PhUSE EU Connect 2018

Paper RW06

Experiences and strategies in Linking Real-World Data from different sources

Berber Snoeijer, Irene Bezemer, PHARMO Institute, Utrecht, The Netherlands

ABSTRACT

Real world data from a given patient is stored in different databases on different locations and in information systems with different structures and content. Patient numbers differ between the databases and crucial characteristics are present in one but not in the other database. To get a complete overview of the clinical history of a patient, it is important to be able to combine all this information. This paper will handle some strategies and pitfalls in linking data from different sources. The linked data can be used to give the caregivers more insight into their own patients and for safety and effectiveness reports towards pharmaceutical companies.

INTRODUCTION

Record linkage means that records from different databases are joined when it concerns the same entity (in our case a patient or subject). Clinical record linkage is more and more used in current practice to fill data gaps, add outcome information, delete duplicates or increase research populations.

At the PHARMPO Institute several databases are used for their studies such as the clinical outpatient pharmacy database, the GP database, the hospitalization database and the national cancer registry. These databases are linked by an independent foundation (STIZON) with regard to all privacy requirements and legislation. There are different methods to perform clinical record linkage and there are also different levels of certainty on how well the linkage is performed. Each linkage process will result in a number of false positives (records mistakenly linked) and false negatives (no link due to missing or inaccurate data). Bohensky et.al studied the validity and completeness of data linkage in 33 published studies and compared the difference in characteristics of unmatched and matched patients which might indicate a bias in the data linkage and thus in the results of the study[1)]. In this article we will give a few examples of the linkage process and we will discuss the difficulties in linkage of the databases that we experienced.

LINKAGE METHODS

There are different methods of record linkage that can be used when different databases are to be joined. First, if both databases use the same numbering you can use deterministic linkage based on this unique number. For this, the national social security number is a good option to be used. However, in practice part of these numbers is still missing and a small part of the numbers prove to contain typing errors. Other identifiers that can be used for linkage are birth date, gender, initials and (maiden) name. Deterministic linkage gives an "all or nothing outcome" [2)]. There are two methods, the first is exact deterministic linkage where an exact match on all identifiers is required and the second is iterative deterministic linkage where an exact match is required on one of several rounds of matching.

Probabilistic matching is more powerful in that it includes the discriminatory power of each identifier. For each combination of records from two datasets it is possible to calculate a probability (linkage score) that they are from the same patient based on all identifiers included. However, considering the complete Cartesian product of combined data files in real world databases, results in large databases which are impractical and will use large quantities of computational resources. Therefore, before probabilistic matching can take place, subsets of data have to be created with equal basic values that are preferably free from error and do not change in time. Variables like birth date and gender are good examples of these equal basic values. An example of a probabilistic method is the PHARMO record linkage. This linkage model includes different time independent and time dependent variables as well as logical related variables to perform probabilistic linkage between outpatient pharmacy dispensing records like the hospital admission records [3)]. However, creating a probabilistic model is time consuming, statistically challenging and in case of a single study for matching data often not more powerful than deterministic linkage. The examples in this paper will therefore handle only two different types of deterministic linkage.

SELECTION OF IDENTIFIERS

The first step in linkage regardless of the method is to go through the databases to identify the patient characteristics that are logically related to both registries. These can be classified as time-independent identifiers, time-dependent identifiers and logically related identifiers. Time independent identifiers are characteristics that will not change over time. Only typing errors or rare events such as gender transition will change these. Time dependent identifiers are characteristics that may change over time but are rather consistent for a shorter period of time. Logical related variables may be disease and the corresponding prescribed drug. If that combination is present then the probability of

a correct match will increase. Or when a patient is admitted in the hospital it is not likely that he received medication in the outpatient pharmacy. So that combination will decrease the probability that it is the same subject.

Which variables to use as matching variables, is dependent on the information available in both databases and the level of completeness of these variables.
After having selected the identifiers to use, one can set up a linkage method. The selection of the model will be based on development time of the model (setting up deterministic linkage models is faster than probabilistic linkage models) and number of possible identifiers to match.

List of possible identifiers by type

Type	**Identifier**
Time independent	Date of birth, Gender, Initials, (maiden) name
Time dependent	GP, Insurance number, Postal code,
Logically related	Start of treatment vs study inclusion Date of hospital inclusion vs certain medical events hospital admission vs prescriptions in local pharmacy

EXAMPLE 1: LINKAGE OF GP AND HOSPITAL DATA: DETERMINISTIC MATCHING WITH DATA ENRICHMENT

This example is about a large united group of GPs who liked to know whether they miss information in their information systems regarding cardiovascular diseases. This can be checked based on the hospitalizations that are reported by the hospitals in their neighborhood. It was decided to do a pilot for this to see how much extra information could be retrieved from the hospital information system such that the GPs could give better and more specific care to their patients. For this purpose, the GP records were to be linked to a hospitalization database. Both databases use different numbering.

The possible identifiers that were used for matching were birth date, gender and postal code. These variables were present in both databases. In addition, the practice number from the GP database was regarded in relation to the GPs that were reported in the hospitalization database. The most frequently occurring GPs in relation to a certain practice after the first matching step on time independent variables were considered as acceptable matches. This was done because the time period we observed was rather long and comprised 7 years. That implies that a number of GPs had changed from practice or had retired. Not including those GPs as correct matches would exclude a lot of links between the databases. To go through all kinds of public available records to find out which GPs had been working for the specific practice would also give a lot of burden. This pragmatic method resulted in a quick overview of all GPs having worked in the specific practice.

STEP 1: CREATION OF UNQIUE PATIENT RECORDS

The first step was to create unique patient datasets for both databases including 1 row per patient. From the GP database we used only those patients that were registered for the specific group of GPs. From the hospitalization dataset we used all information from patients living in the area of the GPs. This included also patients from other GPs in the area.

STEP 2: MATCH ON TIME INDEPENDENT VARIABLES

In the second step both patient records were combined using SAS PROC SQL. Only the time independent variables were included. In this pilot case we regarded the postal code as a time independent variable. This will result in some data not being linked. On the other hand using the most recent hospitalization and GP data per patient will increase the likelihood of a link. With this method the more recently hospitalized patients will have a higher likelihood of correctly being linked. For the purpose of this pilot, this was no problem. However, for a RWE trial it is necessary to have an

even chance of being linked through time. Then considering postal code as a time dependent variable is important.

```
PROC SQl;
       CREATE TABLE MyLinkedData AS
              SELECT a.practiceNo, a.PatIdGP, a.birthdate, a.gender, a.zip AS zipGP,
                     a.gp, a.gp_agb,
                     b.HospNo, b.patIdHosp, b.GPHosp, b.zip as b.zipHosp, b.GPHospName
              FROM PatsGP as a
                     INNER JOIN PatsHosp as b
                            ON a.birthdate=b.birthdate and a.gender=b.gender
                                   and SUBSTR(a.zip,1,4)=SUBSTR(b.zip,1,4)
                     ORDER BY a.practiceNo, a.gp, a.patIdGP;
QUIT;
```

The resulting dataset will contain all possible matches based on birth date, gender and 4 number postal code. It includes the full postal code and information regarding the GP from both databases as well but the linkaged based on these was not yet performed.

STEP 3: DETERMINISTIC MATCH ON TIME DEPENDENT GP VARIABLE

During the linkage process it appeared that in many practices the GPs had changed. This results in other GP numbers and names. Performing an exact match on these GPs would result in the loss of many patients. Therefore, we first made an inventory of the GPs from the hospitalization dataset that were likely to be a match to the particular GP practice.

To do this, based on the dataset of all possible links, we first counted the number of matches per possible combination of practice fro the GP database and GPs from the hospitalization database.

```
PROC SQL;
       CREATE TABLE HA1 AS
              SELECT DISTINCT practiceNo, gp, gp_agb, HospNo, GPHosp, GPHospName,
                                   count(*) as Noccur
              FROM MyLinkData
              WHERE gp <> "" OR gp_agb <> ""
              GROUP BY practiceNo, gp, gp_agb, HospNo, GPHosp, GPHospName
              ORDER BY practiceNo, Noccur desc;
QUIT;
```

Upon viewing the results it was decided that an occurrence of at least 100 matches per combination was enough to regard it as a valid GP for the corresponding practice. Furthermore, if the name corresponded in both databases and it matched at least 20 times then it was regarded as a good match as well. Finally when the GP number, which is unique for a physician, was equal in both databases then it was regarded as a good match in any case.

The information whether it was a good GP match was merged back to the original Linked dataset (MylinkedData).

```
PROC SQL;
       CREATE TABLE MyLinkData2 as
              SELECT a.*, b.HA_OK
              FROM MyLinkData as a
                     LEFT OUTER JOIN HAOk as b ON
                            a.practiceNo=b.practiceNo and a.gp=b.gp and
                            a.gp_agb=b.gp_agb
                            and a.HospNo=b.HospNo and a.GPHosp=b.GPHosp and
                            a.GPHospName=b.GPHospName
              ORDER BY practiceNo, PatIdGP, HospNo, PatIdHosp, firstopname,
lastopname;

QUIT;
```

Because of the differences over time it was likely that for many patients we had more than 1 record matched. In the next step the information of those records was combined and compressed to one row per patient whereby the most complete postal code and whether a correct GP match was found in one of the matches were kept in the final records.

STEP 4: FINAL MATCH SELECTION

In the final selection step, we only selected those patients who had a complete corresponding postal code and a matched GP as defined in the step before.
By using this method 184159 patients, which was 39% of the total number of patients, registered in the practices were linked to the Dutch hospital registry data. This registry included hospitalizations but also visits to the outpatient departments of hospitals.

STEP 5: CHECKING THE RESULTS

After matching, the data was logically checked regarding gender and age category. It appeared that relatively more women were matched (53% linked patients were women vs 50% of all patients) and there were more children (between 0 and 10) and elderly (between 50 and 80) patients linked. It is logical that those patients visited the hospital more often than people between 10 and 50 years.
The final check whether all practices were linked correctly gave a negative results for practices using 2 specific information systems. The GP information from these was not matched correctly.

EXAMPLE 2: LINKAGE OF STUDY DATA AND OUTPATIENT PHARMACY DATA: ITERATIVE DETERMINISTIC MATCHING

Another example is about a study that was performed by Philips research, Radboud UMC and IQ healthcare regarding an adherence control engine that was developed by them. They performed an effectiveness study for this engine to test to what extent they could improve medication adherence for patients by giving them a tailored intervention based on their assessed adherence profile, i.e., their non-adherence risk and the underlying reasons for non-adherence. The medication adherence scores in the intervention group had to be compared to those of the control group that received only standard care.
The medication adherence scores that were to be calculated indicated to what extend the patient had enough medication for the study period. A score of 50%, for example, indicated that the patient received only half of the medication from the pharmacy that he should have gotten based on the dose and dosing regimen. Although we are not really checking that the patient actually took the medication, the receipt of medication is in practice a good approximation of the adherence. This medication adherence is assessable based on outpatient pharmacy data.
For this study we combined the medication adherence scores based on outpatient pharmacy data with the study database. Those included patients all went to a number of different pharmacies which were included in the study. These pharmacies used different information systems indicating that their patient files were not shared. The information from one system was more complete than that of another system and it was not clear which patient was recruited in what pharmacy.

The possible identifiers that we used were:
The time independent match variables were social security number (if available), birth date and gender; the time dependent match variable was gp and the logical related variables were study inclusion date versus start of treatment and study recorded disease area versus kind of started treatment.

STEP 1: UNQIUE PATIENT RECORDS

An unique patient record of all patients included was delivered by the research manager of the study. From the outpatient pharmacy database all data of the included pharmacies were selected. From these datasets only the patients were selected who started on one of the specified drug classes within the specified inclusion period. Because a patient could start on more than one of the interested treatments within the defined period a patient could be more than once in the pharmacy patient dataset.

STEP 2: MATCH ON TIME INDEPENDENT VARIABLES

In the next step the patient records of both sources (study data and pharmacy data) were combined. For a number of pharmacies we had the patients social security number available but not for all. When the social security number matched then this prevailed above all other matches. In this step a match on social security number or a match on birth date and gender was accepted.

```
PROC SQl;
    CREATE TABLE MyMatch as
           SELECT a.*, b.* FROM MyPharmacyPatients as a
                  LEFT OUTER JOIN MyStudypatients as b
                         ON (a.birthdate=b.birthdate and a.gender = b.gender)
                         OR (a.SocNo=b.SocNo and a.SocNo <> .)
           ORDER BY a.PatIdPh, b.PatIdStud;
```

```
QUIT;
```

STEP 3: INDICATORS OF MATCH COMPATIBILITY

In the next step the compatibility of GPs and disease area versus started treatment were stored in ranking variables. If the physician corresponded between study database and pharmacy data then the GPMatch variable got a value 2; if the GP from both datasets worked in the same practice the GPMatch variable got value 1 and otherwise it got value 0. If the disease area corresponded with the started treatment, the corresponding ranking variable (indexmed) got value 1; otherwise it got value 0. In addition also a variable exactmatch indicating a BSN match was added.
The absolute difference between the inclusion date and the start of treatment date was derived and stored as well (adifdat) to indicate the compatibility between the study patients and the GP patients.

STEP 4: ITERATIVE LINKING

Although it is called iterative linking, the linkage in this example was performed in only a few data steps while we actually mimicked 6 iterations of linking. This was done by giving all possible matches a linking score (see MatchOK in the example below). The highest score would indicate a match in the first iteration and so on. In this case, the highest score was 6 indicating that the match would have been made in the first iteration.
After all possible matches received a linking score the data was sorted by study patient number and score. For each study patient the match with the highest score has been taken as the most probable match.

```
DATA Mymatch3;
    SET MyMatch2;
                /* equal social security number: prevails above all other matches
*/
              IF exactmatch=1 THEN MatchOk=6;
                /* equal GP and corresponding inclusion date and corresponding
                index medication */
          ELSE IF GPMatch IN (1,2) AND indexmed=1 and adifat<15 THEN MatchOK=5;
                /* corresponding GP and inclusion date */
          ELSE IF GPMatch IN (1,2) AND adifdat<15 THEN MatchOk=4;
                /* corresponding GP and inclusion data within 2 months */
          ELSE IF GPMatch IN (1,2) AND adifdat<60 THEN MatchOK=3;
                /* inclusion data within 2 months */
          ELSE IF adifdat<60 THEN MatchOK=2;
                /* equal GP */
          ELSE IF GPMatch IN (1,2) THEN matchOK=1;
                /* all other matches */
          ELSE MatchOk=0;
RUN;

PROC SORT DATA=mymatch3;
    BY PatIdStud DESCENDING matchOk;
RUN;

DATA MyMatch4;
    SET MyMatch3;
    BY PatIdStud DESCENDING matchOk;
    IF FIRST.patIdStud THEN OUTPUT;
    KEEP patIdStud patIdPh;
RUN;
```

ISSUES IN THE MATCHING PROCESS

The above method seems rather simple and univocal. However, in practice, a number of practical problems occurred. For example, some patients go to different pharmacies. Especially when you regard a larger period of time, this can cause problems in the matching process. When the second pharmacy without the data of interest records a social security number and the other pharmacy didn't then you will miss the correct link.
Another issue was that we seemed to have a match in a number of cases but the matched patient had started other medication then indicated in the study. Or in some cases we had a match on GP, birth date and gender but the inclusion date deviated largely from the start date of medication. The possible reasons for a non-match or an unlikely match are diverse. The study records as well as the pharmacy records could contain errors. A number of patients were for example included by the pharmacies while the inclusion criteria were not met. Still, the final match and results of

these were good. When disregarding the unlikely matches we had matched 89% of the study patients to the corresponding pharmacy records including adherence scores.

CONCLUSION

There are a number of methods available for matching different databases with each other. This paper describes the practical approach of two deterministic methods. Each combination of databases will have its own challenges and pitfalls which will need to be covered. The programmer should alWays examine the matching results and be aware for possible matching problems and search for solutions.

REFERENCES

1) MA Bohensky et. Al. BMC Health Services Research2010 **10**:346
2) Stacie B Dusetzina et. Al. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]. Rockville (MD): Chapter 4. Agency for Healthcare Research and Quality (US); 2014 Sep. Report No.: 14-EHC033-EF
3) RMC Herings, PHARMO a record linkage system for postmarketing surveillance of prescription drugs in the Netherlands. 1993. ISBN 90-393-0196-4

ACKNOWLEDGMENTS

I would like to thank Philips research, IQ healthcare and the Radboud UMC for using the data of their study for the second example in this paper.

CONTACT INFORMATION (HEADER 1)

(In case a reader wants to get in touch with you, please put your contact information at the end of the paper.)
Your comments and questions are valued and encouraged. Contact the author at:

Berber Snoeijer & Irene Bezemer
Pharmo Institute
Van Deventerlaan 30-40
3528AE Utrecht, The Netherlands
Work Phone: +31 30 7440 800
Email: pharmo@pharmo.nl
Web: www.pharmo.nl