NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Karolinska Institutet

Data quality

Mika Gissler
Research Professor, THL
Visiting Professor, Karolinska Institutet
Adjunct Professor, University of Oulu
mika.gissler@ki.se or mika.gissler@thl.fi

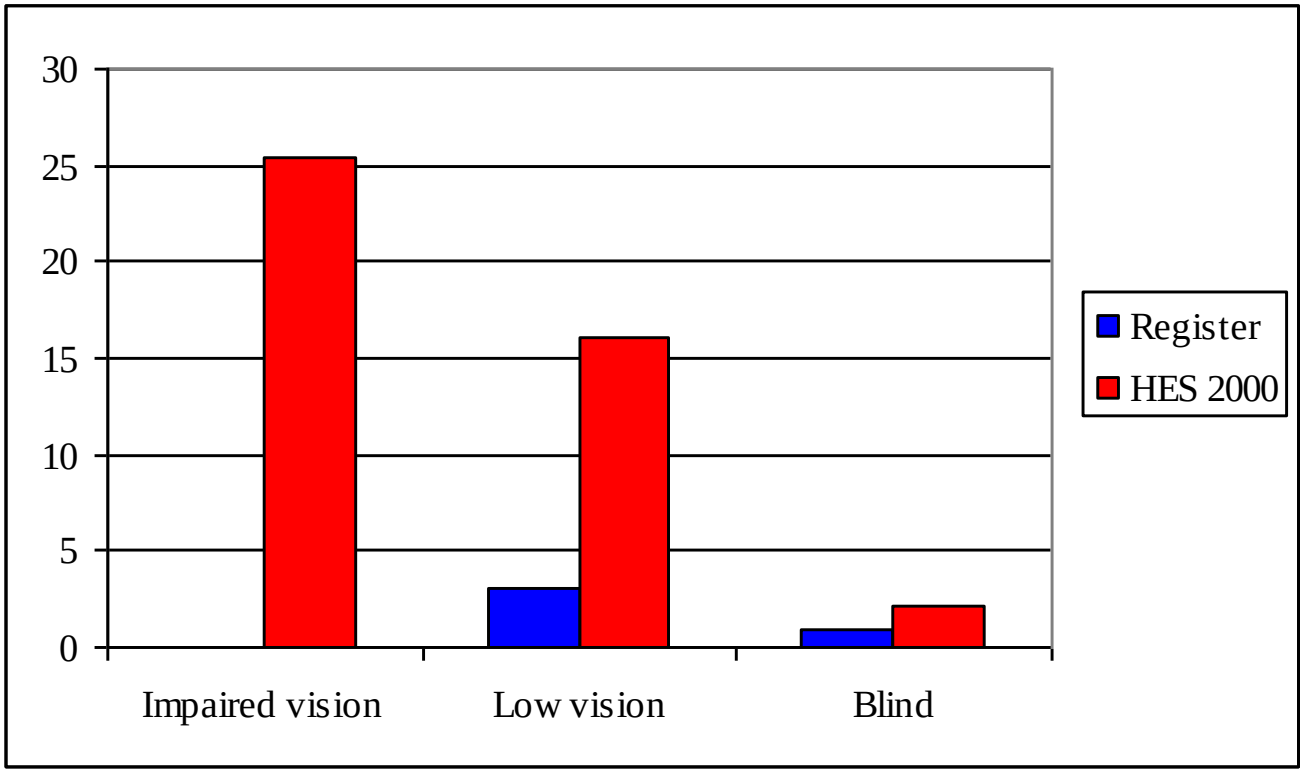
Data quality

- Completeness of records
  - Are all cases included in the register?
- Validity
  - Does the register data reflect reality?
- Study types
  - Internal validation
  - Comparisons to different studies and previous literature
  - Data linkages
  - Comparisons between the original data source and register data

Examples of studies

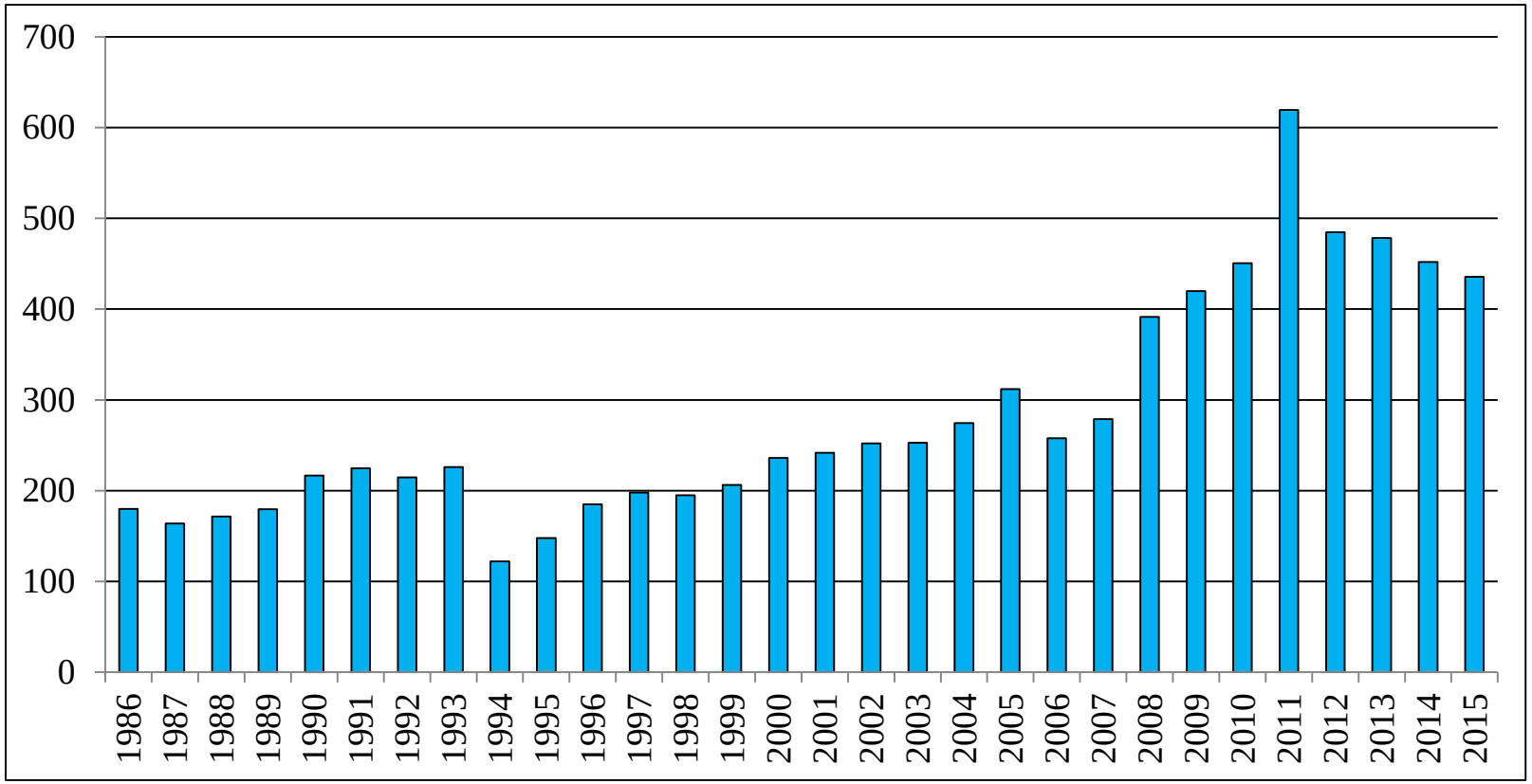
- Internal validation
  - How many diagnosis are correct by age and sex?
- Comparisons to different studies
  - Health examination study and register information
- Data linkages
  - Medical Birth Register and Central Population Register
- Comparisons between the original data source and register data
  - The Finnish Register on Induced Abortions included 483 cases of the 488 cases found in the hospitals = 99%

Prevalence of visual impairments per 1000 population, Finland 2000



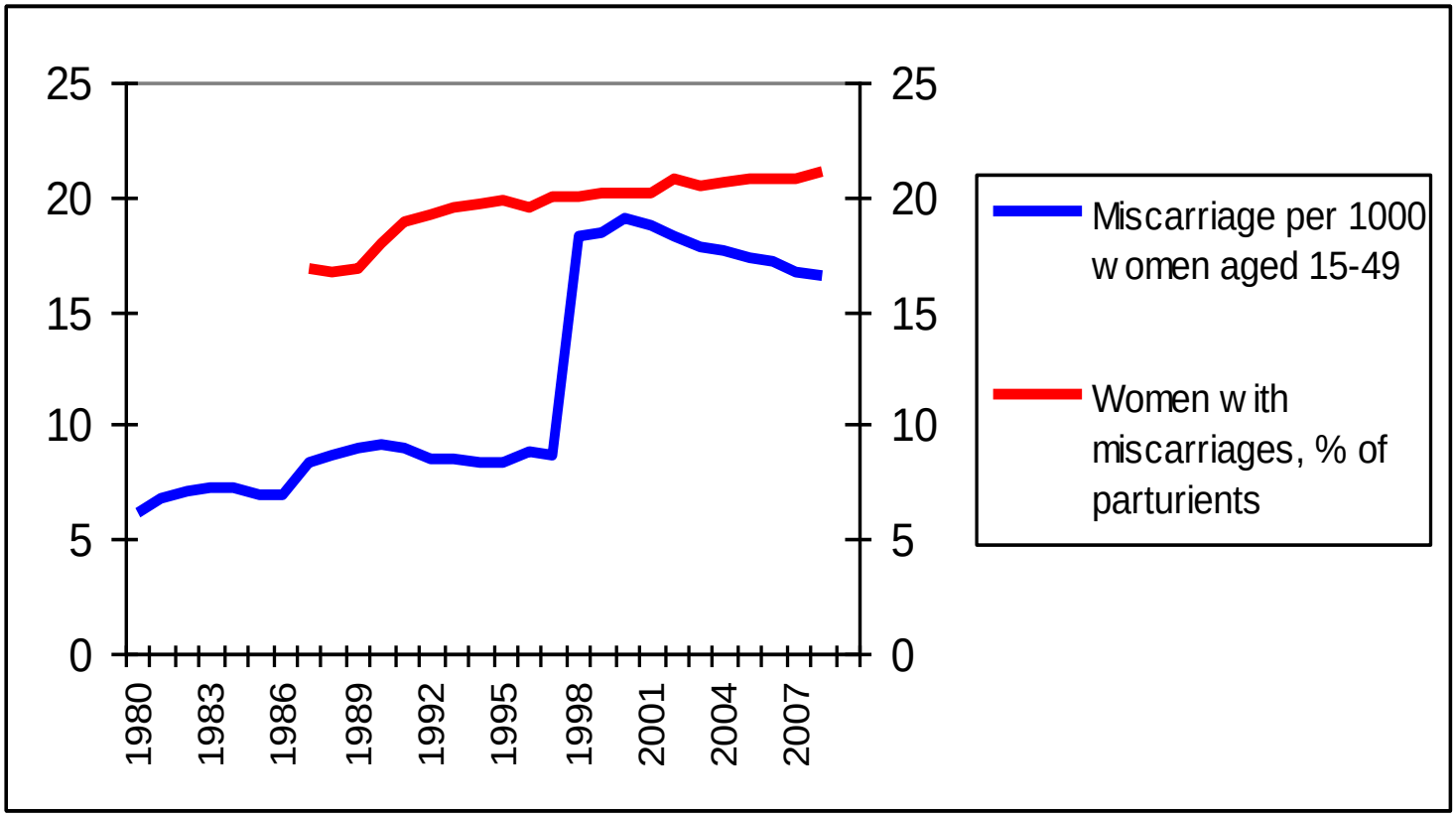
Not covered	19%	43%

Diabetes per 100 000 population Finland 1980-2015



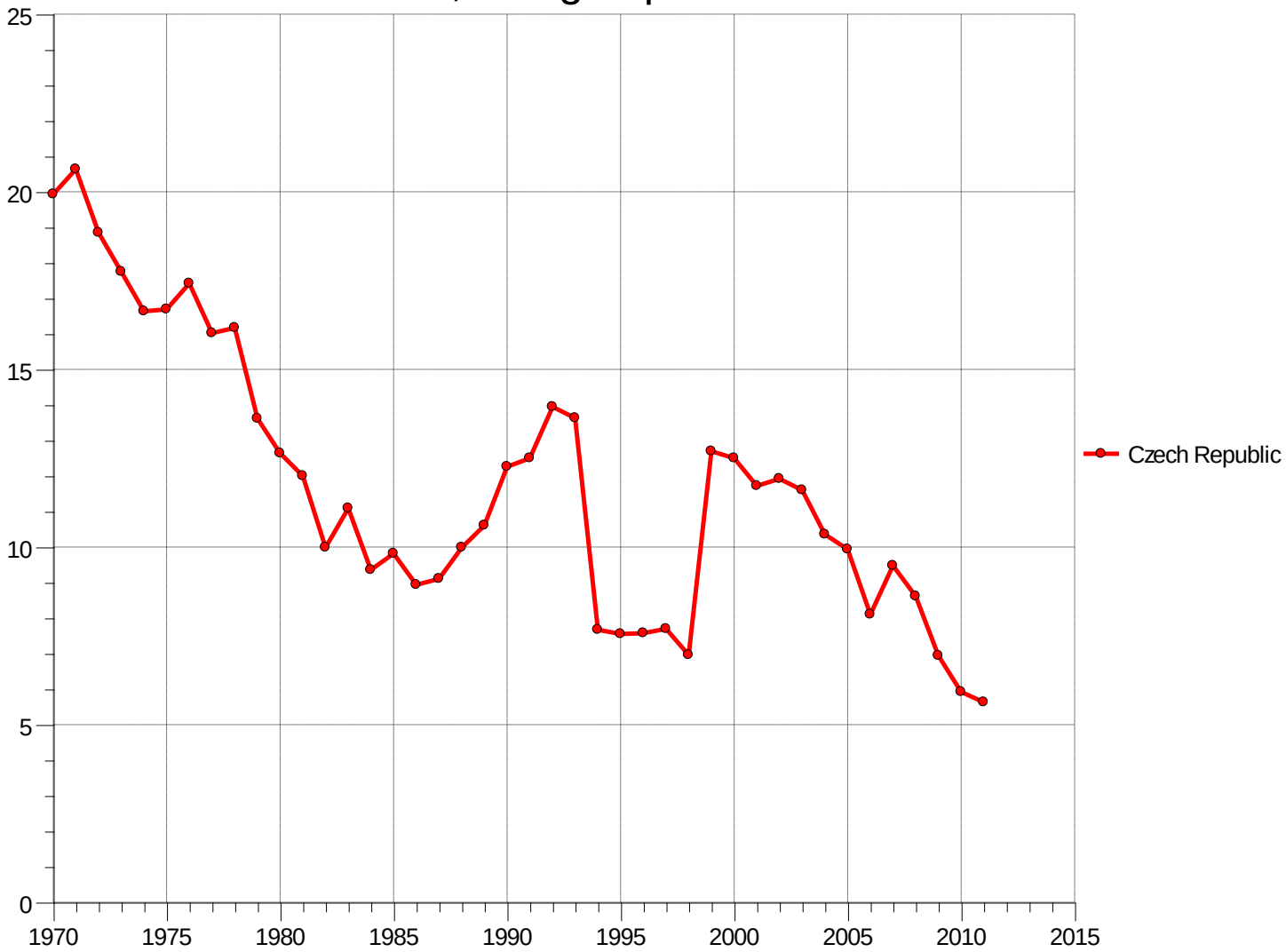
WHO/Euro: HFA statistical database

Miscarriages in Finland 1980-2008

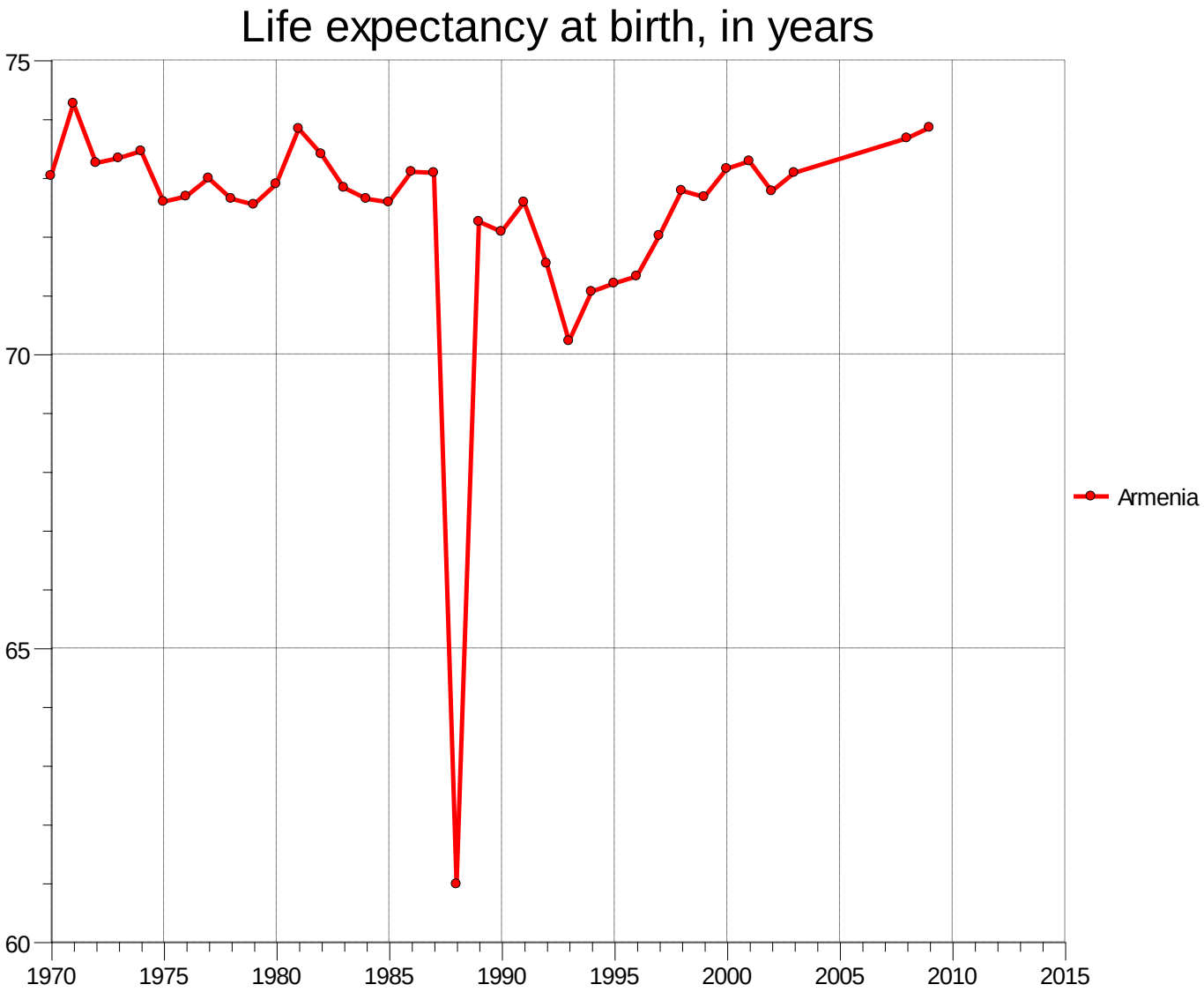


THL: Medical Birth Register, Hospital Discharge Register

SDR, motor vehicle traffic accidents, all ages per 100000

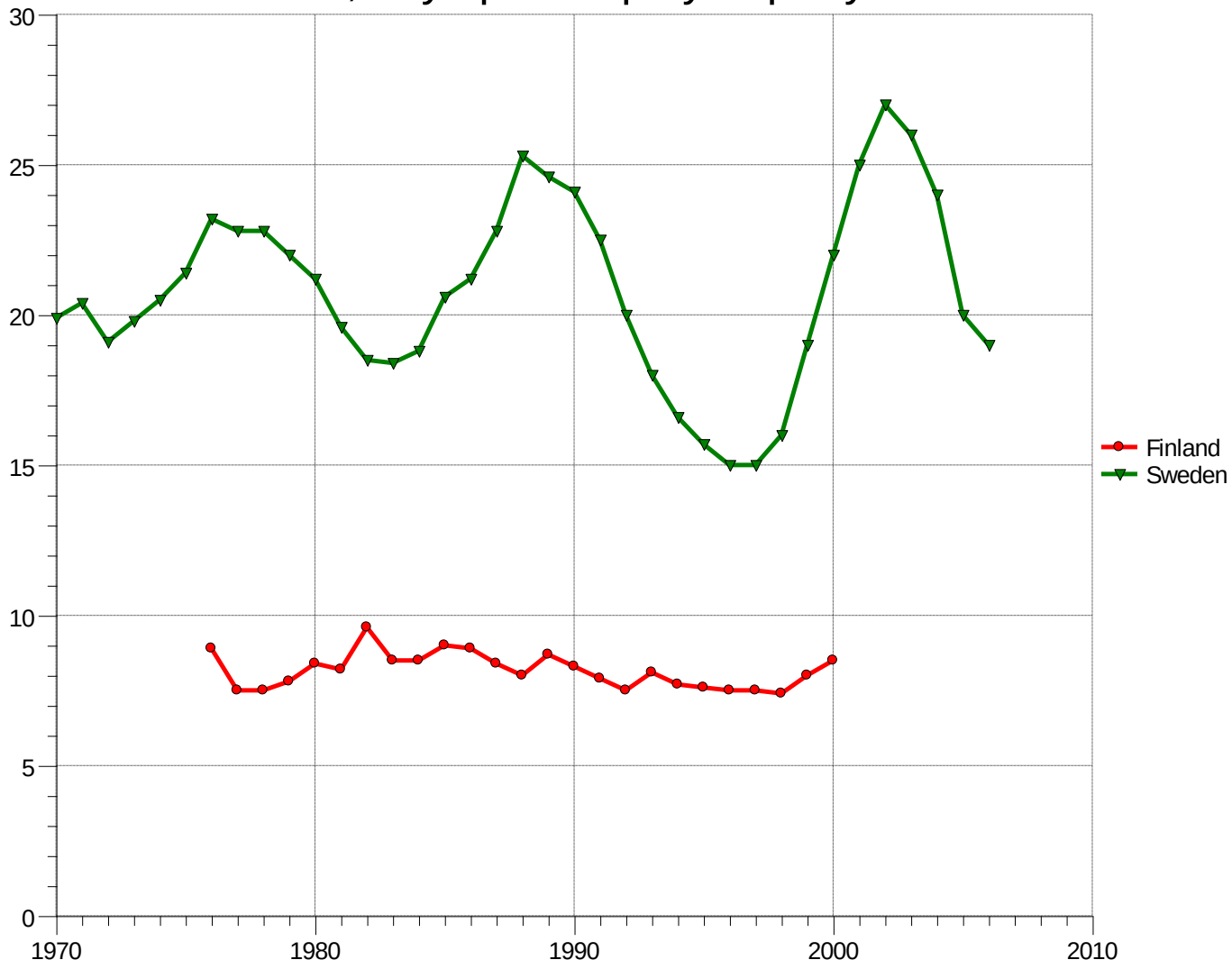


WHO/Euro: HFA statistical database



WHO/Euro: HFA statistical database

Absenteeism from work due to illness, days per employee per year



WHO/Euro: HFA statistical database

Essential to know

- How data is collected in the primary data source?
  - Inclusion/exclusion criteria
  - Definitions, classifications, concepts...
- What affects the data collection:
  - Legislation
  - Care practices
  - Local applications
- Local circumstances

How to measure quality?

Validity index

- If
  - VAL = 1, if $I_{SK} = I_{SR}$ otherwise 0
  - $VAL_e = 1$, if $I_{SK} = I_{SR} \pm$ error otherwise 0
  - where
    - I_{SK} = information in medical records
    - I_{SR} = information in register
    - N = sample size
- Then
  - $VALID_i = 100 * \Sigma VAL_i / N$
  - $VALID_i = 100 * \Sigma VAL_{ie} / N.$

Example (IVF in Finland)

	Source 1		
Source 2	+	-	
+	3296	1088	4384
-	1087	171227	172314
	4383	172315	176698
%			
+	1,9 %	0,6 %	2,5 %
-	0,6 %	96,9 %	97,5 %
	2,5 %	97,5 %	100,0 %

ALL CASES INCLUDED

Correctly reported: 98.8%
Mistakes - Source 1: 0.6%
New cases - Source 2: 0.6%

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Karolinska Institutet

Example (IVF in Finland)

	Source 1		
Source 2	+	-	
+	3296	1088	4384
-	1087	171227	172314
	4383	172315	176698
%			
+	1,9 %	0,6 %	2,5 %
-	0,6 %	96,9 %	97,5 %
	2,5 %	97,5 %	100,0 %

	Source 1		
Source 2	+	-	
+	3296	1088	4384
-	1087		1087
	4383	1088	5471
%			
+	60,2 %	19,9 %	80,1 %
-	19,9 %	0,0 %	19,9 %
	80,1 %	19,9 %	100,0 %

ALL CASES INCLUDED

NON-CASES EXCLUDED

Correctly reported:	98.8%	60.2%
Mistakes - Source 1:	0.6%	19.9%
New cases - Source 2:	0.6%	19.9%

Example (IVF in Finland)

	Source 1		
Source 2	+	-	
	3296	1088	4384
	1087	171227	172314
	4383	172315	176698
%			
+	1,9 %	0,6 %	2,5 %
-	0,6 %	96,9 %	97,5 %
	2,5 %	97,5 %	100,0 %

Capture-recapture -method: 1087 * 1088 / 3296 = 359

It can be estimated that there are 359 (=0.2% of the total population of 171 227) missing cases in this population.

Kappa score/statistics

- Hypothesis
  - cases not related to each other
  - information collected independently
  - both data sources are valid
- $K = (p_o - p_c) / (N - p_c)$
- 95 % CI= $(K - 1.96 * s.e_{(k)}, K + 1.96 * s.e_{(k)})$, where $s.e_{(k)} = [p_o * (N-p_o) / N * (N-p_c)^2]^{1/2}$.
  - p_o = the proportion of cases with identical data
  - p_c = expected value for the proportion of cases where the data for is identical by chance

Methods (III)

- How to interpret K statistics (Landis 1977)?
- Value — Interpretation
  - negative — poor
  - 0.00 - 0.20 — slight
  - 0.21 - 0.40 — fair
  - 0.41 - 0.60 — moderate
  - **0.61 - 0.80** — **substantial**
  - **0.81 - 1.00** — **almost perfect**
- Heavy criticisms recently

Essential to know

- How data is collected in the primary data source?
  - Inclusion/exclusion criteria
  - Definitions, classifications, concepts...
- What affects the data collection:
  - Legislation
  - Care practices
  - Local applications
- Local circumstances

Methodologically sound epidemiological research

- Four principles by McLaughlin 2002:
  - The exposed population should be enumerated completely without selection bias.
  - The exposed population should be tracked without loss to follow-up.
  - The type, occurrence, and severity of health and medical outcomes should be evaluated, and their clinical/public health significance to be determined.
  - Appropriate comparison groups should be established.

Reference: McLaughlin JK: The need for population-based epidemiological studies in the United States. Journal of Long-term Effects of Medical Implants 12 (4): 251-253, 2002.

Problems related to register research

- The data is unavailable
  - primary health care, diseases and conditions not requiring a contact to health care system, self-rated health, opinions, experiences,...
- Data protection: are such studies possible in general?
- Ethically controversial topics:
  - abortion, miscarriage, infertility, malformations, psychiatric disorders, family studies, contact to relatives of a death patient, genetics…
- High data costs: Statistical offices, Central Population Register
- Data overload syndrome
  - Too much data, too little time…?
- Fishing:
  - Easy to find statistically significant results, if the data is large.

Major prerequisites for register linkage studies

- National data protection legislation which enables the use of administrative data in scientific research.
- Administrative data with good quality.
- Possibility to utilise various data sources by using data linkages.
- Good imagination and creativeness helps.
- Patience also sometimes needed!

Reijo Sund: Utilisation of Administrative Registers Using Scientific Knowledge Discovery. Intelligent Data Analysis 7:6, 501-519, 2003

Four questions:

1. Theory: Research and process schema
2. Theory: Understanding the problem / understanding the data
3. Practice: Hip fracture data as an example: do we understand the problem/data?
4. Practice: Hip fracture data as an example: do we understand the results and conclusions?

R. Sund / Utilisation of administrative registers using scientific knowledge discovery

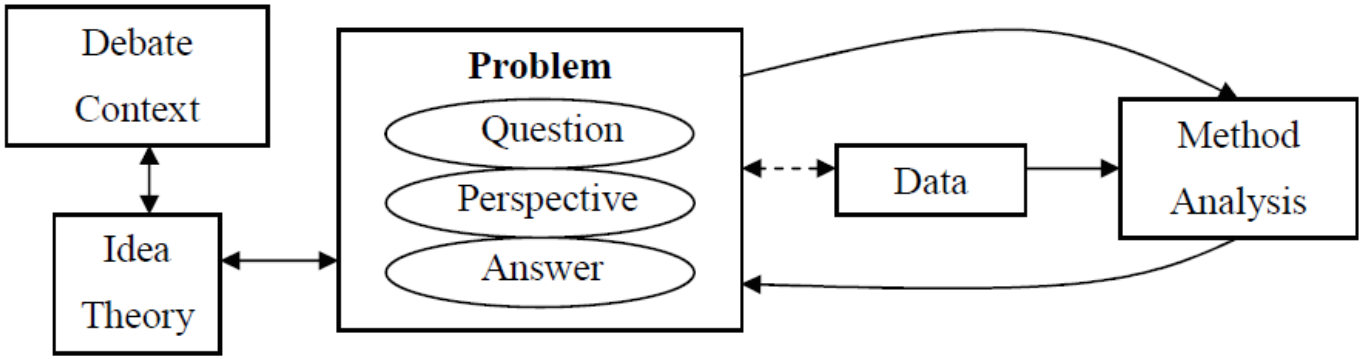


Fig. 1. Research process schema.

R. Sund / Utilisation of administrative registers using scientific knowledge discovery

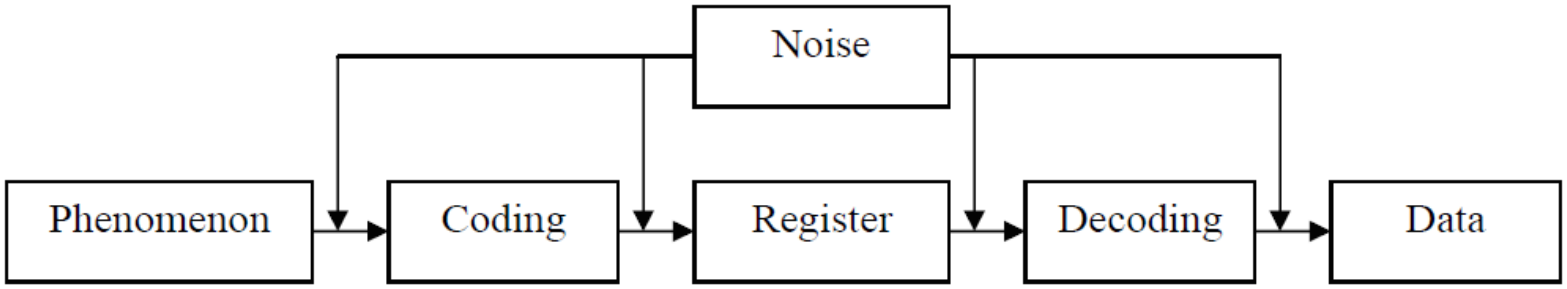


Fig. 3. Schematic diagram of information communication via administrative registers.

Problems related to register research

- The data is unavailable
  - primary health care, diseases and conditions not requiring a contact to health care system, self-rated health, opinions, experiences,...
- Data protection: are such studies possible in general?
- Ethically controversial topics:
  - abortion, miscarriage, infertility, malformations, psychiatric disorders, family studies, contact to relatives of a death patient, genetics…
- High data costs: Statistical offices, Central Population Register
- Data overload syndrome
  - Too much data, too little time…?
- Fishing:
  - Easy to find statistically significant results, if the data is large.

Major prerequisites for register linkage studies

- National data protection legislation which enables the use of administrative data in scientific research.
- Administrative data with good quality.
- Possibility to utilise various data sources by using data linkages.
- Good imagination and creativeness helps.
- Patience also sometimes needed!