dHealth 2019 – From eHealth to dHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-971-3-210

Patient Record Linkage for Data Quality Assessment Based on Time Series Matching

Alphons EGGERTH[a,b,1], Dieter HAYN[a], Karl KREINER[a], Sai VEERANKI[a,b], Heimo TRANINGER[c], Robert MODRE-OSPRIAN[a] and Günter SCHREIER[a,b]
[a]*AIT Austrian Institute of Technology GmbH, Graz, Austria*
[b]*Graz University of Technology, Graz, Austria*
[c]*ZARG Zentrum für ambulante Rehabilitation GmbH, Graz, Austria*

Abstract. Background: Huge amounts of data are collected by healthcare providers and other institutions. However, there are data protection regulations, which limit their utilisation for secondary use, e.g. research. In scenarios, where several data sources are obtained without universal identifiers, record linkage methods need to be applied to obtain a comprehensive dataset. Objectives: In this study, we had the objective to link two datasets comprising data from ergometric performance tests in order to have reference values to free text annotations for assessing their data quality. Methods: We applied an iterative, distance-based time series record linkage algorithm to find corresponding entries in the two given datasets. Subsequently, we assessed the resulting matching rate. The implementation was done in Matlab. Results: The matching rate of our record linkage algorithm was 74.5% for matching patients' records with their ergometry records. The highest rate of appropriate free text annotations was 87.9%. Conclusion: For the given scenario, our algorithm matched 74.5% of the patients. However, we had no gold standard for validating our results. Most of the free text annotations contained the expected values.

Keywords. medical record linkage, data analysis, ergometry, exercise test, cardiac rehabilitation

1. Introduction

Clinical trials are commonly limited to a specific study population, which cannot represent the target population in full detail. Additionally, in clinical trials there is a limited time of observation and a relatively small number of subjects. To get a more comprehensive view, different strategies can be followed. On the one hand, researchers are trying to reuse available datasets from various clinical trials by combining them to bigger datasets ("record linkage"), e.g. using EUPID [1]. On the other hand, routine care increasingly relies on information and communications technology (ICT), which leads to huge amounts of patient data. As they are documenting treatments and their outcomes under real-world conditions, routine data are a highly valuable resource for research studies ("secondary use").[2]

[1] Corresponding Author: Alphons Eggerth, AIT Austrian Institute of Technology GmbH, Reininghausstraße 13, 8020 Graz, Austria, E-Mail: alphons.eggerth@ait.ac.at

Secondary use of healthcare data brings high responsibilities for the research team. The research environment needs to meet legal and ethical requirements, which limit the usage of the data. A very important aspect is the protection of the patients' personal information (e.g. GDPR for Europe [3], HIPAA for the US [4]). Along with providing secure data storage and computing environments, data needs to be cleaned from identifying elements. Thus, names, social security numbers, telephone numbers, etc. must be removed from the datasets. [2]

Once all requirements are met, there is the question for the optimal record linkage algorithm. Usually, several datasets are given, which need to be linked. In an optimal case, a universal identifier or common patient pseudonyms exist, which can be used to directly connect all the records. However, there are situations, when such a universal identifier (e.g. for privacy reasons) or common patient pseudonyms (e.g. due to various origins of the datasets) are not present. To enable linkage of de-identified datasets in such situations, two datasets need to share some of their fields, i.e. some information needs to be present in both datasets (e.g. date of birth, ZIP code, sex), which contain enough information to obtain a unique combination of values for each patient (see k-anonymity concept [5]). While deterministic approaches try to match records through rule-based algorithms, probabilistic approaches rely on statistical methods to calculate weights for the available parameters, which are then applied for estimating matching probabilities [6-9].

In our study, we obtained data from two different sources: a) patient record data from a manually entered database (exported as an Excel file) as well as b) raw data, metadata and manually entered free text annotations of ergometric performance tests recorded with the used ergometers (exported as separate XML files). We tried to link these two data sources to validate the XML files' free text annotations. However, there was neither a universal identifier nor common patient pseudonyms nor was the data format suited for commonly used record linkage algorithms. Thus, we transformed the data and applied a distance based time series record linkage approach, as proposed by J Nin and V Torra [10].

This paper is organised as follows: For the given datasets, we first present the application of the time series record linkage algorithm. Second, we investigate the quality of the XML files' free text annotations.

2. Methods

Pseudonymised ergometry data from the rehabilitation centre *ZARG Zentrum für ambulante Rehabilitation GmbH* were obtained comprising of an Excel file containing manually entered database entries from 1.538 cardiac rehabilitation patients as well as of 29.876 XML files that had been recorded with ergometers and contained data from one ergometric performance test each. In this paper, we use "PAT file" as a notation for the Excel file and "ERGO files" as a notation for the XML files. However, the date ranges of the datasets were just partly overlapping. Thus, for most of our analyses, the PAT file entries after 13.06.2017 were not used. For a detailed description of the source datasets see Table 1. All analyses were conducted using Matlab (The MathWorks, Natick, US).

During pre-processing, two pseudonymised IDs of the ERGO files were dismissed, as they had more than 100 performance test entries, which is very unlikely for a common patient. Furthermore, minor typos (e.g. year = "2217" instead of "2017"), which became obvious due to unexpected outcomes during implementation, were corrected.

As an initial step, the information from the ERGO files was parsed and transformed into a table. After comparing the ERGO and the PAT dataset with each other, we extracted for all available performance tests the date and six parameters providing identical entries in both datasets. We converted all dates and parameter values to integer values and created a separate table for dates and for each of the parameters both for PAT and ERGO: a) for the PAT files we arranged the integers in the columns with one row for each patient (referred to as "PAT tables") and b) for the ERGO files, we arranged the integers in the columns with one row for each ID (referred to as "ERGO tables").

Table 1. Properties of the ergometry data received from the ZARG rehabilitation center.

Property	PAT file (= Excel file)	ERGO files (= XML files)
Content	Four sheets of manually entered patient information and results from ergometric performance tests. Every sheet represented one of four ergometric performance tests, which had been conducted during the cardiac rehabilitation program: • Start of phase 2 • End of phase 2 • Start of phase 3 • End of phase 3	Each ERGO file contained the data of one performance test. The data comprised of raw data (e.g. heart rate curves, workload step profiles, ECGs), metadata and sometimes annotations (e.g. free text entries denoting the reason for the test). Files for the same patient were linked by a pseudonymised ID.
Number of records	1,538 patients	29,876 ergometric performance tests
Origin of the dataset	Cardiac rehabilitation	Cardiac rehabilitation; by order of a physician; as part of a training program.
First performance test	11.02.2013	22.01.2004
Last performance test	13.09.2018	13.06.2017
Identical entries in both datasets	Sex, age, height, weight, maximum workload value of applied step profile, maximum heartrate value during performance test, date of performance test.	

2.1. Record linkage algorithm

We implemented an iterative, distance-based time series record linkage algorithm to match the de-identified patient record data of each cardiac rehabilitation patient from the PAT file with his/her corresponding, pseudonymised ID in the ERGO files, which he/she had obtained during ergometry. The matching was done in up to six iterations (one for each parameter) for each patient, one patient after the other. In every iteration, for each parameter one iteration step was done, which consisted of 5 sub-steps (see Figure 1). For each iteration, the one parameter, which resulted in the minimum number of remaining IDs, was chosen. If more than one ID remained, only these dates and values of the ERGO tables, which were related to the remaining IDs, were used for the next iteration. Already chosen parameters were omitted in further iterations. Within each iteration step, we applied a criterion for testing equality (see Formula 1), and marked all identical date-value pairs. Then, we counted the number of these exact matches for every ID of the ERGO tables and kept the IDs with the maximum number.

The iterations were continued, until finally only one single ID remained, which was then chosen to be the matched ID for the current patient. However, if in the end more than one rehabilitation patient was linked to the same ID, the ID was assigned to a patient, if she/he alone had the highest total number of exact matches during the last iteration.

$$abs(date_{value1} - date_{value2}) + abs(value_1 - value_2) == 0 \qquad (1)$$

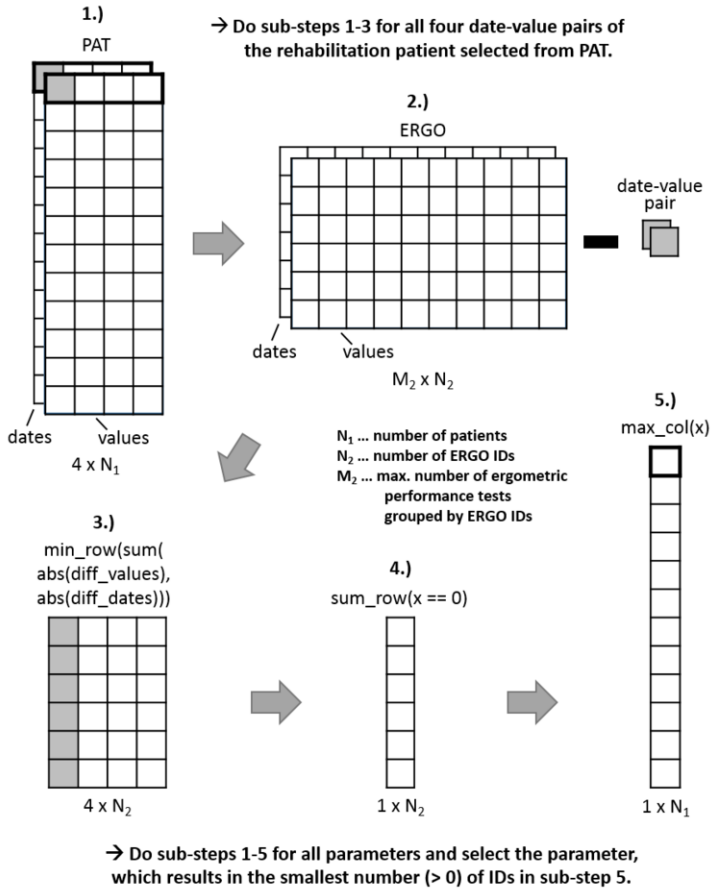

Figure 1. Depiction of one iteration step, which is run several times during an iteration and contains 5 iteration sub-steps. For every iteration step, the date tables of PAT and ERGO are used along with the value tables of PAT and ERGO for the currently evaluated parameter. In the PAT table up to four entries are available for each patient. In sub-step 1, a patient is selected from the PAT tables and one of her/his four date-value pairs is selected. In sub-step 2, the ERGO tables contain dates and values in their columns and each of their rows represents a pseudonymised ID. The date-value pair selected in sub-step 1 is now subtracted from all the values and dates of the ERGO tables. In this way, values that are identical will result in zeros. In sub-step 3, the minimum of the differences from sub-step 2 is calculated for each row and stored (in the same column as in the PAT tables and in the same row as in the ERGO tables). Sub-steps 1-3 are done for the up to four date-value pairs of the selected patient. In sub-step 4, the entries equal to 0 (= exact matches) are counted for each row of the table resulting from sub-step 3 and stored to another table. Now, for the patient selected in sub-step 1, this table shows in each row the number of exact matches between her/his date-value pairs and the date-value pairs of this row from the ERGO tables. In sub-step 5, the IDs from the ERGO tables' rows, which have the maximum number of exact matches, are stored for the selected patient. Sub-steps 1-5 are done for each parameter to finally choose the one, which results in the fewest IDs in sub-step 5.

2.2. Mapping of free text classifications

At ZARG, cardiac rehabilitation patients can participate in two phases of rehabilitation, which are denoted “phase 2” and “phase 3”. At the start and at the end of each phase, a performance test is conducted to track the patient’s performance. While the reason of a performance test can easily be obtained from the PAT file due to the various sheets as “start of phase 2”, “end of phase 2”, “start of phase 3” and “end of phase 3”, the ERGO files merely contain a field with free text entries (“ReasonForStudy”).

Using the unambiguous matching results from the implementation of the record linkage algorithm, we retrieved the entries of the “ReasonForStudy” fields from the ERGO files: First, for the matched IDs, we arranged all their date and “ReasonForStudy” values in a table. Then, we extracted all “ReasonForStudy” entries for the four types of reasons by using the respective dates from the PAT file together with the matched IDs.

3. Results

3.1. Record linkage algorithm

With our record linkage algorithm, we initially obtained 761 matches for the full PAT file, which equals to 49.5%. Removing all patients, which had at least one date value outside the overlapping date range of PAT file and ERGO files, the matching rate was 74.5%. Detailed results can be found in Table 2. For further analyses, the matches of the overlapping date range were used.

Table 2. Results of applying our record linkage algorithm. A “matched patient” is a patient of the PAT file, which can be unambiguously linked to a single pseudonymised ID of the ERGO files. For the column “PAT file until 13.06.2017” all patients with values after 13.06.2017 (= date of the last performance test of the ERGO files) were omitted.

Property	Full PAT file	PAT file until 13.06.2017
Number of patients	1,538	877
Matched patients	761 (49.5%)	653 (74.5%)
IDs matched to more than one patient (thus rejected)	206 (13.4%)	129 (14.7%)
Patients without a matching ID	571 (37.1%)	95 (10.8%)

3.2. Mapping of free text classifications

As described in section 2.2, the free text entries of the ERGO files for the reason of the performance test could be obtained from the “ReasonForStudy” field. Thus, for the overlapping time range, we collected all these free text entries for each of the four reasons. Table 3 shows the obtained free text entries for each of the four performance test reasons together with the number of their occurrence.

For “start of phase 2” 167 free text entries were obtained through the matching IDs. 78.4% of these entries contained the expected string “Erstuntersuchung Phase II”. 15.6% of the records contained an empty string. Thus, only very few unrelated entries remained.

Table 3. Available free text entries from the ERGO files, which could be unambiguously matched to PAT entries for the respective performance test reasons. There can be four different reasons, relating to the current stage of the cardiac rehabilitation program ("start of phase 2", "end of phase 2", "start of phase 3", "end of phase 3"). Only entries of the PAT file, which were within the overlapping time range of both data sources (11.02.2013 – 13.06.2017) were considered for the matching.

Start of phase 2 (167 free text entries obtained from the ERGO files)	
Free text entry from the matched ERGO files	**Number of occurrences**
"Erstuntersuchung Phase II"	131 (78.4%)
" " (empty string)	26 (15.6%)
"Erstuntersuchung Phase III"	3 (1.8%)
"Abschlußuntersuchung Phase II"	1 (0.6%)
"Abschlußuntersuchung Phase III"	1 (0.6%)
"Anfangsuntersuchung ProHeart"	1 (0.6%)
"CAVE!! Hr. [name] [birthdate] Erstuntersuchu" (sic!)	1 (0.6%)
"EU II"	1 (0.6%)
"Pro-Heart 3"	1 (0.6%)
"ZU Proheart"	1 (0.6%)
End of phase 2 (174 free text entries obtained from the ERGO files)	
Free text entry from the matched ERGO files	**Number of occurrences**
"Abschlußuntersuchung Phase II"	153 (87.9%)
" " (empty string)	17 (9.8%)
"AU II"	1 (0.6%)
"Erstuntersuchung Phase II"	1 (0.6%)
"Proheart ZU"	1 (0.6%)
"Zwischenuntersuchung Phase III"	1 (0.6%)
Start of phase 3 (184 free text entries obtained from the ERGO files)	
Free text entry from the matched ERGO files	**Number of occurrences**
"Erstuntersuchung Phase III"	111 (60.3%)
" " (empty string)	49 (26.6%)
"Zwischenuntersuchung Phase III"	6 (3.3%)
"Abschlußuntersuchung Phase III"	5 (2.7%)
"Erstuntersuchung Phase II"	4 (2.2%)
"Pro-Heart 2"	2 (1.1%)
"Abschlußuntersuchung Phase II"	1 (0.5%)
"Anfangsuntersuchung ProHeart"	1 (0.5%)
"EU Phase III"	1 (0.5%)
"Eingangsuntersuchung Phase III"	1 (0.5%)
"Pro Heart"	1 (0.5%)
"Pro Heart ZU"	1 (0.5%)
"Rehaabbruch"	1 (0.5%)
End of phase 3 (194 free text entries obtained from the ERGO files)	
Free text entry from the matched ERGO files	**Number of occurrences**
"Abschlußuntersuchung Phase III"	149 (76.8%)
" " (empty string)	32 (16.5%)
"Pro Heart"	3 (1.5%)
"Erstuntersuchung Phase III"	2 (1.0%)
"10/10/1min"	1 (0.5%)
"Abschlußuntersuchung Phase III Verl."	1 (0.5%)
"Anschlussuntersuchung Phase III"	1 (0.5%)
"Kontrolluntersuchung"	1 (0.5%)
"Pro Heart / Herzverband"	1 (0.5%)
"Rehaabbruch Phase III"	1 (0.5%)
"Vorzeitiger Rehaabbruch/AU PH3"	1 (0.5%)
"Zwischenuntersuchung Phase III"	1 (0.5%)

For “end of phase 2” 174 free text entries were obtained through the matching IDs. 87.9% contained the expected string “Abschlußuntersuchung Phase II”. Only 9.8% of the entries contained an empty string and the total number of remaining unrelated values was 2.3%. Thus, “end of phase 2” showed the highest rate of accurate entries along with the fewest empty strings and the lowest number of unrelated entries. For “start of phase 3” 184 free text entries were obtained. Only 60.3% of the performance tests were tagged with the expected entry “Erstuntersuchung Phase III”. With 26.6%, more than a quarter of the performance tests was annotated with an empty string. Also, the number of unrelated values was the highest in comparison to the other phases, totaling in 13%. The final reason “end of phase 3” showed similar characteristics like “start of phase 2”. Of 194 obtained free text entries, there were 76.8% of expected entries containing the string “Abschlußuntersuchung Phase III” and 16.5% empty strings. The number of unrelated entries was 6.7%.

While “start of phase 2” and “end of phase 3” had a similar rate of accurate entries, the rate of “start of phase 3” was lower and the rate of accurate entries of “end of phase 2” was comparably higher than for the other reasons of the performance tests.

4. Discussion

Looking at the outcome of this study, the applied time series record linkage algorithm achieved a matching rate of 74.5% and the observed free text entries were in accordance with our expectations for up to 87.9% of the entries. However, at this time, we had no gold standard for evaluating the accuracy of our matches. For more reliable analyses of the resulting combined dataset, the datasets should be linked by patient pseudonyms.

Another issue were the different date ranges of the two data sources. While the PAT file contained records ranging from 2013 to 2018, the ERGO files contained records ranging from 2004 to 2017 only. Obviously, no matches outside the overlapping date range were possible and considering the full date range, only half of the patients (49.5%) from the PAT file could be unambiguously matched to their IDs of the ERGO files. Looking at the overlapping date range, 74.5% of the patients could be matched.

For our matching approach, we assumed that no patient had more than one ID in the ERGO files and allowed only one single linkage between PAT file patients and ERGO file IDs. Thus, if a patient would have had two IDs in the ERGO files, one “correct” linkage would have been dismissed.

Even if only patients of the overlapping date range were considered for this study, the gathered knowledge could still be used for the full date range of the datasets. There were up to 87.9% of correctly entered free text entries, which gives the reassurance that these entries are quite reliable.

The proposed record linkage algorithm can be used to combine de-identified datasets to one comprehensive, de-identified dataset, which could be the basis for further insights. However, it is not possible to identify single patients or to recreate personal information.

Formula 1 gives the used criterion for testing equality. It was chosen, because our implemented routine transformed all parameter time series to date-value pairs in integer format for easier handling. For identical values at the same date, this criterion was logically true, which allowed to count these entries as exact matches. For allowing some distance between two values instead of only counting exact matches, adaptations would be needed: The values’ dates would separately need to be checked for an exact match,

while the values themselves would be allowed to diverge within some boundaries (e.g. ± 5 bpm for the maximum heart rate).

The matching results show, that on the one hand, the proposed matching algorithm was suitable for the given scenario, as unrelated free text entries were very rare. On the other hand, the free text entries showed to be very accurate.

5. Conclusion

For the given scenario with two data sources containing identical entries for some of their parameters, our iterative, distance-based time series record linkage algorithm achieved a matching rate of 74.5%. Furthermore, the free text annotation entries in the ERGO files were in accordance with our expectations for up to 87.9% of the entries, which showed that inclusion of the PAT file will be unnecessary for our future analyses of this dataset.

6. Conflict of Interest

To the authors' knowledge, no conflicts of interest were given.

7. Acknowledgement

This work was partly funded by the Austrian Research Promotion Agency (FFG) as part of the project EPICURE under grant agreement 14270859.

References

[1] D. Hayn *et al.*, "IT Infrastructure for Merging Data from Different Clinical Trials and Across Independent Research Networks," (in eng), *Stud Health Technol Inform,* vol. 228, pp. 287-91, 2016.

[2] S. Dusetzina, S. Tyree, A. Meyer, A. Meyer, L. Green, and W. Carpenter, "Linking Data for Health Services Research: A Framework and Instructional Guide.," University of North Carolina at Chapel Hill, Rockville, MD, 2014, vol. AHRQ Publication.

[3] G. Chassang, "The impact of the EU general data protection regulation on scientific research," (in eng), *Ecancermedicalscience,* vol. 11, p. 709, 2017.

[4] R. Nosowsky and T. J. Giordano, "The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research," (in eng), *Annu Rev Med,* vol. 57, pp. 575-90, 2006.

[5] K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity," (in eng), *J Am Med Inform Assoc,* vol. 15, no. 5, pp. 627-37, 2008 Sep-Oct 2008.

[6] A. Sayers, Y. Ben-Shlomo, A. W. Blom, and F. Steele, "Probabilistic record linkage," (in eng), *Int J Epidemiol,* vol. 45, no. 3, pp. 954-64, 06 2016.

[7] G. P. Oliveira, A. L. Bierrenbach, K. R. Camargo, C. M. Coeli, and R. S. Pinheiro, "Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis," (in eng|por), *Rev Saude Publica,* vol. 50, p. 49, Aug 2016.

[8] Y. Zhu, Y. Matsuyama, Y. Ohashi, and S. Setoguchi, "When to conduct probabilistic linkage vs. deterministic linkage? A simulation study," (in eng), *J Biomed Inform,* vol. 56, pp. 80-6, Aug 2015.

[9] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association,* vol. 64, no. 328, pp. 1183--1210, 1969.

[10] J. Nin and V. Torra, "Distance Based Re-identification for Time Series, Analysis of Distances.," in *Privacy in Statistical Databases. PSD 2006. Lecture Notes in Computer Science*, vol. 4302, J. Domingo-Ferrer and L. Franconi, Eds. Berlin, Heidelberg: Springer, 2006.