Michael Væth

EPIDEMIOLOGY FOR BIOSTATISTICIANS

January 8-10 2007

Day 1, 9 – 11:

1. Basic concepts: Terminology, measures of disease occurrence, measures of association, regression models, main designs.

2. The epidemiological study as a measurement exercise: Precision, Validity, Selection problems, Confounding, Information problems. Generalizability. Effect-modification.

3. Independent risk factors.
   The effect of omitting covariates in RCTs.
   Omitting covariates in multiple regression and logistic regression.

1

"Epidemiology is concerned with the patterns of disease occurrence in human populations and the factors that influence these patterns. Epidemiologists are primarily interested in the occurrence of disease as categorized by time, place and person".

Personal characteristics include:

- Demographic factors
- Biological characteristics and genetic factors
- Life style and social and economic factors

Purpose:

- To elucidate the etiology of a disease
- To evaluate the consistency of epidemiologic data with etiological hypotheses
- To evaluate preventive procedures and public health practices

What causes the disease? Can it be prevented?

Here focus on the methods used to answer these questions

2

Problem: epidemiologic studies are typically observational studies. (Causal) interpretation of the results becomes complex.

Terminology

Outcome: The variable or factor that summarizes the disease or health phenomenon under study
Determinants or exposures: Covariates or factors that may influence the outcome
Treatment is often used for exposures that are controlled.

Most epidemiologic methods concern categorical outcomes, typically a dichotomies.

Ex. Presence or absence of a specific diagnosis. Cause of death

Continuous outcomes are occasionally considered

Ex. Blood pressure.

3

Measures of disease occurrence I

Two basic aspects:

- Prevalence: Who have the disease?
- Incidence: Who develop the disease?

Proportions

Prevalence (proportion): the relative frequency of individuals in a population (or sample), who have the disease in question at a given point in time.

Cumulative incidence (proportion): The relative frequency of healthy individuals, who develop the disease in a given period of time

The statistical model should allow estimation of these quantities, either directly as parameter estimates or indirectly as functions of other parameters.

Note: In epidemiology the distinction between the theoretical quantities (probabilities) and their estimates (proportions) is seldom made.

4

Measures of disease occurrence II

Rates
The "time" aspect is an important. A description of how the "instantaneous risk" depends on time is useful

Incidence rate: The relative rate of change of the population size due to the disease

In statistical terms:
The incidence rate is the hazard rate of the time-to-disease distribution or the transition intensity in a simple two state model (healthy, diseased) of the life course of an individual.

To model epidemiologic data important to consider

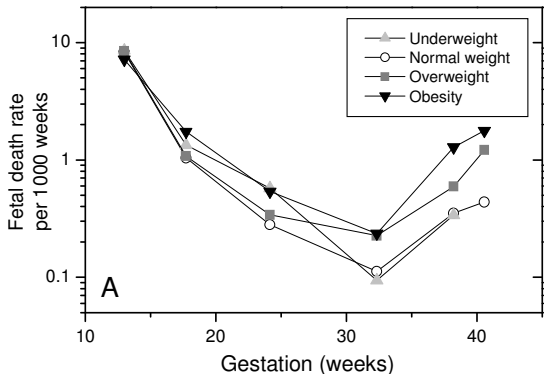
- Choice of time scale – usually age or time since exposure
- Censoring and truncation
- Competing causes

5

Measures of disease occurrence III

Rates: If the hazard rate is (piecewise) constant the maximum likelihood estimate becomes Events/Risktime.

Example: Fetal death by pre-pregnancy BMI categories in DNBC
Features: Left truncation, right censoring, competing causes



6

Exposure-outcome relationships - Measures of association

Two exposure categories: Unexposed (0) and Exposed (1)

Measures based on proportions

Relative risk: $RR = \frac{p_1}{p_0}$

Risk difference: $RD = p_1 - p_0$

Odds ratio: $OR = \frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}$

Measures based on rates

Hazard ratio $HR(t) = HR = \frac{\lambda_1(t)}{\lambda_0(t)}$

Continuous outcomes are usually described by linear models and the "effect" of exposure is the difference between expected outcomes

7

The basic regression models used in epidemiology I

Binomial regression models

Generalized linear model with a link function that correspond to the measure of association

Odds ratio:	logit link (logistic regression)
Relative risk:	log link
Risk difference:	identity link

Log-linear hazard rate models

Data with records on individuals

Proportional hazards regression models (e.g. Cox regression)

Aggregated data (multiway tables of counts and person-years)

Poisson regression (log-linear model with piecewise constant hazard rate)

8

The basic regression models used in epidemiology II

Continuous outcome (Gaussian)

Multiple regression models

Unlike the multiple regression model the binomial regression models are derived from a one-parameter exponential family and the variance is a function of the mean.

Model-based standard errors of estimates may underestimate the uncertainty in the estimates.

Solutions:

- Use "robust" standard errors (GEE methodology).
- Generalized linear mixed models: Introduce additional random component(s) in the model

9

Robust standard errors – Are they always robust?

```
. logistic resp1 covx[fw=number], coef

Logistic regression                     Number of obs   =         70
                                        LR chi2(1)      =       3.01
                                        Prob > chi2     =     0.0827
Log likelihood = -27.202656             Pseudo R2       =     0.0524
------------------------------------------------------------------------------
       resp1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        covx |   .3111931   .1882717     1.65   0.098    -.0578127    .6801988
       _cons |  -1.926355   .3834859    -5.02   0.000    -2.677973   -1.174736
------------------------------------------------------------------------------

. logistic resp1 covx [fw=number], coef robust

Logistic regression                     Number of obs   =         70
                                        Wald chi2(1)    =      13.87
                                        Prob > chi2     =     0.0002
Log pseudo-likelihood = -27.202656      Pseudo R2       =     0.0524
------------------------------------------------------------------------------
             |               Robust
       resp1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        covx |   .3111931   .0835506     3.72   0.000     .1474368    .4749493
       _cons |  -1.926355   .3367578    -5.72   0.000    -2.586388   -1.266322
------------------------------------------------------------------------------
```

10

Populations

Study population: The individuals who are included in the study.

Source population: The population from which the study population is obtained

Target population: The population to which the results are to be applied

The terminology is not used consistently

The study population is the sample(s) – but rarely random sample(s). Inference in the study population may therefore not apply directly to the source population (or the target population).

Important: Investigate to what extend the findings in the study population are valid more generally.

11

The design of epidemiologic studies I

Cohort studies. From causes to effects

Basic form: Disease occurrence in a study population of healthy individuals, who are followed forward in time. Exposure is assessed at entry.

In principle: A **fixed cohort** followed for a fixed time period allows direct estimation of cumulative incidence and the corresponding measures of association

In practice: Censoring by competing causes and loss to follow-up make models based on incidence rates more attractive.

Dynamic cohort: Entry in and exit from the cohort during follow-up.

Lexis diagrams (more later) are useful to clarify more complex designs

Choice of **time scale**
Prolonged exposure: Hormone Replace Therapy and Breast Cancer

12

The design of epidemiologic studies II

Case-control studies. From effects to causes

Problem with cohort studies of rare diseases:
Large studies are required to obtain the number of cases needed.

Alternative approach:
Case-control studies or case-referent studies

Basic idea: Identify the cases in a hypothetical cohort study and compare their exposure histories with those of a suitable sample of referents

Several types of case-control designs reflecting choice of measure of association and/or use of matching between cases and controls

13

The design of epidemiologic studies III

Case-control studies. *Controls as non-cases.*
Measure of association: **Odds Ratio**

Source population:

	case	non-case
E=1	p_{11}	p_{10}
E=0	p_{01}	p_{00}

or

	case	non-case
E=1	P_1	1 - P_1
E=0	P_0	1 - P_0

$$OR_{pop} = \frac{p_{11}p_{00}}{p_{10}p_{01}} = \frac{P_1/(1-P_1)}{P_0/(1-P_0)}$$

Sampling probabilities: π_{ij}

	case	non-case
E=1	$p_{11}\pi_{11}/c$	$p_{10}\pi_{10}/c$
E=0	$p_{01}\pi_{01}/c$	$p_{00}\pi_{00}/c$

$$OR_S = \frac{p_{11}p_{00}}{p_{10}p_{01}}\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}$$

if $\pi_{11} = \pi_{01}$ and $\pi_{10} = \pi_{00}$
i.e. sample fraction independent of exposure then $OR_{pop} = OR_S$

14

The design of epidemiologic studies IV

Case-control studies. Controls = Random sample of Source Population. Measure of association: **Incidence Rate Ratio**

Rate Ratio: $$\frac{Y_1}{T_1} \Big/ \frac{Y_0}{T_0} = \frac{Y_1}{Y_0} \Big/ \frac{T_1}{T_0}$$

only the relative size of the time-at-risk is needed

Basic idea: Use the ratio of exposed to unexposed in a random sample from the source population instead.

If possible: Use sampling probabilities proportional to the contribution to the time-at-risk

The odds ratio in the sample estimates the rate ratio in the source population.

Nested case-control studies
Case-cohort studies
Case-crossover studies
Macthed case-control studies (also counter-matching)

15

The design of epidemiologic studies V

Cross-sectional studies

Basic idea: Disease status and exposure status are determined in a study population in a "given point" in time.

Allow estimation of prevalences and the corresponding measures of association.

The timing of disease and exposure may not be available so causal interpretation is copmplex (at best).

Further design options:
Prevalent cohort: Follow-up of a study population identified in a cross-sectional study

16

Precision and validity in Epidemiologic studies

"One way to formulate the objectives of an epidemiologic study is to view the study as an exercise in measurement"

Precision: Random error

Validity: Bias

- **Internal validity:**
  From the study population to the source population
  - Selection bias
  - Information bias
  - Confounding
- **External validity - Generalizability**
  Beyond the source population

17

Precision

Random variation – lack of precision - unpredictability

Sources of uncertainty
- Sampling variation
- Unexplained heterogeneity
- Measurement errors in key variables

Model-based standard errors of estimates reflect sampling variation only. Alternative, robust, estimates of uncertainty may be more reliable.

To improve precision
- Increase sample size
- Stratified sampling, matching
- Improve measurement of key variables
- Reduce heterogeneity by refining the modeling

18

Internal validity – Selection Bias I

Selection bias: The relation between exposure and disease is different for members of the **study population** (the participants) and for members of the **source population** (the eligible persons).

Selection bias: A result of procedures used to select study subjects and/or factors that influence study participation

Associations observed in the study population represents a mix of forces determining participation and forces determining disease.

Case-control studies are particularly vulnerable to selection biases

Cohort studies: Non-participation, self-selection

Examples:

- Danish National Birth Cohort (DNBC): Approx. 30% of all eligible pregnant women were included in the cohort.
- Life Span Study (A-bomb survivors): Cohort members were alive when the cohort was established in October 1950.

19

Internal validity – Selection Bias II

Cohort studies in occupational epidemiology:
"Healthy worker" effect

Mortality (or disease incidence) in an occupational cohort is **compared to national figures**. Cohort members, typically identified as workers employed at a given date, are followed forward in time.

Selection bias may occur at the time of first employment. and Moreover those still working when the study is initiated are a "survivors".

Internal comparisons (if feasible) may reduce (eliminate?) the healthy worker effect

Example: "Un-healthy worker" effect: Thule cohort

20

Internal validity – Confounding I

The basic problem: A comparison of disease occurrence in a sample of exposed with disease occurrence in a sample of unexposed.

Experimental research:
Randomization, balancing, blinding, etc. ensure a valid comparison

Observational research:
The apparent exposure effect may also reflect the effect of other factors, which differ between the two samples.
Such factors are called confounding factors or confounders

How do we identify confounders?

How do we correct for the influence of confounders when estimating the effect of exposure?

21

Internal validity – Confounding II

Criteria for a confounding factor (definition?)

1. A confounding factor must be a risk factor for the disease (among unexposed).

risk factor = a "cause" or a factor associated with a "cause" (a marker or a surrogate)

2. A confounding factor must be associated with the exposure (in the source population)

3. A confounding factor must not be affected by the exposure or the disease. In particular, it cannot be an intermediate step in the causal path between the exposure and the disease

Note: These criteria require information outside the data

22

Internal validity – Confounding III

Counterfactual definition of confounding: The average counterfactual response in the exposed study population differ from the average response in the unexposed study popuation.

Operational "definition" of confounding:
If the association between exposure and disease is unchanged when adjusting for a factor, the factor is not a confounder.

Fundamental problem: The definition depends on the chosen measure of association.

Practical problem: Due to sampling variation small, "unimportant" changes are expected.

Solution(?): Change-in-estimate criteria: accept changes less than 10% (or 5% or…)

The definition of non-confounding as collapsibility or of confounding as non-collapsibility **differs** from the "classical" definition.

23

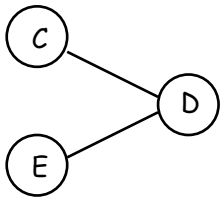
Internal validity – Confounding IV

Non-confounding and collapsibility

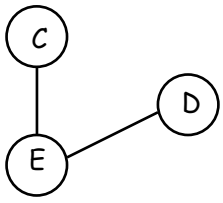
Contingency tables: Three-way table classified by E, D and C.
A measure of association of E and D is collapsible across C if it is constant across the conditional subtables and this constant value equals the value in the marginal (E,D) table.

Collapsibility of the odds ratios in a contingency table (or of the parameter in a logistic regression) :

$C \perp E \mid D$



$C \perp D \mid E$



24

Internal validity – Confounding V

Correction for the effect of confounding factors

In the **design**:

- Matching (used mainly for case-control studies)
- Restriction

In the **analysis**:

- Stratification (Mantel-Haenszel methodology)
- Regression modeling
- Standardization methods (weighting)

If the number of confounding factors is large some of these methods are not feasible. Alternative approach: **Propensity score**

Confounding and intermediate factors
Analyses both with and without correction for a particular factor may provide useful information on the relationships (mediation)

25

Internal validity – Information Bias I

Measurement error and misclassification

Measurement error: E true exposure, F measured exposure

Two basic types:

		Note:
Classical:	$F = E + error_C$	Var(F) > Var(E)
Berkson:	$E = F + error_B$	Var(E) > Var(F)

Consequences in a simple linear model with additive error

Classical model

$$Y = \alpha + \beta_E E + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Observed exposure: $F = E + U, \quad U \sim N(0, \sigma_U^2)$
U independent of the error ε

The regression of Y on F: $\beta_F = \dfrac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}\beta_E$

26

Internal validity – Information Bias II

Residual variation: $Var(Y \mid F) = \sigma_\varepsilon^2 + \frac{\beta_E^2 \sigma_U^2 \sigma_X^2}{\sigma_X^2 + \sigma_U^2}$

Use of observed exposure instead of the true exposure leads to an **attenuation of the slope** and **increased residual variation**

Berkson model

Relation between exposures: $E = F + V, \quad V \sim N(0, \sigma_V^2)$
V independent of the error ε

The regression of *Y* on *F:* $\beta_F = \beta_E$

Residual variation: $Var(Y \mid F) = \sigma_\varepsilon^2 + \beta_E^2 \sigma_V^2$

Use of observed exposure instead of the true exposure leads to **an unbiased slope estimate** and **increased residual variation**

General solution methods: Regression calibration, modeling (e.g. EM algorithm), SIMEX

27

Internal validity – Information Bias III

Misclassification of exposure, disease or confounders

Differential misclassification – the misclassification probabilities dependent on other variables
Non-differential misclassification

Non-differential misclassification in exposure or outcome leads usually to bias against the null value.

Non-differential misclassification in a confounder reduces the degree to which confounding can be controlled. May cause bias in either direction (depending on the direction of the confounding).

Differential misclassification: Can cause bias in both direction.

If validation data are available: Correction for misclassification by modeling the misclassification or by sensitivity analyses

Example: Cause of death in LSS

28

Internal validity – Information Bias IV

Example: Pregregnancy BMI in DNBC and pregnancy outcomes

	DNBC				
NJC	Under-weight	Normal weight	Over-weight	Obese	Total
Underweight	149	24	0	0	173
Normal weight	72	2932	45	0	3049
Overweight	0	173	998	23	1194
Obese	0	3	96	518	617
Total	221	3132	1139	541	5033

Agreement: 91.4%

Consequences (from a sensitivity analysis):

Relative bias in *RR* for obese: ~-5%

29

External validity – Generalizability

Scientific generalizations: Move from a time-place specific results to an abstract "universal" hypothesis, e.g. "radiation causes cancer".

BUT this is a complex problem and insight in the mechanisms is usually required.

Example:
In the LSS cohort both additive and multiplicative models for the excess risk related to radiation may describe the data.

The consequences of these models are very different when the models are "transported" to other settings, since the background cancer rates differ between countries.

Similar problem with projections in time: Non-trivial differences between excess lifetime risk per unit dose derived from different models that fit equally well

30

Effect modification - Interaction

Effect modification: The effect of a risk factor depends on other covariates (effect modifiers).

Statistical terminology: Interaction between the effect of exposure and the effect of the covariate(s).
Effect modification is described by including interaction terms in the regression model.

Usually:

- We want to **eliminate** (reduce) the influence of **confounding factors.**
- We want to **describe effect modification**

But presence of effect modification depend on the measure of association used in the analysis

All other things being equal we may prefer to use a measure of association for which the data can be modeled without interactions terms.

31

Independent risk factors

In experimental studies randomization ensures

- Unbiased estimation of average treatment effect
- Treatment allocation is independent of covariates: Known and unknown risk factors are expected to be balanced
- Valid significance test of the hypothesis of no treatment effect can be based on the radomization distribution

In observational studies exposure is not allocated at random

Exposure may be related to important covariates, i.e. confounding is present.

Adjustment for the confounding factors is necessary. But what about independent risk factors?

32

The effect of omitting an independent risk factor X

	X=1 Y=1	X=1 Y=0		X=0 Y=1	X=0 Y=0		X=1	X=0	
E=1	20	5	25	10	15	25	25	25	50
E=0	15	10	25	5	20	25	25	25	50
	35	15	50	15	35	50	50	50	100

RR	1.33	2.00	
RD	0.20	0.20	
OR	**2.67**	**2.67**	

E and X are marginally independent

The covariate X is a risk factor

OR within exposure categories = 6

Summed over X

	Y=1	Y=0	
E=1	30	20	50
E=0	20	30	50
	50	50	100

RR	1.50
RD	0.20
OR	**2.25**

33

The effect of omitting an independent risk factor X

More general set-up:

Estimation of treatment effect in randomized experiments with non-linear regressions and an omitted covariate

Model	link	Bias against null
Binomial	identity	no
	log	no
	logit	yes
Poisson	identity	no
	log	no
Exponential	log	no
	inverse	yes
Cox reression		yes

The effect of omitted covariates on model-based Score test of no exposure effect has also been investigated

34

Omitted covariates – relation to marginal and conditional models

Implicit assumption: The values within levels of the covariate are "correct" and the value obtained when the covariate is omitted may be biased.

Another perspective: Some important (unknown) covariates are always missing so some heterogeneity is always expected.

The omitted-covariate-results are not biased but correspond to associations on the population level: A marginal model.

The associations within levels of the covariates correspond to conditional models, i.e. given the fixed and random factors that determine the relationship between exposure and disease.

Different answer to different questions!

35

Conventional wisdom – Multiple regression

Consider the two models

Model 1 $E(Y \mid E) = b_0^* + b_1^* E, \qquad Var(Y \mid E) = \sigma_1^2$

Model 2 $E(Y \mid E) = b_0 + b_1 E + b_2 X, \qquad Var(Y \mid E, X) = \sigma_2^2$

Data: Simple random sample. Both models are fitted to the data using method of least squares. Estimators of exposure effect: $\hat{b}_1^*, \hat{b}_1$

Asymptotic relative precision

$$ARP(\hat{b}_1 \; to \; \hat{b}_1^*) = \frac{\left[Var(\hat{b}_1)\right]^{-1}}{\left[Var(\hat{b}_1^*)\right]^{-1}} = \frac{Var(\hat{b}_1^*)}{Var(\hat{b}_1)} = \frac{1-\rho_{EX}^2}{1-\rho_{YX|E}^2}$$

No confounding (collapsibility), i.e. $b_1^* = b_1$ if one or both of the following conditions holds

Condition 1: $\rho_{XE} = 0$

Condition 2: $\rho_{YX|E} = 0 \quad (i.e. \quad b_2 = 0)$

36

Conventional wisdom – Multiple regression

Condition 1 alone: ARP > 1

Desirable to adjust for a predictive covariate in randomized studies

Condition 2 alone: ARP < 1

Undesirable to adjust for a non-predictive covariate which is correlated to the risk factor of interest.

Both conditions holds: ARP = 1

In general (also when confounding is present): The relative size of the two correlations determine whether ARP < 1, = 1 or > 1

37

Logistic regression

Y, E and X dichotomous variables. Consider the two models

Model 1 $\text{logit}(p) = b_0^* + b_1^* E$

Model 2 $\text{logit}(p) = b_0 + b_1 E + b_2 X$

Technical problem: Both models cannot be "true" simultaneously

Data: Random samples of exposed and unexposed individuals. Both models are fitted to the data using maximum likelihood. Estimators of exposure effect: $\hat{b}_1^*, \hat{b}_1$

Asymptotic relative precision

$$ARP(\hat{b}_1 \text{ to } \hat{b}_1^*) = \frac{\left[Var(\hat{b}_1 \mid E)\right]^{-1}}{\left[Var(\hat{b}_1^* \mid E)\right]^{-1}} = \frac{Var(\hat{b}_1^* \mid E)}{Var(\hat{b}_1 \mid E)} \leq 1$$

with equality if and only if X independent of (Y,E)

Proof: Minkowski's inequality

38

Logistic regression

No confounding (collapsibility), i.e. $b_1^* = b_1$ if one or both of the following conditions holds

Condition 1': E and X are independent given Y

Condition 2': Y and X are independent given E (i.e. $b_2 = 0$)

Condition 1' alone or condition 2' alone: ARP < 1

Both conditions hold: X independent of (Y,E) and ARP = 1

Unlike multiple regression, including a "non-confounding" covariate associated with Y leads to loss of precision of the exposure effect.

Does it matter in practice?

The loss in precision may not be negligible!

39