New metrics for assessing linkage quality in deterministic record linkage of health databases

Erwan Drézen, André Happe, Sandrine Kerbrat, Frédéric Balusson, Emmanuel Oger

New metrics for assessing linkage quality in deterministic record linkage of health databases

Erwan Drézen *[1], André Happe[2], Sandrine Kerbrat[2], Frédéric Balusson[2], and Emmanuel Oger[2]

[1] *CUBR*

[2] *EA REPERES, University of Rennes*

Abstract

Background Linking several datasets is becoming increasingly important for epidemiological research. However, assessing linkage quality can be challenging. This paper introduces a deterministic record linkage strategy that focuses on assessing linkage quality using new quality metrics.

Methods We developed a deterministic linkage strategy that systematically considers all combinations of individual identifiers. An exhaustive exploration of all variable combinations makes it possible to compute a new metric, referred to as robustness, and to generate a linkage cartography that precisely summarizes the linked pair characteristics. This cartography is central to our approach and makes it possible for the expert to easily accept/reject groups of linked pairs. The approach was tested on synthetic datasets staging a variety of possible linkage scenarios, and on two real-world studies (a registry database and a clinical trial).

Results Dataset simulations demonstrated very good accuracy with a limited impact of different factors (datasets size/ratio, overlap, and errors), scalability, and encouraging runtimes. Minima were greater than 0.95 for recall and greater than 0.99 for precision, whatever the scenario. Feasibility on real datasets was verified with good results: among 3985 patients from the registry, the algorithm found 3850 single linked pairs and 135 proposals with multiple candidates out of 504,795 candidates. After reviewing the linkage cartography, the expert validated 3783 linked pairs and a manual review of multiple candidates added 20 pairs, reaching a linkage rate of 95.4%. For the trial, only 2 records out of 129 were not linked among 22,426 candidates, as a result of early withdrawal (no information in the trial database), giving a linkage rate of 98.4%.

Conclusions The novelty of our approach is twofold: first, the linkage cartography provides a new way of classifying and comparing deterministic rules from the set of all possible rules and second, the approach is by design resilient to data corruption and can reach better recall than standard deterministic linkage strategies. Finally, good performance and scalability open the door to the linkage of very large datasets.

*correspondence: erwan.drezen@cubr.fr

1 Background

Linkage among healthcare databases, claims, and external data collections such as registries, cohorts or clinical trials is becoming increasingly important for conducting epidemiological research [1]. Access to complementary sources of information provides a comprehensive picture, as each dataset has some missing or insufficiently detailed data [2]. For example, the French National Health Insurance Information System (SNDS) contains claims-based data on dispensed drugs, on out-of-hospital laboratory tests, but without results, and on outpatient medical consultations, but without diagnosis or even the symptoms motivating the patient's consultation [3]; smoking and alcohol-use histories can only be deduced from hospital diagnoses or from specific drug deliveries, which is far from the required accuracy. Registries, cohorts or clinical trials collect medically validated data, but long-term information on patients' conditions and treatments over time is time-consuming and costly, and prone to attrition bias. Healthcare databases and claims are an neat way to extract data at low cost and without attrition. The linkage of the two sources could overcome lack of information and limit confounding biases.

Linkage algorithm accuracy is a key issue, with the implications of a trade-off between missing true matches and identifying false matches [4]. Classic linkage methods fall into two main classes:

- deterministic methods using a set of rules to binary classify matching pairs of records as related to the same individual or not.
- probabilistic linkage methods using a weight or score to estimate the likelihood that two records belong to the same individual [5].

Defining a set of rules is a key point in deterministic record linkage strategies and, when no gold standard is available, assessing linkage quality according to the chosen rules can be difficult. Strategies using a small set of combinations of specific personal identifiers have been assessed using a gold standard; a classification of the set of combinations is provided according to linkage rate and correct linkage rate [6] or sensitivity, specificity and positive predictive value [7, 8]. Deterministic strategies using

combinations of $N-1$ identifiers have also been compared to full deterministic linkage and probabilistic linkage using simulated datasets [9]. Similarly, match-keys are created by putting together pieces of identifiers to create unique keys, each match-key being designed to resolve a particular type of data inconsistency [10]. Reducing the number of match-keys can be achieved by skipping redundant match-keys in the hope of reducing computational load [11]. Actually, a strategy using *all* possible combinations of identifiers in order to classify and compare them does not seem to have been tested so far. It is worth noting that some probalistic linkage methods consider the full set of possible combinations but with different intent (e.g. calculating a suitable threshold setting) [12]. A series of several progressively less restrictive steps generated from combinations of identifiers is often used in deterministic linkage strategies and a link is assigned the step of the algorithm that made the link, known as the match rank [13, 14]. Usually, only links with low match rank are kept since they are expected to be trustworthy. This is however not fully satisfactory because match rank does not tell us if another link was close enough to be acceptable as well; a deterministic linkage strategy assessing *all* link/rank couples could determine "how far" a linked pair is from the other potential links and thus provide to the expert a new quality metric in addition to the match rank.

A known issue concerns the scalability of the record linkage process in the face of ever-growing datasets. Fast matching, in the meantime, is also a concern, since new applications (for example deduplication in identity management systems) require real-time answers rather than batch processing [15]. While scalability to large datasets has been generally addressed by blocking or filtering methods in order to mitigate the calculation burden, these adaptations expose to a reduction in true positive matches. This problem is exacerbated when the two datasets differ greatly in size, since the search space expands quadratically while the number of potential matches only evolves linearly.

Lastly, clear and transparent reporting of data linkage is mandatory. While data linkage evaluation generally rests on traditional measures such as sensitivity, specificity and false positive rates, further aspects such as the completeness and the quality of the data sources and the precise description of the linkage process should also be taken into consideration [16]. In addition, one should be able to verify whether the resulting linked dataset fits the researcher's needs and its intended uses, and also to determine whether the characteristics of the linked population differ or not from the group composed by the unmatched individuals [17].

In light of these challenges, we developed a deterministic linkage strategy which systematically considers all combinations of individual identifiers, while attempting to overcome the complexity inherent in an exhaustive search space of this type. This strategy could fit various situations of dataset size and data quality, while also addressing the two above-mentioned major issues [18]: accuracy and scalability. We evaluated the recall, precision, and computation time of this new algorithm across a variety of record linkage scenarios using simulated datasets and two real-world datasets.

2 Methods

2.1 Rationale

Let us consider a set S of individuals such that x belongs to S which will be called the source database and a set T of individuals such that y belongs to T which will be called the target database. Each individual in the two databases presents characteristics called variables such as $\{V_1, V_2, ...V_p\}$ for x and $\{V'_1, V'_2, ...V'_p\}$ for y. The linkage process consists in finding for each x in the source a single candidate y in the target, x and y sharing some common characteristics. In this case, we say that the process has found a linked pair (x, y).

Since a single common key may be not available between the source and the target databases, the record linkage process often relies on a combination of several variables used as a proxy for a single key, so-called quasi-identifiers (QIDs, see [19]). Intuitively, the more variables are available for linkage, the more likely is their association to provide a specific combination. So when a linked pair (x, y) is found with x and y sharing a large number of linkage variables, we can be more confident in the linkage result than when only very few variables are common to x and y. Moreover, if a linked pair (x, y) shares a set of variables such as $(V_1, V_2, ..., V_n) = (V'_1, V'_2, ..., V'_n)$, removing one variable without losing the uniqueness of (x, y) is a good indicator of the strength of the link between the two individuals.

This observation gives a scheme that is at the heart of our proposal, relying on a twofold mechanism. First, for a given x in the source database and for its given set of linkage variables $(V_1, V_2, ..., V_n)$, we look for a possible single y in the target database that could share the same combination of variables (i.e. perfect match) *or any proper subset of it*. If a y of this nature is found, the second mechanism can quantify "how far" the (x, y) pair is unique by finding how many of the variables used can be removed without losing its uniqueness. This second operation provides a quality metric to estimate the robustness of the linked pair at the most granular level, which completes the overall linking rate. Ultimately, the algorithm explores all possible pairs (x, y) of the Cartesian product SxT and most notably all possible combinations of K variables among the N available linkage variables.

This approach clearly belongs to the deterministic record linkage category and we will see how it overcomes the flaws that a standard deterministic approach can have.

2.2 Notations and definitions

We use the following notations :

- $|E|$ for the cardinality of the set E, i.e. the number of items in E
- $A \subset B$ for A as a subset of B.
- $A \cap B$ for the intersection of the sets A and B
- $\bigcap_{i \in I} A_i$ for the intersection of A_i for i belonging to a set of integers I.
- P_I for the powerset of set I (minus the empty set as a convention), the set of all subsets of I. For instance, the set $I = \{1, 2, 3\}$ has the powerset $P_I = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. Note that $|P_I| = 2^{|I|} - 1$.

A set of N available variables $V = \{V_1, V_2, ...V_N\}$ for the source database S and $V' = \{V'_1, V'_2, ...V'_N\}$ for the target database T are used for the linkage process; they are called *linkage variables*. In this paper, linkage variables entailing only structured data (e.g. ICD-10 diseases codes, ATC drug classes,..) are preferred to any free text entries (e.g. names, drugs labels,..). Indeed, since medico-economic databases like SNDS in France are natively pseudonymized before being granted to researchers, free text directly identifying variables (like names or addresses) cannot be used as linkage variables. In addition, care trajectories abound with temporal information acting as temporal fingerprints characterizing each patient [20]. An individual p in the source or target databases can be characterized by several variables such as V_1="I21 (ICD-10 code for acute myocardial infarction) diagnosed in November 2018" or V_2 ="C09CA03 (ATC code for valsartan) delivered in May 2019". Thus variables here enumerate all possible properties held by this particular individual, including timestamps, and each variable has an associated set of patients (e.g. all the patients diagnosed with acute myocardial infarction in November 2018 for variable V_1). It means that p belongs to the sets represented by the variables the given individual matches. Hereafter, a variable and its associated set of patients will be used interchangeably.

We will note x a patient in the source database S and y a patient in the target database T.

For one patient x or y, we define the *signature* as the set σ that carries the indexes of the variables the patient belongs to. By convention, the signature will be defined by:

- the variables V_i in the source for a patient x

$$\sigma(x) = \{i \in [1..N]; x \in V_i\}$$

- the variables V'_i in the target for a patient y

$$\sigma(y) = \{i \in [1..N]; y \in V'_i\}$$

The signature σ therefore represents the combination of variables matched by one patient. For better readability, it is possible to assign a letter to each variable and thus a signature can be read as a piece of text. For instance, for three variables, the letter "S" can be assigned for a sex code, the letter "B" for a birth date and the letter "H" for a hospital stay; in this context, it is thus possible to write $\sigma(x) = \{1, 2, 3\}$ or $\sigma(x) = \{S, B, H\}$, or simply $\sigma(x) =$ `SBH`. Finally, we define $\overline{\sigma}(y)$ as the difference between $\sigma(x)$ and $\sigma(y)$, i.e. $\overline{\sigma}(y)$ provides the variables available for x but not available for y. For instance, if $\sigma(x) =$ `BDHUG` and $\sigma(y) =$ `B.H.G`, then $\overline{\sigma}(y) =$`.D.U.` (note: the dot means the absence of a variable).

$P_{\sigma(x)}$ is the powerset of $\sigma(x)$ and carries all the combinations of variables given their indexes in $\sigma(x)$. The cardinality of $P_{\sigma(x)}$ is $2^{|\sigma(x)|} - 1$. For instance, if a patient x from the source belongs to the three variables `B,H,G` among the linkage variables {`B,D,H,U,G`}, then the patient's signature is $\sigma(x) =$ `B.H.G` and $P_{\sigma(x)} =$ {`B....`, `..H..`, `....G`, `B.H..`, `B...G`, `..H.G`, `B.H.G`}.

Now, given x and its signature $\sigma(x)$, the linkage algorithm tries to find a single patient y in the target database that "looks similar to" x in a specific way remaining to be defined. We therefore define the linkage function L that associates a non empty combination of variables J (i.e. their indexes) to a set of patients in the target database:

$$L(J) = \bigcap_{j \in J} V'_j$$

For J as a signature of a patient y, we have:

$$L[\sigma(y)] = \bigcap_{j \in \sigma(y)} V'_j = \bigcap_{\{j \in [1..N]; y \in V'_j\}} V'_j$$

so we always have $y \in L[\sigma(y)]$ and $L[\sigma(y)]$ is never empty but can contain more than one patient. We also define $T_{\sigma(x)}$ as the subset of $P_{\sigma(x)}$ that contains signatures of patients in T:

$$T_{\sigma(x)} = \{J \in P_{\sigma(x)}; \exists y \in T; J = \sigma(y)\}$$

In the linkage context, we are interested in finding the combinations of variables that lead to a single patient y. We therefore define a linked pair as follows:

$$(x, y) \textit{ is a linked pair } \iff \begin{cases} \sigma(y) \in T_{\sigma(x)} \\ L[\sigma(y)] = \{y\} \\ |\sigma(y)| = \max\limits_{J \in T_{\sigma(x)}} (|J|) \end{cases}$$

A linked pair is defined by a signature $\sigma(y)$ that is a subset of $\sigma(x)$ maximizing the number of linkage variables used. A linked pair (x, y) will be *unique* if y is the only patient in T whose signature reaches the maximum cardinality in $T_{\sigma(x)}$.

In a conventional deterministic linkage algorithm, one first tries to see if $L(\sigma(x))$ produces a single patient y (i.e. a perfect match). If this is the case, then a linked pair has been found. However in practice, the following can occur:

- $|L(\sigma(x))| > 1$ which means that there is no single candidate because there is not enough available information to reach a single patient.

- not all the available variables given by $\sigma(x)$ will be actually used by the candidate y because there may exist $i \in \sigma(x)$ such that $y \notin V'_i$ although $y \in V'_j$ for all $j \in \sigma(x)$ with $j \neq i$. This could happen in case of missing information in the target database (i.e. y has not the property V'_i) or because there is a difference in value for the property between the source and the target databases (a date of 12/06/2013 in the source and 13/06/2013 in the target). Consequently the cardinality of $L(\sigma(x))$ will be zero and no patient y is found. In this case, the user of a classic deterministic algorithm can choose to drop one or more variables and relaunch the algorithm in the hope of obtaining a single patient with less information (namely $|\sigma(x)| - d$ variables where d is the number of dropped variables), which amounts to trying combinations of variables in $P_{\sigma(x)}$. This multi-step approach is not satisfactory because (i) the list of rules for dropping one or more specific variables among others will be rather empirical and (ii) the user is unlikely to try manually more than a few combinations other than $\sigma(x)$ because of the potential huge number $2^{|\sigma(x)|} - 1$ of combinations.

Our approach proposes to automatically try all combinations of variables of $P_{\sigma(x)}$ without any a priori choice. This process guarantees *a linked pair will be found if there is one to be found with the available information*, which means that the algorithm is likely to produce good recall.

Obviously, this approach implies computing $2^{|\sigma(x)|} - 1$ intersections $L(J)$ of potentially huge datasets for each candidate pair (x, y) of the Cartesian product SxT. So the algorithm complexity is $O(|S|.|T|.2^N)$, N being the number of linkage variables. This major computational pitfall will be addressed later on.

2.3 Quality metrics

For a linked pair (x, y), we define two sets:

$$\begin{cases} Y = \{J \in P_{\sigma(y)}; L(J) = \{y\}\} \\ Z = \{J \in P_{\sigma(y)}; L(J) \neq \{y\}\} \end{cases}$$

Thus, Y is the set of combinations $J \subset \sigma(y)$ that leads to the one patient y, so $|Y|$ gives the number of ways to obtain y; Y cannot be empty because $\sigma(y) \in Y$ by construction. Z includes combinations that tend to lose the uniqueness of y. One can note that $|Y| + |Z| = |P_{\sigma(y)}|$.

Our approach will "reduce" the signature $\sigma(y)$ by removing one or more variables from it and see whether or not the uniqueness of y still stands. We thus define the notion of robustness for a linked pair:

$$robustness(x, y) = \max_{J \in Y}(|J|) - \max_{J \in Z}(|J|) - 1$$

$\max_{J \in Y}(|J|)$ represents the longest combination of variables that gives the single patient y, which is in fact $|\sigma(y)|$ and $\max_{J \in Z}(|J|)$ represents the longest combination of variables that does not give the one patient y. Note that $robustness(x, y) \geq 0$. We can also write:

$$robustness(x, y) = |\sigma(y)| - \max_{J \in Z}(|J|) - 1$$

Thus the robustness provides a quality trust indicator at the granular level of each linked pair (x, y) showing how many variables can be removed from the signature $\sigma(y)$ without losing the uniqueness of y; the higher is the value the more trustworthy is the linked pair. For instance, with a robustness equal to one, we can remove any variable from $\sigma(y)$ without losing the uniqueness of y; a robustness equal to two make it possible to remove any combination of two variables without losing the uniqueness of y, etc... On the other hand, a robustness equal to zero means that we cannot remove an arbitrary variable from $\sigma(y)$ without losing the uniqueness of y. We also defined a global metric as the linkage robustness for the whole linkage process:

$$linkage\ robustness = \frac{\sum_{pair(x,y)} robustness(x, y)}{number\ of\ pair(x, y)}$$

Finally, we introduced a metric concerning the use of the available information. If $\sigma(y) = \sigma(x)$ (which implies $|\overline{\sigma}(y)| = 0$), then all the available information has been used to find y for a patient x, i.e. a perfect match. On the other hand, we may find a signature $\sigma(y)$ that is a proper subset of $\sigma(x)$, meaning that one or more variables V'_i have not been used to find y (i.e. $|\overline{\sigma}(y)| > 0$) because of data discrepancies between the source and the target databases. In order to measure how well the available information from the two databases has been used, we define the following metric:

$$information\ usage = \frac{\sum_{pair(x,y)} |\sigma(y)|}{\sum_{pair(x,y)} |\sigma(x)|}$$

Obviously, the value lies between 0 and 1 and the higher is the value the better is the use of the available information. The value would be 1 with "perfect" data. Poor values however evidence potential issues in the data. We also define the *information missing data* as $(1 - information\ usage)$.

We can easily extend the definition in order to have the information use for a subset I of variables:

$$information\ usage\ (I) = \frac{\sum_{\substack{pair(x,y) \\ \forall i \in I;\ i \in \sigma(y)}} |I|}{\sum_{\substack{pair(x,y) \\ \forall i \in I;\ i \in \sigma(x)}} |I|}$$

For instance, for $I = \{k\}$, we can obtain the information use for the k_{th} linkage variable.

2.4 Linkage cartography

Our approach provides a convenient way to assess the result of a linkage run by grouping linked pairs (x, y) on certain criteria. For instance, a table giving the number

of linked pairs grouped by $[robustness(x,y), \sigma(x), \sigma(y)]$ provides a cartography of the linked pairs, and is very informative for the expert to accept/reject the linked pairs of one group just by evaluating the characteristics of the group. Table 1 displays an example of a linkage cartography with five linkage variables.

The linkage cartography provides a classification of deterministic rules from the set of all possible rules and makes it possible to compare rules with the help of the robustness metric. To our knowledge, such a rigorous approach to selecting matching rules and comparing them represents an improvement upon using standard deterministic linkage.

For a given linkage cartography, here are some possible decisions:

- if the robustness value is high enough (≥ 1), the corresponding groups should be confidently accepted
- with zero robustness and $\sigma(x) = \sigma(y)$ (or equivalently $|\overline{\sigma}(y)| = 0$), no information is missed and the groups can also be accepted.
- when $|\sigma(y)| < |\sigma(x)|$ (or equivalently $|\overline{\sigma}(y)| > 0$) with robustness equal to zero, the knowledge of the data by the expert should be used to accept/reject the corresponding groups; indeed, if $\sigma(y)$ does not hold crucial information in the expert's eyes, the underlying linked pairs should be rejected.

			robustness			
	$\sigma(x)$	$\overline{\sigma}(y)$	0	1	2	total
1	BDHUL	.....	31	1440	50	1521
2	BDHUL	.D...	182	11		193
3	BDHUL	...U.	76	6		82
4	BDHUL	L	28	4		32
5	BDHUL	B....	12			12
6	BDHUL	BD...	5			5
7	BDHUL	..H.L	5			5
8	BDHUL	.D.U.	2			2
9	BDH.L	.....	7	128	1	136
10	BDH.L	.D...	17			17
11	BDH.L	L	9			9
12	BDH.L	B....	5			5
13	B.HUL	.....	1395	63		1458
14	B.HUL	B....	3			3
15	BDHU.	.....	2	10		12
16	BDHU.	B....	1			1
17	BDHU.	.D...	1			1
18	BD...	.....	60			60
19	B.HU.	.....	7			7
		total	1848	1662	51	**3561**

Table 1: Example of a linkage cartography with 5 linkage variables (B birth date, D death date, H hospital stay, U emergency unit, L location). The table describes each group of linked pairs defined by the signature $\sigma(x)$, the missed variables $\overline{\sigma}(y)$ and a robustness value. For instance, there are 11 linked pairs with robustness equal to 1 with `BDHUL` available in the source database and with no match for the `D` variable in the target database. For instance, the expert could choose to reject lines 6 and 14 because of the little information contained in $\sigma(y) =$ `HUL`; on the other hand, line 18 could be accepted despite the low number of variables because birth and death dates can reasonably be considered sufficient to identify a single patient.

The linkage cartography approach provides the expert with material (i) that makes sense in relation to her/his knowledge of the data, for instance signatures, and (ii) this makes it possible to accept/reject certain linked pairs. The expert can keep control over the final result and is not required to blindly accept all the linked pairs generated.

In practice, this cartography of the linked pairs often comprises a few dozens of groups only and requires little time to be read and analyzed.

2.5 Implementation of the approach

As we have seen, our approach provides attractive properties but presents a major drawback: for each x in the source, one has to compute the intersection set $L(J) = \bigcap_{j \in J} V'_j$ for each subset $J \subset \sigma(x)$. Since there are $2^{|\sigma(x)|} - 1$ such subsets and some V'_j can contain many patients for large populations, the computational task can be huge. A simple experiment of the approach written in Python language produced high execution times. For instance, finding a patient x in a $3,000,000$-patient target database with 6 linkage variables took 12.6 seconds and a full record linkage between a $30,000$-patient database and a $3,000,000$-patient database would take more than 4 days. Obviously, if we had used more than $|\sigma(x)| = 6$ variables, execution times would have dramatically increased as a result of the number of combinations to be explored which grows like $2^{|\sigma(x)|}$.

The conclusion of this crude experiment is that a naive implementation of the approach is a dead end in terms of execution time, except for small populations or limited sets of linkage variables (i.e. small $|\sigma(x)|$). However, to go further, we can make the following observations:

1. the algorithm should not directly rely on classic database management systems (like SQL) for database storage because these systems are not well suited to efficiently computing the intersection of several sets. The algorithm *could* be based on statistical systems (SAS, R, ...) but their inner storage systems are finally not so different from database management systems.

2. in order to reduce execution time for computing $L(J)$, one should try to reduce the size of the V'_i sets as much as possible. One way to achieve this goal is to organize the linkage variable so they do not include too many patients. For instance, one could use the variable "men born in 1937" instead of two variables "men" and "born in 1937"; the former is likely to contain fewer patients than the latter. Moreover, it also reduces the number of variables for $\sigma(x)$; since the algorithm complexity contains a $2^{|\sigma(x)|}$ term, it is not a bad thing to reduce $|\sigma(x)|$.

3. the algorithm *should* be implemented with an efficient language like C++ which takes full advantage of the available computer resources (like multicore architecture, SIMD, etc...)

4. the algorithm should not try *to compute* $L(J)$ *for each* $J \in T_{\sigma(x)}$ because this is too costly. After

all, the important values for discovering a potential linked pair are $|L(J)|$ so an algorithm that efficiently computes $|L(J)|$ without computing explicitly $L(J)$ for all combinations J will provide significant improvement compared to the naive implementation; $L(J)$ can be computed explicitly only when the maximum combination J has been found. In other words, the algorithm should first prove the existence of a linked pair and then find it.

Our current implementation of the approach proposes an effective solution for these different points. In order to achieve good performances, the two datasets need to be first transferred from their initial storage (SQL, CSV, ...) to an index. This step needs to be performed only once and the associated runtime is linear to the dataset sizes. After data preparation, the algorithm takes one patient in the source and, given this patient's identifiers, looks for its counterpart in the target database; in other words, it processes the Cartesian product of the source with the target in a sequential manner (one x versus many y) and does not require all potential pairs to be handled at the same time. Our implementation makes it possible to efficiently use multicore architecture of modern computers which can greatly reduce execution times.

3 Study design

The current implementation of the approach has been tested on both synthetic and real datasets.

3.1 Synthetic datasets

In order to test our approach, we designed a variety of possible linkage scenarios with varying dataset sizes. To mimic administrative health system, registry and clinical trial data encountered in research linkage, we created the following basic datasets as combinations of size and type:

- two large datasets representing the French population in general: 3.000.000 records (named H for huge) based on the characteristics of French patients with cancer (2018 yearly report from INCA, Institut National Du Cancer), and 300.000 records (named L for large) based on the frequency distributions of sex, years of birth, and postcodes of the population living in North-Western Brittany in 2015 extracted from the French National Institute for Statistical and Economic Studies.

- two medium datasets representing a specific population group like a registry: 30.000 records (named I for intermediate) and 10.000 records (named M for medium) based on the Brest Stroke Registry statistics

- two small datasets representing a more specific population such as individuals recruited in a trial: 1.000 records (named S for small) based on the description of an nationwide oncological study [21] and 100 records (named T for tiny) based on a blueprint of Artome, a cancer trial enrolling 129 adults

The frequency distributions were used to create the simulation datasets and basic sets containing a single identifier, and the chosen linkage variables were generated (see supplementary material). These linkage variables are close to the typology of linkage variables used in a conventional linkage project involving SNDS; they include socio-demographic variables (sex, birth and death dates, postcodes) and care trajectory information (hospital stays for instance). The synthetic datasets were generated using a Python-based program.

From these dataset definitions, we created four scenarios combining different sizes and ratios :

- scenario 1: Huge / Intermediate (ratio 100:1)
- scenario 2: Large / Medium (ratio 30:1)
- scenario 3: Intermediate / Small (ratio 30:1)
- scenario 4: Medium / Tiny (ratio 100:1)

In order to assess the effect of errors, the datasets were created with different error rates: (i) datasets with clean records and (ii) datasets with 5% of corrupted records, with two types of error: missing values or errors defined as adding two days to some dates; these kinds of error were combined with two variable types (death date or marker event date)

We also created datasets that simulated the absence of entire records. For instance, we built target datasets containing only 80% of a source dataset plus many other records (see Figure 1). The underlying idea was to check that the missing records in the target dataset were not found and that the algorithm was not attracted to certain false positive records. During our tests, we used both an overlap of 100% (i.e. no missing records in the target database) and an overlap of 80% (80% of the source database in the target database).

As the sampling of records for overlap and the application of errors were random processes, we extended the number of simulation datasets to 30 for each data source combination. In the end, 2160 linkages have to be processed corresponding to 2160 different configurations.

The code for creating synthetic datasets is available at https://github.com/erwandrezen/datasimulation.

3.2 Real Datasets

We considered two real-world matching scenarios.

The first study used nine years 2009-2017 contained in the Brest Stroke Registry (4264 stroke events for 3985 patients). The Brest Stroke Registry collects data mostly from three neighboring hospitals, but also out-patient data (from general practitioners, three neurologists in private practice, private radiology centers, and nursing homes) and data for death certificates providing data for fatal stroke among non-hospitalized subjects [22]. Each

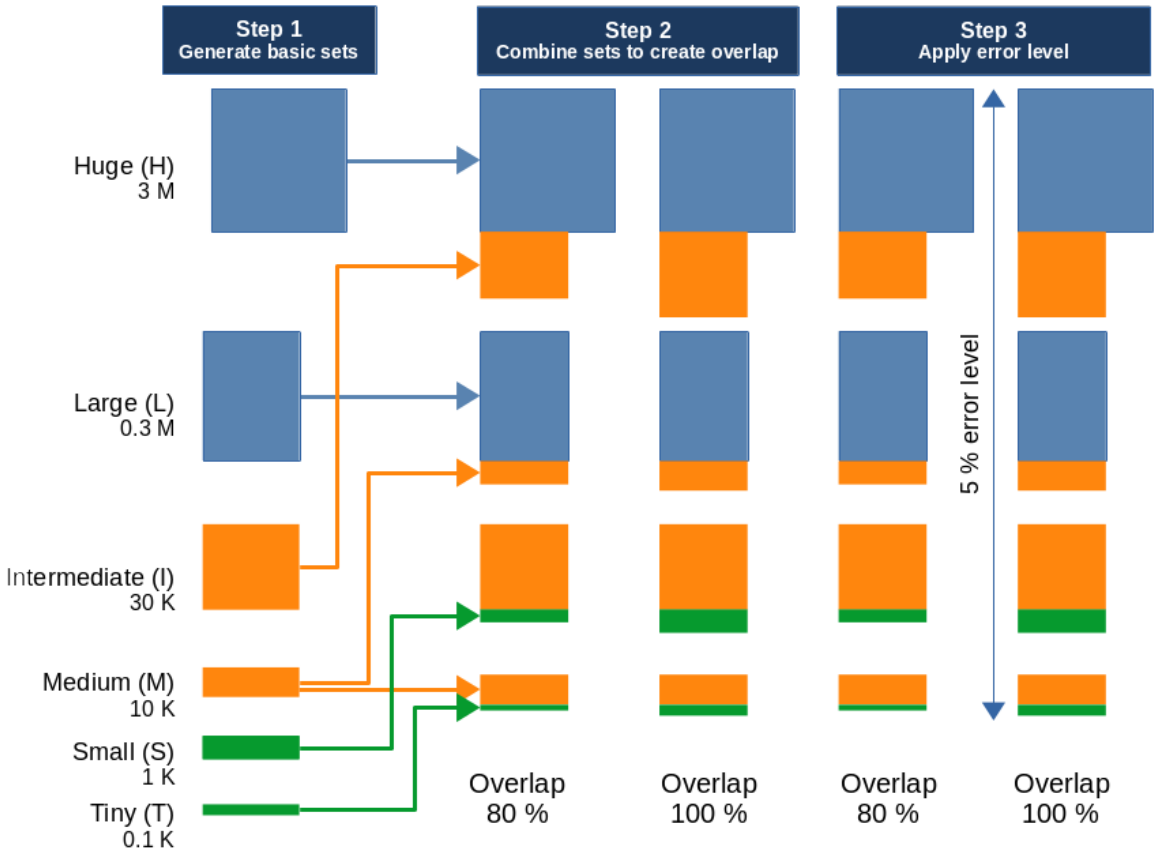

Figure 1: Generation of target synthetic datasets

Characteristics	Brest Stroke Registry	Artome trial
Records number	**4 264**	**129**
Age	$75,6 \pm 13,9$	60.4 ± 7.8
min-max	$16 - 104$	$38 - 78$
Female, n (%)	2266 (53.1%)	17 (13.2%)
Geographical unit	1517 (35.6%)	65 (50.4%)
Top 5 frequency, n (%)	192 (4.5%)	15 (11.6%)
	190 (4.5%)	7 (5.4%)
	158 (3.7%)	6 (4.6%)
	128 (3.0%)	6 (4.6%)
Index event date		
min	01/01/2009	26/06/2013
max	31/12/2013	25/10/2018
Death, n (%)	2440 (57.2%)	29 (22%)

Table 2: Descriptive statistics of real datasets.

hospital admission for stroke yielded a distinct entry. We were provided access to a subset of French National Health Insurance Information System data containing 504,795 adult subjects (> 18 years) living within the Brest area, with at least one health-care reimbursement in 2009-2017. We preprocessed the two datasets to harmonize the variable names and values. Preprocessing entailed dropping variables not used in the linkage runs, and extracting year of birth (YOB), and month of birth (MOB) from DOB. We checked for implausible values and converted implausible values into missing values. We created a new emergency entry date variable equivalent to the hospital entry date with emergency entry mode and we created a new entry date variable equal to the hospital entry date for strokes. After data preprocessing, we proceeded with duplicate checks.

The second study used a cancer trial on oropharyngeal tumours named Artome and run by the Cancer Treatment Centre Eugène Marquis in Rennes. The project was to link this trial (129 patients, mostly men) to a dataset extracted from the French National Health Insurance Information System comprising 22,426 adults selected on the basis of having cancer cared for in one of the 11 participating clinical centers. The same preprocessing as for the Brest Stroke Registry was performed on both the Artome data and the SNDS extracted data.

Table 2 provides descriptive statistics for the two studies and linkage variables are given in the supplementary material.

3.3 Statistical analysis

For the synthetic datasets, we measured for each algorithm run: recall (i.e. sensitivity, the proportion of true matches identified by the algorithm) and precision (i.e. positive predictive value, the proportion of algorithm

matches that were true matches). We also measured execution times in order to assess the scalability of the algorithm. It can be noted that each matching scenario was simulated 30 times, and we calculated the mean and standard deviation of recall and precision across these replicates.

To estimate the impact of the pre-specified factors (size/ratio, overlap, error, ...), we considered regression models for classification probabilities. We used a generalized linear model for binary outcomes (binomial distance and log link), and performed separate models for recall and precision. In addition, we measured the computational performance in terms of their average runtime across the replicates. We used a linear regression model with a log-transformation of runtimes.

For linkage verification between real-world datasets, we compared the characteristics of linked and unlinked records. We used this method of quality appraisal because all of the records in the registry (or oncology trial) were expected to link. We used standardized differences, calculated as the mean difference divided by the standard deviation.

4 Results

4.1 Synthetic datasets

Figure 2 displays recall (true positive rate) across the four linkage scenarios according to event date corruption. A larger dataset size and/or a larger ratio between datasets (100:1 vs. 30:1) and marker event date corruption (error or missing value) were associated with a statistically significant reduction of recall. On the other hand, a reduced overlap (80% vs. 100%) or death date corruption (error or missing value) did not significantly affect recall. Association estimates between pre-specified factors and recall (parameter α_i) or precision (parameter β_i) through generalized linear models (binomial distance and log link) can be found in the supplementary material.

Figure 3 displays precision (positive predictive value) across the four linkage scenarios according to overlap and event date missing data. A larger dataset size, a reduced overlap (80% vs. 100%), and marker event date missing were associated with a statistically significant reduction in precision. On the other hand, a larger ratio between datasets, and death date corruption (error or missing value) did not significantly affect precision.

It can be noted that minima for recall were greater than 0.95 and minima for precision were greater than 0.99 whatever the scenario.

For each scenario, the linkage robustness was stable whatever the overlap or data corruption. Min/Max values are $(0.72, 0.73)$ for scenario 1, $(0.91, 0.95)$ for scenario 2, $(1.73, 1.77)$ for scenario 3 and $(1.00, 1.01)$ for scenario 4. The better value for scenario 3 can be explained by a better discriminant power of the C postcode variable due to an even distribution of geographical units in both source and target datasets.

We measured runtimes for each scenario on an Intel Core i7-4790 with 4 physical CPU cores and 16 GBytes of memory. The measure takes into account only the algorithm runtime and not the time required to prepare the two datasets in a suitable format for the current implementation. The measure takes into account the overall algorithm complexity in $O(|S|.|T|.2^N)$. A larger dataset size and/or a larger ratio between datasets (100:1 vs. 30:1), a reduced overlap (80% vs. 100%), and event date corruption (error or missing value) statistically increased runtimes (see the supplementary material). Association estimates between pre-specified factors and computational performance (log-transformation of runtime) using a multivariate linear regression model can be found in the supplementary material. Overall, runtime was always less than 2 seconds even for scenario 1 (3M vs 30K patients). It can be noted that the current implementation uses all the available CPU cores, so runtimes would decrease on a computer with more CPU cores.

4.2 Real Datasets

4.2.1 Brest Stroke Registry

Among the 3985 patients in the registry, the algorithm found 3850 unique linked pairs and 135 proposals with multiple candidates. After reviewing the linkage cartography, the expert validated 3783 linked pairs. A clerical review of the multiple candidates added 20 linked pairs, so the total number was 3803, which represents 95.4% of the registry. The linkage robustness was 2.36 and the information usage was 95.5% (i.e. 4.5% missing data).

The linkage cartography can be found in the supplementary material. It comprises 58 lines with a robustness ranging from 0 to 5. If we aggregate all the lines with no missed variables (i.e. $|\overline{\sigma}(y)| = 0$), we obtain 3058 linked pairs (see table 3), i.e. 76.7% of the registry. Roughly speaking, the exhaustive exploration of all variable combinations enabled an increase from 76.7% to 95.4% of the linkage rate, i.e. we gained 18.7% more linked pairs than just by looking for patients sharing exactly the same information between the source and the target. If we aggregate all the lines with $|\overline{\sigma}(y)| \leq 1$, we have 3472 linked pairs missing one or zero variables, i.e. 87.1% of the registry and with $|\overline{\sigma}(y)| \leq 2$, we have 3589 linked pairs, i.e. 90.1% of the registry. Clearly, we can see here the advantage of scanning all the variable combinations. All the linkage percentages per number of missed variables can be found in the supplemental material. The runtime was 0.6 second which is consistent with the data volumetry of scenario 2 involving synthetic datasets.

Linkage verification was performed by comparing the distributions of gender, age, type of hospital admission, and death between the resulting linked dataset and the unlinked dataset (see table 4).

4.2.2 Artome trial

For the Artome trial, the two databases to be linked shared many common timestamps (radiotherapy and

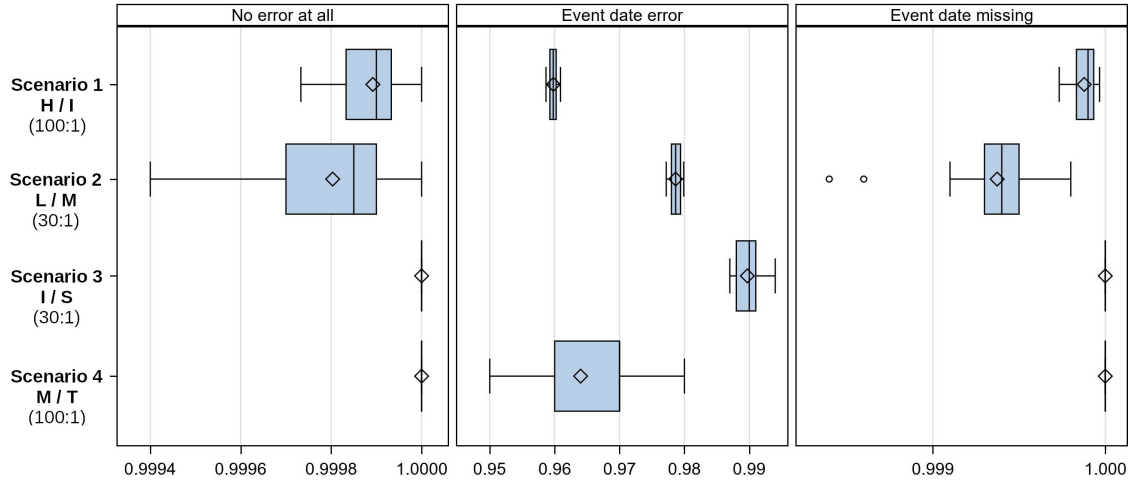

Figure 2: Recall across the four linkage scenarios according to event date corruption.
(H for Huge, I for Intermediate, L for Large, M for Medium, S for Small, T for Tiny)

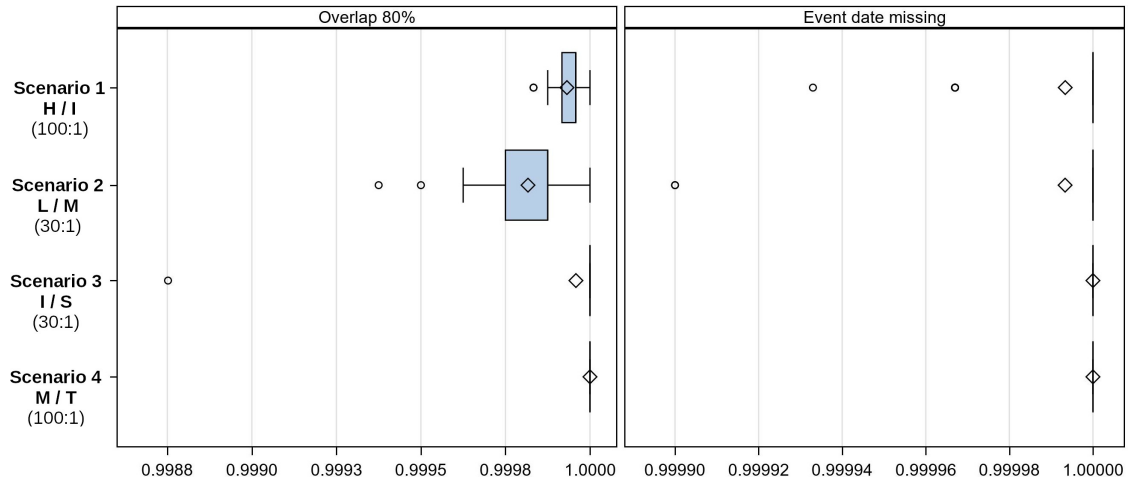

Figure 3: Precision across the four linkage scenarios according to overlap and marker event date missing data.
(H for Huge, I for Intermediate, L for Large, M for Medium, S for Small, T for Tiny)

Project name	**Brest Stroke Registry**	
Patients number		
source	Registry	3985
target	SNDS	504795
ratio		0.007894
Linked pairs, n (%)		
≤ 0 missed var		3058 (76.7%)
≤ 1 missed var		3472 (87.1%)
≤ 2 missed var		3589 (90.1%)
all		**3803 (95.4%)**
Robustness		2.36
min-max		$0-5$
Linkage Variables		
number		9
missing data		4.5%
Runtime (seconds)		0.6

Project name	**Artome trial**	
Patients number		
source	Trial	129
target	SNDS	22426
ratio		0.00575
Linked pairs, n (%)		
≤ 0 missed var		61 (47.3%)
≤ 1 missed var		102 (79.0%)
≤ 2 missed var		120 (93.0%)
all		**127 (98.4%)**
Robustness		4.18
min-max		$2-7$
Linkage Variables		
number		12
missing data		7.8%
Runtime (seconds)		0.01

Table 3: Linkage global indicators for the real-world studies; the associated linkage cartographies can be found in the supplementary material

Characteristics	Linked patients	Unlinked patients	Standardized difference
Female, n (%)	2024 (53.2%)	100 (54.9%)	0.03
Age min (sd)	79.4 ± 12.5	79.2 ± 15.6	0.01
min-max	$18 - 104$	$16 - 102$	
Male, n (%)	1779 (46.8%)	82 (45.1%)	
Age min (sd)	71.5 ± 13.7	66.4 ± 16.4	0.33
min-max	$17 - 98$	$16 - 96$	
Type of admission			1.11
Hospitalised	3907 (95.9%)	106 (56.1%)	
Emergency	68 (1.7%)	4 (2.1%)	
Death certificate	13 (0.3%)	6 (3.2%)	
No hospitalisation (ex: radiology)	87 (2.1%)	73 (38.6%)	
Death, n (%)	2178 (57.1%)	85 (46.7%)	0.21

Table 4: Brest Stroke Registry : Characteristics of linked and unlinked patients. There was a strong imbalance for admission type. Obviously, patients who were not hospitalised but did suffered from stroke had few non missing variables in the stroke registry: date of birth, sex, and the date of stroke has no obvious counterpart in the other database. A patient who died with the diagnosis of stroke on death certificate and who was enrolled as such in the registry would have only a date of death on top of a date of birth and sex. For those cases, we did found at least two candidates for each patient enrolled in the registry, but we could not definitively choose which one was the true candidate.

chemotherapy events, scans, ...), so the record linkage was expected to provide good results. Only 2 records among the 129 records were not linked due to a lack of information in the trial (no treatment date), which provides a linkage proportion of 98.4%. All the other 127 records were linked with good (R=2) to very good (R=7) robustness. A minimum robustness of 2 indicates that the linked pairs are very reliable. The linkage robustness was 4.18.

On the other hand, the information usage was 92.2% (or 7.8% missing data) and not as good as the Brest Stroke Registry project, but since a lot of information was available for Artome, we could afford to use variables that were not as accurate as for the Brest Stroke Registry project. As a result, the number of linked pairs with no missed variable (i.e. with $|\overline{\sigma}(y)| = 0$) was only 61 compared to the total number 127. This can be explained by the large number $N = 12$ of linkage variables with some of them considered as moderately accurate by the expert, which implies that in many cases perfect matches could not be found.

The linkage cartography contains 49 lines and is given in the supplementary material.

Table 3 summarizes global indicators for the two real-world record linkages. The information usage for each linkage variable is also provided in the supplementary material.

5 Discussion

5.1 Main findings

Our study showed that the approach has very good accuracy with a limited impact of some factors (dataset size/ratio, overlap, and errors), scalability, and encouraging runtimes, as demonstrated by the dataset simulations. Feasibility on real datasets with high record linkage and good robustness was also shown on two studies representative of record linkages involving the French National Health Insurance Information System (SNDS). Our results showed that the approach is resilient to data corruption (errors and especially missing data) by way of an exhaustive exploration of the information, and can considerably increase the number of linked pairs compared to a more conservative approach in terms of information exploration.

Of course, good resilience to missing data can occur only if there are enough linkage variables. As we have seen for the Artome project, this may not be so uncommon for oncology trials for instance, since there are many temporal events in care trajectories in this kind of pathology. More generally, the approach could open the way to new habits in deterministic record linkage: using as many linkage variables as possible, the algorithm will deal with them *even if they are not all very accurate*. However, one should take care not to use only dubious variables and to deal with variables of which at least half are reliable rather than less accurate variables, which could be a good tradeoff. As a result, the robustness of each linked pair can be improved even by the less accurate variables. It can be noted that information usage per variable could be used as an estimator of variable reliability. Nevertheless, one must keep in mind that the algorithm has to cope with the exploration of 2^N combinations, so that unrealistically large values for the number N of linkage variables should be avoided. $N = 15$ seems to be a practical maximum choice for keeping reasonable performances. For Artome, we used

$N = 12$ linkage variables which means that $2^{12} = 4096$ combinations had to be explored for each candidate pair (x, y) of the Cartesian product SxT. The observed runtime was low (0.01 second) but essentially because the datasets were small enough. Future work could assess more precisely the impact of N and dataset sizes on performance.

Results on synthetic data show that data errors lead to lower recall than missing data. Indeed, a variable with an erroneous value cannot be found in the target signature of the true positive y but could be found in the target signature of a potentially wrong candidate y' (meaning that $|\sigma(y)|$ lost 1 and $|\sigma(y')|$ gained 1). If y' has enough variables in its signature, it can compete with y, making y lose its uniqueness. On the other hand, missing data will not be found in any target signature and therefore will have less impact on recall than data errors.

5.2 Strengths and limitations

A first strength of the proposed method is its ability to produce quality metrics useful for the expert's decision. For instance, linkage cartography combines information on source and target signatures, missing variable information and also robustness, in a synthetic manner; the expert can then use her/his knowledge of the data to construct rules that make it possible to accept/reject linked pairs and thus have control over decision-making. Moreover, we believe that robustness, as the number of variables that can be dropped without losing uniqueness of a linked pair, is easy to understand for the expert and can be used to assess linkage quality.

A second strength is its ability to reach high linkage rate in a deterministic record linkage context. Indeed, our approach guarantees a linked pair will be found if there is one to be found with the available information. Although the incentive was to design new quality metrics, the definition of the notion of robustness made it clear that the need to explore the whole set of variables combinations would have a positive impact on recall compared to a strict deterministic approach with a pre-defined set of matching rules. This fact is clearly observed on the two real-world studies (see table 3): we gained 18.7% and 51.1% linked pairs compared to a deterministic approach with perfect match, we gained 8.3% and 19.4% linked pairs compared to a deterministic approach with $N - 1$ identifiers, etc... As proposed previously, the intent to use as many linkage variables as possible would be ruined in a strict deterministic approach in case of data corruption. Indeed, the probability of having perfect or nearly perfect matches (i.e. null or low $|\overline{\sigma}(y)|$) would strongly decrease with a larger number of variables with corrupted data. Again, our approach will not be affected by data corruption in this type of context and will still have a chance of finding a linked pair.

A third strength is the ability to quantify "how well" information has been used in the source and target databases. Indeed, information usage tells us how many times linkage variables available in the source database have been actually found in the target database. It should be noted that information usage can be accurately computed because the approach ensures that a linked pair will be found if it can be found. The fact is that poor values for information usage evidence potential issues in the source and/or target databases. For the two real-world studies, information usage seemed good enough with 95.5% and 92.2%. At a more granular level, it is also possible to estimate the information usage for each linkage variable. The expert can compare this information with her/his data knowledge. For instance in the Artome project, it was expected that the "first planification" linkage variable will be hard to map between the source and target databases; its information usage (77.5%, see supplementary files) is indeed the lowest for the set of linkage variables. On the other hand, information usage for the "date of death" linkage variable of the Brest Stroke Registry project was 88.9% and lower than expected; it was eventually explained by the fact that this linkage variable was harder to build in the target database than expected. In summary, the expert can use information usage for a better understanding of the source and target datasets.

A fourth strength is scalability, as the current results show low runtimes even on large datasets. For instance, scenario 1 took less than 2 seconds to link a dataset with 30,000 patients to a dataset with 3,000,000 patients on a 4 CPU core computer. Today, the current implementation can use all the CPU cores of one computer but cannot use the power of a computing grid. However, this implementation could be extended to take advantage of a computing grid for extreme record linkage projects involving very large datasets.

On the other hand, the limitations of our approach are:

- only structured data can be used for linkage variables which precludes the use of free text like names or addresses. As a matter of fact, our studies mainly deal with pseudonymized healthcare databases holding structured data (ICD codes, ATC drug codes, postcodes, etc...) associated to temporal information. We believe that such temporal events in care trajectories are more useful in our specific data linkage context because (i) an event defined by a code and a date provides discriminant power which helps the algorithm to be attracted to the true positive record, and (ii) in case of oncology trials and registries for instance, many temporal events can be found in each care pathway, which can greatly improve robustness.

- our approach does not allow partial agreement on identifiers like ICD codes, ATC drug classes, etc... On the other hand, dates can be compared $\pm d$ days, where d can be defined by the expert.

- our approach is only appropriate for 1:1 linkage, i.e. each record in the source database has at most one

record in the target database. Although it is possible to find multiple target records for one source record, it is no more possible to compute robustness in such a case.

- we used the notion of source and target databases, which creates an asymmetry in the process. The algorithm itself takes one x from the source and looks for its counterpart in the target. The resulting drawback is that the algorithm does not check whether the same y of the target dataset can be linked to a different x in the source dataset. An extra verification should be performed by the expert to check whether any such linked pairs have been found.
- we have currently not fully compared our approach to probabilistic linkage which will be addressed later on in a dedicated paper. It is worth noting that our approach uses an exhaustive exploration of all identifiers combinations and does not require blocking techniques. Preliminary results seem to show an advantage for our approach in terms of execution time or required computer resources when the volume of input data becomes large. In some respects, comparison to a strict deterministic linkage approach with a given set of matching rules is achieved by the fact that our approach generates as a side product how many linked pairs can be found in addition to a strict deterministic linkage approach.
- our synthetic datasets introduced different kinds of error but only on one variable at a time. A more configurable synthetic dataset generation should make it possible to introduce errors on more than one variable in order to assess the effect of these patterns of error on the results. Data are currently randomly corrupted but it could be non-randomly as well. The 5% error rate was low in order to mimic administrative health system used in our real-world studies. However, it would be of great interest to assess recall/precision on much more corrupted data with error rates ranging from 0% to 100%.

5.3 Future work

Following the ideas in [23], it would be interesting to precisely assess the place of our approach among the other deterministic and probabilistic approaches and possibly among other record linkage packages [24]. It was not the primary aim of our paper, but the characteristics of our approach suggest that it could obtain good results in both accuracy and performance. A full simulation study will be conducted in future with improved simulated datasets as described in the previous chapter.

Robustness definition captures only a fraction of the information contained in set Y (combinations of identifiers that lead to a linked pair, see chapter 2.3) and set Z (combinations that lose the uniqueness of that linked pair). Further use of Y and Z could be imagined. For instance, linking the source dataset to itself makes it possible to know the truth since a patient x has to be linked to itself [1]. In this context, (Y, Z) would fully characterize x in the source dataset and could be used as "gold standard information" when processing the source/target linkage. Future work will precisely assess this observation and see how it could be systematically used in our linkage methodology.

Concerning scalability, the current results show low runtimes even on large datasets. For instance, scenario 1 took less than 2 seconds to link a dataset with $30,000$ patients to a dataset with $3,000,000$ patients on a 4 CPU core computer. Future work could test our implementation with much more CPU power, e.g. at least 64 physical cores. It would be interesting to assess the scalability as the function returning linkage runtime for a given number of CPU. Today, the current implementation can use all the CPU cores of one computer but cannot use the power of a computing grid. However, this implementation could easily be extended to take advantage of a computing grid for extreme record linkage projects involving very large datasets.

References

[1] Scailteux LM, Droitcourt C, Balusson F, Nowak E, Kerbrat S, Dupuy A, et al. French administrative health care database (SNDS): The value of its enrichment. Therapies. 2019;74(2):215 – 223. Available from: http://www.sciencedirect.com/science/article/pii/S0040595718302403.

[2] Didier R, Gouysse M, Eltchaninoff H, Le Breton H, Commeau P, Cayla G, et al. Successful linkage of French large-scale national registry populations to national reimbursement data: Improved data completeness and minimized loss to follow-up. Archives of Cardiovascular Diseases. 2020;Available from: http://www.sciencedirect.com/science/article/pii/S187521362030139X.

[3] Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Revue d'Epidémiologie et de Santé Publique. 2017 Oct;65:S149–S167. Available from: https://doi.org/10.1016/j.respe.2017.05.004.

[4] Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020 Feb;47(2):218–226. Available from: https://doi.org/10.1080/03014460.2020.1742379.

[1] In our approach, source/source linkage can produce false negatives but cannot produce false positives.

[5] Fellegi IP, Sunter AB. A Theory for Record Linkage. Journal of the American Statistical Association. 1969 Dec;64(328):1183–1210. Available from: https://doi.org/10.1080/01621459.1969.10501049.

[6] Li B, Quan H, Fong A, Lu M. Assessing record linkage between health care and Vital Statistics databases using deterministic methods. BMC Health Services Research. 2006 Dec;6(1). Available from: https://doi.org/10.1186/1472-6963-6-48.

[7] Setoguchi S, Zhu Y, Jalbert JJ, Williams LA, Chen CY. Validity of Deterministic Record Linkage Using Multiple Indirect Personal Identifiers. Circulation: Cardiovascular Quality and Outcomes. 2014 May;7(3):475–480. Available from: https://doi.org/10.1161/circoutcomes.113.000294.

[8] Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Statistics in Medicine. 2002;21(10):1485–1496. Available from: https://doi.org/10.1002/sim.1147.

[9] 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 May;64(5):565–572. Available from: https://doi.org/10.1016/j.jclinepi.2010.05.008.

[10] 2011 B. Matching Anonymous Data; 2013. Available from: https://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-pbeyond-2011-matchinganonymous\-data--m9-.pdf.

[11] Randall S, Brown AP, Ferrante AM, Boyd JH. Privacy preserving linkage using multiple dynamic match keys. International Journal of Population Data Science. 2019 May;4(1). Available from: https://doi.org/10.23889/ijpds.v4i1.1094.

[12] Brown AP, Randall SM, Ferrante AM, Semmens JB, Boyd JH. Estimating parameters for probabilistic linkage of privacy-preserved datasets. BMC Medical Research Methodology. 2017 Jul;17(1). Available from: https://doi.org/10.1186/s12874-017-0370-0.

[13] Harper G. Linkage of Maternity Hospital Episode Statistics data to birth registration and notification records for births in England 2005–2014: Quality assurance of linkage of routine data for singleton and multiple births. BMJ Open. 2018 Mar;8(3):e017898. Available from: https://doi.org/10.1136/bmjopen-2017-017898.

[14] Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. European Journal of Epidemiology. 2018 Sep;34(1):91–99. Available from: https://doi.org/10.1007/s10654-018-0442-4.

[15] Christen P. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. IEEE Transactions on Knowledge and Data Engineering. 2012 Sep;24(9):1537–1555. Available from: https://doi.org/10.1109/tkde.2011.127.

[16] Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for Information about Linking Data sets†. Journal of Public Health. 2017 Mar;40(1):191–198. Available from: https://doi.org/10.1093/pubmed/fdx037.

[17] Pratt NL, Mack CD, Meyer AM, Davis KJ, Hammill BG, Hampp C, et al. Data linkage in pharmacoepidemiology: A call for rigorous evaluation and reporting. Pharmacoepidemiology and Drug Safety. 2019 Nov;29(1):9–17. Available from: https://doi.org/10.1002/pds.4924.

[18] Vatsalan D, Sehili Z, Christen P, Rahm E. Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. In: Handbook of Big Data Technologies. Springer International Publishing; 2017. p. 851–895. Available from: https://doi.org/10.1007/978-3-319-49340-4_25.

[19] Christen P. Data Linkage: The Big Picture. 12. 2019 Nov;Available from: https://doi.org/10.1162/99608f92.84deb5c4.

[20] Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications. 2019 Jul;10(1). Available from: https://doi.org/10.1038/s41467-019-10933-3.

[21] Blay JY, Honoré C, Stoeckle E, Meeus P, Jafari M, Gouin F, et al. Surgery in reference centers improves survival of sarcoma patients: a nationwide study. Annals of Oncology. 2019 Jul;30(7):1143–1153. Available from: https://doi.org/10.1093/annonc/mdz124.

[22] Grimaud O, Lachkhem Y, Gao F, Padilla C, Bertin M, Nowak E, et al. Stroke Incidence and Case Fatality According to Rural or Urban Residence. Stroke. 2019 Oct;50(10):2661–2667. Available from: https://doi.org/10.1161/strokeaha.118.024695.

[23] Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic

linkage? A simulation study. Journal of Biomedical Informatics. 2015 Aug;56:80–86. Available from: https://doi.org/10.1016/j.jbi.2015.05.012.

[24] Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. Journal of Biomedical Informatics. 2012 Feb;45(1):165–172. Available from: https://doi.org/10.1016/j.jbi.2011.10.006.

6 Declarations

6.1 Ethics approval and consent to participate

The study, analysis and data extraction were approved by the French Data Protection Agency (CNIL, DR-2016-492 and DR-2017-244). Informed consent is waived for the use of these anonymised secondary data, as mentioned in the Social Security Code, Article L161–28-1. All methods were performed in accordance CNIL regulations.

6.2 Availability of data and materials

The data that support the findings of the two real-world studies are available from the French National Health Insurance Information System but restrictions apply to the availability of these data and so are not publicly available. The synthetic datasets generated and/or analysed during the current study are available in the GitHub repository (https://github.com/erwandrezen/datasimulation/tree/master/data_sets)

6.3 Funding

The Brest Stroke Registry is supported by Santé Publique France and INSERM (Institut National de la Santé et de la Recherche Médicale).

6.4 Author's contributions

E.D. performed linkage of synthetic datasets and wrote the paper, A.H. designed the synthetic datasets and wrote the paper, S.K. designed and created the synthetic datasets and performed linkage of one of the two real-world studies. F.B. performed linkage of one of the two real-world studies, E.O. designed the synthetic datasets, performed statistical analysis and wrote the paper. All authors read and approved the final manuscript.

6.5 Acknowledgements

The authors thank the team of the Brest Stroke Registry project and in particular Serge Timsit from the Brest University Hospital and Valérie Olié from Santé Publique France. They also thank the team of the Artome trial project and in particular Boris Campillo-Gimenez from the Cancer Treatment Centre Eugène Marquis in Rennes.