PhUSE 2013

Paper HE04

Cross-sectional and episodic deduplication in Record Linkage

Rients van Wijngaarden, MSc, PHARMO Institute, Utrecht, The Netherlands

ABSTRACT

The Netherlands Perinatal Registry (PRN) is a nationwide registry containing data from all pregnancies from various healthcare professionals. The PHARMO Database Network includes data from multiple healthcare databases such as drug dispensings, hospitalizations, GP's and clinical laboratory measurements and covers app. 20% of the Dutch population. To enable pharmacoepidemiological pregnancy outcome studies both databases need to be linked on patient characteristics. One of the key issues in this linkage is the deduplication of the pregnancy records since one mother may have attended several healthcare professionals and may have had more then one pregnancy, both leading to multiple records belonging to a single patient. We have devised a rule based record linkage technique to perform this deduplication both within pregnancies (cross-sectional) and over time (episodic). We will discuss the technical details through examples and present the results.

INTRODUCTION

As it is not possible to include pregnancies in clinical trials, observational databases are needed to perform pharmacoepidemiological pregnancy outcome studies. Most of these observational databases contain data of a subset of healthcare consumption, so combining various databases through Record Linkage (RL) is essential for this type of research. The main challenge in RL is finding the records in the datasets that belong to the same *entity*. In the pharmacoepidemiological setting this entity is usually a patient.

The Netherlands Perinatal Registry (PRN)[1] is a nationwide registry containing data from all pregnancies. The data is provided by obstetricians, gynecologists and pediatricians that are involved in the perinatal care of the patients.

The PHARMO Database Network[2] includes data from multiple healthcare databases such as drug dispensings, hospitalizations, GP's and clinical laboratory measurements in defined regions and covers approximately 20% of the Dutch population.

Both the PRN database and the PHARMO database are anonymized due to privacy restrictions. A common unique identifier, such as a social security number, is not available for linkage. Therefore an informed decision has to be made whether two records belong to the same entity or not, based on patient characteristics. This informed decision is made by using an *agreement function* that evaluates whether two records belong to the same entity. In the PHARMO database a central patient file is used and medical events are ascribed to individual patients. However the PRN database is an event database where data from various caregivers is collected and linked at the level of the individual pregnancy. A central patient file is therefore not available. This means that two records may belong to the same patient, but this is not explicitly recorded in the database as such. One of the key issues in the successful linkage of the PRN and PHARMO database is the deduplication of the pregnancy records in the former. The focus of this paper is on this process of deduplication.

To successfully deduplicate the PRN database two scenarios have to be taken into account. First a patient can switch from one caregiver to another during a single pregnancy that may result in multiple records in the database. These multiple records belong to the same patient as well as to the same event i.e. a single pregnancy. This part of the deduplication is referred to as *cross-sectional* deduplication. Second a patient may have more than one pregnancy over time also resulting in multiple records in the database. These multiple records belong to the same patient, but they do not belong to the same event. This part of the deduplication is referred to as *episodic* deduplication

Obviously both scenarios can occur for a single patient i.e. a patient can have multiple pregnancies where a switch from primary to secondary care occurs for at least one of these pregnancies. Therefore the cross-sectional and episodic deduplication are performed independently and sequentially.

In the first section the method used for cross-sectional deduplication is described. Next the method used for episodic deduplication is described. The third section describes how the results of both previous steps are combined. Finally some concluding remarks on the usability and constraints of the proposed methods are included.

CROSS-SECTIONAL DEDUPLICATION

The cross-sectional deduplication deals with the switch between caregivers within a single pregnancy. A patient can be referred from primary care (midwife, general practitioner) to secondary care (obstetricians, gynecologists) or a patient can move during a pregnancy and switch from one caregiver (either primary or secondary) to another (also either primary or secondary). This kind of switch is usually already linked internally within the PRN data. However in

some cases a date of birth for the child is missing, which may indicate a switch that has not yet been detected. In order to make an informed decision whether two pregnancy records belong to the same pregnancy and the same mother a set of patient characteristics is necessary. Besides date of birth and gender (predominantly female in this case), which are generally considered to be immutable, several other data items are used. Possible pairs of pregnancy records are made using date of birth and gender and these pairs are scored and evaluated based on the additional data items. This results in the pairs being either a match (i.e. belong to the same patient) or a nonmatch (i.e. belong to different patients).

POSTAL CODE

In the Netherlands a postal code consists of four digits and two characters and the combination with a house number signifies a unique address. In both the databases only the four digit postal codes are known. On average a four digit postal code contains ±2000 households[3], so although the combination of date of birth, gender and four digit postal code is quite specific it's hardly unique. Also one of the cases describes in the cross-sectional deduplication specifically deals with mothers moving house. Therefore a similarity in postal code between a pair of pregnancy records is strong evidence that this pair is a match while dissimilarity should not automatically lead to a nonmatch.

GRAVIDITY AND PARITY

Gravidity refers to the number of times a woman has been pregnant while parity refers to the number of times a woman has given birth to a child. Both data items are recorded in the pregnancy database. As we are trying to identify pregnancy records that belong to a single pregnancy and a single mother it is assumed that gravidity and parity should match. As with the postal code a similarity between a pair of pregnancy records on gravidity and parity is a good indication that the pair is a match. On the other hand a dissimilarity should not automatically lead to a nonmatch.

DATE OF FIRST EXAMINATION AND REFERRAL DATE

In the pregnancy database the date of first examination by the caregiver is recorded. Also, if applicable, the date of referral to another caregiver is recorded. Based on these data items an episode can be constructed. This is either from data of first examination to date of referral or from date of first examination to the end of the pregnancy. If these episodes are subsequent and non-overlapping within a pair of pregnancy records a match is likely. These data items are potentially noisy; therefore a margin of 7 days is used when comparing dates.

EVALUATION

Given the data items described above, a scoring method is used to determine whether a pair of pregnancy records is a match or a nonmatch. First of all pairs of pregnancy records are put side by side in a table for ease of evaluation. This means that every pair is represented as a single row in a table containing all the relevant information of both pregnancy records considered to be a possible pair. Then all the pairs are flagged based on similarity on the various characteristics. The variable `grav_flag` is 1 if the gravidity between a pair of pregnancy records matches and 0 otherwise. The same holds for the variables `par_flag`, `pc_flag` and `epi_flag` representing the data items parity, postal code and episode of care respectively. Whether a pair of candidates or not is determined 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;
```

RESULTS

For this project PRN data for the years 2000-2007 has been used. This dataset contains ±1.6 million records. Of these records ±110 thousand have a missing date of birth for the child. After cross-sectional deduplication ±30 thousand of these records have been paired to another pregnancy record. This means that the cross-sectional deduplication is a minor operation and affects less than 4% of the dataset.

EPISODIC DEDUPLICATION

In the PRN database mothers are identified per individual pregnancy, so a mother with multiple pregnancies within the database will be identified with multiple mother_id's. To successfully link these patients to the PHARMO database later on these multiple mother_id's need to be grouped together under a single identifier. To perform this task pairs of possible matching records are created based on a set of criteria and these pairs are evaluated based on patient characteristics.

CREATING PAIRS

As with the cross-sectional deduplication date of birth and gender are used as the basis for creating pairs of pregnancy records that may belong to a single mother. To limit the number of combinations that are created and need to be evaluated we only look back in time when constructing pairs. This means that for every pregnancy record possible *previous* pregnancies are sought based on date of birth and gender. Three additional constraints are used to

further narrow the number of combinations. First of all the two pregnancies that are being matched can not overlap for obvious reasons. A mother cannot get pregnant while still pregnant. Furthermore the gravidity of the previous pregnancy must be *at least* 1 lower than the gravidity of the current pregnancy. Third, the parity of the previous pregnancy must be *exactly* 1 lower than the parity of the current pregnancy. Based on these criteria as set of possible matches, or pairs, is created. These pairs are classified as either a match or a nonmatch. As with the cross-sectional deduplication the pairs are put side by side in a single row for easy evaluation.

POSTAL CODE

As with the cross-sectional deduplication the 4 digit postal code of patients is used for the evaluation of pairs. Patients may have moved house between pregnancies and therefore a similarity in postal code is a good indication that a pair is a match, but dissimilarity carries little information.

DATE OF PREVIOUS BIRTH

If a birth takes place in a hospital and the previous birth took place in a hospital as well, the date of the previous birth is recorded for the current pregnancy record. So if we take a pair of possible matching pregnancy records and the date of previous birth item in the current pregnancy matches the birth date recorded for the previous pregnancy (± 1 day) then a match is very likely. On the other hand if these items disagree a match is very unlikely. Unfortunately the date of previous birth can only be used as an evaluation criterion for a subset of pairs.

EVALUATION

Given the data items described above a scoring method is used to determine whether a pair of subsequent pregnancy records is a match or a nonmatch. All the pairs are flagged based on similarity on the various characteristics. The variable `prev_flag` is 1 if there is a match between the date of previous birth in the current pregnancy and the recorded birth date in the previous pregnancy. For a match in postal code the variable `pc_flag` is given the value 1, or 0 in the case of differing postal codes. Whether a pair of candidates or not is determined as follows:

```
if prev_flag = 1 then match = 1;
else if pc_flag = 1 then match = 1;
```

While the actual evaluation is simpler than for the cross-sectional evaluation, an additional step has to be made to create a workable dataset. Every pair that is designates as a match contains a mother_id and a mother_id of a previous pregnancy, or p_mother_id. We want to make sure that in the dataset of matches every mother_id is unique and every p_mother_id is unique. This is best explained by a simple example. Given three pregnancies A, B and C with A being the first and C being the last chronologically, a table of matches may look like this:

mother_id	p_mother_id
C	B
C	A
B	A

This table contains redundant information because C is linked to A directly and indirectly through B. If we were to make every instance of mother_id and p_mother_id unique, this problem would be solved and the resulting table would be as follows:

mother_id	p_mother_id
C	B
B	A

RESULTS

For this project PRN data for the years 2000-2007 has been used. This dataset contains ±1.6 million records. Of these records ±325 thousand pairs have been confirmed as matches using the episodic deduplication. This set of matched pairs affects ±610 thousand unique pregnancy records, almost 40%. This makes sense as a lot of people have more than one child.

PUTTING IT ALL TOGETHER

The results of both deduplications need to be incorporated in the pregnancy dataset that will be used for research later on. This is done as described below.

CROSS-SECTIONAL DEDUPLICATION

The result of the cross-sectional deduplication can be incorporated in the pregnancy dataset with relative ease. Two

mother_id's that have been identified as belonging to a single mother can be replaced by a new (unique) id in three steps. If for example we have a table with confirmed matches that looks as follows:

new_id	mother_id	matched_mother_id
1	A	B

Suppose the original pregnancy dataset looks like this:

mother_id	...
A	...
B	...
C	...

Step 1: Join the match set to the original set on mother_id.

new_id1	mother_id	...
1	A	...
.	B	...
.	C	...

Step 2: Join the match set to the original set on matched_mother_id.

new_id2	new_id1	mother_id	...
.	1	A	...
1	.	B	...
.	.	C	...

Step 3: New_id1 and new_id1 are coalesced to form the final identifier. This new identifier replaces the original mother_id.

mother_id	...
1	...
1	...
C	...

EPISODIC DEDUPLICATION

The results of the episodic deduplication are somewhat more complex to incorporate in the pregnancy dataset. The reason for this is that all pregnancies belonging to a single patient may be represented by multiple records within the match set. Therefore the simple three-step approach used for the cross-sectional deduplication will not work for all cases. Instead a recursive approach is used. First we identify the last pregnancy in a chronological *chain* of pregnancies and assign a new id. In the following step this new id is assigned to the previous pregnancy in the chain. This is repeated until every pregnancy in the chain has been assigned with the same id. Take for example a chronological chain of matched pregnancies: A → B → C → D. This chain will be represented in the match set as follows (again with p_mother_id being the identifier of a previous pregnancy):

mother_id	p_mother_id
B	A
C	B
D	C

Step 1: Join the match set with itself flagging all the mother_id's that are also p_mother_id's using the following piece of code:

```
proc sql;
       create table match_set as
       select a.*, b.n as flag
       from match_set as a
       left join
       (select mother_id, p_ mother_id, 1 as n
       from match_set
       where mother_id in (select p_ mother_id from match_set)) as b
```

```
       on a. mother_id = b. mother_id;
quit;
```

This results in the following table:

mother_id	p_mother_id	Flag
B	A	1
C	B	1
D	C	0

Step 2: A new id is assigned to the pair that has not been flagged.

mother_id	p_mother_id	Flag	new_id1
B	A	1	.
C	B	1	.
D	C	0	1

Step 3: This extended match set is again joined with itself were the newly assigned id is assigned to the previous pregnancy in the chain using the following piece of code:

```
proc sql;
       create table match_set as
       select a.*, b.new_id1 as new_id2
       from match_set as a
       left join match_set as b
       on a.mother_id = b.p_mother_id;
quit;
```

This results in the following table:

mother_id	p_mother_id	Flag	new_id1	new_id2
B	A	1	.	.
C	B	1	.	1
D	C	0	1	.

Step 3 is repeated until there is no pair left that has not been assigned a new identifier. In our example the end result is this:

mother_id	p_mother_id	Flag	new_id1	new_id2	new_id3
B	A	1	.	.	1
C	B	1	.	1	.
D	C	0	1	.	.

All the newly assigned id's are coalesced and used to update the pregnancy set, with A, B, C and D now all having the same identifier. Because step 3 is repeated multiple times it is wise to use a macro to avoid repetitive coding.

CONCLUSION

Using the methods described above reasonable results can be achieved with relatively simple agreement score functions. What helps is that the pregnancy is a clearly defined medical episode which is fairly uniform for all women with respect to duration. Consecutive pregnancies cannot overlap in time and a pregnancy will last at most 42 weeks. These constraints are helpful when making pairs of possible matches and constructing agreement rules. While not all medical conditions are as uniform similar kinds of constraints are probably available for many. Therefore medical knowledge is very useful for successful medical record linkage. However, there are some pitfalls. First of all the data used needs to be as clean and noiseless as possible for this kind of linkage. Much effort has to be put in cleaning and shaping the data in a useful way and this project was no exception. Furthermore, in the case of the cross-sectional deduplication, the evaluation relies heavily on matching postal codes. A bias may have been inadvertently been introduced here with an underrepresentation of mother that have moved house during a pregnancy. For the episodic deduplication the same holds for the date of previous birth. Births that take place outside the hospital are much harder to match to others, also leading to a possible bias. These limitations have to be taken into account when the data is used for research.

REFERENCES

[1] http://www.perinatreg.nl/home_english
[2] http://www.pharmo.nl/
[3] http://www.cbs.nl/nl-NL/menu/informatie/beleid/publicaties/maatwerk/archief/2011/111206gemiddeldbesteedbaarinkomenpostcodegebiedmwxls.htm

ACKNOWLEDGMENTS

We would like to thank the people of the Netherlands Perinatal Registry for making this record linkage project possible.

CONTACT INFORMATION

Rients van Wijngaarden, MSc
PHARMO Institute
Van Deventerlaan 30-40
3528 AE Utrecht
Work Phone: +31 30 7440 819
Fax: +31 30 7440 801
Email: rients.van.wijngaarden@pharmo.nl
Web: http://www.pharmo.nl