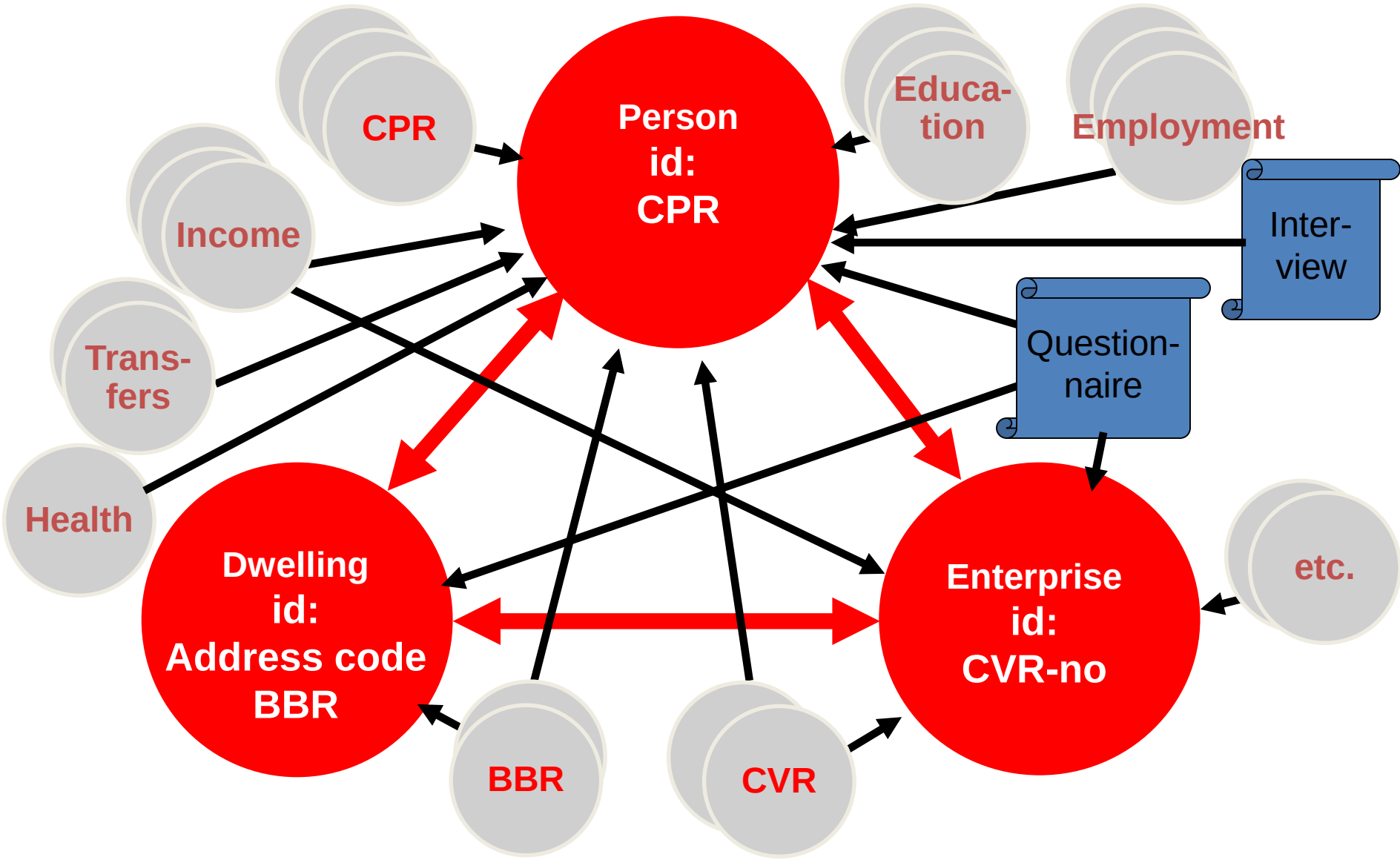
NordForsk PhD course in Register-Based Epidemiology

Introducing information bias, selection bias, and confounding in register-based studies

Lau Caspar Thygesen

The overall statistical system



EPIDEMIOLOGY

When an Entire Country Is a Cohort

Denmark has gathered more data on its citizens than any other country. Now scientists are pushing to make this vast array of statistics even more useful

For years, any woman who got an abortion had to accept more than the loss of her fetus: For some unknown reason, she also faced an elevated risk for breast cancer. At least that was what several small case-control studies had suggested before Mads Melbye, an epidemiologist at the Statens Serum Institute in Copenhagen, undertook the largest effort ever to explore the link. He and his colleagues obtained records on 400,000 women in Denmark's national Abortion Register, then checked how many of the same women were listed in the Danish Cancer Register. Their foray into the two databases led to a surprising result: As they reported in *The New England Journal of Medicine* in 1997, there appears to be no connection between abortion and breast cancer.

Their success underscores the value of a trove of data the Danish government has accumulated on its citizenry, which today totals about 5 million people. Other Scandinavian countries have created powerful database systems, but Denmark has earned a preeminent reputation for possessing the most complete and interwoven collection of statistics touching on almost every aspect of life. The Danish government has compiled nearly 200 databases, some begun in the 1930s, on everything from medical records to socioeconomic data on jobs and salaries. What makes the databases a plum research tool is the fact that they can all be linked by a 10-digit personal identification number, called the CPR, that follows each Dane from cradle to grave. According to Melbye, "our registers allow for instant, large cohort studies that are impossible in most countries."

Beauty in numbers. These Danish twins starred in a variety show at the turn of the 20th century; now it's their medical records, part of a database, that are in demand.

But Melbye and other scientists think they can extract even more from this data gold mine. They argue that not enough money is being spent on maintaining and expanding existing databases, and they say that red tape is hampering studies that require correlation of health and demographic data. The problem is that, while they have unfettered access to more than 80 medical databases maintained by the Danish hospitals, thei databases overs Denmark is tigl mark won't allo its premises dat cedures. For soc unwieldy and e

Statistics D to release data concerns. "Th dence that infc individuals doe stitution," says

Last n ter Bi to bre datab can b told S entifi

W can y the U has s twins tive lifest of Sc Kaar tappe whic twins ing more than older. Christe genes about a man longevity by the unmat the Danish Tw

The health able for prob smaller studi

The Epidemiologist's Dream: Denmark

If the planners of a U.S. study of children's health could work in an ideal world, it might be Denmark. Epidemiologists there finished enrolling a cohort of 100,000 pregnant women into a mother-and-child research project last September and expect to finish collecting data from the children over the next year. The entire survey—which is large for this country of 70,000 annual births—is to be completed in 2005 for about $15 million, a tiny fraction of what the cost would be in the United States.

The Danes didn't design their Better Health for Mother and Child cohort study to answer specific questions or conduct long-term follow-up, as the Americans plan to do (see main text). Instead, they aim to create a databank that generations of researchers can mine and use as a starting point for studies of how medications, infections, nutrition, and even psychological factors affect pregnancy and child health.

Physicians have recruited volunteers among women making their first pregnancy visit. Participants give two blood samples during pregnancy and cord blood when the baby is born. The samples are saved for later use, including possibly for genetic studies. The mothers also answer a detailed questionnaire concerning nutrition; in an 18-month follow-up, they give information on their health and environmental exposures. The public health system is funding the study, with support from private and public foundations.

"Because the Danish population is probably the world's best registered, Denmark is the ideal place for such studies," says epidemiologist Mads Melbye, a steering group member from Statens Serum Institute in Copenhagen. Each citizen has a personal identification number that can be used to track data in centralized health care records, disease registries, and a population registry. Even centralized school records may be used. "It's an epidemiologist's dream," says Mark Klebanoff of the U.S. National Institute of Child Health and Human Development, who says tracking subjects is one of the costliest aspects of long-term U.S. studies.

Ready subjects. Denmark's 18-month-long birth cohort survey will collect data from mothers and newborns for a new database.

Norway, which has a system like Denmark's, is launching a mother-child study that will pool data with the Danish group's. Both benefit from streamlined management. It's difficult to get things done with too many decision-makers, says Melbye: "Running such a large study has taught us many things, but the chief lesson is that it is essential to put a very small group of people in charge."

Results are already beginning to trickle out of the Danish study. For example, one group published an article in *The Lancet* last November that disproved the existing consensus view that a fever early in pregnancy increases the risk for miscarriage. That's just the beginning: Denmark's scientific ethics committee has so far given the green light to more than 70 research protocols based on the mother-child study.

—LONE FRANK

Lone Frank is a science writer in Copenhagen.

Why register-based research

- Easy access to data – utilize existing data
- Large sample size – total population (rare diseases?)
- Population-based studies / real-world data / complete
- Great statistical power
- Follow-up easy
- No need to contact individuals
- No non-response bias (participation, reporting)
- Easy to do due to information technology
- Valuable time has passed – latency analyses
- Administrative data high quality
- Independent data

Selectionbias

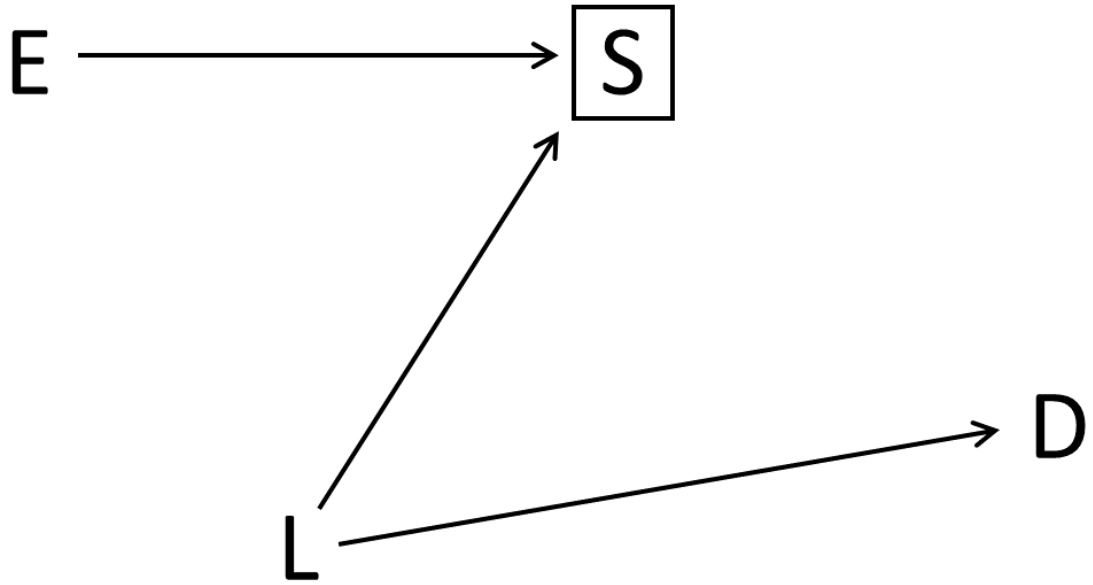
- No self selection bias
- No loss to follow-up / attrition bias

AND

- Nordic population relatively stable and homogeneous demography
- Universal health care system

Minor problem in register-based studies?

No/minimal selection bias



- Minimal non-response bias
- Minimal loss to follow-up (attrition bias)
- Under risk as long as you are residents of the country
- Censor persons when they emigrate from the country
- Assuming censoring is non-informative

Norwegian study:

- How did emigration influence mortality:
- Mortality was high among those who re-immigrated (the Salmon effect)

(Kristensen et al. Eur J Epidemiol 2010;25:155-61)

Exercise 1

- What are the main strengths of the study you planned on Monday?
- How could selection bias have influenced the results if you had not used registers?

Research economy

- All reasons could be formulated as research economy in the broadest sense
- If the registers were not available the costs would have been higher and in some circumstances the quality would have been lower

Exercise

- Please consider limitations of doing register-based research

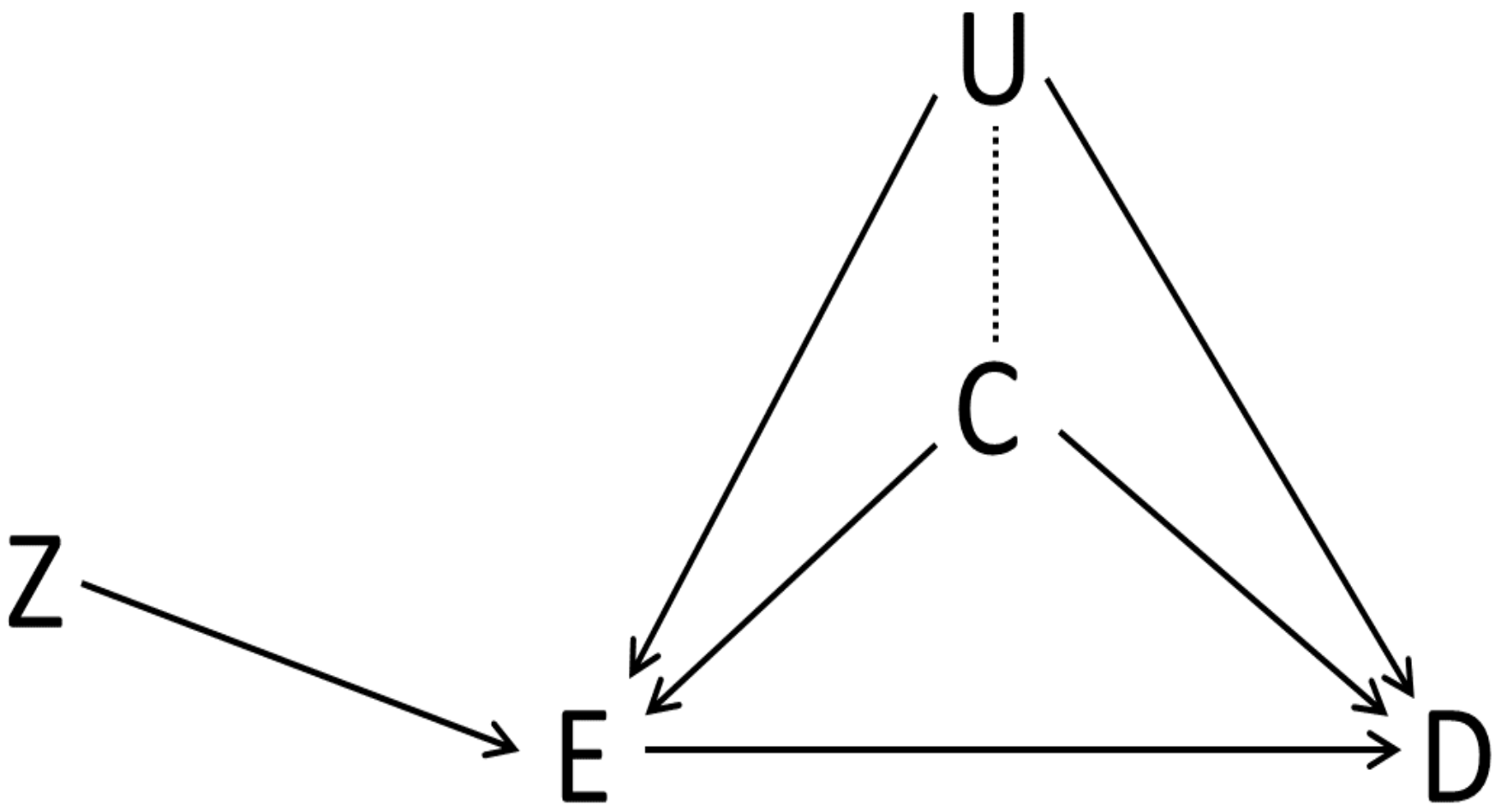
  – compared to cohort or case-control studies where data on exposure, confounders and outcome ae collected from a survey

Bias in register-based studies

- Same bias as in all observational studies
  - Vulnerable to systematic (and random) errors
- Data is predetermined
- Confounding / non-comparability
- Validity / misclassification
- Truncation bias
- Immortal time bias
- Data dredging
- Statistical tests – are they relevant?

Bias in register-based studies

- Same bias as in all observational studies
  - Vulnerable to systematic (and random) errors
- Data is predetermined
- **Confounding / non-comparability**
- **Validity / misclassification**
- **Truncation bias**
- **Immortal time bias**
- Data dredging
- Statistical tests – are they relevant?


U
C
Z
E
D

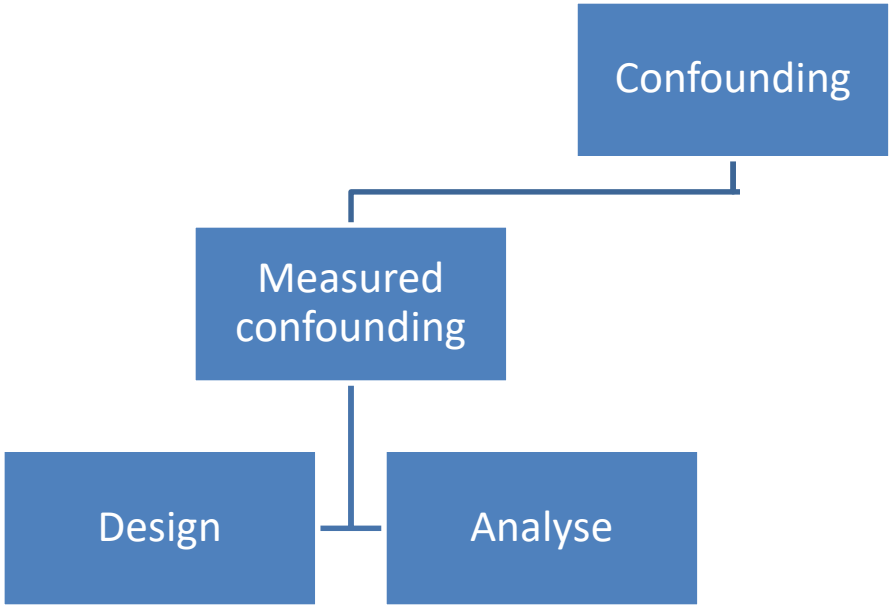
Confounding and unmeasured confounding

- Exposed group not comparable to the unexposed group with regard to some specific factors, e.g.
  - Physicians prescribe drugs based on diagnostic and prognostic information
  - Factors influencing this decision vary by physician and patient
  - Clinical, functional, behavioral characteristics of patients
  - Physician's prescribing preferences
  - Often associated with the outcome
  - Could result in findings that medications appear to cause outcomes they are meant to prevent
- The aim of handling confounding is to obtain comparable groups
- Ideally we wish to construct exposed and unexposed groups similar on all factors except exposure

Strength of RCT

- Groups identical *(at least in large studies)*
- Bias
  - Perfect randomization?
  - Non-compliance and loss to follow-up (ITT)
- Possible in observational studies?
  - Assumption of no unmeasured confounders!!

Methods to adjust for confounding



Restriction

Matching

Standardization

Stratification

Multivariate regression

Schneeweiss et al. Pharmacoepidemiology and drug safety 2006;15:291–303

’Adjustment’

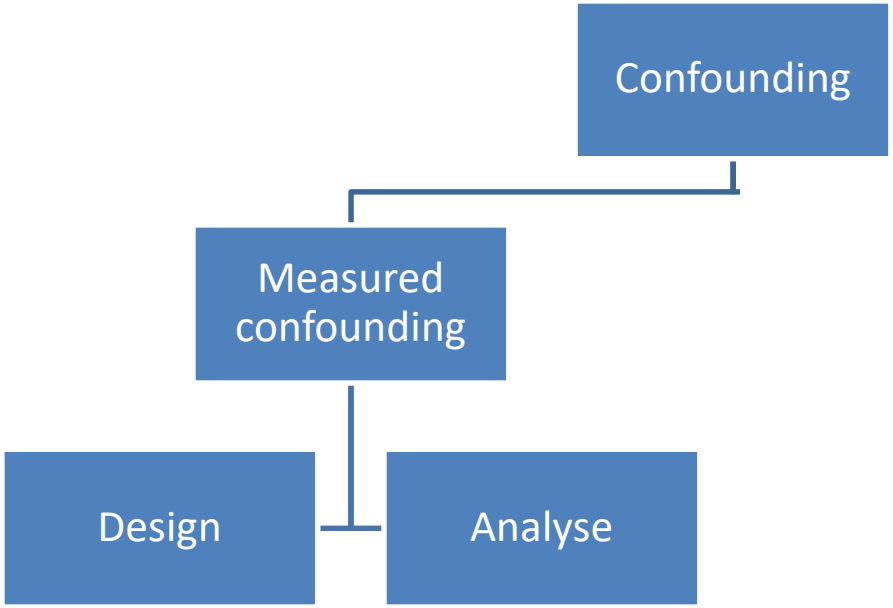
- Confounders that require detailed information on
  - clinical parameters
  - lifestyle
  - over-the-counter medications
- are often not measured in registers

- Causing confounding and residual confounding bias

Register-based studies

- Often few and unspecific confounders
- Combined with great statistical strength finding small effects
- Large risk of confounding bias

Methods to adjust for confounding



Restriction

Matching

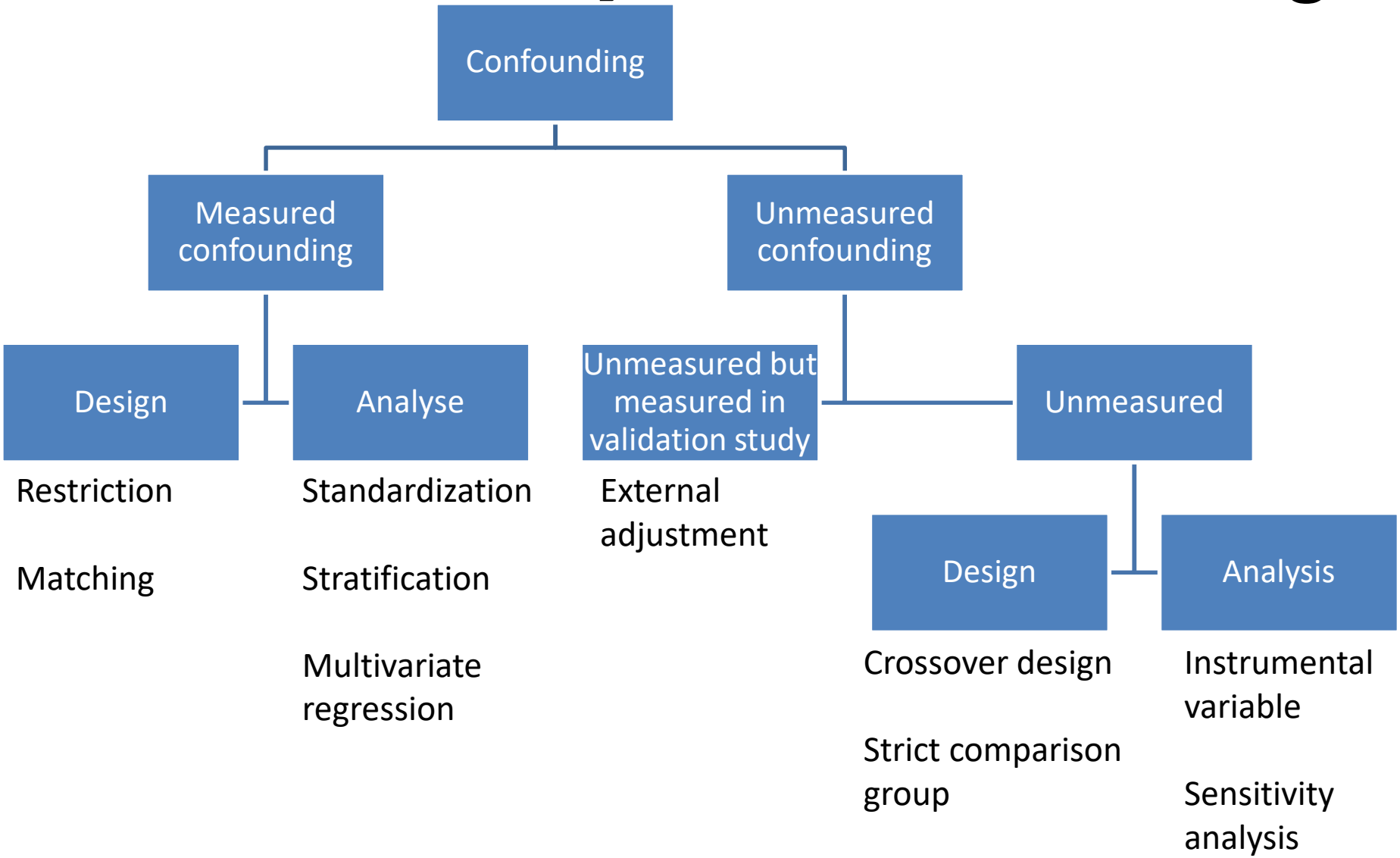
Standardization

Stratification

Multivariate regression

Schneeweiss et al. Pharmacoepidemiology and drug safety 2006;15:291–303

Methods to adjust for confounding



Schneeweiss et al. Pharmacoepidemiology and drug safety 2006;15:291–303

Exercise 2

- Which confounders are most important in the study you planned on Monday?
- Do you have unmeasured confounding?
- Please consider the methods presented in last slide – any of them relevant for your study?

Data collection is predetermined

- Not controlled by the researcher
- Research topic needs to suit the database
- Hard to know exactly how data were generated
- Very difficult to validate

Data collection is predetermined

- Limit the usefulness of coded diagnoses
  - Variation in coding
    - Between persons?
    - Between departments?
    - Institutions?
    - Over time: New coding
- Errors in coding
- Limitation in specificity in the available codes
- Bound to used definitions and administrative practices
  - ‘Administrators view of the world!’
  - Registers contain information on the citizens in relation to public administrators
  - Researchers distant from the actual data collection

Validity

- Misclassification
  - Risk of substantial errors due to many people entering data
  - Variation in coding
- Changes in coding and classifications over time
  - Disease diagnoses (ICD-8 until 1993, ICD-10 1994 onwards)
  - Industrial classification
- DRG taxation (changes in fees for diagnoses and treatments)
- **Validation studies important**

Data quality

Two fundamental concerns:

1. Completeness of registration of individuals

2. Validity of the information
   - Accuracy and degree of completeness of the registered data

Goldberg et al. Epidemiologic Reviews 1980;2:210-20.

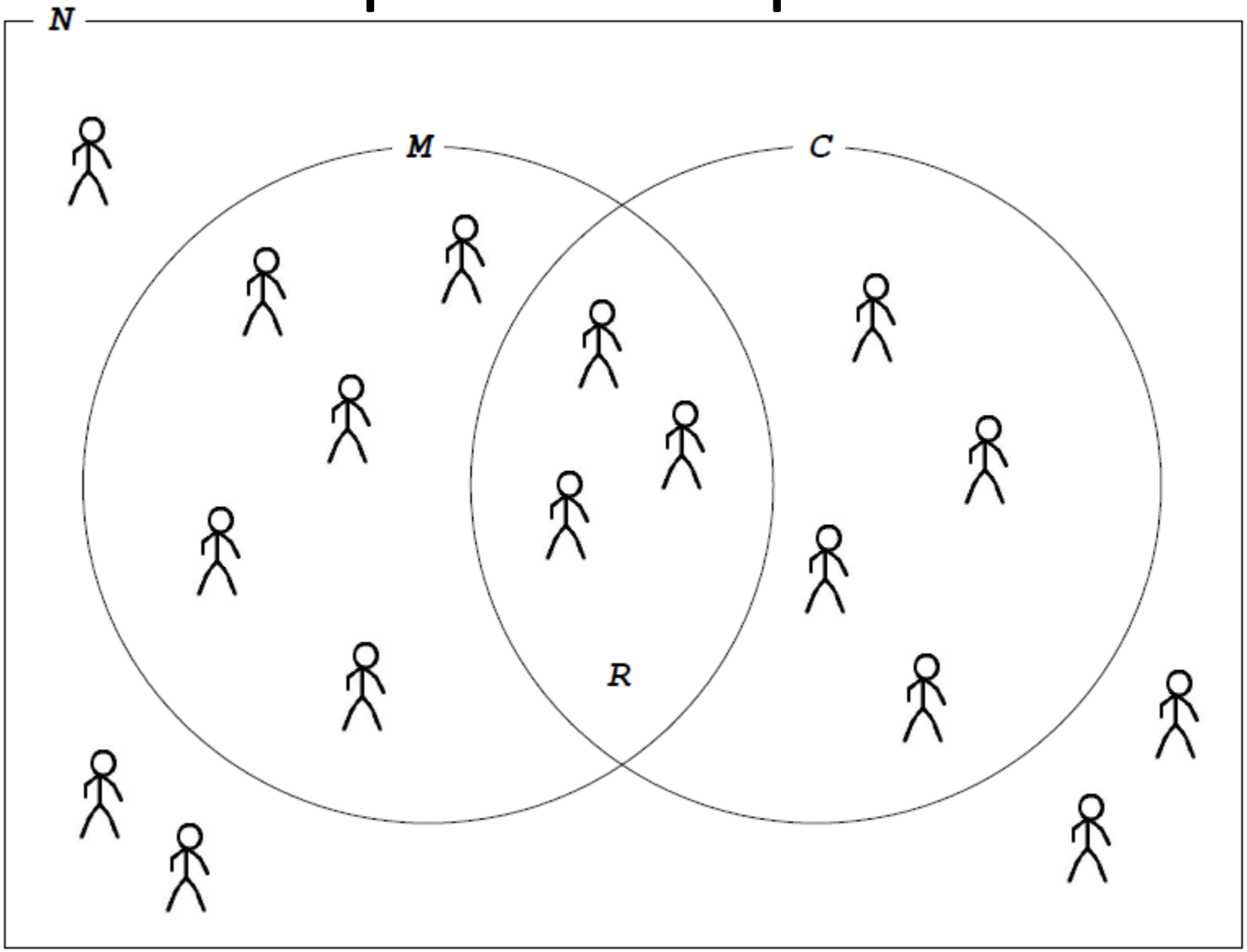
Data quality - completeness

- Completeness: ”The proportion of individuals in the target population which is correctly classified in the register”
- Important to know whether the data source is population-based
  - Or whether it has been through one or more selection procedures (e.g. Medicare)
- Also important to know whether the target population is stable

Methods to evaluate completeness

- Compare sources
- Comprehensive records review
- Aggregated methods
- Capture – recapture

Capture-recapture



Validity

- Often the question: How high is the validity of register data
- Validity is the extent to which a variable measures what it is intended to measure
- Important measures
  - Sensitivity / specificity
  - Positive and negative predictive value

Validity

- Data validity can be categorized
  - Errors in the register may reflect incorrect data entry or lack of available information
  - The original source of information, correctly entered into the register, may itself be inaccurate
- Record review is often used for the validation
  - The ratio between the number of correctly registrered persons and all registered persons is measured

What you need to know

- The total number of
  - True sick and healthy
  - Positive and negative test results

- Often impossible

Clinical Epidemiology

Dovepress

open access to scientific and medical research

Open Access Full Text Article

REVIEW

The Danish National Patient Registry: a review of content, data quality, and research potential

This article was published in the following Dove Press journal:
Clinical Epidemiology
17 November 2015
Number of times this article has been viewed

Morten Schmidt[1]
Sigrun Alba Johannesdottir Schmidt[1]
Jakob Lynge Sandegaard[2]

Background: The Danish National Patient Registry (DNPR) is one of the world's oldest nationwide hospital registries and is used extensively for research. Many studies have validated algorithms for identifying health events in the DNPR, but the reports are fragmented and no overview exists.

Schmidt 2015

- 114 papers, validating 1–40 codes/algorithms each and 253 in total
- PPVs ranged from below 15% to 100%.
- May result from different reference standards used

- Majority: Cross-sectional studies with medical record review as reference standard
- Other reference standards used:
  - Patient interviews
  - Danish Cancer Registry
  - Research database
  - Clinical registries
  - A military conscription system database
  - Danish prescription registries
  - Radiology reports
  - Clinical Laboratory Information
  - Danish National Pathology
  - Hospital pharmacy systems
  - GP verification
  - Autopsy reports

Setting and calendar year

- PPV depends on the prevalence of disease
- Higher PPV in specialized departments
- Calendar year seems to increase quality, given the continuous improvement in diagnostic criteria and procedures used

Schmidt 2015

- The large variation underscores the need to validate diagnoses and treatments before using DNPR data for research
- Validation studies may need updates, as newer diagnostic criteria and procedures may differ from those used in older validation studies

Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

EDITORIAL

Helping everyone do better: a call for validation studies of routinely recorded health data

This article was published in the following Dove Press journal:
Clinical Epidemiology
12 April 2016
Number of times this article has been viewed

Vera Ehrenstein[1]
Irene Petersen[1,2]
Liam Smeeth[3]
Susan S Jick[4]
Eric I Benchimol[5,6]

There has been a surge of availability and use for research of routinely collected electronic health data, such as electronic health records, health administrative data, and disease registries. Symptomatic of this surge, in 2012, *Pharmacoepidemiology and Drug Safety* (PDS) published a supplemental issue containing several reviews of validated methods for identifying health outcomes using routine health data,[1]

What to do next?

American Journal of Epidemiology
Vol. 138, No. 11
Printed in U.S.A.

Use of the Positive Predictive Value to Correct for Disease Misclassification in Epidemiologic Studies

Hermann Brenner[1] and Olaf Gefeller[2]

Misclassification problems of the disease status often arise in large epidemiologic cohort studies in which the outcome is classified on the basis of record linkage with routinely collected error-prone data sources, such as cancer registries or mortality statistics. If the misclassification is nondifferential, i.e., independent of the exposure status, this leads to bias toward the null in estimates of relative risk. A variety of methods have

What to do next?

Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2005;**34**:1370–1376

doi:10.1093/ije/dyi184

A method to automate probabilistic sensitivity analyses of misclassified binary variables

Matthew P Fox,[1,2]* Timothy L Lash[2,3] and Sander Greenland[4]

Accepted 9 August 2005

Background Misclassification bias is present in most studies, yet uncertainty about its magnitude or direction is rarely quantified.

Methods The authors present a method for probabilistic sensitivity analysis to quantify likely effects of misclassification of a dichotomous outcome, exposure or covariate. This method involves reconstructing the data that would have been observed had the misclassified variable been correctly classified, given the sensitivity and specificity of classification. The accompanying SAS macro implements the method and allows users to specify ranges of sensitivity and specificity of misclassification parameters to yield simulation intervals that incorporate both systematic and random error.

Exercise 3

- Do you have validation problems with your exposure or outcome?
- How do you think it will influence your results?
  - Non-differentially or differentially?
- Do you have validation problems with your (most important) confounders?
- Do you think this could influence your results?

Documentation / metadata

- Statistical metadata is descriptive information or documentation about statistical data

- Statistical metadata facilitates the sharing, querying, and understanding of statistical data over the lifetime of the data

- Increasing demand
  - The need for metadata in the statistical production has been increasingly evident
  - Most statistical offices are striving to introduce metadata systems, or improve existing ones

Why register-based research

- Easy access to data – utilize existing data
- Large sample size – total population (rare diseases?)
- Population-based studies / real-world data / complete
- Great statistical power
- Follow-up easy
- No need to contact individuals
- No non-response bias (participation, reporting)
- Easy to do due to information technology
- Valuable time has passed – latency analyses
- Administrative data high quality
- Independent data

Bias in register-based studies

- Same bias as in all observational studies
  - Vulnerable to systematic (and random) errors
- Data is predetermined
- Confounding / non-comparability
- Validity / misclassification
- Truncation bias
- Immortal time bias
- Data dredging
- Statistical tests – are they relevant?