A new hybrid record linkage process to render epidemiological databases interoperable: application to the GEMO and GENEPSO studies involving BRCA1 and BRCA2 mutation carriers

YUE JIAO
Institut Curie https://orcid.org/0000-0001-9872-6034

Fabienne Lesueur
Institut Curie

Chloé-Agathe Azencott
Institut Curie

Maïté Laurent
Institut Curie

Noura Mebirouk
Institut Curie

Lilian Laborde
Institut Paoli-Calmettes

Juana Beauvallet
Institut Curie

Marie-Gabrielle Dondon
Institut Curie

Séverine Eon-Marchais
Institut Curie

Anthony Laugé
Institut Curie

GEMO Study Collaborators
Institut Curie

GENEPSO Study Collaborators
Institut Paoli-Calmettes

Catherine Noguès
Institut Paoli-Calmettes

Nadine Andrieu
Institut Curie

Dominique Stoppa-Lyonnet
Institut Curie

Sandrine M. Caputo (✉ Sandrine.caputo@curie.fr)
Institut Curie https://orcid.org/0000-0001-5338-9388

Research article

Keywords: Record linkage, hybrid process, probabilistic linkage, supervised machine learning, human-in-the-loop

Posted Date: September 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-64751/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on July 29th, 2021. See the published version at https://doi.org/10.1186/s12874-021-01299-6.

A new hybrid record linkage process to render epidemiological databases interoperable: application to the GEMO and GENEPSO studies involving *BRCA1* and *BRCA2* mutation carriers

Authors

Yue Jiao[1,2,3,4,5], Fabienne Lesueur[2,3,4,5], Chloé-Agathe Azencott[2,3,4,6+], Maïté Laurent[1,2+], Noura Mebirouk[2,3,4,5], Lilian Laborde[7], Juana Beauvallet[2,3,4,5], Marie-Gabrielle Dondon[2,3,4,5], Séverine Eon-Marchais[2,3,4,5], Anthony Laugé[1,2], GEMO Study Collaborators[2,3,4,5], GENEPSO Study Collaborators[8], Catherine Noguès[8,9+], Nadine Andrieu[2,3,4,5+], Dominique Stoppa-Lyonnet[1,10,11+], Sandrine M. Caputo[1,2]*

* corresponding author

+ these authors contributed equally to this work

1- Department of Genetics, Institut Curie, Paris, France
2- Paris Sciences-Lettres Research University, Paris, France
3- Inserm, U900, Paris, France
4- Institut Curie, Paris, France
5- Mines ParisTech, Fontainebleau, France
6- Mines ParisTech, PSL Research University, CBIO-Centre for Computational Biology, Paris, France
7- Institut Paoli-Calmettes, Centre de Traitement des Données IPC-PACA, Département de la Recherche Clinique et de l'Innovation, Marseille, France
8- Institut Paoli-Calmettes, Département d'Anticipation et de Suivi du Cancer, Oncogénétique clinique, Marseille France Inserm, U830, Université Paris Descartes, Paris, France
9- Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques et Sociales de la Santé & Traitement de l'Information Médicale, Marseille, France
10- Paris University
11- Inserm, U830, Paris, France

Corresponding author: Sandrine M. Caputo, Institut Curie - Site Paris

Service de Génétique, 26 rue d'Ulm, 75248 Paris cedex 05

sandrine.caputo@curie.fr 0033-0172389367

ABSTRACT

Background:

Linking independent sources of data related to same individuals enable innovative epidemiological and health studies but requires a robust record linkage approach. We describe a hybrid record linkage process to link databases from two independent ongoing national studies, GEMO (Genetic Modifiers of *BRCA1* and *BRCA2*), which focuses on the identification of genetic factors modifying cancer risk of *BRCA1* and *BRCA2* mutation carriers, and GENEPSO (prospective cohort of *BRCAx* mutation carriers), which focuses on environmental and lifestyle risk factors.

Methods:

To identify the maximum individuals participating in the two studies but may not be registered by a common number, we combined Probabilistic Record Linkage (PRL) and supervised Machine Learning (ML). This combined linkage was named "PRL+ML". We built the ML model using a first version of the two databases as a training dataset on which matching status was assigned by PRL followed manual review.

Results:

The Random Forest (RF) algorithm showed a highest sensitivity (0.985) among six widely used ML algorithms: RF, Bagged trees, AdaBoost, Support Vector Machine, Neural Network.

Therefore, RF was selected to build the ML model since our goal was to identify the maximum of true matches. Our combined linkage PRL+ML showed a higher sensitivity (range 0.988-0.992) than either PRL (range 0.916-0.991) or ML (0.981) alone. It identified 2,068 individuals participating in both GEMO (6,375 participants) and GENEPSO (4,925 participants).

Conclusions:

Our hybrid linkage process represents an efficient tool for linking GEMO and GENEPSO. It may be generalizable to other epidemiological studies involving other databases and registries.

Keywords:

Record linkage, hybrid process, probabilistic linkage, supervised machine learning, human-in-the-loop

Background

BRCA1 and *BRCA2* genes testing has become part of routine clinical practice in European countries and North America since the identification of the two genes in the 90's, which greatly improved recommendations about breast and ovarian cancer risk management treatments. Nonetheless, both retrospective and prospective studies on large datasets of *BRCA1* and *BRCA2 (BRCA1/2)* mutation carrier families are very much needed to refine individual cancer risk estimates by using different cancer risk factors such as genetic factors, lifestyle/environmental factors, family history and breast pathology and also to better understand the correlation between mutant *BRCA1/2* alleles and phenotype.

GEMO (Genetic Modifiers of *BRCA1* and *BRCA2*) [1] and GENEPSO (prospective cohort of *BRCAx* mutation carriers) [2] are two independent ongoing nation-wide studies involving *BRCA1/2* carriers, with unconnected databases and whose individuals may not be registered by a common number. These two resources provide an overview of a well-characterized sample of counseled Hereditary Breast and Ovarian Cancer (HBOC) families in France.

Through GEMO study, blood DNA from *BRCA1/2* mutation carriers is available to perform genetic epidemiological projects aiming at identifying and characterizing genetic factors modifying breast and ovarian cancer risk. In the prospective cohort GENEPSO, which aims at assessing environmental and lifestyle risk factors, *BRCA1/2* mutation carriers are followed over time to observe prospectively characteristics of subjects who are developing either primary or secondary cancer.

GEMO and GENEPSO were set up at different time by two different coordinating centers and investigators involved in the Genetics and Cancer Group[3] (GCG, UNICANCER), a French multicenter group composed of clinicians, molecular geneticists and scientists. Participants in both studies undergo genetic counseling and they are invited to participate in GEMO and/or GENEPSO through the family cancer clinics if tested positive for a mutation in *BRCA1* or *BRCA2*. About 26% of index cases carrying such a mutation (*i.e.* the first individual tested in the family) are included in GEMO, and 21% in GENEPSO [4]. Therefore, it is essential to identify the overlap between participants of the both studies by linking the two data sources, which will allow to better understand the cancer risk for people carrying a *BRCA1* or *BRCA2* mutation by studying simultaneously genetic and non-genetic factors. Studies conducted in subjects enrolled in both studies will also allow, for instance, assessment of whether it is possible to predict response to treatment according to *BRCA1/2* mutation status and other genetic variant profile.

Record linkage is the process of identifying the individual records from different sources refer to as the same entities[5], by determining the matching status of a record pair as match (same individual) or non-match (distinct individual). Record linkage consists of data pre-processing, record pair comparison and linkage. Data may be recorded in different formats and data items may be missing or contain errors, hence data pre-processing is essential before linkage. The record pair comparison could be computational expensive, as the number of all possible record pairs is the product of numbers of records in each dataset. To reduce the number of possible comparisons, blocking that splits the datasets into blocks is commonly performed. Only records agreeing on one or several blocking variables are then compared. When two datasets are to be linked, a unique person identifier may be unavailable for linkage. In such a case, linkage has to be performed by comparison of shared matching variables between the two datasets. The records comparison involves the comparison of values in each matching variable, this is based on comparison function (*i.e.* similarity function) that returns an indication of how similar two values of matching variables are (*i.e.* similarity score). These indications are used for linkage methods that classify record pairs into matches and non-matches. Then the linkage performance is assessed based on the confusion matrix[6]. Finally, the record linkage results may have two types of errors: False Positive (FP), *i.e.* true non-match classified as match, and False Negative (FN), *i.e.* true match classified as non-match.

Deterministic and probabilistic linkages are the two main types of linkage methods[5, 7, 8]. Deterministic record linkage involves the exact matching on a single unique matching variable or a set of matching variables. Matching status can be assessed in a single step or in a stepwise manner. If data are of very good quality (*i.e*. no more than 5% of missing data or errors in any matching variable), the deterministic linkage can have a satisfying linkage quality. Otherwise, it will produce a large number of FP[9]. By contrast, Probabilistic Record Linkage (PRL) can have a greater linkage capacity than deterministic linkage when the data are not of good quality[10]. PRL is also able to take into account the difference in discriminatory power of each matching variable. Indeed, the more frequent a value of matching variable is, the less discriminative for linkage this value is. Fellegi and Sunter[7]

(1969) first proposed a formal probabilistic record linkage model, many extensions have been proposed later[11]. The Fellegi-Sunter (FS) model produces match likelihood scores based on estimating the true positive matches probability and false positive matches probability. Without relying on the assumptions of these probabilities, the likelihood score can also be calculated by EpiLink[12], a simple approach within the scope of the FS model. To note that in PRL, the determination of likelihood score thresholds that classify the matches and non-matches is critical and eventually alters the relative numbers of FP and FN[13]. Although the threshold could be estimated by minimizing the number of possible links with given error rates for FP and FN[7], the most common practice is to specify it by examining the observed distribution of scores. Then the choice of the threshold depends on the study objective for limiting FN or FP.

From a machine learning point of view, record linkage can be considered as a classification task. Each record pair is represented by a comparison vector containing, for each matching variable, the similarity score between both records. The supervised machine learning (ML) algorithm can learn a model that takes such a comparison vector as input and returns matching status as output, based on a training set in which the matching status of record pairs are known. Various ML algorithms have been applied to record linkage, such as Classification Tree (CT), Support Vector Machines (SVM), Neural Networks (NNET), Random Forest (RF)[14–19], etc. However, their application is usually limited by the need of a training set.

Distinct FP and FN can occur in both PRL and ML due to errors in data entry and linkage process[20]. Hence in this study, in order to minimize FP and FN in linkage process, we propose a human-in-the-loop hybrid record linkage process. We combine the PRL and ML, which means the candidate matches classified by both linkage methods are combined and manually reviewed. In order to overcome the lack of training dataset for ML, we propose to build the ML model from a training dataset (the whole initial databases) whose matching status was assigned by PRL followed manual review. We applied this hybrid linkage process to data from GEMO and GENEPSO studies.

Materials and Methods

Data

In September 2016, 4,688 and 3,339 participants had been enrolled in GEMO and GENEPSO, respectively. This initial dataset (dataset 1) was used for selecting ML algorithms and evaluating an optimal linkage method. The selected linkage method was applied on the updated dataset (dataset 2) of the two studies as of December 2019, which corresponded to 6,375 participants for GEMO (*i.e.* 1,687 new participants), and 4,925 participants for GENEPSO (*i.e.* 1,586 new participants).

Name and address of individuals were not available here due to privacy policy and confidentiality. Ten matching variables shared between GEMO and GENEPSO were used for comparison (Table 1): *BRCA1* mutational status (BRCA1), *BRCA2* mutational status (BRCA2), mutation description using the HGVS nomenclature (MUT_HGVS), gender (GENDER), consultation center number (CTR), family number (NUMFAM), individual number (SUJID), year of birth (Yob), month of birth (Mob) and day of birth (Dob).

Data pre-processing

Record linkage is highly sensitive to data quality. Therefore, we performed data cleaning and standardization[21–23], such as removing duplicates, deleting spaces in strings, standardizing format of linkage variables, converting mutation descriptions to standard Human Genome Variation Society (HGVS) nomenclature[24], and splitting dates of birth into month, day and year in order to compare respectively each of them and give credit for partial agreement.

Record pair comparison

Let X and Y be two databases and $x \in X$ and $y \in Y$ two arbitrary records in form of a *d*-dimensional vector, *i.e.* $x = [x_1, \dots x_d]$ and $y = [y_1, \dots y_d]$. In our illustration, $d = 10$. The space of comparison is the Cartesian product $X \times Y$ which contains of all possible record pairs (x, y). All matching variables are discrete numerical values except MUT_HGVS which is a string. A similarity vector $s = [s_1, \dots, s_d]$ is then computed as $s_i = sim(x_i, y_i)$ where x_i, y_i are the *i*-th matching variables and $sim(\cdot,\cdot)$ is a measure of similarity given by the Jaro-Winkler similarity sim_{JW} for the string matching variable (MUT_HGVS) and by the binary similarity sim_B (*i.e.* extract agreement) for the others (Supplementary Data).

Probabilistic Record Linkage (PRL)

A weighted sum is computed on the similarity vector S in order to obtain a score which is assessed as the probability of match for the record pair(x, y):

$$S(x,y) = \frac{\sum_i w_i \, sim(x_i, y_i)}{\sum_i w_i} = \sum_{i=1}^{d} w_i^0 s_i \qquad \text{Equation 1}$$

where $w = [w_1, \dots w_d]$ is the vector of unnormalized weight and w^0 the normalized one. For the matching variable i, the weight is calculated based on the EpiLink approach[12] as

$$w_i = log_2\,(1 - e_i)/f_i$$

where f_i denotes the average frequency of values in the matching variable and e_i the estimated error rate. We assumed $e_i = 0.01$ for all matching variables[12]. Since most software packages performed similarly[25], we implemented the record linkage here by RecordLinkage R package[26]. An example is shown in Table 1.

After examining the observed distribution of scores, we specified the threshold t which separates potential matches ($S(x, y) >$ t) and non-matches ($S(x, y) \leq$ t), then manually reviewed these potential matches[27] and also pairs whose scores below and near the threshold. In this study, we also evaluated the performances of PRL with varying likelihood score thresholds.

Supervised machine learning (ML) linkage

We first used blocking to reduce the number of possible record pair comparisons. Missing data in blocking variables (BRCA1, BRCA2, GENDER and Yob) were tolerated here. After blocking, the imputation of missing data could be then performed. The missing data in similarity for MUT_HGVS (numeric) were imputed by Bayesian linear regression and those for other categorical matching variables were imputed by logistic regression.

Next, the labeled record pairs were randomly partitioned into two sets: the training dataset (60%) on which we trained ML models, and the test dataset (40%) on which we evaluated the predictive performance of the trained models. We employed six broadly used ML algorithms (CT, Bagged trees, AdaBoost, RF, SVM and NNET) (Supplementary Data). In order to find the optimal parameters for each of these algorithms, we carried out a 5-fold cross validation on the training set. Finally, we compared the performance of the 6 models to that of a naïve baseline, consisting in a Bernoulli model that randomly classifies a record pair as matching or non-matching.

Hybrid record linkage process

We first used initial dataset 1 for which the matching status assignments were made by PRL followed manual review (Figure 1a). The likelihood score threshold of PRL was designated low enough. We then manual reviewed all pairs whose scores exceed the threshold and the pairs whose scores below and near the threshold, in order to minimize the FP and FN in PRL. Based on this dataset with known matching status, we aimed to: (1) select a ML algorithm by testing the performances of 6 ML algorithms (Figure 1b); (2) evaluate three linkage methods by 5-fold cross validation: PRL, ML and PRL+ML – the latter means that the candidate matches classified by both PRL and ML are combined and manually reviewed (Figure 1c); (3) build a ML model from the whole dataset 1 (Figure 1d).

To achieve (1) and (2), all record pair comparisons were partitioned into 5 equal size groups. Of the 5 groups, a single group (dataset B) was retained for comparing the predictive performances of three linkage methods. The remaining four groups (dataset A) served itself for choosing an optimal ML algorithm (Figure 1b) then building a ML model which could be used by ML and PRL+ML linkage methods on dataset B (Figure 1c). The whole process was repeated five times with each of the 5 groups used only once as dataset B.

From updated dataset 2, we then applied the selected linkage method to identify new true matches (Figure 1e).

Performance measures

The predictive performance of all algorithms was assessed. In record linkage, the data is imbalanced, meaning that the two classes are not represented equally. Indeed, there are far more non-matches than matches. In such case, instead of standard accuracy, the precision and sensitivity are commonly used for evaluating the linkage quality. We calculated these performance metrics based on the confusion matrix described in Table S1.

The precision is the proportion of classified matches that are true matches.

$$precision = \frac{TP}{TP + FP}$$

The sensitivity is the proportion of true matches that have been classified correctly.

$$sensitivity = \frac{TP}{TP + FN}$$

A good linkage algorithm will typically have values of precision and sensitivity greater than 0.95[28].

Results

Matching status assignment by PRL followed manual review

Up to September 2016, 4,688 and 3,339 individuals had been enrolled in GEMO and GENEPSO, respectively. After data pre-processing, 15,653,232 record pairs were built as the Cartesian product of the two databases in dataset 1. The similarity score of each record pair was computed from Equation 1 of PRL (see Methods). The score distribution is given in Figure 2 and Table S2. We observed a large peak of low scores, corresponding to a large number of non-matches, and a small peak of high scores, corresponding to a small number of matches. This bimodal distribution suggests that PRL worked as expected. We designated a lower threshold 0.6 and performed a stage of manual review. 1,257 pairs were classified as matches, and 15,651,938 pairs were classified as non-matches (Figure 1a). 37 pairs whose matching status were not sure after verification were excluded.

Selection of a supervised machine learning algorithm

The average predictive performances of the Bernoulli model and of the ML models, as assessed on dataset A (Figure 1b), are presented in Table 2. The six supervised ML models outperformed the Bernoulli model, and their performance values were all higher than 0.97 suggesting that they performed good linkage prediction. The RF algorithm showed the highest sensitivity (0.9853) whereas the NNET algorithm showed the highest precision (0.9843). Here our objective was to have an optimal sensitivity, which means to identify as many true matches as possible. In addition, the RF algorithm required less tuning and was therefore more likely to generalize better. Therefore, we chose the RF algorithm as our ML algorithm.

Evaluation of three linkage methods

In the 5-fold cross validation step, the averaged performance of three linkage methods (PRL, RF, PRL+RF) was assessed on dataset B (Figure 3, Figure 1c). PRL and PRL+RF methods had score thresholds varying from 0.6 to 0.8 in PRL. As expected, increasing score threshold introduced fewer FP which led to an increasing precision, but more FN which led to a decreasing sensitivity. The PRL+RF method with varying threshold showed a higher sensitivity than that of RF (Figure 3a), while the sensitivity of PRL decreased significantly. The sensitivity of RF was similar to that of PRL at threshold 0.65. RF showed a higher precision than that of PRL and PRL+RF, suggesting fewer FP, while PRL and PRL+RF showed a similar precision which increased significantly with thresholds (Figure 3b).

In conclusion, the PRL approach was very sensitive to the threshold and did not perform better than RF, except for the measure of sensitivity at threshold 0.6, which, naturally, comes as the cost of a lower precision. Conversely, RF had a high precision but a modest sensitivity. PRL+RF had very high sensitivity and decreasing slightly with increasing thresholds, and had similar precision as PRL. Since the goal of our study here was to minimize FN, we chose the combined linkage method PRL+RF with lower threshold 0.6 which had the highest sensitivity.

After having selected the PRL+RF as the optimal linkage method and blocking on the whole dataset 1, 107,599 out of 15,653,195 record pairs were trained for a RF model in PRL+ML (Figure 1d).

Linking records in updated GEMO and GENEPSO databases

As of December 2019, 1,687 and 1,586 new *BRCA1* or *BRCA2* mutation carriers had been enrolled in GEMO and GENEPSO, respectively. These updated GEMO and GENEPSO samples constitute dataset 2. The combined linkage method PRL+RF was applied on this dataset to identify the new true matches (Figure 1e). Linkage was first performed by PRL and by RF model separately. The RF model here was trained on record pairs of dataset 1 (*i.e.* using GEMO and GENEPSO participants enrolled before September 2016) (Figure 1d). RF and PRL predicted 819 and 1,268 candidate matches, respectively (Figure 4a). Besides the 772 matches that were common between these two linkage methods, RF had 47 additional candidate matches and PRL had 496 additional candidates. Those 1,315 candidate matches (772 + 496 + 47) were then manually reviewed. The PRL approach was correct for 57.3% (727/1,268) records; while the RF was correct for 87.3% (715/819) records (Figure 4b). The cost of manual reviews of PRL+RF was 47 more than that of PRL, but PRL+RF allowed us to identify 12 more true matches than RF alone.

Finally, the combined linkage method PRL+RF suggested 1,315 candidates, out of which 738 were true new matches. All those true matches were among those identified by either PRL or RF, but the PRL+RF approach allowed identifying the maximum true matches. With the

PRL+RF method, we obtained 1.5% (11/738) additional gains as compared to the PRL method alone, and 3.1% (23/738) additional gains as compared to the RF method (Figure 4 and Table S3). This confirmed also that the PRL+RF method had a higher sensitivity than PRL and RF alone.

Hence, in December 2019, GEMO included 6,375 participants and GENEPSO included 4,925 participants, and our hybrid record linkage identified 2,068 *BRCA1/2* mutation carriers from 1,693 families that had been enrolled in both studies.

Discussion

In this paper, we propose a human-in-the-loop, hybrid record linkage process which involves both PRL and ML approaches (Figure 5). The hybrid process was performed using datasets from GEMO and GENEPSO which are two independent ongoing nation-wide epidemiological studies involving *BRCA1/2* mutation carriers. PRL and ML were combined to classify the record pairs into matches and non-matches, in which the ML model was built from PRL followed by manual review.

PRL has a lower computational cost but the linkage quality is impacted by likelihood score determination. PRL with lower threshold leads to more FP whereas higher threshold PRL leads to more FN. The ML approach can have higher precision and fewer manual reviews. However, the blocking step can generate FN if the data contain errors in blocking variables. We found that the PRL+ML combined method having the highest sensitivity compared to other two methods alone can improve the linkage by identifying as many true matches as possible without paying more additional manual review than PRL.

If a high precision with lower manual review cost is required and some missing true matches are tolerated, the ML approach could be an interesting option. If more matching variables had been shared between GEMO and GENEPSO been available, matching variables selection by feature selection algorithms could have been performed to identify the highest discriminating matching variables for linkage. On the other hand, if fewer matching variables had been shared, it could have been difficult to build a ML model and in such a case, PRL would have been preferred.

Previously, Elfeky *et al.*[29] described a hybrid technique for record linkage, combining both supervised and unsupervised machine learning methods. The class assignments were made to a data sample through unsupervised clustering, and the resulting data was then used as a training dataset for a supervised model[6]. However, this technique was not suitable here since two unsupervised machine learning methods (K-means and bagged k-means) showed independently poor performance (Supplementary Data, Table S4), probably due to our imbalanced data.

Conclusions

It is important to note here that in our hybrid process, as in the PRL and ML stand-alone approaches, each linkage step needs human-in-the-loop that can correct the classification errors and improve the models by paying more time, cost and expertise. Especially, PRL errors may succeed in construction of ML model. Hence, an exhaustive manual review is necessary in PRL.

GEMO and GENEPSO are ongoing studies and their respective databases are continuously updated. About 730 and 590 new subjects are included each year in GEMO and GENEPSO respectively. Hence, we will apply on regularly basis our hybrid approach on the updated versions of the two databases to identify the new true matches in order to increase the power of the research projects. Our hybrid record linkage process was driven by the need of a specific epidemiological question and may be generalizable to other epidemiological or translational studies involving other databases and registries.

List of abbreviations

Abbreviations are shown in Table S5.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The GEMO dataset is included in the published article describing the GEMO resource[1]. The GENEPSO dataset is not publicly available but has been used in a number of studies[30–36]. Data are however available from the authors upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

CANSOP was supported by the French National Institute of Cancer (INCa) [grant 2013-1-BCB-01-ICH-1, D. Stoppa-Lyonnet]. GEMO is currently supported by the INCa [grant 2013-1-BCB-

01-ICH-1, D. Stoppa-Lyonnet and grant SHS-E-SP 18-015, F. Lesueur] and the Fondation ARC pour la recherche sur le cancer [grant PJA 20151203365, F. Lesueur], and the contract grant sponsor: Association "Le cancer du sein, parlons-en!". GENEPSO is currently supported by a grant from Institut National du Cancer (INCa) as part of the European program ERA-NET on Translational Cancer Research (TRANSCAN) [grant TRANSCAN-JTC2012, n°2014-008, N. Andrieu], Ligue Nationale Contre le Cancer.

Authors' Contributions

SMC, FL, CN and DS-L conceived of the presented idea and supervised the project. YJ performed the computations in consultation with C-AA and SMC. YJ wrote the manuscript with support from FL, C-AA, M-GD, NA, JB, SE-M and SMC. AL and ML helped implement the interface. FL and DS-L coordinated the GEMO study. NM managed the DNA samples, managed family and clinical data in GEMO study. CN and LL coordinated the GENEPSO study. All authors read and approved the final manuscript.

Acknowledgements

We thank the patients and the participants in the contributing studies.

The Genetic Modifiers of Cancer Risk in *BRCA1/2* Mutation Carriers (GEMO) study is a study from the National Cancer Genetics Network «UNICANCER Genetic Group», France. We wish to pay a tribute to Olga M. Sinilnikova, who with Dominique Stoppa-Lyonnet initiated and coordinated GEMO until she sadly passed away on the 30th June 2014. The team in Lyon (Olga Sinilnikova, Mélanie Léone, Laure Barjhoux, Carole Verny-Pierre, Sylvie Mazoyer, Francesca Damiola, Valérie Sornin) managed the GEMO samples until the biological resource center was transferred to Paris in December 2015 (Noura Mebirouk, Fabienne Lesueur, Dominique Stoppa-Lyonnet). We want to thank all the GEMO collaborating groups for their contribution to this study. **Coordinating Center:** Service de Génétique, Institut Curie, Paris: Muriel Belotti, Ophélie Bertrand, Anne-Marie Birot, Bruno Buecher, Sandrine M. Caputo, Chrystelle Colas, Anaïs Dupré, Emmanuelle Fourme, Marion Gauthier-Villars, Lisa Golmard, Claude Houdayer, Marine Le Mentec, Virginie Moncoutier, Antoine de Pauw, Claire Saule, Dominique Stoppa-Lyonnet, and Inserm U900, Institut Curie, Paris: Fabienne Lesueur, Noura Mebirouk. **Contributing Centers:** Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon - Centre Léon Bérard, Lyon: Nadia Boutry-Kryza, Alain Calender, Sophie Giraud, Mélanie Léone. Institut Gustave Roussy, Villejuif: Brigitte Bressac-de-Paillerets, Olivier Caron, Marine Guillaud-Bataille. Centre Jean Perrin, Clermont–Ferrand: Yves-Jean Bignon, Nancy Uhrhammer. Centre Léon Bérard, Lyon: Valérie Bonadona, Christine Lasset. Centre François Baclesse, Caen: Pascaline Berthet, Laurent Castera, Dominique Vaur. Institut Paoli Calmettes, Marseille: Violaine Bourdon, Catherine Noguès, Tetsuro Noguchi, Cornel Popovici, Audrey Remenieras, Hagay Sobol. CHU Arnaud-de-Villeneuve, Montpellier: Isabelle Coupier, Pierre-Olivier Harmand, Pascal Pujol, Paul Vilquin. Centre Oscar Lambret, Lille: Aurélie Dumont, Françoise Révillion. Centre Paul Strauss, Strasbourg: Danièle Muller. Institut Bergonié, Bordeaux: Emmanuelle Barouk-Simonet, Françoise Bonnet, Virginie Bubien, Michel Longy, Nicolas Sévenet. Institut Claudius Regaud, Toulouse: Laurence Gladieff, Rosine Guimbaud, Viviane Feillel, Christine Toulas. CHU Grenoble: Hélène Dreyfus, Dominique Leroux, Magalie Peysselon, Christine Rebischung. CHU Dijon: Amandine Baurand, Geoffrey Bertolone, Fanny Coron, Laurence Faivre, Vincent Goussot, Caroline Jacquot, Caroline Sawka. CHU St-Etienne: Caroline Kientz, Marine Lebrun, Fabienne Prieur. Hôtel Dieu Centre Hospitalier, Chambéry: Sandra Fert-Ferrer. Centre Antoine Lacassagne, Nice: Véronique Mari. CHU Limoges: Laurence Vénat-Bouvet. CHU Nantes: Stéphane Bézieau,

Capucine Delnatte. CHU Bretonneau, Tours and Centre Hospitalier de Bourges: Isabelle Mortemousque. Groupe Hospitalier Pitié-Salpétrière, Paris: Florence Coulet, Florent Soubrier, Mathilde Warcoin. CHU Vandoeuvre-les-Nancy: Myriam Bronner, Sarab Lizard, Johanna Sokolowska. CHU Besançon: Marie-Agnès Collonge-Rame, Alexandre Damette. CHU Poitiers, Centre Hospitalier d'Angoulême and Centre Hospitalier de Niort: Paul Gesta. Centre Hospitalier de La Rochelle: Hakima Lallaoui. CHU Nîmes Carémeau: Jean Chiesa. CHI Poissy: Denise Molina-Gomes. CHU Angers: Olivier Ingster. CHRU de Lille: Sylvie Manouvrier-Hanu, Sophie Lejeune.

GENEPSO Centers: the Coordinating Center: Institut Paoli-Calmettes, Marseille, France: Catherine Noguès, Lilian Laborde, Pauline Pontois and the Collaborating Centers: Institut Curie, Paris: Dominique Stoppa-Lyonnet, Marion Gauthier-Villars; Bruno Buecher, Institut Gustave Roussy, Villejuif: Olivier Caron; Hôpital René Huguenin/Institut Curie, Saint Cloud: Catherine Noguès, Emmanuelle Mouret-Fourme; Centre Paul Strauss, Strasbourg: Jean-Pierre Fricker; Centre Léon Bérard, Lyon: Christine Lasset, Valérie Bonadona; Centre François Baclesse, Caen: Pascaline Berthet; Hôpital d'Enfants CHU Dijon – Centre Georges François Leclerc, Dijon: Laurence Faivre; Centre Alexis Vautrin, Vandoeuvre-les-Nancy: Elisabeth Luporsi; Centre Antoine Lacassagne, Nice: Marc Frénay; Institut Claudius Regaud, Toulouse: Laurence Gladieff; Réseau Oncogénétique Poitou Charente, Niort: Paul Gesta; Institut Paoli-Calmettes, Marseille: Catherine Noguès, Hagay Sobol, François Eisinger, Jessica Moretta; Institut Bergonié, Bordeaux: Michel Longy, Centre Eugène Marquis, Rennes: Catherine Dugast; GH Pitié Salpétrière, Paris: Chrystelle Colas, Florent Soubrier; CHU Arnaud de Villeneuve, Montpellier: Isabelle Coupier, Pascal Pujol; Centres Paul Papin, and Catherine de Sienne, Angers, Nantes: Alain Lortholary; Centre Oscar Lambret, Lille: Philippe Vennin, Claude Adenis; Institut Jean Godinot, Reims: Tan Dat Nguyen; Centre René Gauducheau,

Nantes: Capucine Delnatte; Centre Henri Becquerel, Rouen: Annick Rossi, Julie Tinat, Isabelle Tennevet; Hôpital Civil, Strasbourg: Jean-Marc Limacher; Christine Maugard; Hôpital Centre Jean Perrin, Clermont-Ferrand: Yves-Jean Bignon; Polyclinique Courlancy, Reims: Liliane Demange; Clinique Sainte Catherine, Avignon: Hélène Dreyfus; Hôpital Saint-Louis, Paris: Odile Cohen-Haguenauer; CHRU Dupuytren, Limoges: Brigitte Gilbert; Couple-Enfant-CHU de Grenoble: Dominique Leroux; Hôpital de la Timone, Marseille: Hélène Zattara-Cannoni.

References

1. Lesueur F, Mebirouk N, Jiao Y, Barjhoux L, Belotti M, Laurent M, et al. GEMO, a national resource to study genetic modifiers of breast and ovarian cancer risk in BRCA1 and BRCA2 pathogenic variant carriers. Front Oncol. 2018;8. doi:10.3389/fonc.2018.00490.

2. Lecarpentier J, Noguès C, Mouret-Fourme E, Buecher B, Gauthier-Villars M, Stoppa-Lyonnet D, et al. Breast Cancer Risk Associated with Estrogen Exposure and Truncating Mutation Location in BRCA1/2 Carriers. Cancer Epidemiol Prev Biomark. 2015;24:698–707.

3. UNICANCER - Le Groupe génétique et cancer (GGC). http://www.unicancer.fr/recherche/les-groupes-recherche/groupe-genetique-et-cancer-ggc. Accessed 9 Nov 2018.

4. INCa. Oncogénétique en 2016. 2016. www.e-cancer.fr/Professionnels-de-sante/L-organisation-de-l-offre-de-soins/Oncogenetique.

5. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic Linkage of Vital Records. Science. 1959;130:954–9.

6. Christen P, Goiser K. Quality and complexity measures for data linkage and deduplication. In: Quality measures in data mining. Springer; 2007. p. 127–151.

7. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc. 1969;64:1183–1210.

8. Newcombe HB. Handbook of record linkage: methods for health and statistical studies, administration, and business. Oxford University Press, Inc.; 1988.

9. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. J Biomed Inform. 2015;56:80–6.

10. Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat Med. 2002;21:1485–1496.

11. Winkler WE. Overview of record linkage and current research directions. In: Bureau of the Census. Citeseer; 2006.

12. Contiero P, Tittarelli A, Tagliabue G, Maghini A, Fabiano S, Crosignani P, et al. The EpiLink record linkage software. Methods Inf Med. 2005;44:66–71.

13. Guillet F, Hamilton HJ. Quality measures in data mining. Springer; 2007.

14. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv CSUR. 1999;31:264–323.

15. Cochinwala M, Kurien V, Lalk G, Shasha D. Efficient data reconciliation. Inf Sci. 2001;137:1–15.

16. Verykios VS, Elmagarmid AK, Houstis EN. Automating the approximate record-matching process. Inf Sci. 2000;126:83–98.

17. Wang F, Wang H. Record Linkage Using the Combination of Twice Iterative SVM Training and Controllable Manual Review. In: Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2016 IEEE 14th Intl C. IEEE; 2016. p. 31–38.

18. Pixton B, Giraud-Carrier C. Using structured neural networks for record linkage. In: Proceedings of the Sixth Annual Workshop on Technology for Family History and Genealogical Research. 2006.

19. Kim K, Giles CL. Financial Entity Record Linkage with Random Forests. In: Proceedings of the Second International Workshop on Data Science for Macro-Modeling. ACM; 2016. p. 13.

20. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017;46:1699–710.

21. Clark DE. Practical introduction to record linkage for injury research. Inj Prev. 2004;10:186–191.

22. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques. Springer Science & Business Media; 2007.

23. Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Eng Bull. 2000;23:3–13.

24. Callenberg KM, Santana-Santos L, Chen L, Ernst WL, De Moura MB, Nikiforov YE, et al. Clinical Implementation and Validation of Automated Human Genome Variation Society (HGVS) Nomenclature System for Next-Generation Sequencing–Based Assays for Cancer. J Mol Diagn. 2018.

25. Karr AF, Taylor MT, West SL, Setoguchi S, Kou TD, Gerhard T, et al. Comparing record linkage software programs and algorithms using real-world data. PLOS ONE. 2019;14:e0221459.

26. Sariyar M, Borg A. The RecordLinkage Package: Detecting Errors in Data. R J. 2010;2.

27. Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. John Wiley & Sons; 2015.

28. Dusetzina SB, Tyree S, Meyer A-M, Meyer A, Green L, Carpenter WR. An overview of record linkage methods. 2014.

29. Elfeky MG, Verykios VS, Elmagarmid AK. TAILOR: A record linkage toolbox. In: Proceedings 18th International Conference on Data Engineering. IEEE; 2002. p. 17–28.

30. Andrieu N, Goldgar DE, Easton DF, Rookus M, Brohet R, Antoniou AC, et al. Pregnancies, Breast-Feeding, and Breast Cancer Risk in the International BRCA1/2 Carrier Cohort Study (IBCCS). JNCI J Natl Cancer Inst. 2006;98:535–44.

31. Pijpe A, Andrieu N, Easton DF, Kesminiene A, Cardis E, Noguès C, et al. Exposure to diagnostic radiation and risk of breast cancer among carriers of BRCA1/2 mutations: retrospective cohort study (GENE-RAD-RISK). BMJ. 2012;345. doi:10.1136/bmj.e5660.

32. Phillips K-A, Milne RL, Rookus MA, Daly MB, Antoniou AC, Peock S, et al. Tamoxifen and Risk of Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. J Clin Oncol. 2013;31:3091–9.

33. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips K-A, Mooij TM, Roos-Blom M-J, et al. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. JAMA. 2017;317:2402–16.

34. Schrijver LH, Olsson H, Phillips K-A, Terry MB, Goldgar DE, Kast K, et al. Oral Contraceptive Use and Breast Cancer Risk: Retrospective and Prospective Analyses From a BRCA1 and BRCA2 Mutation Carrier Cohort Study. JNCI Cancer Spectr. 2018;2. doi:10.1093/jncics/pky023.

35. Mavaddat N, Antoniou AC, Mooij TM, Hooning MJ, Heemskerk-Gerritsen BA, Noguès C, et al. Risk-reducing salpingo-oophorectomy, natural menopause, and breast cancer risk: an international prospective cohort of BRCA1 and BRCA2 mutation carriers. Breast Cancer Res. 2020;22:8.

36. Li H, Terry MB, Antoniou AC, Phillips K-A, Kast K, Mooij TM, et al. Alcohol Consumption, Cigarette Smoking, and Risk of Breast Cancer for BRCA1 and BRCA2 Mutation Carriers: Results from The BRCA1 and BRCA2 Cohort Consortium. Cancer Epidemiol Prev Biomark. 2020;29:368–78.

	BRCA1	BRCA2	CTR	GENDER	NUMFAM	SUJID	MUT_HGVS	Yob	Mob	Dob	PRL Score
Individual GEMO$_{\_5789}$	1	0	1	2	17455	0001	$\mathrm{c.}3403\mathrm{C} > \mathrm{T}$	1959	08	05	
Individual GENEPSO$_{\_01082300001}$	1	0	1	2	08230	0001	$\mathrm{c.}3481\_3491\mathrm{del}$	1958	08	05	
Similarity $\boldsymbol{s}$	1	1	1	1	0	1	0.7825	0	1	1	
f	0.3333	0.3333	0.02272	0.5000	0.00025	0.0018	0.0006	0.01098	0.07692	0.03125	
w	1.57	1.57	5.45	0.99	11.95	9.1	10.69	6.49	3.68	4.99	sum(w)=56.48
$w * s$	1.57	1.57	5.45	0.99	0	9.1	8.36	0	3.68	4.99	sum(w*s)=35.71
score $\boldsymbol{S}$											0.6322

Table 1. An example of a record pair comparison and its likelihood score calculation. Ten matching variables were used to identify record pairs: *BRCA1* mutational status (BRCA1), *BRCA2* mutational status (BRCA2), mutation description using the HGVS nomenclature (MUT_HGVS), gender (GENDER), consultation center number (CTR), family number of the consultation (NUMFAM), individual number in family (SUJID), year of birth (Yob), month of birth (Mob) and day of birth (Dob). BRCA1 and BRCA2 matching variable: 1: “carrier of a *BRCA1/2* mutation”, 0: “noncarrier of a *BRCA1/2* mutation”. GENDER matching variable: 1: male, 2: female. The similarity vector $\boldsymbol{s}$ in the third row is used as input in the machine learning approaches. The PRL score $\boldsymbol{S}$ is calculated from the weight $\boldsymbol{w}$ and the similarity $\boldsymbol{s}$.

Models	Atest dataset			
	Sensitivity		Precision	
	M	*SD*	*M*	*SD*
Bernoulli	0.01172	0.00079	0.01139	0.00096
CT	0.9841	0.016	0.9779	0.0059
Bagged trees	0.9809	0.012	0.9826	0.0080
AdaBoost	0.9839	0.011	0.9828	0.0075
RF	**0.9853**	0.011	0.9824	0.010
SVM	0.9821	0.017	0.9789	0.0068
NNET	0.9823	0.012	**0.9843**	0.0078

Table 2. Mean of the performance for the different tested algorithms on Atest dataset. Six machine learning algorithms were tested: Classification Tree (CT), Bagged trees, AdaBoost, Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NNET). *M*= mean, *SD* = standard deviation. Highest mean value among the different algorithms are in bold. We used here 4 digits to the right of the decimal points after rounding, because the differences among the values of sensitivity and precision are small.

Figure Legends

Figure 1 Elaboration of hybrid record linkage process and main steps. (a) Assignment the matching status by PRL followed manual review. We manual reviewed all the pairs whose scores exceed the threshold and also the pairs whose scores below and near the threshold, in order to minimize the FP and FN in PRL. (b) Selection of the supervised machine learning algorithm. (c) Selection of the combined linkage method PRL+RF. (d) Building of RF model on initial databases. (e) Application of PRL+RF combined linkage method to classify the updated data.

Figure 2 Score distribution of 15,653,232 record pairs in dataset 1. (a) Whole score distribution. (b) Zoom on the distribution for the highest scores.

Figure 3 Performance of three linkage methods: PRL (Probabilistic Record Linkage), RF (Random Forest) and PRL+RF. PRL has thresholds varying from 0.6 to 0.8. (a) Comparison of their sensitivities. (b) Comparison of their precisions.

Figure 4 Comparison of candidate matches predicted by the RF and PRL models for the updated databases. (a) Before manual review, RF model and PRL predicted 819 and 1,268 new candidate matches, respectively; 772 candidate matches were found by both approaches. (b) After manual review, PRL+RF identified 738 true matches, in which 727 true matches were identified by PRL and 715 true matches were identified by RF. 704 true matches were identified by both approaches. 23 true matches were identified only by PRL, and 11 true matches were identified only by the RF model.

Figure 5 General overview of the hybrid record linkage process. (a) Probabilistic record linkage (PRL) followed by a stage of manual review is first applied to build a machine learning (ML) model. (b) The PRL+ML combined linkage is then used to classify the updated datasets (Record pair comparison from Database X' and Database Y'). The ML model obtained in (a) is used (dotted arrow) for the prediction in (b).

Figures

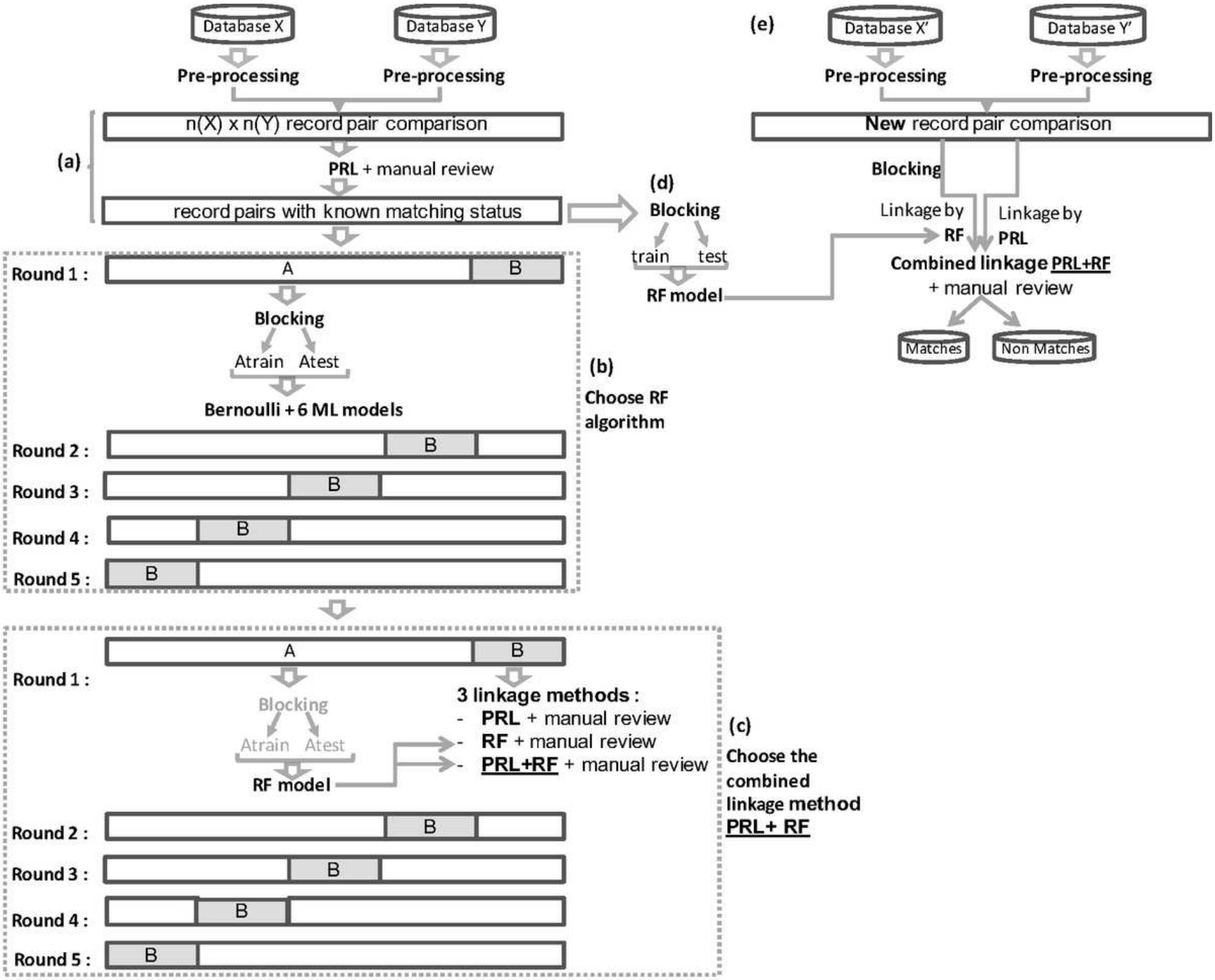

Figure 1

Elaboration of hybrid record linkage process and main steps. (a) Assignment the matching status by PRL followed manual review. We manual reviewed all the pairs whose scores exceed the threshold and also the pairs whose scores below and near the threshold, in order to minimize the FP and FN in PRL. (b) Selection of the supervised machine learning algorithm. (c) Selection of the combined linkage method PRL+RF. (d) Building of RF model on initial databases. (e) Application of PRL+RF combined linkage method to classify the updated data.

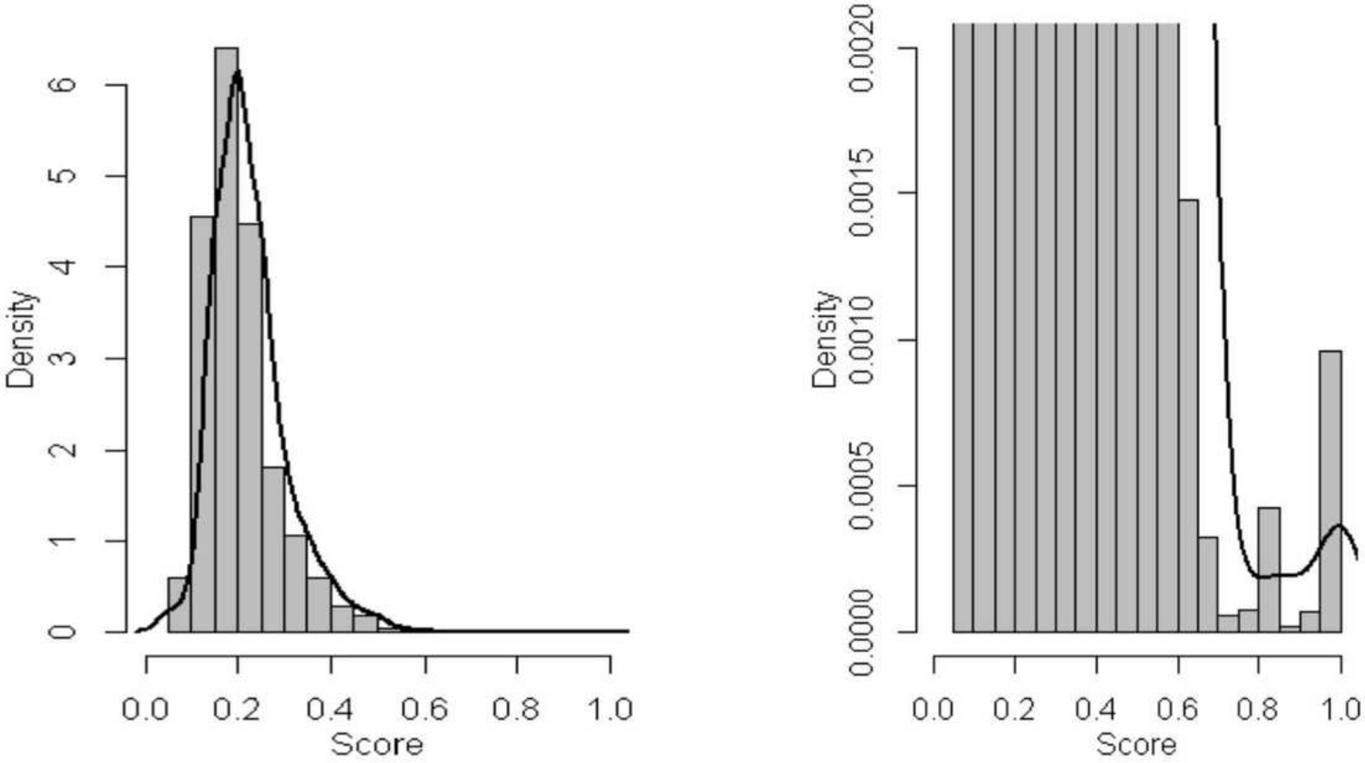

Figure 2

Score distribution of 15,653,232 record pairs in dataset 1. (a) Whole score distribution. (b) Zoom on the distribution for the highest scores.

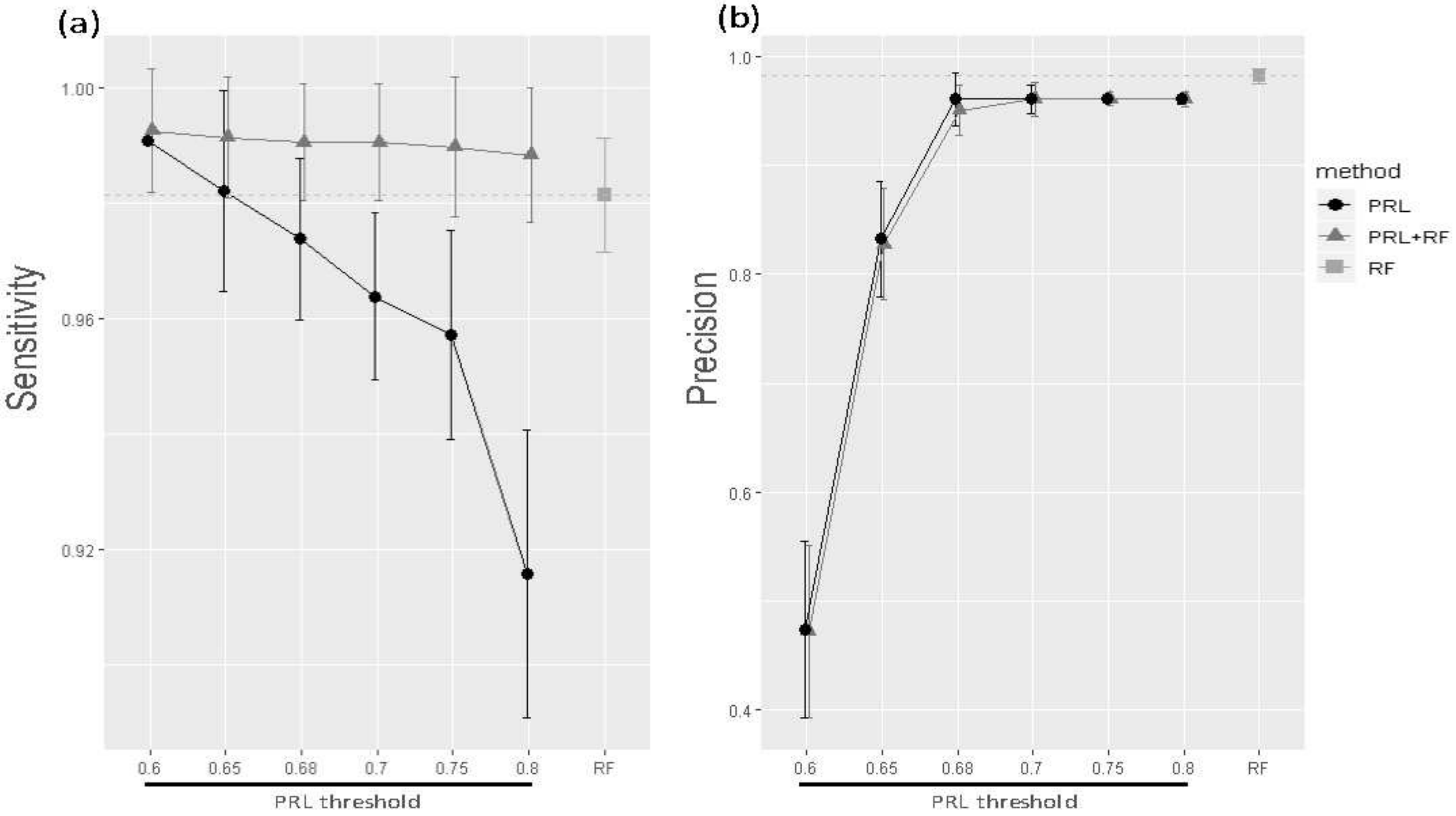

Figure 3

Performance of three linkage methods: PRL (Probabilistic Record Linkage), RF (Random Forest) and PRL+RF. PRL has thresholds varying from 0.6 to 0.8. (a) Comparison of their sensitivities. (b) Comparison of their precisions.

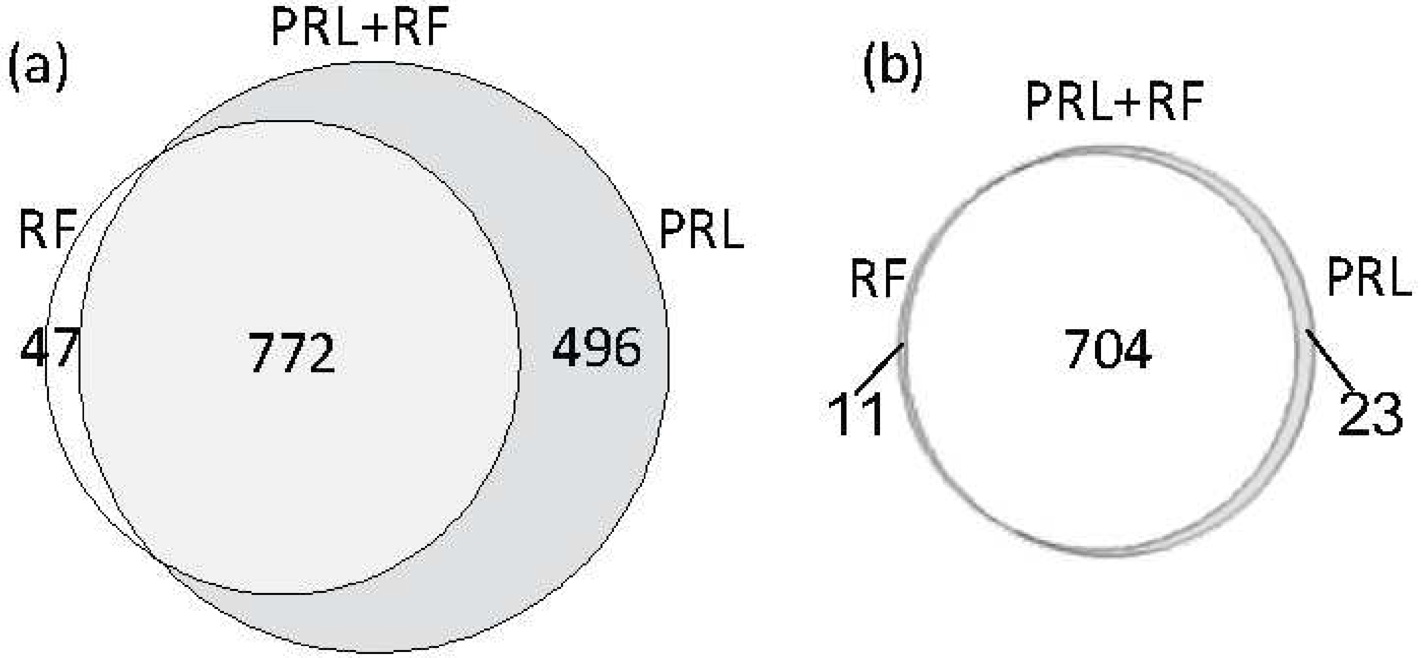

Figure 4

Comparison of candidate matches predicted by the RF and PRL models for the updated databases. (a) Before manual review, RF model and PRL predicted 819 and 1,268 new candidate matches, respectively; 772 candidate matches were found by both approaches. (b) After manual review, PRL+RF identified 738 true matches, in which 727 true matches were identified by PRL and 715 true matches were identified by RF. 704 true matches were identified by both approaches. 23 true matches were identified only by PRL, and 11 true matches were identified only by the RF model.

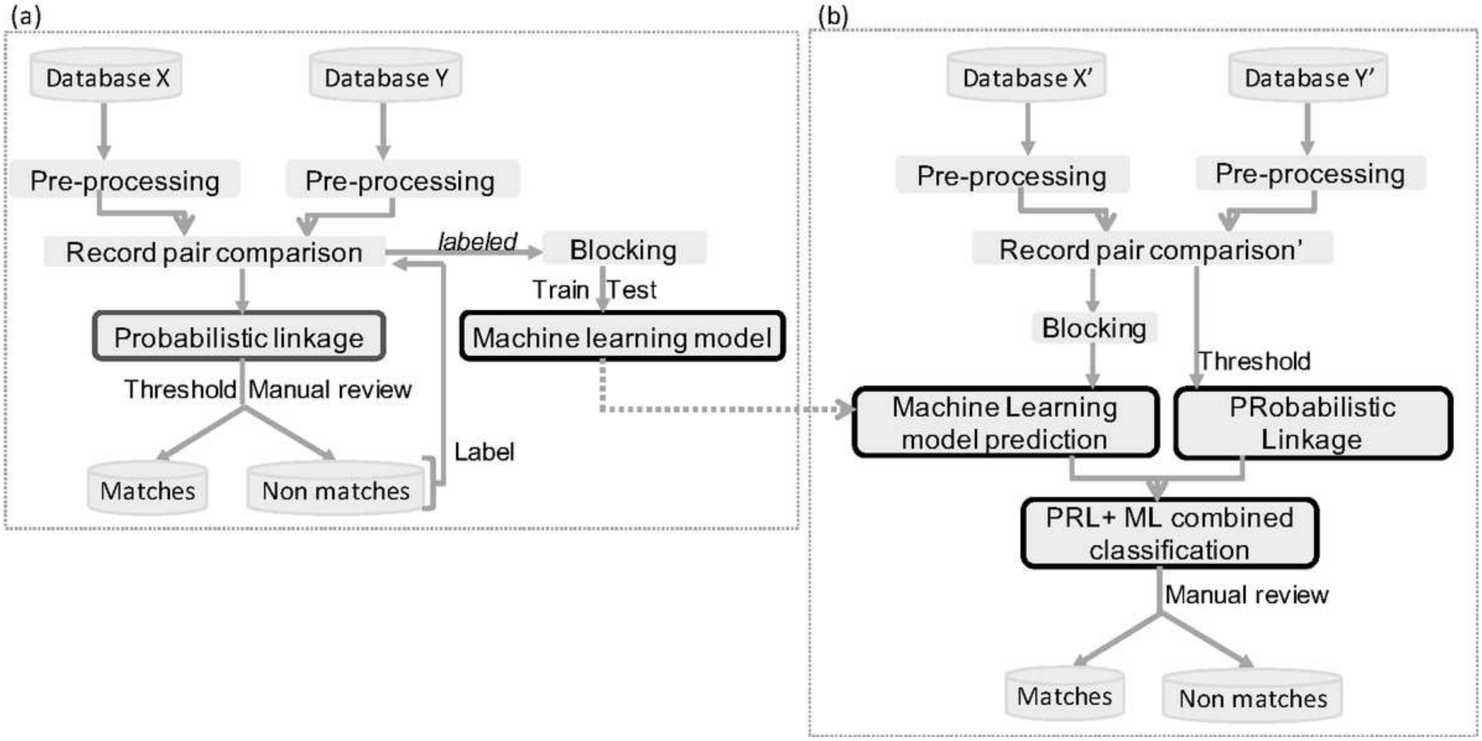

Figure 5

General overview of the hybrid record linkage process. (a) Probabilistic record linkage (PRL) followed by a stage of manual review is first applied to build a machine learning (ML) model. (b) The PRL+ML combined linkage is then used to classify the updated datasets (Record pair comparison from Database X' and Database Y'). The ML model obtained in (a) is used (dotted arrow) for the prediction in (b).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData20200807.docx