Probabilistic Record Linkage System for Maternal Mortality

Saw Yu Yu Wai, Thin Thin Htike
University of Computer Studies, Pathein
sawyuyuwai88@gmail.com

Abstract

Computerized record linkage is increasing used in health information systems. Medical record linkage expresses the concept of collecting health care records into commutative personal file, starting with birth and ending with death. Probabilistic linkage methods produce more accurate, dynamic, and robust matching results than rule-based approaches, particularly when matching patient records that lack unique identifiers. Theoretically, the relative frequency of specific data elements can enhance the probabilistic record linkage method, including minimizing the false-positive or false-negative matches. This system estimates the maternal mortality ratio (MMR) using probabilistic record linkage method. This system uses three information systems on mortality data, live birth data and another hospital data, in year 2009. Finally, the maternal mortality ratio is calculated and information on declared maternal deaths was obtained. From these data, the Mortality Information System is probabilistically linked with the Live Birth Information System and Hospital Information System, with a multiple-step blocking strategy. For paired records, the diagnoses and hospital procedures brought together by the best known criteria for severe maternal morbidity were detailed. A total of 33 maternal deaths are recorded in 2009. The official and adjusted maternal mortality ratio is 9192.2(deaths per 100,000 live births). By correlating with data from mortality and hospital systems the most frequent age for maternal death was between 30 and 34.

1. Introduction

Probabilistic record linkage is commonly used in other country for health research. Record linkage is the methodology of finding a unified record from two or more records that are in different files and belong to the same entity [7]. Record linkage methods can be deterministic or probabilistic or a combination of both. Deterministic linkage is used when there is a unique identifier or if variables used for comparison are error-free and highly discriminatory, whereas probabilistic linkage takes into account the uncertainty that can exist in comparing variables used for comparison in both files. Sex, for example, induces a two fold partition of a file: males and females, and if records agree on sex, we cannot say with a high degree of confidence that they belong to the same person.

Maternal mortality is difficult to measure, even in developed countries with good systems for recording vital statistics, despite the low rate of underreporting deaths [1]. Errors in attributing the cause of death may occur, thus leading to underreporting of maternal mortality. It is even more complicated to obtain reliable estimates in developing countries, where the vital records generally have low coverage and there are also high rates of underreporting of specific causes of death [10].

The maternal mortality indicator that is most used today is the maternal mortality ratio (MMR), which is obtained as the quotient between the number of maternal deaths and the number of live births over a given period, multiplied by 100,000 [2]. The number of live births (LB) is given by Live Birth Information System. The main problem in calculating the MMR is the difficulty in identifying a maternal cause as a clearly recognized and recorded cause of death, especially in places where the vital record system does not exist or is faulty.

There are various methods for estimating maternal mortality [1, 2]. Among these are the reproductive Age Mortality Survey (RAMOS) and the Sisterhood method. These systems do not cover the whole country; the lack of complete counting and the low trustworthiness of the causes of death limit the continuous monitoring of maternal mortality.

In addition to the fact that these systems do not cover the whole country, the luck of complete counting and low trustworthiness of the cause of death limit the continuous monitoring of the maternal mortality. This limitation on integration between the different health databases has been overcome using a procedure of probabilistic record linkage to identify the same subjects in these different information sources.

2. Related Works

There has been a large body of work on record linkage. In this section, we brief present some of the research literature.

The classical probabilistic record linkage approach is developed by Fellegi and Sunter (1969) has been improved in recent years mainly through application of the expectation-maximization (EM) algorithm for better parameter estimation in record pair classification (Winkler 2000), and through the use of approximate string comparisons to calculate partial agreement weights when attribute values have typographical variations (Christen 2006, Winkler 2006) [9].

Linking or matching databases is becoming increasingly important in many data mining projects, as linked data can contain information that is not available otherwise, or that would be too expensive to collect manually. A main challenge when linking large databases is the classification of the compared record pairs into matches and non-matches. In traditional record linkage, classification thresholds have to be set either manually or using an EM-based approach. More recently developed classification methods are mainly based on supervised machine learning techniques [4] and thus require training data, which is often not available in real world situations or has to be prepared manually.

The California Automated Mortality Linkage System (CAMLIS), established in 1981 to facilitate the conduct of follow-up studies in the State of California, employs a combination of deterministic and probabilistic linkage decision criteria to perform the death clearance function [3]. The system was evaluated against four traditional death clearance procedures and the performance of each procedure measured in terms of measures of sensitivity and specificity.

We previously developed a deterministic record linkage algorithm demonstrating sensitivities approaching 90% while maintaining 100% specificity. Substantially better performance has been reported using probabilistic linkage techniques; however, such methods often incorporate human review into the process. To avoid human review, we employed an estimator function using the Expectation Maximization (EM) algorithm to establish a single true-link threshold [5].

Record linkage is the computation of the associations among records of multiple databases. It arises in contexts like the integration of such databases, online interactions and negotiations, and many others. The autonomous entities who wish to carry out the record matching computation are often reluctant to fully share their data. In such a framework where the entities are unwilling to share data with each other, the problem of carrying out the linkage computation without full data exchange has been called *private record linkage* [12]. Previous private record linkage techniques have made use of a third party.

3. Proposed System Design

This system is to describe the maternal mortality ratio (MMR) according to the mortality information system (SIM), in relation to data corresponding to these records in other systems (Hospital Information System (HIS) and Live Birth Information System (LB)) as shown in Figure 1. The theory utilized for linkages between the systems was divided into three sequential stages: database standardization; linkage, subdivided into blocking and matching; and combination of the files and manual review.

After standardization of data, all records of women in the age group from 18 to 49 whose basic cause of death was in category "Os" were initially selected. Categories "O96" and "O97" were excluded: these refer, respectively, to "death due to any obstetric cause occurring between 45 days after the delivery". From LB, all the records of live births were considered, while from SIH there was an initial selection of women between 18 to 49 years of age.

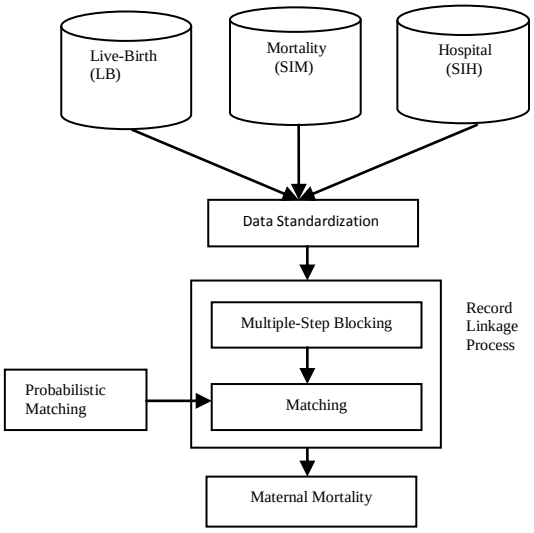

Figure1. Proposed System Design

The principal fields utilized for linking the SIM and LB data were the mother's name and age. Other fields utilized in manual selection for confirming the match were the data of death in SIM versus the date of birth in LB (the date in SIM should be the same as or subsequent to the date in LB) and the address when available.

For matching between SIM and SIH data, the principal fields were the name and date of birth and the auxiliary fields were the mother's date of death in SIM versus the date of discharge in SIH, and the age and address. It is emphasized that in the two linkages above, the principal fields were used as the references in the manual review and confirmation of true matches.

When more than one hospitalization record was found for the same person, they were evaluated manually and the reference record was the one corresponding to the date of death in SIM, or when they were not coincident, the closest one to this date. Thus, information on re-hospitalization of the same woman was maintained in the same database line (record).

3.1. Data Standardization

Data standardizing refers to the process of standardizing the information represented in certain fields with some special contents. This is used for information that can be stored in many different ways in different data sources and must be converted to a uniform representation before record linkage process starts. The basic ideas of standardization are (1) to replace the many spelling variations of commonly occurring words with standard spellings such as a fixed set of abbreviations or spelling and (2) to use certain key words that are found during standardization as hints for paring subroutines.

In our system, we firstly remove U and Daw from name. Secondly, we converts different format of date into uniform format (mm/dd/yy). Moreover, our system checks whether the format of NRC-number is correct or not. Table 1 shows example of data standardization.

Table 1.Example of data standardization

Original Data		Standardized Data	
Name	Date	Name	Date
U Kyaw Kyaw Tun	12-Jan-09	Kyaw Kyaw Tun	1.12.2009
Daw Myo Myo Aye	3/25/2009	Myo Myo Aye	3.25.2009
U Zaw Min Lwin	22,6,09	Zaw Min Lwin	6.22.2009
Thidar Cho	30-6-09	Thidar Cho	6.30.2009

3.2. Probabilistic Record Linkage

Probabilistic record linkage methods recommended over traditional deterministic methods (i.e. exact matching). Probabilistic record linkage uses information on greater number of matching variables, and allows for the amount of information provided by any (dis)agreement on matching variables. At the heart of probabilistic record linkage are *u* probabilities and *m* probabilities. *u* probabilities are the probabilities that a matching variable agrees given that the comparison pairs being examined is a non match (i.e., the probability that variables agree purely by chance among matches). *m* probabilities are the probabilities that a matching variable agrees given that the comparison pairs being examined is a match.

For example, *u* probability for matching variable "for month of birth" is about 1/12 =0.083.The *m* probability is less than1.0.The value of *m* probability is estimated (sometime iteratively) during the specification of the record linkage based upon prior information and the proportion of agreements among the comparison pairs accepted as links. In this example, assume the *m* probability was 9.5.These *m* and *u* probabilities are used to determine frequency ratios or (dis)agreement weights.

Each variable (field) has an agreement and disagreement weight associated with it. The agreement weight is *log (m/u).*The disagreement weight is *log ((*1-*m)/ (*1-*u)).* Logarithms are to the base two. Despite a sound theoretical basis for probabilistic record linkage, it is often regarded as a best method in comparison to exact linkage.

Table 2. Example of Agreement and Disagreement Frequency Ratios and Weights for Comparison by the Matching Variable 'day of birth'

Comparison Outcome	Proportion/Frequency among: True Link	Non-Link	Frequency Odds	Weight
Agreement	0.95 (m)	0.03 (u)	32/1 (m/u)	4.98 [ln(m/u)/ln(2)]
Disagreement	0.05 (1-m)	0.97 (1-u)	1/19 (1-m/1-u)	-4.28 [ln(1-m/1-u)/ln(2)]

The devisor, ln (2), transforms the natural algorithm to a base 2 algorithm

In this example, a comparison pair that agreed on month of birth would be assigned a weight of 3.51 and a comparison pair that disagreed on month of birth would be assigned a weight of -4.20 (shown in table 2). The setting of *m* and *u* probabilities and the corresponding weight is repeated for all matching variables, and possibly for all values of each or some of the matching variables. The total weight for a given comparison pair is simply the sum of the (dis)agreement weights for each matching variable. The total weight will be a large positive number if all/most matching variables agree or a large negative number if all/most matching variables disagree.

3.2.1. Blocking

The reliability and efficiency of matching is very dependent on the way in which the initial grouping or blocking is carried out [7]. Blocking is used to reduce the number of comparisons of record pairs by bringing potentially linkable record pairs together. Efficiency is improved by comparing records on two files. A good attribute variable for blocking should contain a large number of attribute

values that are fairly uniformly distributed and such an attribute must have a low probability of reporting error.

Blocking is defined as a partition of the file into mutually executive blocks. Blocking can reduce the number of comparisons between two files, but also reduces the sensitivity and increase the number of positive predictive value.

This system uses the multiple step blocking strategies for two linkage processes (SIM vs. LB and SIM vs. SIH). The blocking keys were:(1) phonetic code (Soundex) for the first and last name and the initials of the middle names; (2) phonetic code for the first and last names together; (3) phonetic code for the first name; (4) phonetic code for the last name; (5) same ages; (6) date of death in SIM the same as date of birth in LB and date of death in SIM is the same as the date of birth in HIS.

Soundex Code. The Soundex code has been widely used in medical record systems despite its disadvantages [13]. It is used principally, for the transformation of groups of consonants within names, to specific combinations of both vowels and consonants. The purpose of the Soundex code is to cluster together names that have similar sounds. The Soundex code of a name consists of one letter followed by three numbers. The letter is the first letter of the name. Disregarding the remaining vowels, as well as the letter W, Y and H, the number is assigned to the first three letters following the first letter according to Table 3. For example, the Soundex code for "John" is J500.

Table 3. Soundex Code Guide

Letter	Number	Letter	Number
B,F,P,V	1	C,G,J,K,Q, S,X,Z	2
D,T	3	L	4
M,N	5	R	6

3.2.2. Probabilistic Matching

The most rigorous mathematical framework to formalize the record linkage problem has been proposed by [Fellegi and Sunter, 1969] as an extension of the early work of [Newcombe and Kennedy, 1962]. In our system, we use the concepts defined in the theory of Fellegi and Sunter (FS).

Record linkage algorithms work on two sets (or files) of records denoted as A and B; we will use lowercase letters to indicate records belonging to each set, a ∈ A and b ∈ B. The available information regarding one record is denoted with α (a) and β (b), respectively.

The comparison record set $A \times B$ is partitioned into the two subsets.

$$M = \{(a,b) \in A \times B \mid a = b\} \quad (1)$$

$$U = \{(a,b) \in A \times B \mid a \neq b\} \quad (2)$$

$$m(\gamma) = p[\gamma_1, \gamma_2, ..., \gamma_i, ..., \gamma_k,]^T \quad (3)$$

Where γ =comparison vector which is a vector function of the record pairs $(a,b) \in A \times B$.

$$\gamma_i = \{0,1\}$$

From (1), (2) and (3),

$$m(r) = P(r \mid (a,b) \in M) = P(r \mid M) \quad (4)$$

$$u(r) = P(r \mid (a,b) \in U) = P(r \mid U) \quad (5)$$

In agreement case, γ^k tends to one and γ^k tends to zero. In disagreement case, γ^k tends to zero and γ^k tends to one.

$$w^k(\gamma_k) = \log m(\gamma^k) - \log u(\gamma^k) \quad (6)$$

$$w^k(\gamma_k) = \log_2 \frac{m(\gamma^k)}{u(\gamma^k)} \quad (7)$$

$$w(\gamma) = w^1 + w^2 + ... + w^k \quad (8)$$

If the field agrees, w_i is equal to ….

$$w_i = \log_2 (\frac{m_i}{u_i}) \quad (9)$$

In cases where two records disagree on a specified field, w_i is equal to ….

$$w_i = \log_2 (\frac{1-m_i}{2-u_i}) \quad (10)$$

For example, we use probabilistic matching to decide whether two records are match or not in Figure 2.

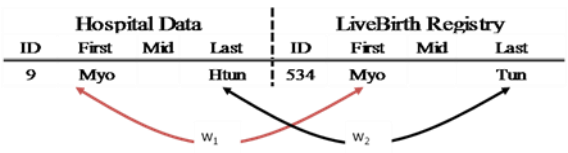

Figure2. Example for matching records

For agreement case, firstly we calculate m_i. m_i for first name is .98, or 98% of the time, if it's a correct match, the first names will agree and then u_i. u_i for Myo is .00001 is the probability of randomly getting two first names that are Myo as shown in Figure 2.

$$w_{i1} = \log_2(\frac{m_i}{u_i}) = \log_2(\frac{.98}{.00001}) = 16.58049$$

For disagreement case, m_i for first name is .96, or 96% of the time, if it's a correct match, the last names will agree. u_i for Htun is .00001 is the probability of randomly getting two last names that are Htun.

$$w_{i2} = \log_2(\frac{1-m_i}{1-u_i}) = \log_2(\frac{1-.96}{1-.00003}) = -4.64381$$

The composite weight, w_t is calculated for each pair of records using Eq.8. The sum of weight across all fields used in linkage is

$$w_t = \sum_{i=1}^{k} w_i$$

$$w_{it} = 16.58049 - 4.64381 = 11.93668$$

Larger w_t suggest a correct match. Smaller w_t suggest an incorrect match. Therefore, we suggest the two records are match.

4. System Implementation

In our system, we only use records from three information systems (SIM, LB and SIH) for Myanaung Township within year 2009. The data from SIM are available from Township Development Council in Myanung Township. The information from LB and SIH is available from Myanaung Township's Hospital. We use 500 death records which were death occurs during 2009. In 2009, the Hospital Information System recorded 550 persons in Myanaung Township. This system used with a sample of 400 live births in ordered to calculate the official MMR.

This system allows users to login. After login, user must standardize the different format of data. Later, the user must do the record linkage process. Probabilistic record linkage process is divided into multiple-steps blocking and matching. Finally, our system shows number of maternal death and the user can calculate maternal mortality ratio. The equation used for calculating MMR is

$$MMR = \frac{NumberofMaternalDeath}{NumberofLiveBirth} \times 100{,}000$$

This system display maternal death shown in Figure 3. In 2009, 33 maternal deaths were recorded as the basic cause in SIM as shown in Figure 3, and 359 Live-Birth were recorded, recorded in LB, thus resulting in an official MMR of 9192.2/100,000 LB. Among the 33 maternal death 72.7% of the case are of women aged less 35 years and 75% of these occurred in hospital. By correlating with data from mortality and hospital systems the most frequent age for maternal death was between 30 and 34. Percentage distribution of declared maternal death is shown in Table 4.

Table4. Percentage distribution of declared maternal death in Myanaung Tsp, 2009. N=33

age	N(number of maternal death)	%
≤19	0	0
20-24	5	15.1
25-29	7	21.2
30-34	12	36.6
35-39	6	18.1
≥40	3	9.0

No	Name	Sex	Age	Address
1	Su Su Aung	f	23	No(13), 12th St
2	Nu Nu Hlaing	f	30	Shwekaung St
3	Shwe Zin	f	34	Moemya Qt
4	Nu Nu Yee	f	23	Ywatthit St
5	Sein Sein	f	35	15 th St
6	Moe Moe Aung	f	28	Zay Thit St
7	Khin Thuzar	f	27	Kaungkyii St
8	Moe Moe Soe	f	28	Yadanar St
9	Nwe Ni Zaw	f	34	Lan Thit St

Figure 3. Example result for Maternal Death

5. Performance Evaluation

The performance of the system depends on true match,true non-match,false match , false non match and precision and recall. We can estimate the performance of the system by calculating precision and recall. Figure 4 shows the performance result of the system according to precision and recall. Probabilistic record linkage system increases the number of precision and recall.

In our system, we get 86% of precision and 100% of recall.

	Match	**Non-Match**
Linked	a (true positive)	b (false positive)
Unlinked	c (false negative)	d (true negative)

Positive Preductive Value(precision) = a/(a+c)
Sensitivity(recall) = a(a+b)

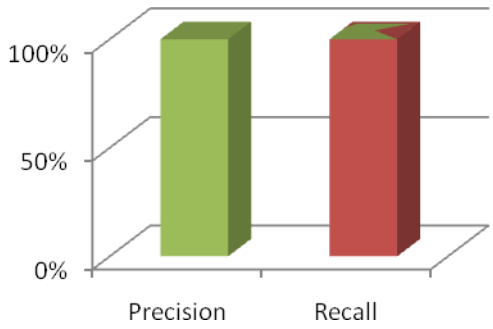

Figure 4. Performance result

6. Conclusion

Record linkage is an important technique in the development, production, analysis and evaluation of statistical data. It is an important tool for the creation of statistical data, particularly in relation to census taking, health research and in survey taking for social and economic statistics.

Probabilistic linkage methods produce more accurate, dynamic, and robust matching results than rule-based approaches, particularly when matching patient records that lack unique identifiers. Probabilistic linkage method uses multiple-steps blocking, and so it can reduce number of comparison between two files. Moreover, it can reduce sensitivity. Positive predictive value is increase using probabilistic method.

Probabilistic record linkage allows the assembling of information from different data sources. Probabilistic linkage is a feasible method to combine death and birth records. This system produces the possible birth-death records pairs to declared maternal death and estimate Maternal Mortality Ratio (MMR). This system can approximately estimate the MMR by calculating the probability of match weight and so it can easily determine the cause of death which commonly occurs in pregnancy within a year.

In 2009, 33 maternal deaths were recorded as the basic cause in SIM and 359 Live-Birth were recorded, recorded in LB, thus resulting in an official MMR of 9192.2/100,000 LB. Among the 33 maternal death 72.7% of the case are of women aged less 35 years and 75% of these occurred in hospital. Among the 33 maternal deaths, 12 maternal deaths (36.6%) are between 30 year and 34 year of age.

7. References

[1] Maria Helena de Sousa, José Guilherme Cecatti, Ellen Elizabeth Hardy, Suzanne Jacob Serruya. Declared maternal death and the linkage between health information systems, Rev Saúde Pública 2007; 41(2);

[2] Maternal Mortality Measurement Resources, (available at http://www.maternal_measurement.org.)

[3] M. G. Arellano, Ma, Ms, G.R. Petersen, D.B. Petitti, R.E. Smith, "The California Automated Mortality Linkage System (CAMLIS)", *American Journal of Public Health 0090-0036/84*, 1984.

[4]. Peter Christen, "A Two-Step classification Approach to unsupervised Record Linkage"

[5] Shaun J.Grannis M.D., M.S.,J. Marc Overhage M.D. Ph.D., Siu Hui Ph.D.,Clement J. McDonald M.D. "Analysis of a Probabilistic Record Linkage Technique without Human Review", Registrief Institute and Indiana University School of Medicine, Indianapolis. *INAMIA 2003 Symposium Proceedings.260-265*

[6] Carla Jorge Machado, Kenneth Hill, "Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem", Johns Hopkins University, Baltimore, U.S.A.

[7] Tony Blakey and Clare Salmond, "Probabilistic record linkage and a method to calculate the positive predicative value", *International Journal of Epidemiology*.

[8] I.P. Fellegi and A. B. Sunter, *A Theory for Record Linkage*, Journal of the American Statistical Association 64(1969), no.328, 1183-1210.

[9] Federico Maggi."A Survey of Probabilistic Record Matching Models, Techniques and Tools", Advanced Topic in Information System B, Scientific Report TR-2008-22

[10] H.Newcombe, H.B., Kennedy, J. M. Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954-959.

[11] Carig A.Mason,Ph.D, "Probabilistic Linkage:Issues and Strategies", (availiable at http://www.umit.maine.edu)

[12] Mohamed Yakout, Mikhail J. Atallah, Ahmed Elmaharmid, "Efficinet Private Record Linkage", Department of Computer Sciences, Purdue University, West Lafeyette, IN 47907, USA.