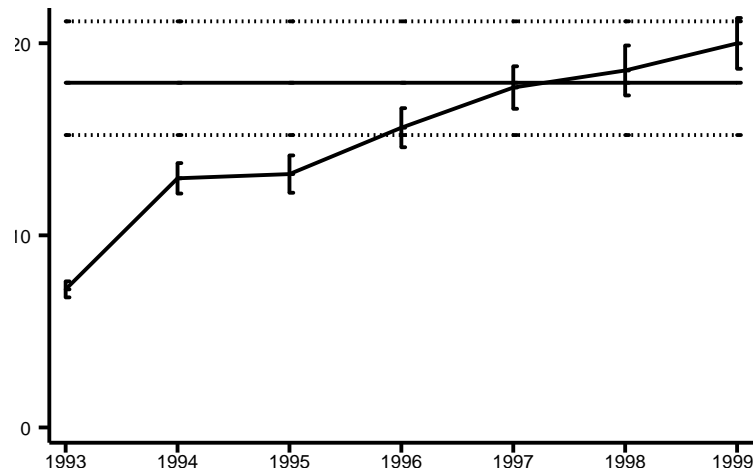


(d) Females

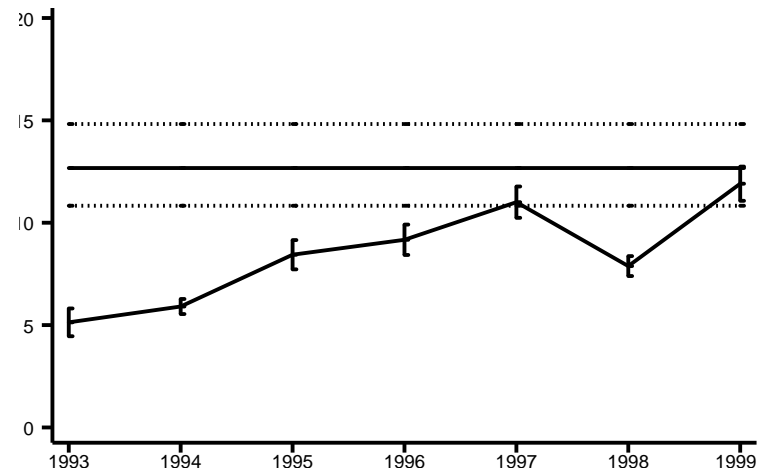


(e) Males

Figure 4: Prevalence per thousand of use of antidepressants among immigrants (horizontal lines) and non-immigrants.



(a) Females



(b) Males

Figure 5: Incidence per thousand person years of use of antidepressants among immigrants (horizontal lines) and non-immigrants.

14 Summary

- Formulated an intuitive idea as a statistical model
- Obtained model, which
 - allows estimation without run-in or cut-point
 - * can be used at start of database
 - * does not exclude immigrants
 - applicable to follow-up on cross-section (FLUKS)
 - incorporates dependent censoring
 - allows maximum likelihood estimation
 - allows construction of diagnostic procedures
 - robust to misspecification

- Problems with the approach
 - Static, ie. ignores interrelation of incidence, prevalence and mortality
 - Relies on a finite time interval containing at least one event for all prevalents
 - Parameters linked to length of observation window
 - Parametric model with intended robustness, but difficult to justify theoretically
 - Only uses first event time in observation window

- Perspectives
 - Non-chronicity
 - Non- or semiparametric modeling
 - Dynamic approach
 - Frailty

- Kierkegaard:

Do not engage with science. You are left defenseless with no control at all. The Scientist immediately begins distracting with details, now to Australia, now to the moon, now to a subterranean cave – all for a tapeworm... Who can stand this?^a

^aAuthors' translation, censored for explicit language.