MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-562

Deterministic Record Linkage versus Similarity Functions: a Study in Health Databases from Brazil

Kátia Mitiko Firmino Suzuki[a], Carlos Humberto Porto Filho[a], Luís Fernando Cozin[a], Lucas Calabrez Pereyra[b], Paulo Mazzoncini de Azevedo Marques[b]

[a] *School of Medicine of Ribeirao Preto (FMRP), University of Sao Paulo (USP), Brazil*
[b] *Medical Center at the School of Medicine of Ribeirao Preto (HCFMRP), Brazil*

Abstract

The record linkage is a strategy that allows linking different databases of information from patient records. Adopting the deterministic method and similarity functions (Dice, Jaro, Jaro-Winkler and Levenshtein) for the integration of heterogeneous databases aimed at different levels of health care Brazilian (primary, secondary and tertiary). The sensitivity of deterministic method was 54.5% (95% CI: 50.4 to 58.5). The best result obtained with the dissent of only one variable (mother's name) was 80.6% (95% CI: 77.2 to 83.6) and the best result obtained using the similarity function Jaro-Winkler was 91.8% (95% CI: 89.4 to 93.9). The deterministic method has high specificity but sensitivity can be reduced by the existence of spellings and typing errors in the databases. Thus, the step-by-step approach where there was disagreement in at least one of the relationship variable can increase the sensitivity of the method and the use of similarity functions.

Keywords:
Information systems, database record linkage, deterministic record linkage, similarity function.

Introduction

The relationship database process has been widely used to combine information from individuals or entities from varied sources, using record linkage [1].

Starting from the basic idea that database relationships can be formalized, the term "record linkage" is defined as the process of comparing two or more records which contain identifying information to determine whether these records refer to the same entity [2]. Although the concept may seem simple, there are many interesting and challenging technical problems, due to the lack of unique identifiers in some databases, the definition of the criteria for selecting the appropriate variables to perform the binding, the existence of duplicate records and the occurrence of misspellings in string variables that must be resolved to permit linking large-scale data.

There are two main strategies to accomplish the record linkage: the deterministic and the probabilistic. The *deterministic record linkage* (DRL) strategy uses a unique identifier or identifiers common to both databases, uniquely identifying a record and ranking it as a peer or non-peer. It is simple to understand and implement. The *probabilistic record linkage* (PRL) strategy is based on statistical theory developed by Fellegi and Sunter [3] and it is suitable when the database does not contain a unique identifier common to the databases under examination. Additionally, the combined record of the classification can vary between full agreement (exact) and total disagreement, spanning various intermediate levels of agreement [4].

In the deterministic method, the challenge is to create a model that performs the appropriate comparison classifying records as equal and different. To accomplish this, relationship variables must be chosen with care and discretion. The best comparison model is one that relates the maximum number of true pairs (specificity) with the smallest number of erroneous pairs (sensitivity). However, if two non-equal records are related or paired, this fact is deemed a "false positive." Similarly, two identical records not paired or related are deemed a "false negative" [5].

In Brazilian healthcare databases, social security numbers (SSN) or “Cadastro de Pessoas Físicas” (CPF) use as a unique identifier is vitually non-existent. In fact, no other identification code exists in the national health records of patients that enables the unique identification of individuals. Thus, the use of relationship databases is presented as a viable alternative to integrate the various databases on the network of health services (primary, secondary and tertiary) in municipal areas, state and federal. The integration of databases, within this context, allows monitoring cohort studies, creation of health history, enabling improvement of information quality and consistency, preparation of records for studying diseases, and following cohorts to determine vital status of the individual stories and genealogical study or historical [6-9].

The aim of this study was to investigate the performance of the deterministic method as a strategy for forming relationships between the databases of the municipal-level state primary and tertiary care level in exact form. It will also perform step-by-step agreement on three relationship variables identified between the study databases, in addition to applying various similarity comparison metrics: *Dice* [10], *Jaro* [11], *Jaro-Winkler* [12], and *Levenshtein* [13].

Materials and Methods

To carry out the proposed study, we selected a random sample of the municipal healthcare database, consisting of 1,100 patient records residing in the city of Ribeirao Preto / SP, Brazil who attended the health service in primary level during the years January 2006 to August 2008. The other database came

from the tertiary care records of the Medical Center at the School of Medicine of Ribeirao Preto (HCFMRP) University of Sao Paulo – HCFMRP-USP, and consisted of all records that met the same requirement as to the place of residence of the patient, yielding 375,370 records.

The process of analyzing the databases identified 27 variables common variables, with only 12 variables having more than 80% of the rows filled. These were classified as relationship variables: patient's name, mother's name, date of birth and gender.

The data sets were exported to the management database system MySQL©. The study was approved by the Research Ethics HCFMRP according to the process nº 4635/2010 and CSE-Cuiaba in its 57th meeting held on March 13, 2007.

The DRL was performed using the software tool *REcord Linkage At Istat* - RELAIS 2.0, developed by the Italian National Institute of Statistics (*Istituto Nazionale li Statistica*), and the similarity metrics as: *Dice*, *Jaro, Jaro-Winkler* e *Levenshtein*.

Deterministic Record Linkage Strategy

DRL strategy employing similarity metrics was employed to compare all records from the primary and HCFMRP databases using relationship variables (patient's name, mother's name, gender and date of birth). With the deterministic method, step-by-step comparison is widely used. The first step is to combine all variable relationships removing all pairs formed; the next step allows one or more variables disagree to increase the number of pairs found [14-17].

Table 1 - Number of Discordant Couples in each strategy, and percentage classification error.

Strategy	Matched	Number of Pairs (disagree in a variable)	Error Rate (%)	Classification Error
DRL (Exact)	334	0	0.00	---
Total		0	0.00	
DRL (N - S)	335	1	0.30	Divergence in gender
Total		**1**	**0.30**	
DRL (N - D)	343	3	0.87	Divergence on the day of birth
		2	0.58	Divergence on the month of birth
		3	0.87	Divergence on the year of birth
		1	0,29	Unfilled
Total		**9**	**2.62**	
DRL (N - M)	495	4	0.61	Invalid character
		4	0.81	Divergence in surname
		65	13.21	Misspelling
		5	1.02	Error in midlle and last names
		1	0.20	Cases of twins
		11	2.24	Incomplete name
		1	0.20	Use of "Ignorada" (“Unknown”)
		43	8.74	Use of abbreviation
		27	5.49	Using of married name
Total		**161**	**32.53**	
DRL (N - N)	411	19	4.81	Misspelling
		3	0.76	Incomplete name
		1	0.25	Cases of twins
		7	1.,77	Incomplete name
		1	0.25	Double register on the HCFMRP database
		1	0.25	Use of “Óbito” ("Death") as part of the name
		28	7.09	Use of "RN" (“newborns”) on the basis HCFMRP-USP
		13	3.29	Using of married name
Total		**73**	**18.73**	

The following steps were performed comparing only records not matched in the previous step, using the concordance in three variables and similarity metrics for the following variables: patient's name and mother's name. The strategy usied only three variables for comparison agreement, classified as N-1 that resulted in four different combinations of variables: N-S (disagree on gender), N-D (disagree on date of birth), N-M (disagree on mother's name) e N-N (disagree on patient's name).

A strategy based on similarity metrics aims to measure the "similarity" between two fields of data type string. The similarity metrics between strings have been widely used in several areas of study [18]. When applied to a particular word, the similarity values may vary in the range [0,1], where 1 ("one") represents equal words. These scales are adopted by most of the scientific community, although there are authors who use different scales. In the RELAIS software, the adopted scale is the interval [0,1] and similarity metrics have been implemented using the *Java© String Metrics* package. The main similarity metrics used were: *Dice, Jaro, Jaro-Winkler,* and *Levenshtein* with threshold values of 0.9 and 0.8 in the fields patient's name and mother's name. For fields birthdate and gender, the rule adopted was to compare for equality without a similarity metric.

The pairs formed in each stage of the deterministic method were classified as true or false, according to the comparison's gold standard. To evaluate the accuracy of the method, we used the sensitivity, specificity, positive and negative predictive value. The gold standard was obtained by manual review of the sample of 1,100 records. In this sample there are 617 true pairs and 483 false pairs.

Results

Regarding the quality of information, it should be noted that 100% of the records in both databases had gender populated; date of birth and mother's name had 99% population. The HCFMRP database had its gender variable encoded with the following field values: F = Female, M = Male, D = Unknown.

Table 1 demonstrates the error rate for each variable and its classification. The most frequent errors observed were divergent information, typographical errors, use of abbreviation, change of surname and name, the use of the word "RN" for the records of newborns in the HCFMRP database. The gender variable yielded an error rate of 0.30%, 2.62% date of birth, patient's name 18.73% and 32.53% mother's name.

Figure 1 (columns 2-6) shows the results of the techniques of the deterministic exact method (DRL), deterministic disagreement with a variable: N-S (disagree on gender), N-D (disagree on date of birth), N-M (disagree on mother's name) e N-N (disagree on patient's name). It can be observed that the number of pairs found when compared to the gold standard is low 334 pairs true ("Pares Verdadeiros" or blue bars on the graph) for the comparison of four variables relationship (DRL). That number increased when using the disagreement in a variable, but there was no significant increase for the discrepancy of gender and date of birth (335 and 343 true pairs), because of its low error rate (0.30% and 2.62%).

The occurrence of false pairs ("Pares Falsos" or red bars on the graph) disagreement was found for the variable name of the patient (28 pairs false) and using the similarity functions *Levenshtein, Jaro,* and *Jaro-Winkler* (see Figure 1, columns 5, 7, 10, 11,12 and 14). The advantage of applying similarity functions in DRL is that false negatives due to spelling errors, changes of surname, and use of abbreviations may be minimized by increasing the amount of true pairs (Fig. 1, columns 7-14). Moreover, there is an increased possibility of false positives pairs, resulting in increased sensitivity with decreasing specificity.

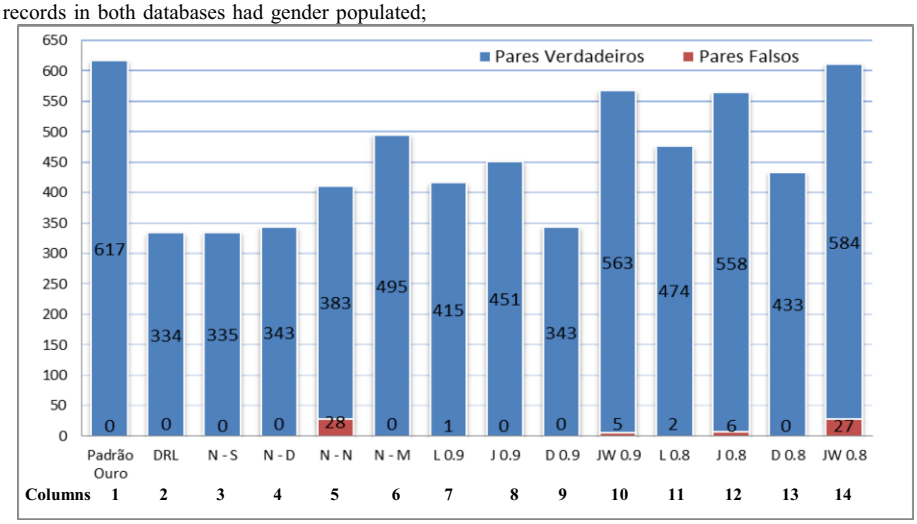

Figure 1 – Performance methods: exact deterministic, deterministic disagreement with a variable relationship. (S=Gender, N= date of birth, N= name e M= mother's name) and deterministic method with function Similarity (L= Levenshtein, D=Dice, J=Jaro, and JW=Jaro-Winkler) with threshold value 0.9 and 0.8 on the gold standard.

In the comparison of accuracy of the methods in Tables 2 and 3 show the sensitivity and specificity of each technique deterministic method and negative and positive predictive values. The results suggest a low sensitivity for the methods: DRL (54.49%), DRL NS (54.65%), DRL ND (55.95%), DRL NN

(62.48%), but show a high sensitivity to the DRL NM (80.59%). The deterministic record linkage method is characterized by presenting high values of specificity, which can be verified by the values shown in Table 2.

Table 2 – Accuracy of deterministic record linkage method.

	DRL		DRL (N-D)		DRL (N-M)		DRL (N-N)		DRL (N-S)	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Sensibility	54.13	50.1 - 58.1	55.59	51.6 - 59.6	80.23	76.9 - 83.3	62.07	58.1 - 65.9	54.29	50.3 - 58.3
Specificity	100	99.2 - 100	100	99.2 - 100	100	99.2 – 100	94.2	91.7 - 96.1	100	99.2 – 100
VPN	63.1	59.5 - 66.5	63.8	60.3 - 67.2	79.8	76.4 - 83.0	66.0	62.4 - 69.6	63.1	59.6 - 66.6

Labels: CI – Confidence Interval; VPN – Negative Predictive Value; DRL – Deterministic Record Linkage; (N-D) variable - Date of Birth; (N-M) variable - Mother's name; (N-N) variable - Name; (N – S) – variable - Gender.

When evaluating the accuracy of deterministic strategy with the use of similarity functions, it was observed that its use increases the sensitivity of matching when compared to previous strategies (Table 3). The similarity function that showed higher sensitivity was Jaro-Winkler (91.8%), followed by Jaro (73.6%), Levenshtein and Dice (both with 67.7%) to the threshold value 0.9 and remained the same order as the measure of sensitivity to a threshold of 0.8. Ten patients are known to have not been matched by any of the functions, due to a record disagreement as to gender and nine records with varying birth dates. For these variables we adopted the criterion of equality.

Table 3 – Accuracy of deterministic methods with similarity function.

	DICE		LEVENSHTEIN		JARO		JARO-WINKLER	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Threshold value 0.9								
Sensibility	55.6	51.6 - 59.6	67.3	63.4 - 71.0	73.1	69.4 - 76.6	91.3	88.7 - 93.4
Specificity	100	99.2 - 100.0	99.8	98.9 - 100.0	99.6	98.5 - 99.9	99.0	97.6 - 99.7
VPN	63.8	60.3 - 67.2	70.5	66.9 - 73.9	74.3	70.8 - 77.7	89.8	87.0 - 92.3
Threshold value 0.8								
Sensibility	70.0	66.2 - 73.6	76.8	73.3 - 80.1	90.4	87.8 - 92.6	94.7	92.6 - 96.3
Specificity	100	99.2 - 100.0	99.4	98.2 - 99.9	98.8	97.3 - 99.5	93.4	90.8 - 95.4
VPN	72.3	68.7 - 75.7	77.0	73.5 - 80.3	89.0	86.0 - 91.5	93.2	90.6 - 95.3

Discussion

Although there is a significant increase in the use of techniques creating relationships records in Brazil, there are few studies that use data relationships for outpatient and hospital records [19,20], especially for database integration of health services in primary and secondary level to tertiary level. The record linkage applied to these health databases enables following patients through the registered information in primary and secondary and tertiary data sources.

The main strategies employed to accomplish database record linkage are deterministic and probabilistic. This study utilizes the deterministic technique and its variations to assess the accuracy of methods creating record linkage of the study's databases, comparing the pairs classified as true or false with a gold standard, in terms of sensibility, specificity and negative predictive value.

The DRL technique was applied with the following approaches: deterministic exact agreement with four variables, deterministic approach in exact step-by-step, with concurrence of three variables, and finally with the deterministic similarity metrics. The exact DRL has been used in other studies and has proved easy to use and offered good results, especially in cases where manual inspection of pairs formed was feasible [21]. The step-by-step process examining discordance in at least one variable method is also a very popular technique and was easy to apply; however, due to the error rate in existing databases used in this work, an increase in the number of pairs found was not significant for the variables gender and date of birth. The disagreement observed in the variables of mother's name and patient's name increased assay sensitivity and specificity decreased, establishing a strong relationship with the power of discrimination and the rate of spelling errors, error in surname and use of abbreviations of these variables.

Existing databases with little or no errors (error rate) and in which the variable of relationship has high discrimination power, the DRL strategy yields good performance and results. Overall, the variables formed by strings like patient's name and mother's name have a large number of possible values (categories) and high discriminating power, but are also likely to create a greater number of errors. Thus, the use of similarity metrics can reduce errors and make the variables most appropriate to use them in the record linkage [22,23].

The results presented here indicate that the use of the method of record linkage accompanied by the use of similarity metrics is a good option for integrating databases with high volume and low quality, especially with string variables. These results still need to be compared with applications that use probabilistic methods, since articles such as Tromp et al. [22] suggest that the use of the such methods are the best option for linkage procedures due to limited sensitivity displayed by deterministic methods.

Conclusion

The use of the deterministic technique with similarity functions proved to be a viable alternative to record linkage using

the patient's name and mother's name as variables, which often have spelling errors, change of name, or even incomplete data. The Jaro-Winkler algorithm showed the highest sensitivity, but is susceptible to loss of specificity. Still, with the results obtained, it can be concluded that the strategy of linkage with similarity functions, especially the Jaro-Winkler algorithm, can be employed with good performance in studies relating multiple databases.

Acknowledgments

This work was partially supported by The National Council for Scientic and Technological Development (CNPq) - grants 409493/2006-6 and 573714/2008-8 (INCT/INCoD).

References

[1] Gomatam S, Carter R. A computerized stepwise deterministic strategy for linkage. Technical Report. 1999.

[2] Howe GR. Use of computerized record linkage in cohort studies. Epidemiologic Reviews. 1998, Vol. 20, 1, pp. 112-21.

[3] Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc, 1969;64(328): 1183-1210.

[4] Christen P, Churches T. Secure health data linkage and geocoding: current approaches and research directions. In: National E-health provacy and security Symposium. 2006.

[5] Gill LE. Methods for automatic record matching and linkin in their use innational statistics. Office for National Statistics. 2001, Vol. 25.

[6] Smith ME. Record - Keeping and data preparation practives to facilite record linkage. In:Kilss, B.; Alvey, W. 1985, pp. 321-26. Available at: <http://www.fcsm.gov/working-papers/1367.pdf>.

[7] Goldacre MJ. Implications of record linkage for health services management. [ed.] J. A. In: Balwin, E. D. Acheson e W. J. Graham. Textbook of medical record linkage. 1987, pp. 305-317.

[8] Gill LE. e Baldwin JA. Methods and technology of record linkage: some practical considerations. In: Acheson, E. D.; Graham, W. J. 1987, pp. 39-54.

[9] Jensen KP. Probabilistc methodology for record linkage determining robustness of weights. 2004. A project submitted to the faculty of Brigham Young University in partial fulfillment of the requirements for the degree of Master of Science.

[10]Kondrak G, Marcu D, Knight K. Cognates can improve Statistical Translation Models. Proceedings of HLT-NAACL 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Lingistics. 2003, pp. 46-48.

[11]Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association. 1989, Vol. 84, pp. 414-420.

[12]Winkler WE. The state of record linkage and current research problems. Statistics of Income Division, Internal Revenue Service Publication R99/04. 1999.

[13]Levenshtein VL. Binary codes capacle of correcting spurious insertions and deletions of ones. Problemy Peredachi Informatsii. 1965, Vol. 1, pp. 12-25.

[14]Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat Med 2002;21:1485e96.

[15]Haas JS, Brandenburg JA, Udvarhelyi IS, Epstein AM. Creating a comprehensive database to evaluate health coverage for pregnant women: the completeness and validity of a computerized linkage algorithm. Med Care 1994;32:1053e7.

[16]Li B, Quan H, Fong A, Lu M. Assessing record linkage between health care and Vital Statistics databases using deterministic methods. BMC Health Serv Res 2006;6:48.

[17]Oberaigner W. Errors in survival rates caused by routinely used deterministic record linkage methods. Methods Inf Med 2007;46(4):420e4.

[18]Chávez E, Navarro G, Baeza-Yates R, MAarroquín JL. Searching in metric spaces. ACM Computing Surveys. v.33, n.3, pp. 273-321, 2001.

[19]Silva JPL, Travassos C, Vasconcellos MM, Campos LM. Revisão sistemática sobre encadeamento ou linkage de bases de dados secundários para uso em pesquisa em saúde no Brasil. Cadernos Saúde Coletiva, v. 14, n. 2, p. 197-224, 2006.

[20]Magalhães VCL, Costa MCE, Pinheiro RS. Perfil do atendimento no SUS às mulheres com câncer de mama atendidas na cidade do Rio de Janeiro: relacionando os sistemas de informações SIH e APAC-SIA. Cadernos Saúde Coletiva, v. 14, n. 2, p. 375-398, 2006.

[21]Bronhara BR, Conde WL, Liciardi DC, França-Junior I. Vinculação Deterministica de Banco de Dados sobre Mortalidade por AIDS. Revista Brasileira de Epidemiologia. v. 11, 4, pp. 709-13, 2008

[22]Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. Journal of Clinical Epidemiology 2011. v. 64, pp.565-572.

[23]Suzuki KMF, Cozin LF, Azevedo-Marques, PM. Applying different deterministic approaches for health electronic databases linkage. In: Conferencia Latinoamericana de Informatica em Saúde, 2011, Guadalajara. INFOLAC'2011, 2011.

Address for correspondence

Kátia Mitiko Firmino Suzuki - Ph.D
School of Medicine of Ribeirao Preto, University of Sao Paulo
Av. dos Bandeirantes 3900, Campus da USP
Monte Alegre. CEP 14049-900. Rib. Preto – SP/Brasil
Phone: + 55 16 3602 0605
Email:kmsuzuki@fmrp.usp.br