# Overview of Data Linkage Methods for Integrating Separate Health Data Sources


**Ana Kostadinovska, Muhammad Asim, Daniel Pletea, and Steffen Pauws**

## 1 Introduction

Access to high-quality care partially determines the overall health of an individual. Environmental, socioeconomic, behavioral, and genetic factors are even larger determinants of health outcomes. Consequently, effectively managing the health of an individual requires full commitment and coordination of care professionals inside and outside of hospital walls, including community and social care, payers, local governments, and wellness and healthcare service providers.

Data is key toward understanding an individual's health, but unfortunately, data related to these different health determinants mostly reside in siloed systems managed by different players in the ecosystem across the health continuum. Usually these datasets contain information about the same patient. Lastly, governmental organizations or quangos ("quasi-autonomous nongovernmental organization") collect census, register, and survey data on health data such as outcomes and utilization, societal data, and economic data such as population count, income, education, employment, and religion.

The datasets need to be brought together in order to generate better insights about the health status of individuals. The process of bringing together those records that are perceived to belong to the same individual, entity, location, or event is called

A. Kostadinovska (✉) · M. Asim · D. Pletea
Philips Research, Eindhoven, Netherlands
e-mail: ana.kostadinovska@philips.com; muhammad.asim@philips.com; daniel.pletea@philips.com

S. Pauws
Philips Research, Eindhoven, Netherlands

Tilburg University, Tilburg, Netherlands
e-mail: steffen.pauws@philips.com; S.C.Pauws@uvt.nl



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_8

data linkage. Linked and extended datasets from various services across the health continuum lead to more insights in comparison to a single dataset individually [24].

The data linkage can be performed exactly and faultlessly if at every source identical uniquely identifiable information is associated with all data elements. The process is more daunting if such information is not available; identifiers are used that are not necessarily unique such as patient names and demographic information. Unfortunately in many practices unique identifiers are missing. To make the situation more difficult, the available non-unique linking variables very often contain errors due to coding errors, spelling variations, or transcription mistakes. These factors threaten the quality of the linked data as records can be missed or wrong records can be linked, which can result in biased analysis of the linked data.

This chapter provides a state-of-the-art survey in data linkage technology within healthcare. It will give a tutorial overview of the various methods in data linkage including deterministic and probabilistic approaches, a discussion on the challenges of using data linkage in healthcare and a synthesis of a healthcare use case in which data linkage is essential.

## 2 Overview of Data Linkage Methods

Data linkage is a process in which the same entities (individuals, location, and events) should be identified in record pairs among two or more different datasets. This section gives an overview (shown in Fig. 1.) of the steps in data linkage.

### *2.1 Data Delivery*

Data delivery is the first required step for linkage. Data can be provided in various schemes such as simple structured data (e.g., pairs of files) or semistructured


**Fig. 1** Data linkage process steps

data (e.g., a pair of XML documents) [1]. Before data can be delivered, consent for sharing or processing the data needs to be in place. Data might need to be anonymized before processing. In addition, different data regulations may apply in different geographies, and data may not be allowed to leave a specific premise. The data owner can constrain the processing to a specific computing environment with strict security access for specific individuals only. The legal and regulatory aspects of data delivery will be further discussed in Sect. 3.1.

## *2.2 Data Cleansing and Standardization*

The data cleansing and standardization process can be quite labor intensive, so it is recommended to assess whether the costs of labor are paid off by the benefits of a cleansed dataset [22]. The process can be broadly divided into six steps: (1) handle different input file formats, (2) handle unstructured data, (3) handle data heterogeneity, (3) handle typographical errors, (4) handle missing data, (5) handle data overlap, and (6) parse identifiers into separate pieces of information [9, 29, 32]. Further explanation of each step is given below.

### 2.2.1 Handle Different Input File Formats

In practice, input files can arrive in different formats such as csv or xlsx. Especially when it concerns longitudinal data, data can be stored in a wide or long format. In a wide format, all data collected over time for each entity (or individual) are in a single row. In a long format, each row is one time point per entity or individual. Variables that do not change over time will have the same value in all rows. Sizes of files can differ substantially. Long formats can grow out of proportions as it stacks redundant data (i.e., variables that do not change over time). It is recommended to convert all files to a single standard format allowing to compare and match corresponding columns containing candidate variables for linkage.

### 2.2.2 Handle Unstructured Data

When data arrives in an unstructured form such as nursing notes, it first needs to be made searchable and retrievable. Natural language processing tools are essential to fit unstructured freeform text into a predefined data record scheme. In particular, named entity recognition (NER) in text identifies and annotates person and organization names, geographical locations, events, and expressions of time, date, and amounts in text that can act as a linking variable value [21, 25].

**Table 1** Data heterogeneity example

| First name | Last name | Date of birth | Address | Enrollment date and time |
|---|---|---|---|---|
| Jessica | ADAMS | 10-8-1985 | 4257 Bart Ave | 25-05-2017 12:05 |
| THERESA | Pratt | December 4, 1965 | West Davison 8100, 48238 | 12/1/2017 08:05:00 |

### 2.2.3 Handle Data Heterogeneity

The coding of linking variables can differ across input files [8]. For instance, they can differ in their data type (e.g., an age variable can be of type integer or represented as a string), in their format (e.g., dates can have many different formats as YMD, DMY, and MDY with various separator signs, digits, and spellings of months). Variables should comply in representation for matching.

Table 1 is an example of data heterogeneity. The table contains two (synthetically created) records containing identifiable information of two patients. First and last names, date of birth, enrollment date and time, and address are variables that differ in their format, type, case, and content.

### 2.2.4 Handle Typographical Errors

Input files might contain typographical errors in the linking variables such as transposed digits and misspellings. Table 2 shows some commonly found variations that should be taken into account. Variation in spelling in proper names or geographical locations can be unintended misspellings but also due to transliterations or transcription from one alphabet (Cyrillic, Chinese, Japanese, Korean, Arabic, Greek, Hebrew, and Latin) to the other. Transliteration is the use of conversion rules for each symbol from the source alphabet to a symbol of the target alphabet. Transcription is the writing down the sound of the name or location in the source language as accurately as possible in the target language. As an example, Oeladzislau Smjahlikau and Vladislav Smjaglikov refer to one and the same person (a boxer) from Belarus though the spelling of the person name is obtained via transliteration and transcription, respectively, from the Cyrillic script. Due to migration, person names in health data can come from various geographical locations, languages, and cultures.

Special language technology tools are developed for overcoming variation in spelling [17, 30], for which Soundex [9, 31] is a commonly used method. Soundex is a system for coding and indexing family (proper) names by transcription. Another solution for handling typographical errors can be done by comparing strings using edit distance techniques to determine the minimum number of operations (e.g., insertions, deletions, and transpositions) to get from string A to string B.

**Table 2** Common variations found in selected linkage identifiers [9] *FIPS* federal information processing standards, *SSA* social security administration

| Field | Type | Examples |
|---|---|---|
| Names | Case | John Smith \| JOHN SMITH |
| | Nicknames | Charles \| Chuck |
| | Synonyms | William \| Bill |
| | Prefixes | Dr. John Smith |
| | Suffixes | John Smith, II |
| | Digits | John Smi9th |
| | Punctuation | O'Malley \| Smith-Taylor \| Smith, Jr. |
| | Initials | JA \| J.A. \| Jessica Adams |
| | Transposition | Jessica Adams \| Adams Jessica |
| | Transliteration and transcription | Oeladzislau Smjahlikau \| Vladislav Smjaglikov |
| Addresses | Abbreviations | RD \| Road \| DR \| Drive |
| Dates | Format | 01012013 \| 01-01-2013 \| 01JAN2013 |
| | Invalid values | Month = 13 \| Day = 32 \| Birth year = 2020 \| Date = 29FEB2013 |
| Social security number | Format | 999999999 \| 999-99-9999 \| 999 99 9999 |
| Geographical location | Abbreviations | NC \| North Carolina |
| | ZIP codes | 99999 \| 99999-99999 |
| Sex | Format | Male/Female \| M/F \| 1/2 |

### 2.2.5 Handle Missing Data

Input files might contain a large number of missing values in linking variables or other variables that can affect the correctness of the data linkage. After investigating a plausible reason for missing data, imputation is a method to fill in values for the missing data [12]. Missing data can happen for various reasons. It is recommended to use imputation only if missing data happen at random (MCAR or MAR). If missing data is due to an informative reason, data cannot be imputed:

- *Missing completely at random* (MCAR) is due to administrative errors or unfortunate incidents during measurement or collection. A missing value is unrelated to any individual/center characteristics or outcome.
- *Missing at random* (MAR) is due to patient characteristics, time, place, or outcome. The probability of a missing value depends on values of other variables. For instance, a patient is too sick to perform a test, which may result in missing values for the test at high severity of the disease.
- *Missing not at random or informative missing* (IM) is due to the value of the variable itself, the hospital data collection protocol, or the de-identification

procedure. For instance, a hospital may not order particular blood tests. This kind of missing is hard to resolve.

Yuan [34] defines several multiple imputation methods depending on the type of missing data pattern. For monotone missing data patterns (a dataset has monotone missing pattern when a missing variable $Xi$ implies that all subsequent variables $Xj$, when $j$ is greater than $i$, are as well missing for one individual), either a parametric regression method or nonparametric one can be used. For an arbitrary missing data pattern, a Markov chain Monte Carlo (MCMC) method is appropriate. An overview of the methods, together with their basic concepts and applications, can be found in [34].

#### 2.2.6 Handle Data Overlap

Input files can contain multiple records that refer to the same entity in the real world. Also, input files can contain referential overlap. For example, a zip code and a house number refer to the same home as a full address, so there is full referential overlap. A zip code and a city name, though referring to different entities, do have some referential overlap as the geographical area of the ZIP code is contained in the city referred to by the city name. If these overlaps are not excluded from the input files, the credit assigned for links on these overlaps is redundant. Referential overlap in data is helpful in iterative linking methods; in a first pass, an exact match can be established on ZIP code to be extended on counties when ZIP codes do not match exactly.

#### 2.2.7 Parse Identifiers into Separate Pieces of Information

Some of the linking variables should be split into multiple parts. This allows the linkage process to get the most out of all parts of available information. For example, a street variable can contain street name and street number. Due to typographical errors, a street or address number can be incorrect, while the street name is without error. In this case, it is better to split the street variable into two variables: street name and street address. Another example, personal information, can change over time, such as a name change after marriage or an address change after a move. In such cases, linking on the separate parts allows for partial agreement, when combined with other information, which may provide evidence that the records being compared refer to the same person.

### *2.3 Searching Data*

Searching entails identifying the pairs of records from two datasets that have a high probability of matching with each other on the basis of the linking variables. In this

search, a compromise is sought between the number of record pairs to be evaluated for matching and the number of true links needed. Evidently, it should exclude the pairs that do not match from further comparison [31]. Searching can be done by *blocking*, *sorted-neighborhood method*, *bigram indexing*, and *canopy clustering*. We elaborate more on the first two as most prominent searching methods. More information on the latter ones can be found here [2].

### 2.3.1 Blocking

Blocking consists of partitioning the two datasets into mutually exclusive subsets and searching for links matching pairs within these subsets. These subsets are called blocks. Typically, blocking is based on a blocking variable on which the partitioning takes place. It limits the number of pairs being evaluated for matching. Without blocking a Cartesian product of all pairs of records need to be evaluated.

A disadvantage of the blocking is that true links are potentially missed out as they can end up in different blocks. A common remedy is to keep the block sizes relatively small and run multiple blocking passes using different blocking variables [20, 29, 31]. The best blocking variables are those that have an almost uniform value distribution on records, are error-free, do not miss values, and do not change over an individual's lifetime. For example, month of birth is an example of such a variable that would result in fairly even number of records in each block [9, 29, 31]. According to Baxter et al. [2], the blocking method trades off pairs' completeness with reduction of the record pairs to be compared as the number of blocks increases. More smaller blocks result in less comparisons but more true match pairs are missed.

### 2.3.2 Sorted-Neighborhood Method

Sorted-neighborhood method starts with sorting the records of the input files. Sorting is done using a sorting key made out of one or several existing variables that have only few records with the same value. Then, comparison of pairs of records is done on records that fall into a sliding fixed-sized window. If the size of the window is $w$ records, then every new record entering in that window is compared with the previous $w - 1$ records. Hence, the number of comparisons is reduced from $n^2$ to $w*n$ (where $n$ is the size of the input files). After the comparison, a transitive closure step is performed; if two records r1 and r2 are found to be similar, and records r2 and r3 are found to be similar, then r1 and r3 are also marked as similar. This allows for a small window size, hence low time complexity but with an invariant accuracy of the result [1].

Due to the various possible types of errors in the input files, some records might be sorted out of the window boundaries from those records with which they should be compared to. Running this method on a single sorting key (i.e., a single-pass) usually does not produce the best results. Therefore, a multi-pass approach can be

used, where a number of sorting keys with small windows sizes are used. The results from the independent passes are then combined to provide the final set of linking records [1, 31]. According to Baxter et al. [2], this method avoids the extremes in performance of blocking, and its behavior changes predictably as the window size w is increased. With larger windows, pairs' completeness results improve, but the number of record pairs to be compared increases.

## *2.4 Matching/Linking Data*

The matching of record pairs can either be done deterministically or probabilistically, dependent on the purpose and research question underpinning the data linkage, time and effort available, and the quantity and quality of the linking or identifiable variable available.

In situations in which identifiable variables are not released for inspection and processing due to privacy concerns, a linkage on encrypted identifiers may be employed. Identifiers are first encrypted by using cryptographic hash functions and then shared with researchers for linkage and processing, without compromising privacy [9]. Manual inspection of encrypted linked results cannot be done for review. A discussion on encrypted methods can be found in Sect. 3.1.

### 2.4.1 Deterministic Algorithm (Single-Pass Strategy)

A deterministic algorithm decides whether a pair of records agrees or disagrees in a given set of linked or identifiable variables on the basis of an exact match comparison. The outcome of the comparison is of binary nature, "all-or-nothing" [9] and can be calculated in one or multiple passes.

A single-pass deterministic algorithm, better known as the "exact deterministic method" [9], compares all pairs of records (within a block) at once using the entire set of linking variables. A pair of records is classified as a match if the two records agree on all variables and are uniquely identified. Note that two records are uniquely identified if no other record in the input files matches on the same values of the linking variables. A pair of records is classified as a non-match if the records disagree on at least one linking variable or if the record pair is not uniquely identified.

This algorithm is of straightforward use if the input files contain unique identifiers of high quality without missing values; it has limitations in use for data containing errors or missing values.

### 2.4.2 Iterative Deterministic Algorithm (Multi-Pass Strategy)

A multi-pass strategy consists of records being linked using criteria for different linking variables in multiple successive passes. Record pairs that do not link in one pass are forwarded to a next pass. If a record pair meets the criteria in any of the passes, the pair is classified as a match. Otherwise, it is classified as a non-match. The method still requires an exact match in any of the passes. It is also known as "approximate deterministic algorithm" [9].

The iterative deterministic approach can be used when the single-pass method provides unsatisfactory results or if no single uniquely identifiable and complete variable in the two input files is available. However, it still requires an exact match and high-quality linking variables.

### 2.4.3 Probabilistic Approach

The deterministic approach does not take into account possible erroneous values of linking variables as it is based on finding an exact match. If linking variables happen to agree partially due to errors (e.g., misspellings), the record pair is registered as a non-match. In addition, the deterministic approach also ignores that linking variables and their values can have differential discriminatory power which expresses to what extent variables are able to discern records to represent the same entity (i.e., patient) or different entities. As defined by Blakely and colleagues, probabilistic linkage is "record linkage of two (or more) files that utilizes the probabilities of agreement and disagreement between a range of linking variables" [3]. It is able to assess (1) the discriminatory power of each linking variable and (2) the likelihood that two records are a true match based on whether they agree or disagree on the various linking variables [5].

A probabilistic method is a good option, if linking variables are available but incomplete, fraught with typographical errors, or imperfectly measured, or when no unique identifiers are available. In these scenarios it can outperform deterministic methods, albeit with more time and resources required for running the method.

#### Calculating and Summing Up Probabilities as Weights

The record pairs identified in the search phase are compared on each linking variable for producing an agreement pattern for their values [20]. Weights for each value of the linking variable for every record pair are calculated to measure the contribution of each linking variable to the probability of making a correct matching judgment. The weight assigned to each linking variable is considered a likelihood ratio comparing the proportion of agreements with the proportion of disagreement for that linking variable. The weight compares two probabilities, $m$ and $u$, associated with every linking variable [5, 9].

The *m probability* is the likelihood that the values of a linking variable agree on a pair of records, given that the records refer to the same entity. It is calculated as 1 minus the error rate of the linking variable. With fewer errors in its values, the linking variable will be more reliable which is expressed by a larger *m probability* [20]. For example, if gender disagrees 10% of the time due to a typographical error, or due to being misreported, then the *m probability* for this field is $1-0.1 = 0.9$. The estimates for the *m probability* can be based on prior knowledge or experience or through a supervised training procedure with data containing true links as ground truth data. Estimation is usually done by using the EM (expectation-maximization) algorithm [29] or the EpiLink algorithm [6].

The *u probability* is the likelihood that the values of a linking variable agree on a pair of records, given that the two records refer to different entities. It is a measure of the likelihood that the values of linking variables of any two records will agree by chance. The *u probability* is often estimated by *1/n* (where *n* is the number of possible values of the linking variable). For instance, the probability that false matches randomly agree on month of birth (*u probability*) is 8.3% (1/12).

Using the *m* and *u* probabilities, we can estimate how closely the linking variables agree on each record pair being compared. If a record pair agrees on a linking variable, an *agreement weight* is calculated by $\log_2(m/u)$, which is most often a positive value. When a record pair disagree on an identifier, the *disagreement weight* is calculated by $\log_2((1-m)/(1-u))$, which is most often a negative value.

For each possible record pair, the various agreement and disagreement weights are summed over all linking variables to produce a composite score referred to as the total weight score. The larger the total weight score, the more likely that both records refer to the same entity and thus should be linked. The *m probability* must always be greater than the *u probability*. If this is not the case, then the linking variable does not aid in discriminating matched from non-matched record pairs and should be discarded [20].

### Determining Links Based on Cut-Off Threshold

The distribution or histogram of the total weight score is generally bimodal, as shown in Fig. 2. Since most pairs of record are non-matched pairs, or true non-links, the left-hand mode represents low total weight scores (also called the *U* region). The other mode represents the larger total weight scores for the matched pairs, or true links (also called the *M* region).

An optimal cut-off threshold needs to be calculated to determine which record pairs should be treated as links (matches) and which pairs as non-links (non-matches). Various manual and automated methods exist to determine the threshold value based on the distribution of the weights. One way to calculate the cut-off value is using the relationships between file sizes, identifiers, and match weights [5]. To determine whether a pair of records should be consider a match or not, the total weight score of that pair is compared with the cut-off threshold value. If the total weight score is above the cut-off, the record pair is considered a match. Otherwise,

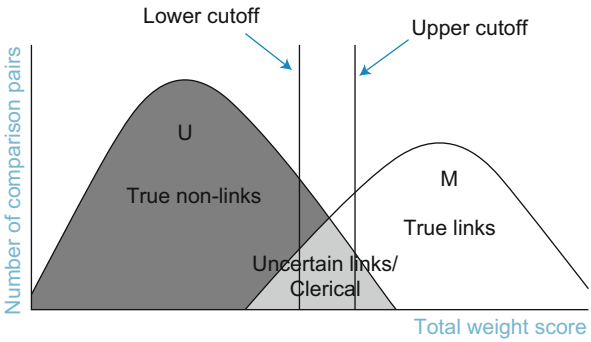


**Fig. 2** Histogram of total weight scores for all comparison pairs [27]

it is not. Sometimes an upper and a lower cut-off threshold value are used, as shown in Fig. 2.

The intersection of the $U$ and $M$ regions represent pairs that are seen as matches but are in fact non-matches, or vice versa, for which a clerical review is required [20]. Clerical review is discussed in Sect. 2.5.2.

### Cut-Off Threshold from File Sizes, Identifiers, and Match Weights

By looking into the relationships between the sizes of the input files, expected number of links, and desired probability of true links, we are able to quantify the cut-off threshold needed to probabilistically link two files. Moreover, we can quantify the extent of information in various linking variables in order to choose which ones are at least necessary to reach a desired linkage performance [5].

The relationship between the input file sizes, the expected number of links, and the desired probability of true links is expressed as

$$wt = \log_2(p/(1-p)) - \log_2(E/(A*B-E)) \quad (1)$$

where $wt$ is a match weight representing the log odds for a true link corrected for finding a true link by chance: $p$ is the desired probability of true links; $A$ and $B$ denote the size of the first and second input files, respectively; and $E$ is the expected number of true links. The match weight $wt$ can act as a cut-off threshold to tell which records match with a probability of at least $p$ of being a true link. For example, if $A$ and $B$ are input files that count 1000 records each, where every record in $A$ uniquely matches a record in $B$ (hence $E$=1000), and with desired probability of selecting true match $p$ equals 0.9, then the match weight is 13.13. This means that at least weight of 13.13 is needed to overcome the current odds and produce matches with probability of at least 0.9 being correct.

### 2.4.4 Hybrid Solutions

A hybrid solution entails combining the advantages of deterministic and probabilistic algorithms into a single one. A deterministic algorithm might miss out some truly linked record pairs due to errors in the linking variables. A hybrid solution tries to reduce this by conducting a probabilistic linkage on the record pairs that are considered non-matches in the deterministic pass. Fewer pairs will be processed and additional pairs will be linked during the probabilistic linkage phase; a hybrid solution is deemed to be more efficient with better outcome than a probabilistic or a deterministic method alone [9].

### 2.4.5 Other Data Linkage Algorithms

In Table 3, we summarize the different matching methods on their advantages, disadvantages, and applicabilities. Probabilistic linkage (or a hybrid solution) is recommended if exact agreement between linking variables cannot be established. A disadvantage of probabilistic linkage is that it requires estimates on weights and thresholds from data where true link status is available as ground truth. *Machine learning (ML)* can be used to arrive at these estimates in which supervised learning takes place on labeled ground truth data to obtain a model. Bayesian methods including Naive Bayes are ML methods that arrive at good estimated models [28, 33]. This model can then be used to discern the links from the non-links using unseen, unlabeled data [1]. However, this training requirement is time-consuming, requires ground truth data, and needs to take place for every new domain. Therefore, new probabilistic techniques known as *scaling methods* try to arrive at these estimates without the need of such a supervised training phase [13].

Missing out links (false negatives) can underestimate the number of truly linked pairs, also in probabilistic methods. A reason is the so-called *entity heterogeneity* problem that appears when the same entity (e.g., patient) is known under different identifiers in the datasets to be linked. A *Bayesian approach* is seen as a solution to that problem by using a distance-based measure in order to express the similarity between the referred entities [8].

Another disadvantage is that probabilistic linkage chooses at most a single matched link for any pair of records that has maximum weight above threshold while ignoring all other potential matches with a lower weight, which may bias linked datasets. By using *multiple data imputation methods*, we can allow for several potentially matched links for record pairs in a subsequent analysis instead of only the maximum one or no one which leads to unbiased and more efficient analyses [12].

**Table 3** Comparison of various matching methods

| Method | Advantage | Disadvantage | Applicability |
|---|---|---|---|
| Single-pass deterministic | Straightforward | Limitations of use in erroneous and missing data | High-quality data requiring exact match |
| Iterative deterministic | Multiple linking criteria. More resource and time efficient than probabilistic approach if the linking identifiers are available | Limitations of use in erroneous and missing data. Less time and resource efficient than single-pass approach | High-quality data requiring exact match If no single unique linking identifier is available, but multiple high-quality attributes are available, this approach would fit better than the single-pass approach |
| Probabilistic | Better coping with erroneous data. Can handle data that is ignored in the deterministic algorithm and classified as a non-link. Can outperform deterministic methods in information-poor scenarios. Compared to the deterministic (both single-pass and iterative) approach, a better combination of variables can be selected by assigning weights and linkage score. | Requires more time, effort, and technical resources to implement than the deterministic algorithm. | No exact agreement due to incomplete data or no unique identifiers: if identifiers are available but incomplete, fraught with typographical errors, or imperfectly measured, or when no unique identifiers are available, the probabilistic approach comes into place |
| Hybrid | Combining advantages of deterministic and probabilistic approaches. Fewer pairs will be processed in the resource-intensive linkage phase, so it can be more efficient than only applying a deterministic or probabilistic algorithm | | After applying the deterministic algorithm, a large number of record pairs are incorrectly classified as non-links due to errors in the input files. |

## *2.5 Evaluating Data Linkage*

This section explains how to assess the quality of data linkage by means of metrics, clerical review, and quality reporting.

### 2.5.1 Metrics

In evaluating data linkage algorithms, an identified match in a pair of records can either be a true link or a false link, and an identified non-match can either be a true

non-link or a missed link. Linkage errors expressed by false and missed links can result in biases in the analyses for which the linkage was established [23].

1. A Type I linkage error occurs when a true non-link is identified as a match, which is called a false positive or false match. This implies that the linked dataset will contain linked information that should not have been linked.
2. A Type II linkage error occurs when a true link is identified as a non-match, which is called a false negative or a missed link. This implies that the linked dataset misses out information that should have been linked.

Four metrics are commonly used to evaluate the performance of a linkage algorithm: sensitivity (recall), specificity, positive predictive value (PPV) (precision), and negative predictive value (NPV) [9]. These metrics measure the ability of the algorithm to correctly classify true links as identified matches and true non-links as identified non-matches. Sensitivity or recall is the fraction of true links that have been identified as match. Specificity is the fraction of true non-links that have been identified as a non-match. Precision is the fraction of true links among the identified matches. In practice, a trade-off between recall and precision takes place. An algorithm can act liberally to find more matched pairs, resulting into high recall and low precision. It can also act more conservatively in finding fewer non-matched pairs, resulting into high precision and low recall. Greater recall produces more true links identified at the cost of more non-matches. Greater precision leads to fewer true links identified but also fewer non-matches [1]. To investigate the effect on precision and recall, sensitivity analyses can be done by performing the linkage on different sets of linking variables.

When data linkage is done for analyzing a rare disease, meaning that relatively few individuals have the diagnosis, a high recall is preferred as we do not want to miss out any diagnosis in the linked dataset. In case a common disease is subject to the analysis, it is preferred to increase precision so we are assured that every match identified is a true link [9].

To demonstrate the trade-off between precision and recall, one of them is often displayed while fixing the other one. The F-measure, introduced by Christen and Goiser [4], combines the two in a single metric; it represents the harmonic mean of precision and recall. Although there is no absolute criterion, a data linkage algorithm that is typified as well-performing should be able to report an F-measure of at least 95% [9].

### 2.5.2 Manual and Clerical Review

Manual or clerical review (i.e., human judgment) is usually performed to identify opportunities to refine the linkage algorithm by accounting for complex cases, such as ties, unforeseen erroneous data, or uncertainty about matches. Reviewing a random sample of the linked dataset is a common method to perform a manual review [9, 27]. A review of the full linked dataset is far too time-consuming and resource intensive.

For instance, ties are multiple pairs of records that have similar values for the linking variables; so ties are all candidates for a link. Additional data may be consulted to resolve these ties. One option is to generate all possible ties or pairs of matched records in a single overview and pick out the ones that are true links [9].

As shown in Fig. 2, uncertainty about matches refers to a midrange of record pairs which can be either a match or a non-match on the basis of how a cut-off threshold is positioned [27].

#### 2.5.3 Quality Reporting

Estimates on algorithmic performance on specific datasets should be reported to characterize the validity and reliability of the linked dataset. It should be transparent how and for what reason one metric (e.g., recall) is prioritized over another one (e.g., precision) and reflected in optimizing the algorithms in its parameter settings. Besides the standard metrics on sensitivity, specificity, precision, and NPV, it is useful to report a tie statistics expressed as the number (or proportion) of records that are linked with more than one record, a non-match statistics expressed as the number (or proportion) of records that are not linked, and a cleansing factor telling the number (or proportion) of records that can be linked before and after the step of data cleansing.

When reporting results, it is also useful to conduct a subgroup analysis of the linked records and non-linked records. Individuals with linked records may differ in characteristics, such as diagnoses, demographics, or outcome, from individuals with no linked records. Propensity analysis can be helpful in estimating the effect of the linkage by accounting for all variables in the datasets (not only the linking variables) that explain all linked records. Differences and commonalities (i.e., linkage bias) between the original uncoupled dataset and the newly linked dataset can be essential to understand what information has been added through the linkage.

## 3 Data Linkage Use Cases in Healthcare

This section is devoted to discuss the challenges of using data linkage in healthcare and to draw up use cases in healthcare in which data linkage is required.

### *3.1 Legal and Privacy Challenges*

One challenge when linking data in healthcare is to address privacy concerns and restrictions. Privacy concerns are justified and necessary to protect individuals. However, information governance for researchers can be overly complicated and disproportionate to the risks involved in protecting patient data. Understanding and

negotiating the legal, ethical, and governance frameworks and requirements may be a barrier to data access for researchers unfamiliar with using linked datasets.

When data is collected, it is usually limited to a single purpose. On the other hand, accessing linked data for a broader purpose would be more efficient and hypothesis-agnostic (though there are regulatory limits to the breadth of consent that can be given under the forthcoming General Data Protection Regulation—GDPR) [10]. The easiest way to deal with such privacy concerns is to inform the patients about the intention to link data and the intended use of the linked data, along with any associated risks, and to ask for permission to use their data for these secondary purposes.

**Getting Patients' Approval for Data Linkage** A patient's informed consent provides language to allow an institute to have access to the patient's data that are captured under strict and well-defined conditions and purposes. Such consent does not necessarily approve for linking the patient data to other data sources. Therefore, either patient's informed consent should contain language to include data linkage as a purpose or the contract for data usage should be specified in terms to cover data linkage as well.

**Performing the Linkage** Data linkage is based on coupling personal data residing in different data sources. In most cases, the data linkage cannot be done by the researchers since they are not allowed to access identifiable information of patients. Hence, dedicated persons usually do the data linkage, who are persons authorized to view identifiable data. In some cases, patient representatives (e.g., a nurse) are asked to do the linkage. Lastly, a third trusted party can do the linkage (in the Netherlands, i.e., ZorgTTP).

**Transferring Data From One Location to Another** Different regulations on legal and privacy aspects apply and should be considered. Some example regulations that outline restrictions on disclosure of personal or sensitive data are (1) the Data-Matching Program Act in Australia [14], (2) EU General Data Protection Regulation (GDPR, effective May 25, 2018) in Europe [10], and (3) Health Insurance Portability and Accountability Act (HIPAA) in the USA [18].

- When data is transferred across the EU borders, adherence to the GDPR rules attached to the data is required. Sufficient guarantees need to be implemented regarding appropriate technical and organizational measures to ensure data linkage is compliant with the GDPR requirements.
- A similar approach is taken for personal data collected in the USA, which is HIPAA applicable. The HIPAA regulation puts limits and restrictions on uses and disclosures without patient authorization. This requires data to be treated (de-identified) before disclosing and or using data for secondary uses, or when it is transferred outside the USA. Depending on the contracts in place, data linkage can only take place after creating a limited dataset [7] or de-identifying a dataset. HIPAA de-identification can be done in two ways: safe harbor which consists of removal of HIPAA 18 identifiers [19] and using an expert determination method

where the data is proven statistically to have a low reidentification risk attached to it.

According to the GDPR, pseudonymization is a method of encrypted data protection, and it may be used in acquiring consent for secondary purposes (e.g., research purposes). Pseudonymization is part of the de-identification process and is performed by replacing real identifiers with pseudo-identifiers. This can be done using a cryptographic hash function (e.g., SHA-256) using a secret key or a lookup table. The use of only a "cryptographic hash function" (e.g., SHA-256 (Name+Surname+DateOfBirth)) is not secure because the generated pseudo-identifiers can be linked back to a pool of people. The option of using "cryptographic hashing function with a secret key" is secure with the main requirement that the key should be kept secure. The use of a lookup table is the most secure because the generated pseudo-identifier is independent of the real identifiers. GDPR also defines "anonymous information" as information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This can be achieved using a reidentification risk assessment (e.g., HIPAA expert determination method), but it is highly dependent on context. De-identification and anonymization are methods which are enabling data usage for secondary purposes.

An example of secondary purpose is research. Data linkage based on two or more original datasets was explained in previous sections, but the resulting linked dataset needs to be also de-identified. Data linkage on two de-identified datasets is another challenge because the identifiers (direct and quasi) were replaced, removed, or generalized. In this case probabilistic methods can be used for performing the record linkage, but starting from de-identified datasets does not guarantee also a de-identified linked dataset. Therefore, additional de-identification actions may be needed. As we have seen in previous sections, probabilistic linkage can produce linkage errors that can result in biases in the data analysis. The additional de-identification step mentioned above may remove additional outlying data, which may add to the bias of the analysis results. This depends on the nature of the data and whether the data analysis is focused on outliers or not.

Data linkage within a single organization does not generally involve privacy and confidentiality concerns. It is usually permitted if the patient consented the secondary purpose for which the data is linked. An example application is the deduplication of a customer database by a business using data linkage techniques for conducting effective marketing activities. In this case the secondary purpose is "marketing." However, in many countries data linkage across several organizations, as required in the above example, might not allow the exchange or the sharing of database records between organizations due to laws or regulations. When data linkage across organizations is needed, the informed consent should allow explicit data linkage across organizations. Alternatively, the patient can be asked retrospectively for consent of sharing the data with the new organization or system.

Bringing data together and analyzing it is not always possible, even if patient consent is provided. Several heath organizations are reluctant in sharing their

anonymized data with third parties, either because they fear that their data could be de-anonymized or for proprietary reasons. Federated analysis techniques like secure multiparty computation (SMC) could potentially help in overcoming such issues[11]. In SMC, the objective is to jointly compute a function from the private input of each party, without revealing such input to the other parties. That is, at the end of the computation, all parties learn exclusively the output. This problem is solved using secure data transfer protocols that also apply to the privacy-preserving distributed computation[26].

### *3.2 Linking Data from Homecare Services*

We demonstrate a use case of the data linkage process using two datasets from homecare services. One homecare service is a personal emergency response service (PERS) which enables subscribers at home to summon help from a 24/7 call center after a personal incident that potentially require emergency transport to a hospital. The other homecare service is a telehealth service which remotely manages patients with a long-term condition at home, while there is clinical back office for close watch and triage of patients. Data linkage of the homecare services can help in improving the quality of service to those patients who use both services at the same time.

Since the datasets contain de-identified data, we purposefully synthetically created the identifiable information for which we know the truth and errors introduced. One dataset contains 2729 records whereas the other one includes 369 records. Along with the non-identifiable data, these two datasets contain information for the zip code and the gender of the patients. Additional five variables are synthetically created in order to have identifiable information: first name, last name, address (address name and address number), age, and date of birth. For the purpose of introducing errors to the data, we created several functions that cover misspellings and typographical errors: (1) add a new character in a string, (2) remove the last character of a string, (3) remove random character from a string, (4) swap two characters in a string, and (5) swap values of two variables.

Following the relationship between file sizes, identifiers, and match weights, we defined several test cases. For every test case, we used a probabilistic and deterministic method to link the datasets. The test cases are shown in Table 4. For every test case, we used the dataset with 396 records. Different input files are created by using subsets of the second dataset counting 2729 records. Depending on the errors introduced and the size of the subsets, the number of true links in every test case varies. The true link status is known from ground truth data from the medical record number of the patients involved in both datasets. We chose zip code as a blocking variable and first name, last name, address, age, gender, and date of birth as identifiers. The percentage of errors introduced in every test case is equal though it reflects actual error levels occurring in practice [12].

**Table 4** Test case details

| Probabilistic approach | | | | | | Deterministic approach |
|---|---|---|---|---|---|---|
| | # of record datasets 1 and 2 | # of true links | # of classified true links | # of classified false links | Accuracy | # of classified links |
| Test case 1 | 396 & 2729 | 365 | 364 | 1 | 0.9999 | 182 |
| Test case 2 | 396 & 1000 | 121 | 121 | 0 | 1 | 68 |
| Test case 3 | 396 & 396 | 40 | 40 | 1 | 0.996 | 23 |
| Test case 4 | 396 & 396 | 40 | 40 | 1271 | 0.4682 | 23 |
| Test case 5 | 396 & 396 | 40 | 23 | 0 | 0.9929 | 23 |

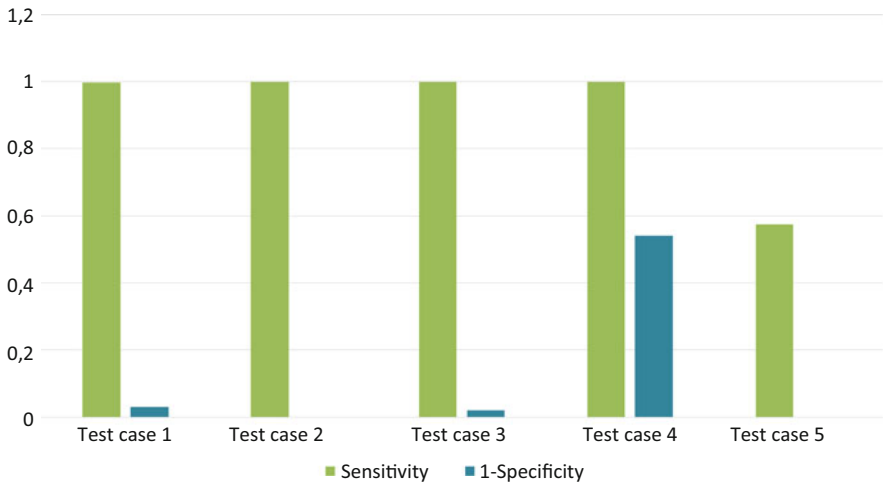


**Fig. 3** Histogram of the test cases

In Fig. 3, a histogram is shown where every test case is represented with a green and blue bar, indicating the sensitivity and 1-specificity of the probabilistic approach.

Test case 4 has the same settings as test case 3, but instead of installing an optimal cut-off threshold, we used a significantly lower value. Lowering the threshold results in more pairs to be wrongly classified as links and thus in lower accuracy. On the other hand, if a threshold is set higher than its optimal value as in test case 5, record pairs will be missed out as true links as can be observed by a higher Type II error level and *no* Type I error though there is still high accuracy.

The test cases demonstrate a clear difference in the results of the deterministic and probabilistic algorithm. For every test case, the deterministic algorithm reveals about 50% of all true links, whereas the probabilistic algorithm reveals more than 99% of the true links. Hence, the probabilistic algorithm outperforms the

deterministic algorithm if data quality is poor due to typographical errors introduced in the data.

## 4 Conclusion

Accessing and coupling data sources for combined analyses has proven itself to be challenging [15]. First, various data sources containing data of the same individual, event, or location need to be brought together under the appropriate regulatory conditions, consent, and infrastructure. Second, data can amount to staggering volumes which requires data linkage to be entrusted to computerized methods allowing only little manual or clerical review. Third, a lack in unique and corresponding identifiers across data sources can hamper linkage accuracy. Fourth, data sources can come with incomplete and erroneous data that need to be cleansed before linkage. Lastly, the actual linked result can contain errors which may bias analyses of the linked datasets.

In this chapter, we have provided a brief overview of the state of the art on deterministic and probabilistic methods for data linkage. Deterministic linkage requires exact agreement of a specified set of unique identifiers between datasets, either via a single step or successive incremental steps. It works best when identifiers are complete and accurate. If a match for any pair of records has been identified, it is typically a true link as a set of identifiers is unlikely to exactly match on all identifiers at chance level. However, due to (spelling) errors in the identifiers, true links might be missed if no precautions in data cleansing are taken.

Probabilistic linkage computes a weight for each pair of records on the basis of its matching identifiers, expressing the likelihood that this pair is a true link. Whether any pair of records is considered a link is based on a cut-off threshold on the weights that is aimed at balancing false links with missed links.

Data linkage poses privacy concerns due to the possibility of misuse of patient data and therefore should be allowed by patient consent. Consent for use of data for secondary purposes is enough when data is linked within an organization, with the condition that the linked dataset is de-identified. When data is linked across organization, the record linkage must be explicit in patient's consent. In both cases, the data protection regulations that apply to the data (e.g., when transferred from one location/jurisdiction to another one) are the ones applicable in the countries where data was collected. Hashing can be used for data linkage, and it should be done using a secret which is complex enough and stored in a secure way.

Future research on data linkage should be focused on identifying the bias and impact on combined analyses due to linkage error in various healthcare domains [16] and new algorithms that minimize linkage error either by better and efficient probabilistic weight estimates [13] or by imputing the potential matches of record pairs [12].

## References


1. Batini, C., Scannapieco, M.: Data and Information Quality. Data-Centric Systems and Applications, Chapter 8. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-24106-7_8
2. Baxter, R., Christen, P., Churches, T.: A comparison of fast blocking methods for record linkage. In: First Workshop on Data Cleaning, Record Linkage and Object Consolidation, CMIS Technical Report 03/139, KDD 2003, Washington DC, 24–27 Aug 2003
3. Blakely, T., Salmond, C.: Probabilistic record linkage and a method to calculate the positive predictive value. Int. J. Epidemiol. **31**, 1246–1252 (2002)
4. Christen, P., Goiser K.: Quality and complexity measures for data linkage and deduplication. In: Guillet, F.J., Hamilton, H.J. (eds.) Quality Measures in Data Mining. Studies in Computational Intelligence, vol. 43, pp. 127–151. Springer, Berlin (2007)
5. Cook, L.J., Olson, L.M., Dean, J.M.: Probabilistic record linkage: relationships between file sizes, identifiers and match weights. Methods Inf. Med. **40**, 196–203 (2001)
6. Contiero, P., Tittarelli, A., Tagliabue, G., Maghini, A., Fabiano, S., Crosignani, P., Tessandori, R.: The EpiLink record linkage software: presentation and results of linkage test on cancer registry files. Methods Inf. Med. **44**(1), 66–71 (2005)
7. Definition of limited data set. https://www.hopkinsmedicine.org/institutional_review_board/hipaa_research/limited_data_set.html. Accessed 26 Jan 2016
8. Dey, D., Sarkar, S., De, P.: Entity matching in heterogeneous databases: a distance-based decision model. Institute of Electrical and Electronics Engineers Computer Society (1998). https://www.computer.org/csdl/proceedings/hicss/1998/8251/07/82510305.pdf. Accessed 21 Jan 2019
9. Dusetzina, S.B., Tyree S., Meyer, A.-M., Meyer, A., Green, L., Carpenter, W.R.: Linking Data for Health Services Research: A Framework and Instructional Guide. The University of North Carolina at Chapel Hill, Rockville (MD)/Agency for Healthcare Research and Quality (US), report no.: 14-EHC033-EF (2014)
10. General Data Protection Regulation (GDPR) http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf. Accessed 26 Jan 2016
11. Goldreich, O., Warning, A.: Secure multi-party computation (1998)
12. Goldstein, H., Harron, K., Wade, A.: The analysis of record linked data using multiple imputation with data value priors. Stat. Med. **31**(28), 3481–3493 (2012)
13. Goldstein, H., Harron, K., Cortina-Borja, M.: A scaling approach to record linkage. Stat. Med. **36**, 2514–2521 (2016). https://doi.org/10.1002/sim.7287
14. Government data-matching: Office of the Australian Information Commissioner—OAIC. https://www.oaic.gov.au/privacy-law/other-legislation/government-data-matching. Accessed 26 Jan 2018
15. Harron, K., Goldstein, H., Dibben, C. (eds.): Methodological Developments in Data Linkage. Wiley, Chichester (2015)
16. Harron, K., Doidge, J.C., Knight, H.E., Gilbert, R.E., Goldstein, H., Cromwell, D.A., Van der Meulen, J.H.: A guide to evaluating linkage quality for the analysis of linked data. Int. J. Epidemiol. **46**(5), 1699–1710 (2017)
17. Hendriks, P., Reynaert, M., van der Sijs, N.: Transcriptor, language and speech technology technical report series, Radboud University, Nijmegen (2016)
18. HIPAA for Professionals. https://www.hhs.gov/hipaa/for-professionals/index.html. Accessed 26 Jan 2016
19. HIPAA PHI: List of 18 Identifiers and Definition of PHI. https://cphs.berkeley.edu/hipaa/hipaa18.html. Accessed 21 Jan 2019
20. Jaro, M.A.: Probabilistic linkage of large public health data files, Match Ware Technologies. Stat. Med. **14**, 491–498 (1995)
21. Jiang, R., Rafael, E., Li, B., Li, H.: Evaluating and combining named entity recognition systems. In: Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, Berlin, 12 August 2016, pp. 21–27

22. Krewski, D.A., Wang, Y., Bartlett, S., et al.: The effect of record linkage errors on risk estimates in cohort mortality studies. Surv. Methodol. **31**(1), 13–21 (2005)
23. Kum, H.-C., Krishnamurthy, A., Machanavajjhala, A., et al.: Privacy preserving interactive record linkage (PPIRL). J. Am. Med. Inform. Assoc. **21**, 212–220 (2014)
24. Linking social care, housing & health data, Data linking: social care, housing & health: Paper 1, Data Linkage literature review (2010)
25. Marrero, M., Sánchez-Cuadrado, S., Lara, J.M., Andreadakis, G.: Evaluation of named entity extraction systems. In: Advances in Computational Linguistics, Research in Computing Science, pp. 41–47 (2009)
26. Mendes, R., Vilela, J.: Privacy-preserving data mining: methods, metrics, and applications. IEEE Access. **5**, 10562–10582 (2017). https://doi.org/10.1109/ACCESS.2017.2706947
27. Queensland Data Linkage Framework, Published by the State of Queensland (Queensland Health) (2014)
28. Sadinle, M.: Bayesian estimation of bipartite matchings for record linkage. J. Am. Stat. Assoc. **112**(518), 600–612 (2017). https://doi.org/10.1080/01621459.2016.1148612
29. Statistical Data Integration involving Commonwealth Data, National Statistical Service, Australian Government. https://toolkit.data.gov.au/index.php/Statistical_Data_Integration. Accessed 21 Jan 2019
30. Van der Sijs, N., Hendriks, P.: Al-Kadafi and Tsjechov: Waarom de spelling van namen ertoe doet. Onze Taal **11**, 10–14 (2017)
31. Verykios, V.S., Elmagarmid, A.K., Moustakides, G.V.: Cost optimal record/entity matching. Purdue e-Pubs, Purdue University, report number: 01-014 (2001)
32. Winkler, W.E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, Bureau of the Census* Statistical Research Division, Rm 3000-4, Washington, DC 20223 (1990)
33. Winkler, W.E.: Methods for record linkage and Bayesian networks. In: Proceedings of the Section on Survey Research Methods, pp. 3743–3748. ASA, Boston (2002)
34. Yuan, Y.C.: Multiple imputation for missing data: concepts and new development. In: Statistics and Data Analytics. SAS Institute, Rockville, Paper 267-25 (2000)