FMS/DSBS autumn meeting 2014

Challenges in design and analysis of large register-based epidemiological studies

Caroline Weibull & Anna Johansson

Department of Medical Epidemiology and Biostatistics (MEB)
Karolinska Institutet
Stockholm, Sweden

caroline.weibull@ki.se
anna.johansson@ki.se

Outline

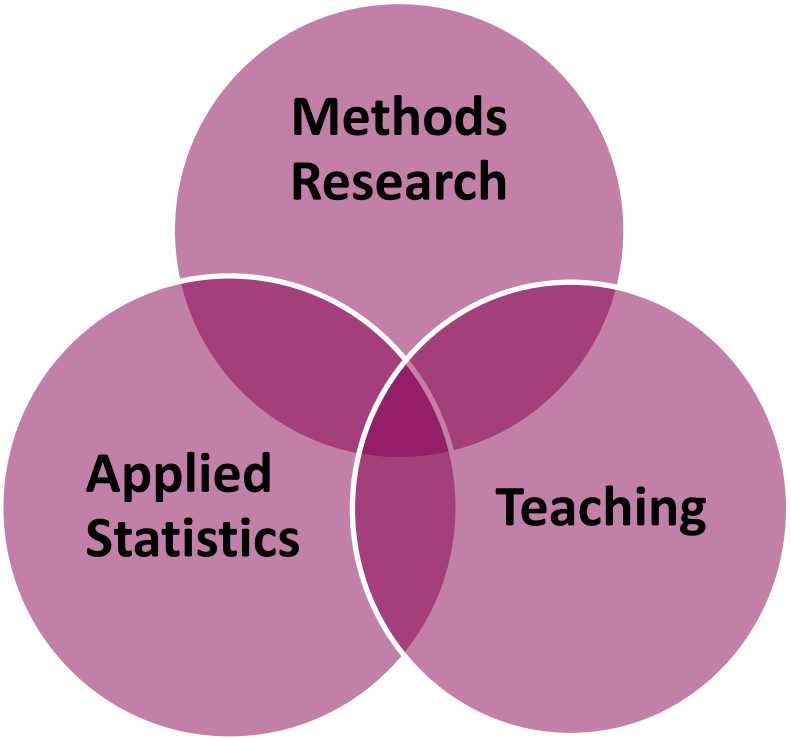
- **Who are we**
  - Karolinska Institutet and our department
- **Register-based research**
  - Data sources, data linkages, some unique registers
  - Some statistical problems with register data
  - Design of register studies: Classical designs and other sampling strategies
- **Example**
  - Parkinson disease and cancer: A family design
- **Final remarks**

Who are we

- **Karolinska Institutet (KI)**
  - A medical university
  - Research and education

- **Department of Medical Epidemiology and Biostatistics (MEB)**
  - Cancer epidemiology (e.g. breast cancer, prostate cancer)
  - Psychiatric disorders (e.g. ADHD, schizophrenia)
  - Neurological and aging related diseases (e.g. dementia, Alzheimer)
  - Pediatric and reproductive epidemiology (e.g. asthma)
  - Genetic and molecular epidemiology
  - Swedish Twin Register
  - KI Biobank

Biostatistics group at MEB



- **Largest biostatistics group among universities in Sweden (n ≈ 40)**
  - Faculty including four professors
  - No “water tight boundaries”

- **Methods Research:**
  - Statistical methods for register-based research and epidemiology
  - Study design and sampling (e.g. developments of cohort and case-control designs)
  - Twin and family modelling
  - Causal inference
  - Predictive modelling
  - Cancer patient survival analysis
  - High-throughput data analyses and statistical genetics

Register-based research, data sources and linkages

- **Register-based epidemiology**
  - Uses population-based registers as the primary data source

- **Population-based register**
  - Encompassing the total population in a geographic region (e.g. Sweden)
  - Data collected via routine systems, e.g. health services, tax office
  - Reporting mandatory by law
  - Register holders are typically authorities, e.g. Statistics Sweden, National Board of Health and Welfare (Socialstyrelsen)
  - Registers hold millions of individuals

Examples of registers used in health research

Register	Including	Start
Multi-Generation Register	Links all Swedish residents to their mother and father, including birthdates	1961 (born 1932)
Swedish Cancer Register	All newly diagnosed cancer cases	1958
Cause of Death Register	All deaths in Sweden	1961 (1952)
Medical Birth Register	All births in Sweden	1973
Patient Register	All in-patient care in Sweden All out-patient care in Sweden	1987 (1964) 2005
Prescribed Drug Register	All dispensed drugs in Sweden	1999

Register-based research, data sources and linkages

- **Special registers**
  - Quality registers (www.kvalitetsregister.se): e.g. Swedish Hip Fracture Register, Swedeheart, National Prostate Cancer Register (opt out)
  - Special cohorts: e.g. Twin Register, clinical cohorts (informed consent)
  - Population-based?

- **Why are register-based studies useful**
  - When RCTs are ethically or logistically unfeasible
  - When an outcome is rare and cases need to be accrued over time (=historical data collection)
  - When it is possible to link several registers together (=enriching information from multiple sources)
  - We can enumerate the whole Swedish population

Register-based research, data sources and linkages

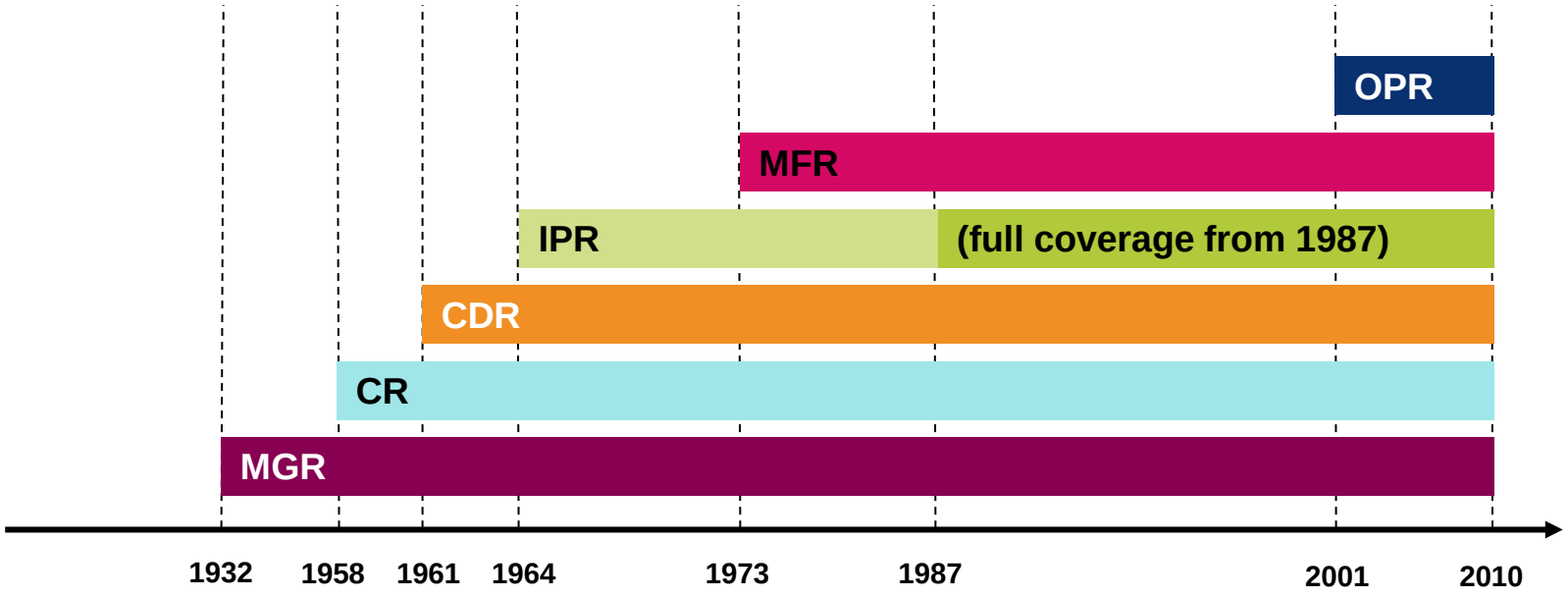
- **Linkages between registers:**
  - Nordic countries is a paradise for an epidemiologist!
  - Possible to use the PIN (=personal identification number) assigned to all citizens to link between registers
  - Huge possibilities to design register-based studies by combining information from multiple sources
  - Not possible in other parts of the world – others have difficulties to link data!
  - Nordic countries = small populations, but still competitive!

- **Financial effort from government to boost register-based research in Sweden**
  - Funding 2012-2016, including Register Service support and infrastructure
  - SIMSAM/Vetenskapsrådet and other directed efforts towards universities

Some statistical problems with register data

- **Data is not collected for research purposes**
  - "What they collect is what you get"
- **Coding of variables has changed over time**
  - Demands knowledge of registers and history
- **Observational data**
  - Confounding
  - Subject knowledge necessary
- **Truncation**
  - Coverage → selection bias
- **Clustering and correlated data**
  - Family data: nuisance and/or advantage

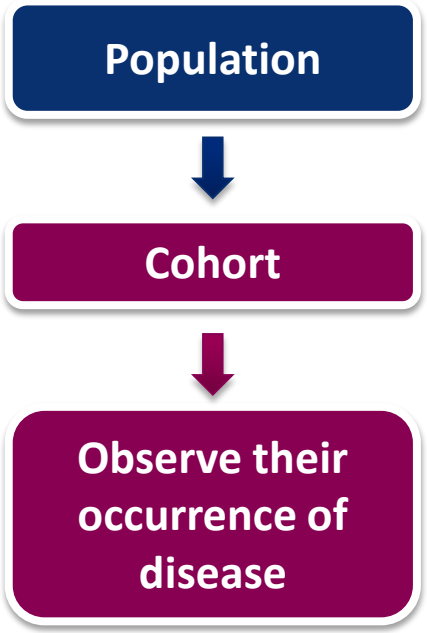
Truncation due to start year of registers



Design of register studies: Classical designs

Cohort design

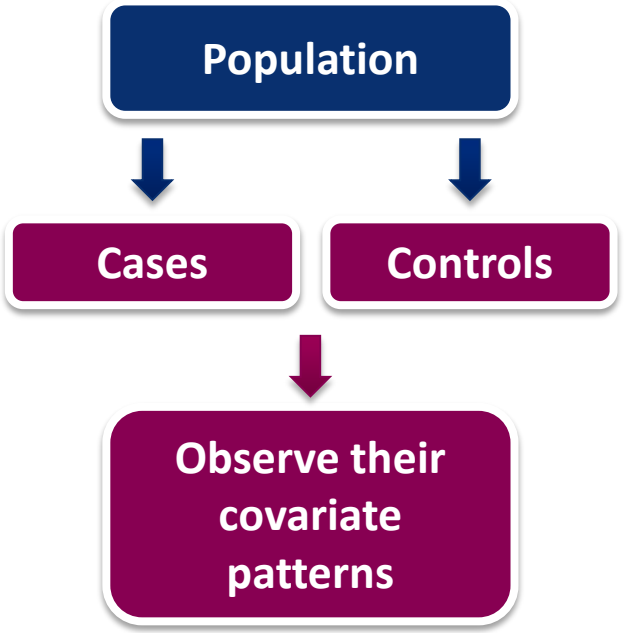
Rare exposure, common outcome



Estimate risks, relative risks (RR)

Case-control design

Common exposure, rare outcome

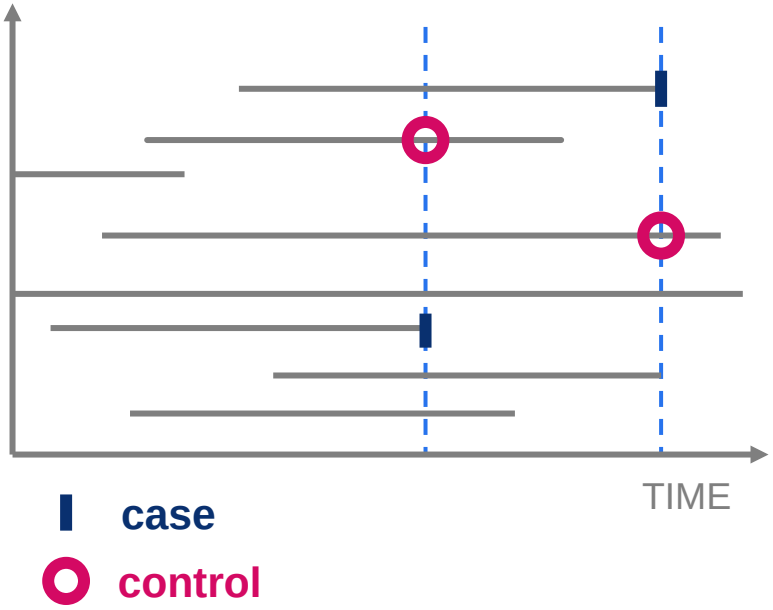


Estimate odds ratios (OR), as measure of relative risks

Design of register studies:
Other sampling designs (variants of the classical designs)

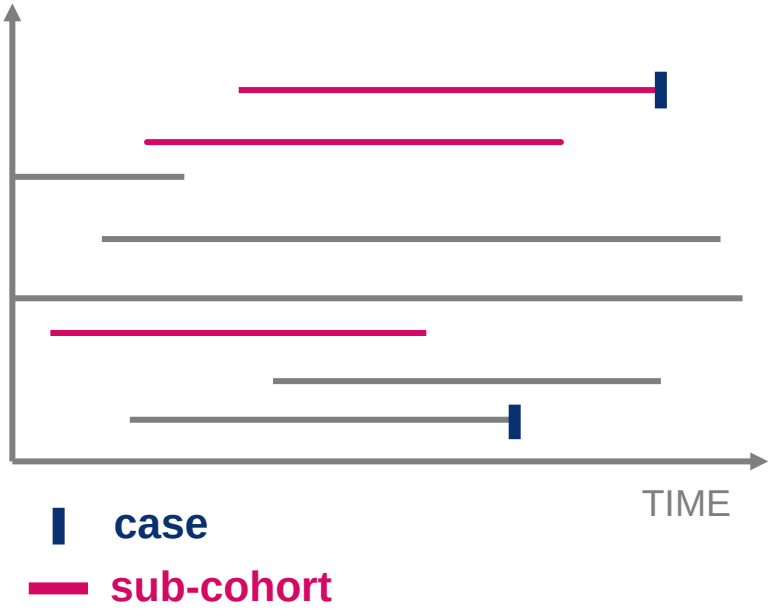
Nested case-control

Can estimate same things as in a cohort



Case-cohort

Can estimate same things as in a cohort

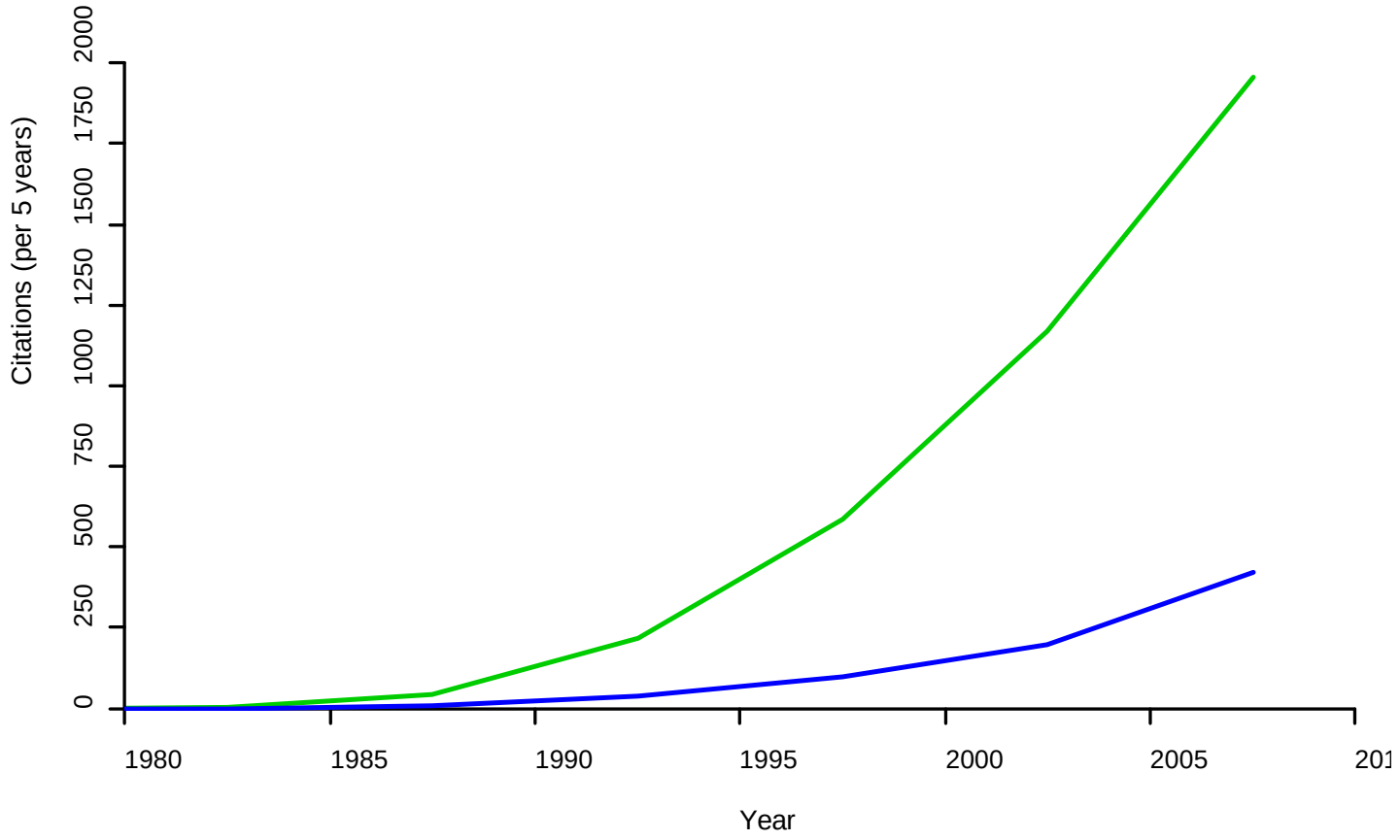


+ Other matched designs - Improve the statistical efficiency, i.e. same power with fewer subjects (e.g. matched cohort study, matched case-control study)

Popularity of these designs has increased

- **References to nested case-control and case-cohort in Web of Science**

(Ørnulf Borgan acknowledged)



Example

American Journal of Epidemiology

Vol. 179, No. 1
DOI: 10.1093/aje/kwt232
Advance Access publication:
October 18, 2013

Original Contribution

Parkinson's Disease and Cancer: A Register-based Family Study

Karin Wirdefeldt*, Caroline E. Weibull, Honglei Chen, Freya Kamel, Cecilia Lundholm, Fang Fang, and Weimin Ye

* Correspondence to Dr. Karin Wirdefeldt, Department of Medical Epidemiology and Biostatistics, Box 281, Karolinska Institutet, SE-17177 Stockholm, Sweden (e-mail: karin.wirdefeldt@ki.se).

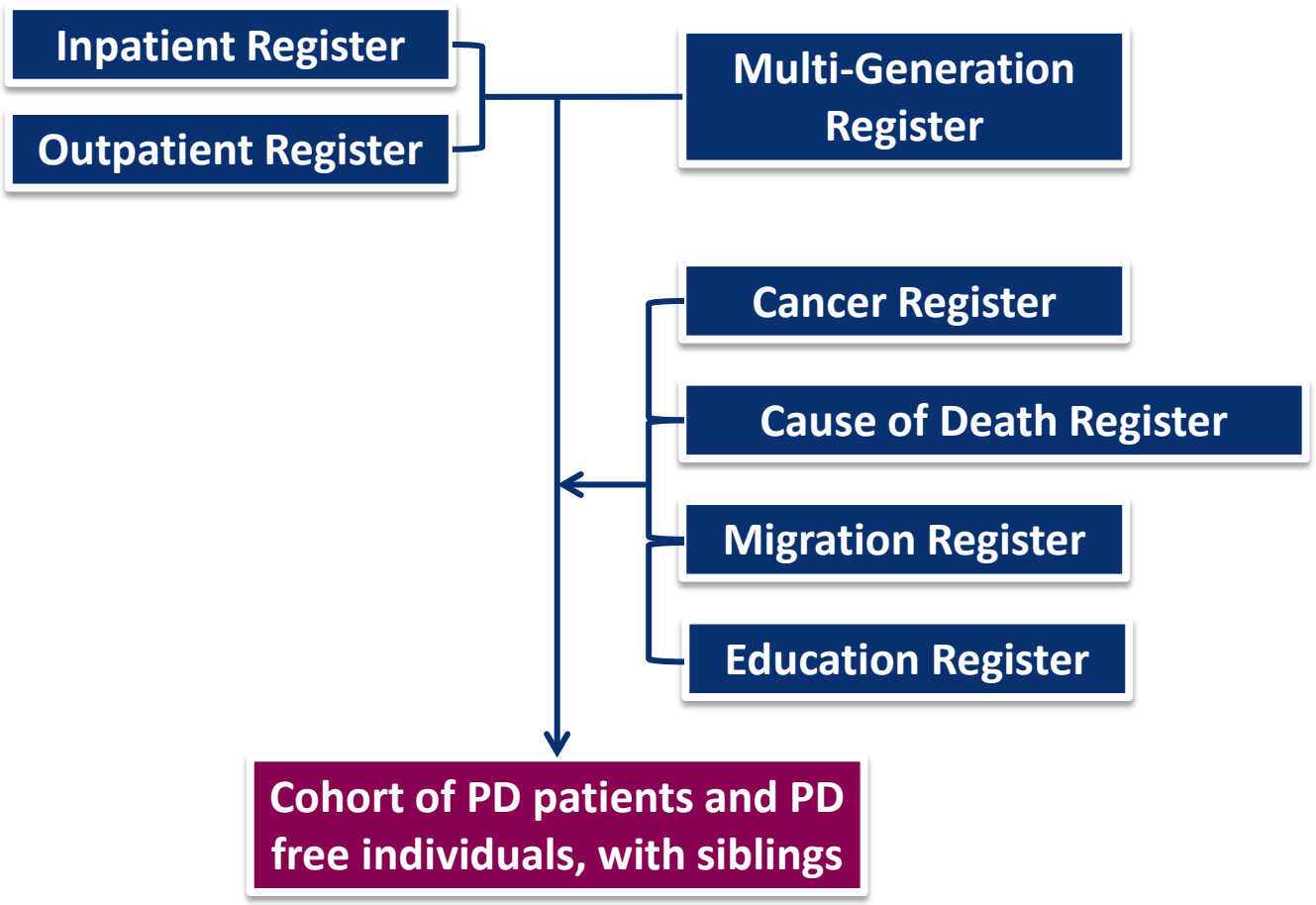
Background and aims

- Observed comorbidity between PD and cancer:
  - Melanoma ↑
  - Smoking-related ↓

- Aims:
  1. Study association between PD and adulthood cancer(s) in the Swedish population.

  2. Assess whether possible associations might be due to familial factors.

Registers used



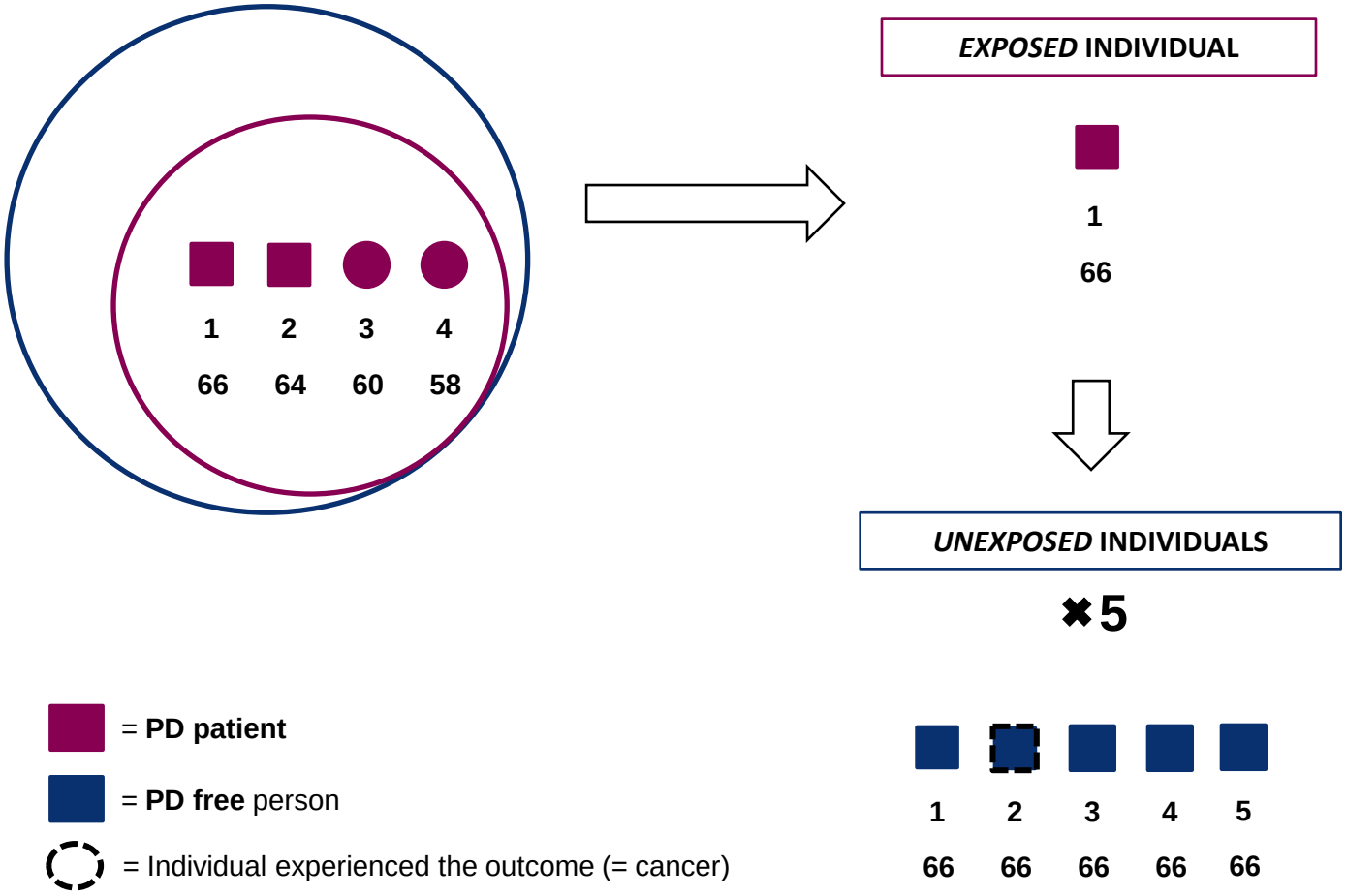
PD → CANCER in individual

- **Exposure:** PD diagnosis in registers (Yes/No)
- **Outcome:** Cancer diagnosis (Yes/No)

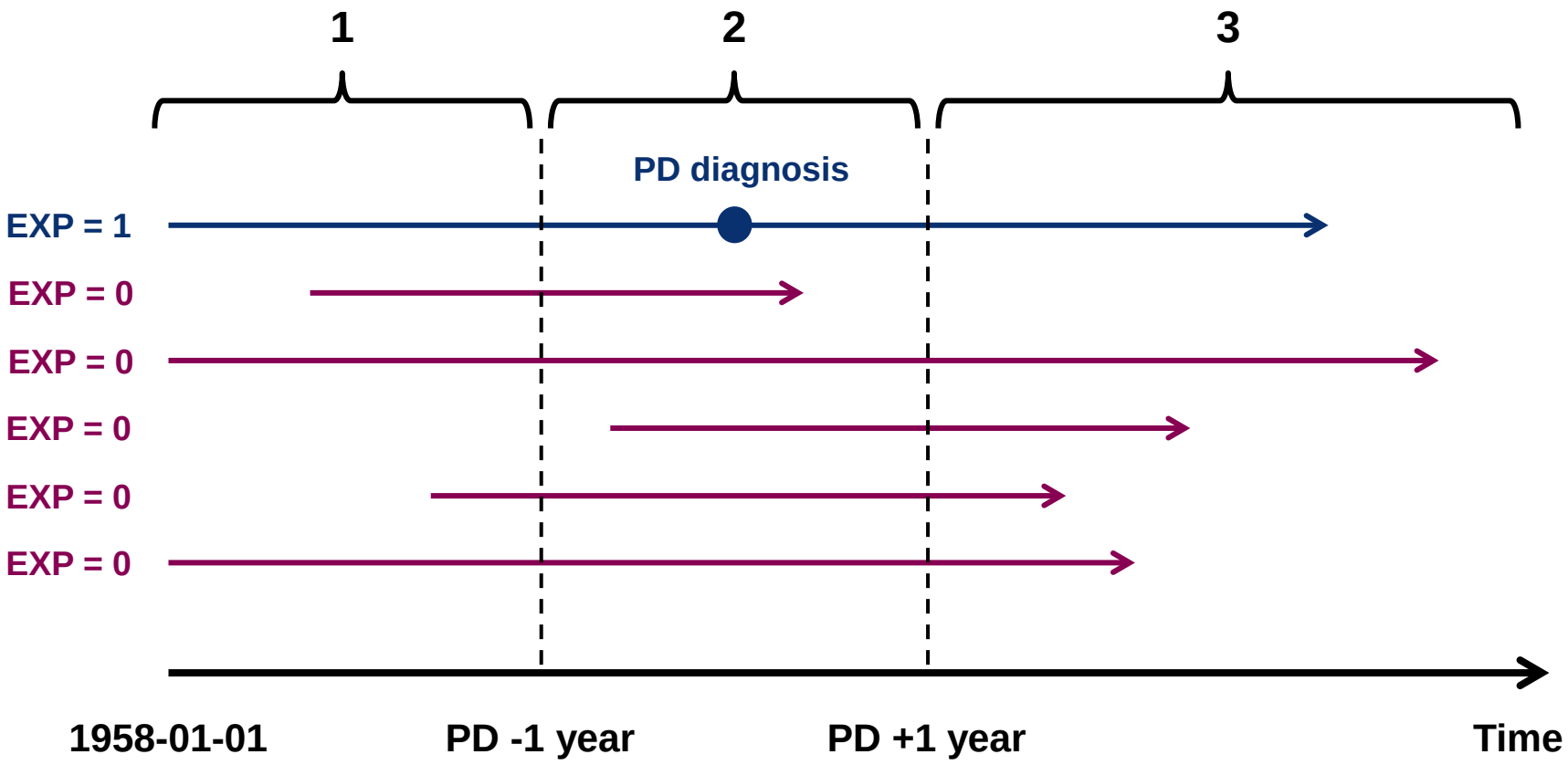
- **Matched cohort design (1:5)**
- Matching variables:
  - Birth year, sex, being alive and in Sweden when PD patient gets diagnosis

- Survival analysis using stratified Cox regression
  - In each strata: 1 PD patient + 5 PD free persons
  - Time scale: Attained age
  - Adjust for highest achieved education level

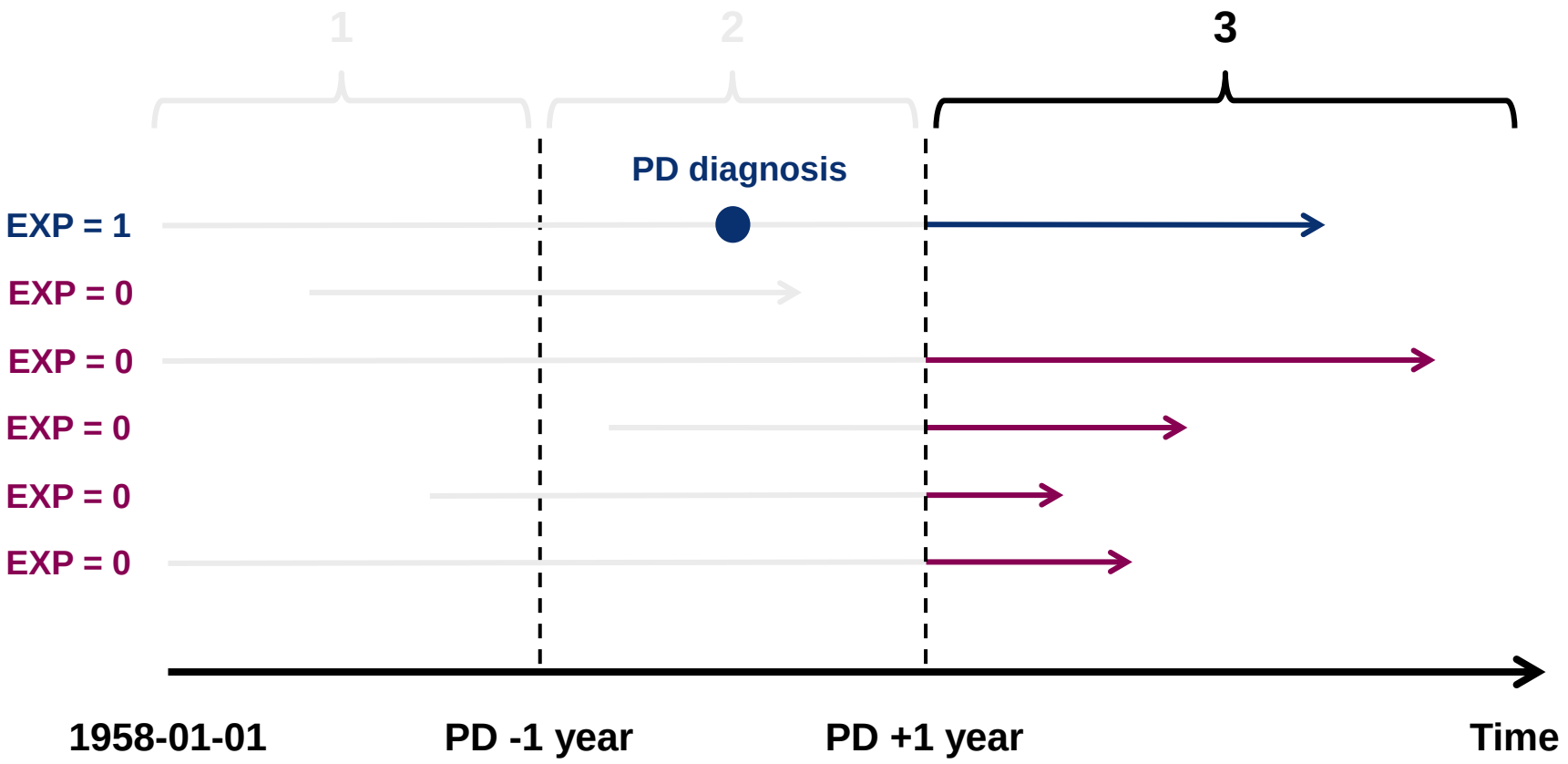
PD → CANCER in individual



PD → CANCER in individual



PD → CANCER in individual



PD → CANCER in individual

Main results (11,786 PD patients):

Cancer site	HR (95% CI)
• All sites combined	0.87 (0.79 – 0.96)
• Smoking related sites	0.70 (0.56 – 0.87)
• Lung cancer	0.40 (0.24 – 0.66)
• Melanoma	1.46 (1.01 – 2.10)

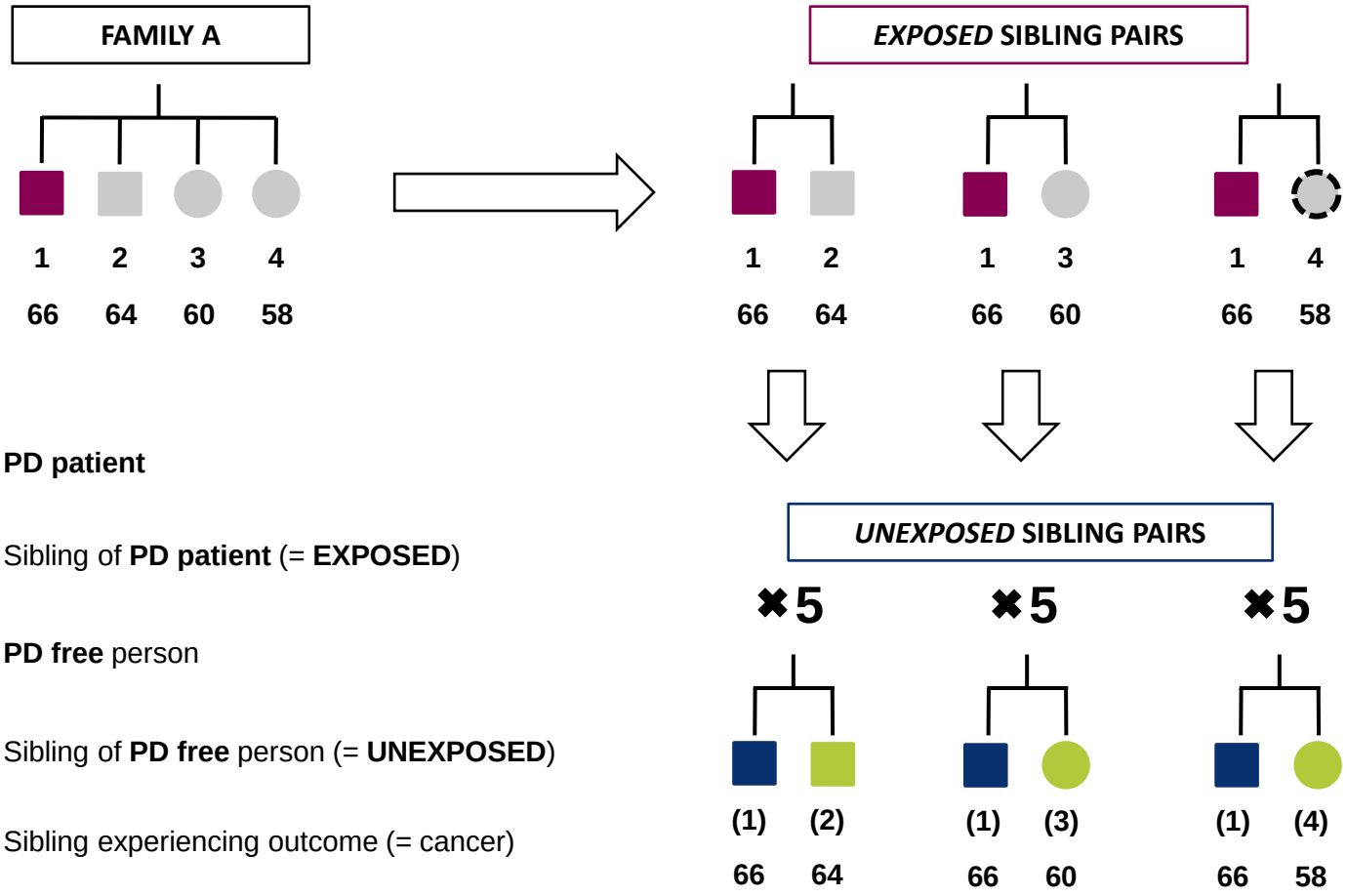
PD → CANCER in sibling

- **Exposure:** PD diagnosis in registers (Yes/No)
- **Outcome:** Cancer diagnosis ***in sibling*** (Yes/No)

- **Matched design (1:5)**
- Matching variables (PD patients/free):
  - Birth year, sex, being alive and in Sweden when PD patient gets diagnosis
- Matching criteria (their siblings):
  - Birth year, sex, sib ship

- Survival analysis using stratified Cox regression
  - In each strata: 1 exposed sibling + 5 unexposed siblings
  - Time scale: Attained age of sibling
  - Adjust for highest achieved education level

PD → CANCER in sibling



PD → CANCER in sibling

Results (16,841 siblings to PD patients):

Cancer site	HR (95% CI)	HR (95% CI)
• All sites combined	0.99 (0.95 – 1.03)	0.87 (0.79 – 0.96)
• Smoking related sites	0.93 (0.86 – 1.00)	0.70 (0.56 – 0.87)
• Lung cancer	0.90 (0.73 – 1.10)	0.40 (0.24 – 0.66)
• Melanoma	0.88 (0.73 – 1.08)	1.46 (1.01 – 2.10)

Limitations

- **Quality in register data**
  - Completeness and coverage
  - Date of onset vs. date of PD diagnosis
- **Definition of being exposed in sibling analysis**
  - Exposed if any sibling has PD?
  - Analyze whole families instead of pairs?
  - Only include one random sib pair per family?
- **Unmeasured confounding**
  - E.g. smoking status
- **Matching strata small and many, which can increase SE**

Final remarks

- **Computing time is an issue!**
  - Large databases require smart designs – the luxury of having too much data!
- **Importance of sensitivity analysis**
  - Observational data: evaluate bias assuming best/worse scenarios
- **Causality**
  - Not so much

Thank you!