

Examining the quality of record linkage process using nationwide Brazilian administrative databases to build a large birth cohort

CURRENT STATUS: UNDER REVIEW

BMC Medical Informatics and Decision Making  BMC Series

Daniela Almeida
CIDACS

David Gorender
CIDACS

Maria Yury Ichihara
CIDACS

Samila Sena
CIDACS

Luan Menezes
CIDACS

George C. G. Barbosa
CIDACS

Rosimeire L. Fiaccone
Universidade Federal da Bahia

Enny Paixao
London School of Hygiene and Tropical Medicine

✉ npaixaoenfo@yahoo.com.br *Corresponding Author*
ORCID: <https://orcid.org/0000-0002-4797-908X>

Robespierre Pita
CIDACS

Mauricio L. Barreto
CIDACS

DOI:

10.21203/rs.3.rs-15927/v1

SUBJECT AREAS

Medical Informatics

KEYWORDS

Examining the quality of record linkage process, using nationwide Brazilian administrative databases to build a large birth cohort

Abstract

Introduction Research using linked routine population-based data collected for non-research purposes has increased in recent years because they are a rich and detailed source of information. The objective of this study is to present an approach to prepare and link data from administrative sources in a middle-income country, to estimate its accuracy and to identify potential sources of bias by comparing linked and no-linked case.

Methods We linked two administrative datasets with data covering the period 2001 to 2015, using maternal attributes (maternal name, age, date of birth, and municipality of residence) from Brazil: live birth information system and the baseline of the 100 Million Brazilian Cohort (created using administrative records from over 114 million individuals whose families applied for social assistance via the National Register for Social Programmes) implementing an in house developed linkage tool CIDACS-RL. We then estimated the accuracy of the linkage and examined the characteristics of missed-matches to identify any potential source of bias.

Results A total of 27,699,891 live births were recorded of those, 16,447,414 (59.4%) were linked with SINASC. The sensitivity of the linkage ranged from 39.3% in 2001 to 82.1% in 2014. A substantial improvement in the linkage sensitivity after the introduction of maternal date of birth attribute, in 2011, was observed. Our analyses indicated a slightly higher proportion of missing data among missed matches and a higher proportion of people living in an urban area and self-declared as Caucasian among linked pairs when compared with non-linked sets.

Discussion We demonstrated that CIDACS-RL is capable of performing high quality and accurate linkage even with a limited number of common attributes, using indexation as a blocking strategy in large routine databases from a middle-income country. However, residual records occurred more among people under worse living conditions. The results presented in this study reinforce the need of evaluating linkage quality and when necessary to take linkage error into account for the analyses of any generated dataset.

Introduction

Research using routine population-based data collected for social, financial, and clinical purposes has

increased in recent years because they are a rich and detailed source of information available at a relatively low cost ¹. Record linkage (process used to bring together information recorded in different sources about the same individual) ² of multiples databases can further enhance the ability to answer scientific questions of isolated databases. Particularly on maternal and infant health, where administrative linked data can increase the availability of information on maternal health, social, and economic trajectories before and during pregnancy ³. The use of linked high-quality administrative datasets provides a unique opportunity to examine factors that might result in long-term and rare child and maternal outcomes over time, with the additional advantage of using large samples, little loss to follow-up, high level of external validity and a great deal of applicability for policymaking⁴⁻⁶. Record linkage can be conducted using two main methods: deterministic and non-deterministic. Deterministic linkage usually uses a unique identifier or a set of several attributes present in all the databases to be linked⁸. Non-deterministic record linkage solutions are suitable when there is not a shared key to identify univocally an individual across disparate data sources⁹. This situation is frequent in different countries, in particular in low and middle-income ones. To perform this procedure, we have to submit the most reliable and discriminative variables present in both databases to calculate similarity scores representing the likelihood that two records belong to the same person. The non-deterministic approach tolerates some variations between records, such as missing data, and it can link records with errors in the linking fields, and it has facilitated many studies using datasets without a unique identifier ⁹. The similarity score is used to classify records as links, non-links, and uncertain links based on one or more thresholds. The choice of threshold needs to balance the risk of "false-matches" (records from different individuals that are mistakenly linked) and "missed-matches" (records from the same individual that fail to link) ¹⁰. Some extensions of linkage error in administrative data are expected and inevitable due to the imperfect and transient nature of the attributes. However, even a small amount of linkage error can lead to biased results, diluting real association, or creating spurious ones¹¹. Measures of sensitivity,

specificity, positive and negative predictive values are commonly used to estimate the linkage accuracy. Nevertheless, these measures might not evidence in which extend the results of analyses using the linked data could be biased. Therefore, it is essential to combine these measures with alternative methods to evaluate linkage quality⁴.

We aimed to use Brazilian nationwide administrative databases to build a birth cohort originated by the intercept of live births and the baseline of the 100 Million Brazilian Cohort (created using administrative records from over 114 million individuals whose families applied for social assistance via the National Register for Social Programmes). Linkage aimed to enhance the live birth data with socioeconomics information. There is a large overlap between the baseline of the 100 Million Brazilian Cohort and the live birth databases. In this scenario, we were able to measure the linkage error. This study presents an approach to prepare and link data from administrative sources in a middle-income country, estimating its accuracy and identifying potential sources of bias by comparing link and no-links.

Methods

In this section, we will describe all the methodology to integrate two majors nationwide databases, namely the Live Birth Information System (SINASC) and the United Registry for Social Programmes (CadUnico) from 2001 to 2015.

Datasets

SINASC (Sistema de Informação Sobre Nascidos Vivos/ Live Birth Information System)

The Brazilian Ministry of Health defines live births as the complete expulsion or extraction from the body of the pregnant woman of a product of conception, independent of the duration of pregnancy, who, after the separation, breathes or shows any other signs of life, such as heartbeat, umbilical cord pulsation, or definite movement of voluntary muscles, whether or not the cord is cut and whether or not the placenta is attached. SINASC records live births in Brazil, and this system is updated using the registration of live birth. It is a compulsory document, completed by a health professional who assisted the delivery. This form is divided into eight blocks. I -characteristics of the newborn; II- identification of the place of birth; III- characteristics of the mother; IV- identification of the father; V-

characteristics of pregnancy and delivery; VI- characteristics of congenital anomalies: this block should be filled in when congenital anomalies are identified at birth using the ICD-10 code. VII- identification of the professional completing the notification. VIII- registry office identification ¹².

The baseline of the 100 Million Brazilian Cohort

The Cadastro Único has become the main instrument used by the Brazilian government to assess the inclusion criteria of potential beneficiaries of social programs. To be enrolled in CADU, one person in the family must provide information and required documents of all family members to an interviewer. This person must be at least 16 years old and, preferably, be a woman. The information is renewed periodically as long as the person is a candidate to receive one of the several Brazilian government social protection programs¹³. The Centre for Data and Knowledge Integration for Health - CIDACS has the custody of several snapshots of CADU. Each snapshot file refers to a year backup from 2001 to 2015. The efforts to build the 100 Million Brazilian Cohort were concentrated in three main steps. The first was the harmonization of attributes with a scheme or meaning divergence on some attributes across three different versions of CADU. Second, the data cleansing to ensure the standardization of the categories. The third step aims to find the first appearance of each record over a disparate CADU backup file. This single register for social programs is an instrument that identifies and characterizes low-income families applying for any social protection program, that also allows to improve the understanding of the social reality of this population group. It contains information on social, environmental, and economic features on named individuals grouped into families.

The process of linking

Data pre-processing

During the data pre-processing phase, first, we searched automatically for invalid names (e.g., "unknown" or "newborn"), by comparing the recorded name with a standardized list of possible Brazilian names. All names considered invalid are submitted to a clerical review to confirm that they cannot be used in the linkage process, then this attribute is excluded. We removed punctuation, deleted consecutive spaces; middle initials, prefixes, and suffixes were maintained as recorded to retain the discriminatory power of the name variable.

Blocking/ Indexing

The complexity of the record linkage task is quadratic. We have to find the best candidate, on database B, for each record in database A, $|A| \times |B|$. To enable the record linkage promptly when massive datasets are involved, we need to resort to methods capable of avoiding unnecessary comparisons, keeping the accuracy, once, the total number of pairwise comparisons between SINASC and CadUnico would be prohibitively high $44,485,267 \times 114,007,705 = 5,07166e15$. To meet these challenges, we use the CIDACS-RL ¹⁴; a novel record linkage tool developed to link big administrative datasets at the CIDACS.

The CIDACS-RL applies the combination of indexing and searching algorithms implemented in Apache Lucene solution as the blocking strategy to reduce the number of comparisons during the linkage. The indexation strategy allows the CIDACS-RL to search the most similar records from the Indexed baseline of the 100 Million Brazilian Cohort for each record in SINASC and submit them to the pairwise comparisons step, instead of restricts the comparison group as an ordinary blocking step. This search was performed in two ways, (i) using the mothers' name, municipality, and mothers date of birth records as attributes, from 2011 to 2015 (ii) using mothers name and municipality, from 2001-2010, because the mothers' date of birth was not registered before 2011. This search strategy uses a mixture of exact, semi fuzzy and fuzzy queries to return the 1000 best candidates from the indexed baseline of the 100 Million Brazilian Cohort. The exact queries return only records with equal attributes in every querying, while the semi-fuzzy and fuzzy approaches permit more flexibility by retrieving candidates where one (semi-fuzzy) or more attributes differ (fuzzy). In cases were certain uncertainty is included in the name variable, the Damerau-Levenshtein distance is used as a string comparator, and values above 0.5 are considered ¹⁴.

Pairwise Comparison

The most discriminant variables available on the live birth database to identify a child are a mother's name, municipality, and age. For those records from 2011 to 2015, the mothers' date of birth attribute becomes available, and its filling increases gradually across the years. For 2001-2010, where

the mothers date of birth is not available, we proceeded with the search using only two attributes (mothers name and municipality) then, we create a new variable by subtracting the date of birth of the child information recorded in SINASC from the date of birth of the mother recorded in baseline of the 100 Million Brazilian Cohort, and this value was compared with the age of the mother registered in SINASC, only the candidates with exacted same value were considered as possible candidates and submitted to the pairwise comparison step. This step was also executed for records from 2011 to 2015 with missing values in the mothers' date of birth.

Figure 1 describes the two different approaches for each set of available variables. Then CIDACS-RL set weights according to the discriminatory power of the attributes (name of the mother: 1 maternal age or date of birth: 1 state of birth: 0.008, municipality of birth: 0.16). At that moment, a combined scoring and query modules are used to perform the record linkage.

The similarities between names recorded in SINASC and the 1000 best candidates from the baseline of the 100 Million Brazilian Cohort were compared using the Jaro-Winkler string comparator ¹⁵. The Jaro-Winkler string comparator¹⁵ counts the number of common characters between two strings and the number of transpositions of these common characters, producing similarity values varying between 0 and 1 (perfectly similar). To compare the date attributes, we applied the Hamming distance ¹⁴, which measures the minimum number of substitutions required to change one string into the other. Then a linkage score is generated, and the function returns all pairs matched along with the score obtained.

Selection of the threshold

Candidate linking records were ordered by the scores achieved; only the comparison pair with the highest score is retained as a potential link. All remaining candidate records are discarded. Then a sample of 2000 pairs stratified in three categories of linkage score (high score - above 0.95, intermediate score - values between 0.90 and 0.95, and low score - bellow 0.90) is evaluated manually, and the records pairs are classified as likely true pairs or likely false pairs. Based on the training dataset of 2000, the receiver operating curve (ROC) is built to choose the best cut off point,

and calculating the area under the curves (AUC), balancing between sensitivity and specificity values. Records were therefore classified as links or non-links based on a single threshold. The software R is used to generate accurate results.

Evaluation of the linkage error

Since we expected that all births registered in the baseline of the 100 Million Brazilian Cohort overlapped with the births existing at SINASC databases, we were able to identify the number of missed matches (records from the same mother-baby pair that failed to link) and to estimate the sensitivity (true links among the matches) of the linkage. We then examined which characteristics were associated with missed matches. We examined race, sex, place of residence, sewage treatment, water supply, garbage collection.

Results

A total of 27,699,891 live births were recorded in the baseline of the 100 Million Brazilian Cohort dataset from 2001 to 2015. Of those, 16,447,414 (59,4%) were linked with SINASC dataset. However, the proportion of linked pairs were not similar over the years (Table 1). In general, the sensitivity of the linkage improved over the years. It ranged from 39.3% in 2001 to 82.1% in 2014. The greatest improvement was observed from 2010 to 2011 when the proportion of links increased by 10% (Table 1).

In general, missed-matches had a higher proportion of missing data in some living conditions variables such as water supply, sewage treatment, garbage collection, compared with linked pairs. According to the socio-demographic's characteristics, the linked group was more likely to live in an urban area and self-declared as Caucasian when compared with non-linked pairs (Table 2).

Discussion

We have implemented the linkage tool CIDACS-RL (REF) developed in house in a dataset with a known number of expected matches and consequently were able to quantify measures of linkage accuracy. We demonstrated that CIDACS-RL is capable of performing high quality and accurate linkage even with a limited number of common attributes, using indexation as a blocking strategy in a large routine dataset from a middle-income country. Our study showed that the improvement of data quality,

characterize by the addition of one more identifier (mother date of birth), lead to a significant improvement in the linkage quality, which increased the sensitivity in more recent years, reaches more than 80%. Our comparison of missed-matches indicates a slightly higher proportion of missing data among missed matches and a higher proportion of people living in an urban area and self-declared as Caucasian among linked pairs when compared with non-linked sets.

An essential consideration of this linkage is the massive amount of data, which increases the technical complexity to perform the linkage process in a scalable and accurate way. The innovative of the CIDACS-RL is the use of the search engine indexing as a blocking strategy¹⁴. A traditional blocking strategy is applied to reduce the number of potential records comparisons that likely not match and avoid waste of computational resources. However, this strategy can result in linkage error if true matches were separated in different blocks¹⁶. To avoid linkage error without compromise the linkage scalability, CIDACS-RL implemented a dynamical search function that uses all linkage attributes for searching, which avoids computational waste similar to traditional blocking strategy without compromise the linkage accuracy since it prevents linkage errors by non-separating in blocks potential matches in the process.

The use of a classical record linkage approach, as proposed by Fellegi and Sunter²⁰, was unfeasible given the unavailability of the matches and unmatched expected to the pair of disparate data sets involved in the merge procedure. Without these weights, frequently provided by a gold-standard, one cannot fit the probability-based classification model. The main difference between the CIDACS-RL method to the classical approach is the implementation of a similarity-based linkage that outputs the best pair of records and its similarity. Whereas the lack of a probability to separate the matches from unmatched, a cut-off point must be calculated through a time- and cost-intensive clerical review. On the non-deterministic linkage approach, the choice of thresholds is not straightforward, and it is going to impact directly on linkage quality. Decisions about linkage are usually based on linkage scores of the complete dataset⁴. However, due to the massive amount of data, manual review for the complete dataset of comparison pairs was not possible. Therefore, it was selected a stratified sample

size of 2000. The size of the sample was decided based on reasonability for manual revision that exhibited the same characteristics of the complete dataset on score distribution. The next step will be increasing the sample size and vary the characteristics of the sample and the linkage threshold to evaluate the linkage quality further.

Although linkage to enhance the same individual information can accomplish high sensitivity rates ¹⁷, the process of link information of two different people (in this case, mother and baby) has been considered a more problematic task, due to the limited number of shared identifiers within datasets^{3,7}. Which directly impacts on sensitivity results, which tend to be lower. In our study, the proportion of missed-matched records varied from 61% to 18%. In the first years of the study, our sensitivity was much lower than identified in similar studies in high-income countries. However, after the inclusion of the mother date of birth attribute, the proportion of missed-matches was similar to studies developed in the US States of Georgia ¹⁸ and New Jersey ¹⁹. Another similarity with those studies was the higher proportion of vulnerable populations among residual records (rural, and worse living conditions).

This study has several limitations. A weakness that must be discussed concerning each specific research question to be answered using CIDACS birth cohort is restricted to the people enrolled at CadUnico (half poorest of the Brazilian population). The main limitation inherent to the linkage process is the low sensitivity in the first years before the introduction of the mother's date of birth. This information is highly valuable because when using our cohort it could be decided to use only those years that have achieved the highest sensitivity results. More important than the accuracy of linkage in terms of sensitivity, the linkers have to guarantee that the linkage error did not introduce bias in the final analyses. Although the difference in some living conditions variables and socio-demographic's characteristics between the linked and non-linked groups were less than 10% percent, even small amounts of linkage error can result in substantially biased results. Therefore, we recommend further studies to evaluate if these small differences can introduce bias and to take this in consideration in any future analyses using our birth cohort.

An essential step of the linkage process is to estimate the linkage accuracy and to identify potential sources of bias that can be introduced in the results of analyses using the linked data. The linkage involving two nationwide large Brazilian databases evaluated here showed sensitivity value for more recent years comparable with previous finds in developed countries^{18, 19}. Although before the introduction of maternal date of birth in SINASC form, the proportion of missed match was much higher. The results presented in this study reinforce the need to evaluate linkage quality and to take linkage error into account as a preliminary step in the analyses of the linked datasets.

References

1. Casey, J. A., Schwartz, B. S., Stewart, W. F. & Adler, N. E. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu. Rev. Public Health* (2016) doi:10.1146/annurev-publhealth-032315-021353.
2. Sayers, A., Ben-Shlomo, Y., Blom, A. W. & Steele, F. Probabilistic record linkage. *Int. J. Epidemiol.* (2016) doi:10.1093/ije/dyv322.
3. Harron, K., Gilbert, R., Cromwell, D. & van der Meulen, J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. *PLoS One* **11**, e0164667 (2016).
4. Harron, K. *et al.* Challenges in administrative data linkage for research. *Big Data Soc.* (2017) doi:10.1177/2053951717745678.
5. Walker, J. R., Hilder, L., Levy, M. H. & Sullivan, E. A. Pregnancy, prison and perinatal outcomes in New South Wales, Australia: A retrospective cohort study using linked health data. *BMC Pregnancy Childbirth* (2014) doi:10.1186/1471-2393-14-214.
6. Hockley, C. *et al.* Linking Millennium Cohort data to birth registration and hospital episode records. *Paediatr. Perinat. Epidemiol.* (2008) doi:10.1111/j.1365-3016.2007.00902.x.
7. Paixão, E. S. *et al.* Evaluation of record linkage of two large administrative databases in a middle income country: stillbirths and notifications of dengue during pregnancy in Brazil. *BMC Med. Inform. Decis. Mak.* (2017) doi:10.1186/s12911-017-0506-5.

8. Newcombe, H. B., Kennedy, J. M., Axford, S. J. & James, A. P. Automatic linkage of vital records. *Science* (80-.). (1959) doi:10.1126/science.130.3381.954.
9. Clark, D. E. Practical introduction to record linkage for injury research. *Injury Prevention* (2004) doi:10.1136/ip.2003.004580.
10. Harron, K. A guide to evaluating linkage quality for the analysis of linked data.
11. Rentsch, C. T. *et al.* Impact of linkage quality on inferences drawn from analyses using data with high rates of linkage errors in rural Tanzania. *BMC Med. Res. Methodol.* (2018) doi:10.1186/s12874-018-0632-5.
12. S.Paulo, S. M. de S. de S. Manual de Preenchimento da Declaração de Nascido Vivo Prefeito do Município de São Paulo. 1-24 (2011).
13. Barros, R. P. de, Carvalho, M. de & Mendonça, R. Sobre as utilidades do Cadastro Único. *Texto para discussão no 1414* (2009).
14. Ali, M. S. *et al.* Administrative Data Linkage in Brazil: Potentials for Health Technology Assessment. *Front. Pharmacol.* **10**, 1-20 (2019).
15. William E. Yancey. Evaluating string comparator performance for record linkage. *Stat. Res. Div.* 3905-3912 (2005).
16. Steorts, R. C., Ventura, S. L., Sadinle, M. & Fienberg, S. E. A comparison of blocking methods for record linkage. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2014). doi:10.1007/978-3-319-11257-2_20.
17. St Sauver, J. L. *et al.* Linking medical and dental health record data: A partnership with the Rochester Epidemiology Project. *BMJ Open* (2017) doi:10.1136/bmjopen-2016-012528.
18. Reichman, N. E. & Hade, E. M. Validation of birth certificate data: A study of women in New Jersey's healthstart program. *Ann. Epidemiol.* (2001) doi:10.1016/S1047-

2797(00)00209-X.

19. Adams, M. M. *et al.* Constructing reproductive histories by linking vital records. *Am. J. Epidemiol.* **145**, 339–348 (1997).
20. FELLEGI, Ivan P.; SUNTER, Alan B. A theory for record linkage. *Journal of the American Statistical Association*, v. 64, n. 328, p. 1183-1210, 1969.

Abbreviations

SINASC: Live Birth Information System; CadUnico: Single Register for Social Programs; CIDACS: Centre for Data and Knowledge Integration for Health; ROC: receiver operating curve; AUC: area under the curves

Declarations

Ethics approval

The CIDACS maintains a linkage system for social and health-related data following all ethical, legal, privacy, and confidentiality requirements. The study protocol was reviewed and approved by the Instituto of Public Health Ethics Committee at the Federal University of Bahia (CAAE registration number: 18022319.4.0000.5030).

Availability of data and material

The identified data used to conduct this study is highly sensible and confidential, because they include patient personal information that can be traced back to individual. They are obtainable in the Brazilian Ministry of Health but restrictions apply to the availability of these data, which were used under license, and so are not publicly accessible. However de-identified linked data can be accessed upon reasonable request for researchers who meet the criteria for access to confidential data.

Consent to publish

Not applicable

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank the CIDACS data processing team for all the intense work.

Funding

CIDACS received core support from Health Surveillance Secretary, Ministry of Health, Brazil; Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB); Wellcome Trust (Grant number 202912 / Z / 16 / Z); Financiadora de Estudos e Projetos-FINEP; Secretary of Science and Technology of the State of Bahia-SECTI. ESP is funded by the Wellcome Trust (grant number 213589/Z/18/Z) However the funder of this study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Author's contributions

DA, DG, LM, GCGB carried out the analysis and interpretation. ESP, SS, RP wrote the first draft of the article. MB, ESP, RLF, MYI conceived the study. All authors revised the manuscript and approved the final version.

Tables

Table 1. Number and percentage of linked records by year, Brazil, 2001-2015.

| Year | Total | Linked | |
|--------------|-------------------|-------------------|--------------|
| | | N | % |
| 2001 | 2,448,609 | 961,605 | 39.27 |
| 2002 | 2,319,071 | 1,175,223 | 50.68 |
| 2003 | 2,224,872 | 1,179,781 | 53.03 |
| 2004 | 2,165,661 | 1,144,809 | 52.86 |
| 2005 | 2,161,484 | 1,183,292 | 54.74 |
| 2006 | 2,050,534 | 1,271,179 | 61.99 |
| 2007 | 1,961,446 | 1,087,254 | 55.43 |
| 2008 | 1,936,675 | 1,077,781 | 55.65 |
| 2009 | 1,855,919 | 1,052,394 | 56.70 |
| 2010 | 1,778,515 | 1,067,417 | 60.02 |
| 2011 | 1,765,211 | 1,249,492 | 70.78 |
| 2012 | 1,662,414 | 1,251,251 | 75.27 |
| 2013 | 1,505,476 | 1,227,162 | 81.51 |
| 2014 | 1,271,156 | 1,043,499 | 82.09 |
| 2015 | 592,848 | 475,275 | 80.17 |
| Total | 27,699,891 | 16,447,414 | 59.38 |

Table 2: Associations between the characteristics of the cohort and the accuracy of the linkage

| Characteristics | 2001 | | | | 2002 | | | |
|------------------------|--------|-------|------------|-------|--------|-------|------------|-------|
| | Linked | | Non-linked | | Linked | | Non-linked | |
| Water supply | | | | | | | | |
| Missing | 11610 | 0.78 | 11610 | 0.78 | 6737 | 0.57 | 10704 | 0.94 |
| Public supply | 982902 | 66.10 | 982902 | 66.10 | 840166 | 71.49 | 718413 | 62.81 |
| Well | 361618 | 24.32 | 361618 | 24.32 | 237348 | 20.20 | 306442 | 26.79 |
| Other | 130874 | 8.80 | 130874 | 8.80 | 90972 | 7.74 | 108289 | 9.47 |
| Sanitary sewage | | | | | | | | |
| Missing | 8188 | 0.85 | 22896 | 1.54 | 10698 | 0.91 | 22498 | 1.97 |
| Public collection | 378673 | 39.38 | 616471 | 41.46 | 542173 | 46.13 | 413379 | 36.14 |
| | 158983 | 16.53 | 207798 | 13.97 | 169977 | 14.46 | 174661 | 15.27 |

| | | | | | | | | |
|------------------------------------|--------|-------|---------|-------|--------|-------|--------|-------|
| Septic tank Rudimenta ry Pit | 253761 | 26.39 | 371292 | 24.97 | 275583 | 23.45 | 314883 | 27.53 |
| Ditch | 143119 | 14.88 | 237890 | 16.00 | 154700 | 13.16 | 193111 | 16.88 |
| Other | 18881 | 1.96 | 30657 | 2.06 | 22092 | 1.88 | 25316 | 2.21 |
| Waste destinati on | | | | | | | | |
| Missing | 5518 | 0.57 | 11616 | 0.78 | 6746 | 0.57 | 10707 | 0.94 |
| Collected | 698035 | 72.59 | 1035356 | 69.63 | 895313 | 76.18 | 760128 | 66.45 |
| Burnt / Buried | 173667 | 18.06 | 294548 | 19.81 | 185235 | 15.76 | 255937 | 22.38 |
| Landfill | 75287 | 7.83 | 127226 | 8.56 | 76279 | 6.49 | 102559 | 8.97 |
| Other | 9098 | 0.95 | 18258 | 1.23 | 11650 | 0.99 | 14517 | 1.27 |
| Education | | | | | | | | |
| Missing | 33774 | 3.51 | 64030 | 4.31 | 30566 | 2.60 | 26766 | 2.34 |
| Pre-school | 149013 | 15.50 | 199390 | 13.41 | 159451 | 13.57 | 144989 | 12.68 |
| Literacy | 63098 | 6.56 | 84264 | 5.67 | 50419 | 4.29 | 59410 | 5.19 |
| Elementar y school | 204054 | 21.22 | 455525 | 30.63 | 258841 | 22.02 | 321182 | 28.08 |
| High school | 956 | 0.10 | 2062 | 0.14 | 810 | 0.07 | 1369 | 0.12 |
| College education | 52 | 0.01 | 108 | 0.01 | 40 | 0.00 | 55 | 0.00 |
| Illiteracy | 510658 | 53.10 | 681625 | 45.84 | 675096 | 57.44 | 590077 | 51.59 |
| Race/colo ur | | | | | | | | |
| Missing | 21525 | 2.24 | 25275 | 1.70 | 7842 | 0.67 | 5756 | 0.50 |
| Caucasian | 312717 | 32.52 | 474201 | 31.89 | 416619 | 35.45 | 317729 | 27.78 |
| Black | 56836 | 5.91 | 78562 | 5.28 | 64200 | 5.46 | 58459 | 5.11 |
| Asian | 3465 | 0.36 | 4667 | 0.31 | 3179 | 0.27 | 3413 | 0.30 |
| Brown | 562957 | 58.54 | 890725 | 59.90 | 678971 | 57.77 | 746025 | 65.22 |
| Indigenous | 4105 | 0.43 | 13574 | 0.91 | 4412 | 0.38 | 12466 | 1.09 |
| Sex | | | | | | | | |
| Male | 491672 | 51.13 | 763571 | 51.35 | 601587 | 51.19 | 585191 | 51.16 |
| Female | 469933 | 48.87 | 723433 | 48.65 | 573636 | 48.81 | 558657 | 48.84 |
| Zone | | | | | | | | |
| Missing | 123 | 0.01 | 405 | 0.03 | 138 | 0.01 | 236 | 0.02 |
| Urban | 724567 | 75.35 | 1079682 | 72.61 | 920713 | 78.34 | 796832 | 69.66 |
| Rural | 236915 | 24.64 | 406917 | 27.36 | 254372 | 21.64 | 346780 | 30.32 |

Figures

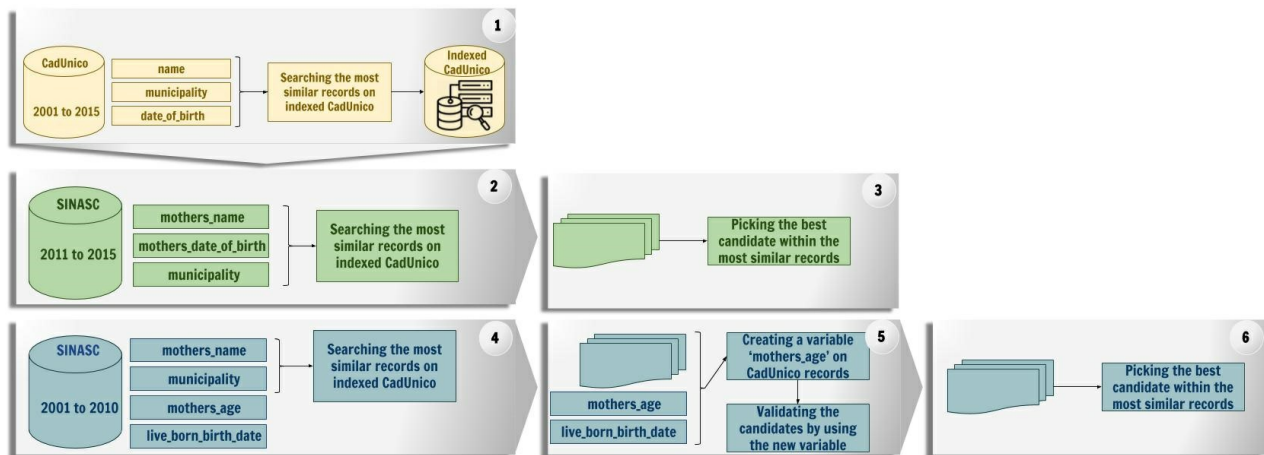


Figure 1

describes the two different approaches for each set of available variables. Then CIDACS-RL set weights according to the discriminatory power of the attributes (name of the mother: 1 maternal age or date of birth: 1 state of birth: 0.008, municipality of birth: 0.16). At that moment, a combined scoring and query modules are used to perform the record linkage.