



Cross-sectional and episodic deduplication in Record Linkage

Outline

- Introduction
 - What is Record Linkage
 - Why use Record Linkage
- Data sources used
 - The Netherlands Perinatal Registry
 - The PHARMO Database Network
- Cross-sectional deduplication
 - Method
 - Results
- Episodic deduplication
 - Method
 - Results
- Conclusion
- Questions?

Introduction

- What is Record Linkage?
 - “the task of finding *records* in a data set that refer to the same *entity* across different data sources” (Wikipedia)
 - In our case:
 - Which records belong to the same *patient*?
 - We have no common unique identifier such as social security number.
 - We need to make an *informed decision* based on *patient characteristics*.
 - We use an *agreement score function* to make this decision.

Introduction

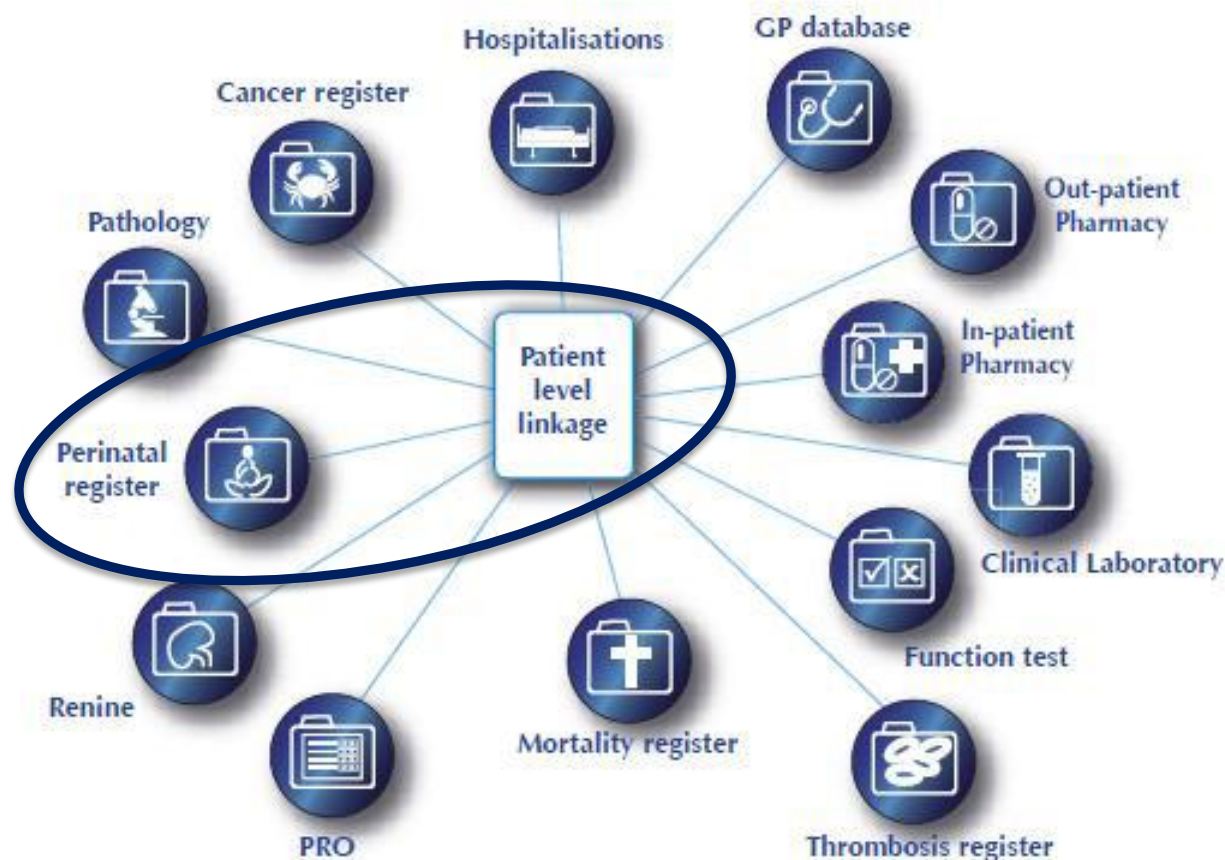
- Why use record linkage?
 - To perform pharmacoepidemiological outcome studies using observational databases.
 - Specifically, pregnancies cannot be included in clinical trials.

The Netherlands Perinatal Registry

- Nationwide registry containing data from all pregnancies.
- Health records regarding perinatal care of patients
- Data provided by:
 - Obstetricians
 - Gynecologists
 - Pediatricians
- PRN data is linked on the level of individual pregnancies
- <http://www.perinatreg.nl/>



PHARMO Database Network (NL)



www.pharmo.nl

“...to use patient-centric data to derive real-life insights into value of medicines for individual patients.”

Cross-sectional deduplication

- Some records are missing a date of birth of the child.
- This means that a mother switched to a different caregiver.
- For example if a mother moved house.
- We want to try to deduplicate these records belonging to a single pregnancy.

Cross-sectional deduplication

- Make an informed decision on patient characteristics.
- What characteristics are available?
 - Date of birth (DoB)
 - Gender (predominantly female)
 - Postal code
 - Gravidity (number of pregnancies)
 - Parity (number of births)
 - Date of first examination
 - Date of referral (optional) } episode

Cross-sectional deduplication

- Make pairs of possible matches, based on DoB and gender.
- Judge pairs to be either a *match* or a *nonmatch*.
- Make a table with every pair as a single record:

ID1	DoB	Gender	Etc.	ID2	DoB	Gender	Etc.
1056	15/1/80	F	...	4517	15/1/80	F	...

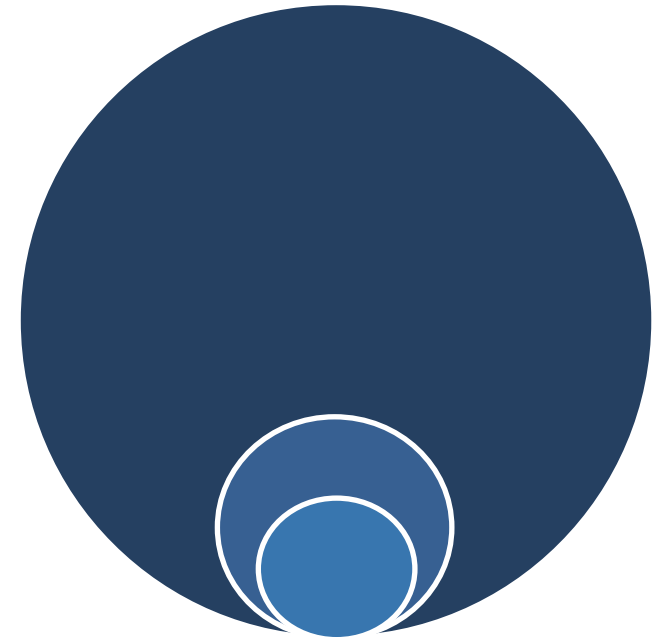
- Flag every record according to matches on characteristics.
- Make the final decision using the agreement score function.

Cross-sectional deduplication

- Results

- We used PRN data from 2000-2007:

- $\pm$ 1.6 million records
 - $\pm$ 110 thousand records with missing DoB for the child
 - $\pm$ 30 thousand pairs are judged to be matches



Episodic deduplication

- PRN data linked on the level of individual *pregnancies*.
- We are interested in *consecutive pregnancies* of individual *mothers*.

Episodic deduplication

- Make an informed decision on patient characteristics.
- What characteristics are available?
 - Date of birth (DoB)
 - Postal code
 - Gravidity (number of pregnancies)
 - Parity (number of births)
 - Date of previous birth
 - Only available if current and previous births took place in a hospital

Episodic deduplication

- Make pairs of possible *previous* pregnancies, based on DoB and gender.
- Further constrain these pairs as follows:
 - Pairs cannot overlap in time
 - Gravity of the previous pregnancy must be *at least* 1 lower than of the current pregnancy
 - Parity of the previous pregnancy must be *exactly* 1 lower than of the current pregnancy

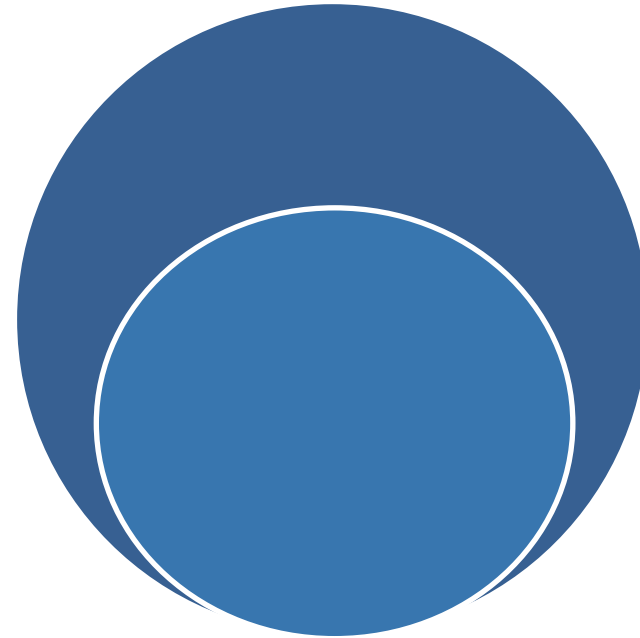
Episodic deduplication

- Judge pairs to be either a *match* or a *nonmatch*.
- Make a table with every pair as a single record:

ID1	DoB	Gender	Etc.	ID2	DoB	Gender	Etc.
1056	15/1/80	F	...	4517	15/1/80	F	...

- Flag every record according to matches on characteristics.
- Make the final decision using the agreement score function.

Episodic deduplication



- Results
 - Again PRN data from 2000-2007:
 - ± 1.6 million records
 - ± 325 thousand pairs confirmed as a match

Conclusion

- A relatively simple agreement function can achieve reasonable results.
- A clearly defined medical event (such as a pregnancy) is very helpful in creating agreement functions.
- However, medical knowledge is very important.
- Furthermore, data needs to be cleaned for this approach. This is very time consuming.
- This approach may be susceptible to biases.

Questions?

Questions?

Cross-sectional deduplication

- Make pairs of possible matches, based on DoB and gender.
- Judge pairs to be either a *match* or a *nonmatch*.
- Make a table with every pair as a single record:

ID1	DoB	Gender	Etc.	ID2	DoB	Gender	Etc.
1056	15/1/80	F	...	4517	15/1/80	F	...

- Flag every record according to matches on characteristics.
 - *grav_flag* is 1 if gravidity corresponds, 0 otherwise
 - *par_flag* is 1 if parity corresponds, 0 otherwise
 - *pc_flag* is 1 if postal code corresponds, 0 otherwise
 - *epi_flag* is 1 if episodes are non overlapping, 0 otherwise

Cross-sectional deduplication

- Code used to evaluate is as follows:

```
if grav_flag = 1 and par_flag = 1 then do;  
  if pc_flag = 1 then match = 1;  
  else if pc_flag = 0 then do;  
    if epi_flag = 1 then match = 1;  
    else match = 0;  
  end;  
end;
```

Episodic deduplication

- Judge pairs to be either a *match* or a *nonmatch*
- Make a table with every pair as a single record:

ID1	DoB	Gender	Etc.	ID2	DoB	Gender	Etc.
1056	15/1/80	F	...	4517	15/1/80	F	...

- Flag every record according to matches on characteristics
 - *prev_flag* is 1 if previous birth date corresponds with the current birthdate, -1 if it does not correspond and 0 otherwise
 - *pc_flag* is 1 if postal code corresponds, 0 otherwise

- **Code:**

```
if prev_flag = 1 then match = 1;
else prev_flag = 0 then do;
    if pc_flag = 1 then match = 1;
end;
else match = 0;
```