Optimizing Record Linkage Parameters with Genetic Algorithms for Health Data Integration

Pablo L. Pita[2], Katie Harron[3], Islame F. C. Fernandes[1], José Gomes[2], Gabriel F. Rocha[1,2], Pablo I. P. Ramos[2], Samila Sena[2], Roberto Carreiro[2], Fernanda Eustáquio[2], George Cardozo[2], Carlos C. Santos[1,2], Leandro Sá Ribeiro[2], Sara Sena[2], Bethania Almeida[2], Mauricio L. Barreto[2], Robespierre Pita[1,2]

[1] Institute of Computing – Federal University of Bahia (UFBA)

[2]Center for Data and Knowledge Integration for Health (CIDACS)

[3]Institute of Child Health – University College London (UCL)

pablolemos2004@hotmail.com, k.harron@ucl.ac.uk, islame.felipe@ufba.br,

{joseaugusto.duarte,gabriel.frocha,pablo.ramos}@fiocruz.br,

{roberto.carreiro,fernanda.eustaquio,samila.sena}@fiocruz.br,

{george.cardozo,carlos.jsantos,leandro.sa}@fiocruz.br,

{sara.sena,bethania.almeida,mauricio.barreto}@fiocruz.br,

robespierre.pita@ufba.br

Abstract. *Record linkage is widely used to integrate administrative health databases, but its performance depends on appropriate parameterization and decision thresholds. We propose optimizing CIDACS-RL parameters using genetic algorithms to efficiently explore the parameter space. Experiments linking Brazilian live birth (SINASC) and mortality (SIM) records using a labeled dataset reduced false positives by about 90% while maintaining high recall. Precision increased from 0.70 to 0.96 and accuracy from 0.89 to 0.99, with consistent improvements across all southern Brazilian states analyzed. These results suggest that parameter optimization can improve linkage quality and the reliability of large-scale health data integration.*

1. Introduction

Extracting knowledge from administrative data in Brazilian health information systems often requires combining data from multiple sources through record linkage (RL) solutions. Integrating health-related records with supplementary demographic and socioeconomic data is essential for addressing complex analytical and research questions, such as estimating disease risk or assessing the impact of public policies on specific populations [Ali et al. 2019]. However, the absence of unique identifiers for individuals routinely collected by different health services makes RL a critical but error-prone task.

The CIDACS-RL is a score-based probabilistic linkage tool designed to support the integration of high-dimensional administrative data. It has been used to link national health information systems to large-scale cohorts, including the 100 Million Brazilian

Cohort [Barreto et al. 2022] and a national birth cohort [Paixao et al. 2021]. CIDACS-RL relies on a composite similarity score computed from pairs of quasi-identifiers (QIDs). These QIDs correspond to sets of non-unique attributes that, when combined, can be used to identify the same individual [Harron et al. 2016].

A successful data integration procedure is highly dependent on selecting appropriate weights and penalties for the composite similarity score, which is a non-trivial task that has a substantial impact on matching quality. However, including a comprehensive grid-search strategy in the CIDACS-RL pipeline for large data collections, as detailed in Section 3, is impractical due to the substantial computational demands of large-scale data processing and the extensive execution time required.

To address these limitations and enable cost-efficient parameter discovery, this study proposes using genetic algorithms to automatically optimize CIDACS-RL's parameters. We hypothesize that defining a record linkage quality metric, such as accuracy, as the objective function can reduce linkage errors by systematically guiding parameter selection. Our main contribution is the introduction of a decision-layer optimization pipeline for record linkage, which improves linkage quality and potentially mitigates bias propagation in downstream analytical and epidemiological studies.

2. Related Work

Recent developments in score-based probabilistic record linkage software have introduced support for modular pipelines and for parametrizing pairwise comparisons. For example, traditional solutions, such as the Record Linkage Toolkit [De Bruin 2022] and Splink [Linacre et al. 2022], allow users to configure indexing, pairwise comparison, and classification stages, as well as to experiment with similarity measures tailored to date, numeric, textual, and categorical attributes. However, reducing reliance on expert-driven parameter tuning remains a challenging problem. Most recent advances have been reported in the context of privacy-preserving record linkage (PPRL) [Gkoulalas-Divanis et al. 2021, Yu et al. 2020], where automation typically relies on machine learning models or heuristic strategies to guide Bloom filter-based encoding under privacy constraints.

The application of genetic algorithms in record linkage has predominantly focused on selecting or combining similarity functions within score-based probabilistic frameworks [Shaikh and Ragkhitwetsagul 2008, WAYKOLE and SHINDE 2014]. Alternative optimization strategies have also been explored. For example, [Nelson et al. 2023] proposed a Bayesian optimization framework to refine matching decisions in a probabilistic record linkage setting. However, this approach depends heavily on the availability of large amounts of previously integrated and labeled data for training, which may limit its applicability in large-scale administrative contexts where curated ground truth is scarce. This study addresses the lack of alternatives optimization of decision-layer parameters in score-based systems, introducing a genetic algorithm to the CIDACS-RL workflow.

3. CIDACS-RL and the Parameter Optimization Problem

Figure 1 presents the seven-step CIDACS-RL integration workflow for administrative health data. The description below focuses on the components highlighted in gray, which correspond to the original CIDACS-RL pipeline. Firstly, the system indexes the

larger data source into an Elasticsearch cluster to enable distributed approximate nearest-neighbor search over the indexed structure. By retrieving the K most similar candidates for each record in the smaller dataset, the system reduces the number of pairwise comparisons from $|A| \times |B|$ to $|A| \times |K|$, where $|A|$ and $|B|$ denote the sizes of the smaller and larger datasets, respectively.

The following step involves constructing an exact query using a subset of highly discriminative QIDs, including name, mother's name, and date of birth. When no candidates are retrieved at this stage, the workflow proceeds to a non-exact query in the third step. This query relies on should operators and fuzzy clauses to relax matching constraints. It may incorporate additional attributes, such as municipality of residence, ethnicity, sex, or nationality. This two-stage retrieval strategy balances precision during the exact phase with recall during the relaxed phase, limiting candidate explosion while preserving linkage sensitivity.

In the fourth step, a function computes a composite similarity score for each candidate record returned by the querying stages. This scoring procedure assigns attribute-specific weights according to their relative importance and applies penalties to record pairs that contain missing values. These parameters together produce an interpretable ranking of candidate matches. The fifth step specifies the linkage criteria, selects the candidate with the highest similarity score, and assigns it as the link for each record in the smaller dataset.

Given the absence of ground truth and known match status, the subsequent workflow stages focus on validation. In the sixth step, the process constructs a reference dataset by manually reviewing a score-stratified sample of record pairs. During this review, the reviewers may assign a link status to each pair, with '1' indicating a link and '0' indicating a non-link. The labeled dataset supports the construction of a receiver operating characteristic (ROC) curve and enables the estimation of an optimal similarity threshold using Youden's index. In the seventh step, the system applies this threshold to perform the final classification. Pairs with similarity scores above the threshold correspond to true positives (TP), whereas pairs below the threshold correspond to true negatives (TN) when they represent non-links. Misclassifications occur when non-links exceed the threshold, leading to false positives (FP), or when true links fall below the threshold, leading to false negatives (FN).

Let us suppose a pair of data sets, namely $A_{|A|\times m}$ $B_{|B|\times m}$, where p is the number of attributes. Here, a data point $\mathbf{Ax}_{ij}$ corresponds to the j^{th} attribute from the i^{th} observation, and therefore, $\mathbf{Ax}_{ij} = (x_{i1}, ..., x_{im})_{1\times m} \in \mathbf{A}$ is the vector with values of a given record of $\mathbf{A}$. Also, let $\mathbf{D} = (d_1, ..., d_m)_{1\times m} \in \mathbf{T}$ denote the attribute-type vector, where specifies the domain of the j^{th} attribute, assuming types in $\mathbf{T} = \{textual, categorical, date, boolean\}$. In the CIDACS-RL workflow, the overall similarity between two records in disparate sources, expressed as $S(\mathbf{Ax}_{i.}, \mathbf{Bx'}_{i'.})$, must apply an appropriate function for each domain.

The mapping between domains and similarity measures enables a domain-aware comparison framework. Let us consider a hypothetical record linkage task using CIDACS-RL, with six QIDs provided as input to the composite similarity. We define the comparison functions $c_j(.,.)$ according to the attribute type. Specifically, c_1 and c_2

implement the Jaro-Winkler similarity for textual attributes (name and mother's name). The system applies the Hamming similarity to structured numeric attributes (date of birth and municipal codes) in c_3, c_5 and c_6. Finally, c_4 corresponds to the overlap function for the binary attribute sex. In this context, using attribute-specific weights and penalties can contribute to the composite similarity score by reflecting the discriminatory contribution of each attribute and addressing uncertainty from missing values.

We define $\mathbf{w} = (w_1, .., w_6)$ and $\mathbf{p} = (p_1, .., p_6)$ as the vectors of weights and penalties, respectively. The composite similarity score between two records $\mathbf{Ax_{i.}}$ and $\mathbf{Bx'_{i'.}}$ is defined as the following weighted sum of attribute-level similarity

$$S_{w,p}(\mathbf{Ax_{i.}}, \mathbf{Bx'_{i'.}}) = \sum_{j=1}^{6} w_j \times c_j(\mathbf{Ax_{i.}}, \mathbf{Bx'_{i'.}}) - \sum_{j=1}^{6} p_j \times m_j(\mathbf{Ax_{i.}}, \mathbf{Bx'_{i'.}}), \tag{1}$$

where $m_j(.,.)$ flags missing values in the pair of attributes involved in the comparison, assuming zero if both have information present.

From a methodological perspective, the linkage criteria depend on both the score parameters in Equation 1 and the classification threshold, computed in the sixth step from CIDACS-RL's workflow. Identifying an optimal configuration, therefore, corresponds to solving

$$\max_{w,p} M(S_{w,p}), \tag{2}$$

where $M(S_{w,p})$ refers to an arbitrary performance metric estimated from labeled data, such as accuracy.

Let us assume a naive grid-searching task of the best parameters in the linkage example provided before, with the weight values w_{name}, $w_{mothersname}$, $w_{dateofbirth} \in [0, 3]$, and $w_{sex} \in [0, 1]$. The penalty values should be p_{name}, $p_{mothersname}$, $p_{dateofbirth}$, $p_{sex} \in [0, 1]$. Even under moderate discretization (step size of 0.05), the parameter space comprises approximately 6.9×10^{12} possible configurations and linkage runs, making the exhaustive search of solutions for Equation 2 computationally infeasible. Therefore, the parameter selection demands solutions capable of addressing combinatorial and potentially non-convex optimization problems, which motivates the adoption of evolutionary methods.

4. Genetic Algorithm for Parameter Optimization

Our proposal aims to perform parameter tuning in CIDACS-RL, searching for a vector that maximizes the linkage quality using labeled data as input, while avoiding repeated execution of computationally expensive linkage runs. The extension illustrated in Figure 1, highlighted in green, comprises five additional steps. The first four of these additional steps (i.e., a) to d) in Figure 1) focus on constructing a reference dataset, whereas the final step e) applies a genetic algorithm to optimize the decision-layer parameters.

The first step a) of our proposed extension consists of getting a sample from the smaller dataset. Ideally, the sampling strategy must comply with a stratification by geographic region and temporality-related variables, such as state and year of registry, ensuring representativeness across heterogeneous subpopulations. In the second and third steps b) and c), the system executes Elasticsearch queries to retrieve candidate matches from the

indexed source. It then assigns candidates from different ranking positions to each sampled record according to a predefined proportion of potential matches and non-matches. Although top-ranked candidates returned by Elasticsearch are more likely to correspond to true links within the sampled data, the fourth step e) requires manual review.

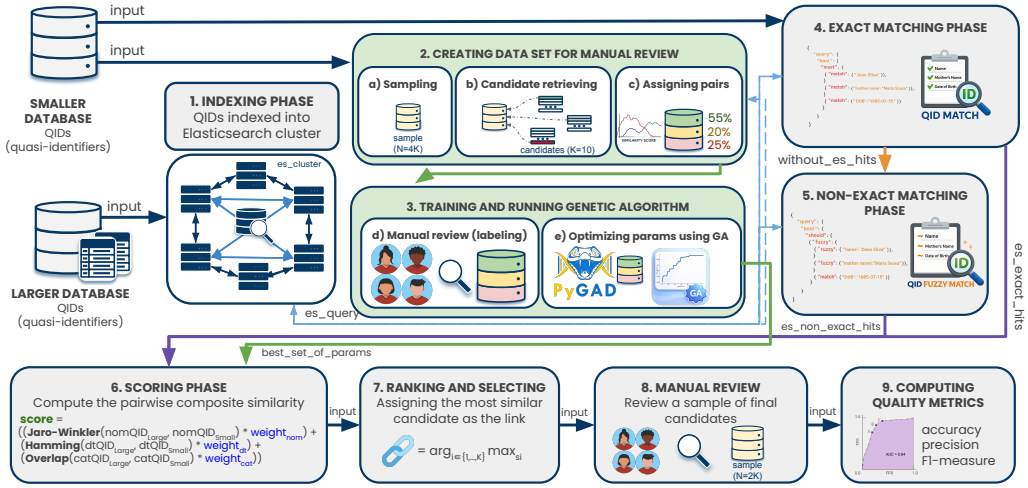

Figure 1. Overview of the complete CIDACS-RL pipeline. The figure highlights the proposed extension in green

A trained team of specialists examines each record pair and assigns a binary label, where '1' denotes a link and '0' denotes a non-link. Now, we define the labeled $\mathbf{L} = (w, p, C) \in R^{n\times 4} = (w_1, \ldots, w_m, p_1, \ldots, p_m, TP, FP, TN, FN)$ as the input data for the genetic algorithm, where $C \in R^{n\times 4}$ is the confusion counts with the columns (TP, FP, TN, FN), and $y = |p| + |w| + 4$.

Algorithm 1 GA-based parameter tuning with kNN surrogate (PyGAD)

Require:
Evaluated configurations $\mathbf{L} \in R^{n\times d}$;
confusion counts $C \in R^{n\times 4}$ with columns (TP, FP, TN, FN);
k (number of neighbors);
metric $\mathcal{M}(\cdot)$;
gene spaces;
population size N;
generations G;
GA operators (selection/crossover/mutation)
Ensure: Best parameter vector $z^\star$
Define $\mathrm{kNNsurrogate}(z, \mathbf{L}, C, k)$ that returns $\widehat{C}(z) = (\widehat{TP}, \widehat{FP}, \widehat{TN}, \widehat{FN})$ for a candidate z
Define fitness $F(z) \leftarrow \mathcal{M}(\widehat{C}(z))$
Initialize population $\mathcal{P}^{(0)} = \{z_1^{(0)}, \ldots, z_N^{(0)}\}$ by sampling from the gene spaces
Set elitism size $E \leftarrow \lceil 0.05N \rceil$
for $t = 0$ to $G - 1$ **do**
    Evaluate $F(z)$ for all $z \in \mathcal{P}^{(t)}$
    Sort $\mathcal{P}^{(t)}$ by decreasing fitness and set elites $\mathcal{E}^{(t)} \leftarrow \{z_{(1)}^{(t)}, \ldots, z_{(E)}^{(t)}\}$
    Select parents from $\mathcal{P}^{(t)}$ according to the chosen selection operator
    Generate offspring via crossover and mutation until obtaining $N - E$ individuals
    Form next population $\mathcal{P}^{(t+1)} \leftarrow \mathcal{E}^{(t)} \cup$ Offspring
    Optional: if early-stopping criterion holds (e.g., target fitness or stagnation), **break**
end for
$z^\star \leftarrow \arg\max_{z \in \cup_t \mathcal{P}^{(t)}} F(z)$
return $z^\star$

Algorithm 1 describes our implementation of the genetic algorithm used to search the linkage parameter space using the labeled dataset previously constructed. Let $L \in R^{n \times d}$ denote the matrix of evaluated parameter configurations and $C \in R^{n \times 4}$ the corresponding confusion-count matrix with columns (TP, FP, TN, FN).

For a candidate parameter vector z, the procedure estimates $\widehat{C}(z)$ using a k-nearest neighbors surrogate defined over the configurations in $\mathbf{L}$. The fitness of each solution is computed as $F(z) = \mathcal{M}(\widehat{C}(z))$, where $\mathcal{M}(\cdot)$ is the chosen linkage quality metric. The genetic algorithm evolves a population of size N for G generations using selection, crossover, and mutation, while preserving the top 5% of individuals through elitism.

5. Experimental Setup

Our experimental setup aims to evaluate the genetic algorithm for parameter tuning in the CIDACS-RL workflow. Therefore, we selected the Live Births Information System (SINASC) and the Mortality Information System (SIM) as the most challenging pair of health-related data sources linked to date. Previous works report on how researchers have tackled the difficulties posed by data quality [Paixao et al. 2021] and provide insights from the integrated data [Rebouças et al. 2024]. We understand that advancing the linkage rate between these two data sources may support several relevant unfolding in future research.

We used the PyGAD (Gad, 2024) to implement our model in the CIDACS-RL workflow. Table 3 systemizes the parameters used to build the model in the experimental setup. The following subsections provide details on the validation of the linkage between two real-world health-related datasets hosted on the CIDACS data platform.

5.1. Data sets

In our experiments, we link SINASC and SIM records to identify mortality events among children under 5 years of age in the southern states of Brazil during 2018–2020. Table 1 summarizes the main characteristics of the SINASC and SIM databases used in the linkage experiment. Although most identifying attributes exhibit high completeness, the SINASC database shows substantial missingness for the newborn's name (83.9%), while the SIM database shows lower levels of missingness for some attributes, particularly the municipality of birth. These characteristics illustrate the challenges of accurately linking records across the two systems, as incomplete or inconsistent identifiers can make identifying true matches difficult and increase the risk of linkage errors.

We constructed two auxiliary datasets from the SIM and SINASC datasets to enable parameter optimization via a genetic algorithm. The first is a labeled reference dataset obtained through stratified sampling of SIM records according to year of occurrence and state. We retrieved 2,806 records from all possible comparisons on SIM and SINASC using Elasticsearch queries, as detailed in the Section 4. The proportion of raking positions was defined as $(position1 : 55\%, position2 : 20\%, position3 : 10\%, position4 : 7\%, position1 : 4\%, position6 : 2\%, position7 : 1\%, position8 : 1\%)$. The second is the set of evaluated runs of CIDACS-RL using the optimal cutoff point determined from the ROC curve (using Youden's index) on the labeled data. To tackle the computational complexity, we sampled 3,000 records from the entire discrete search space described in Section 3.

Table 1. Descriptive analysis of the databases involved in the validation

Live Birth Database (SINASC)							
	name	**mothers name**	**date of birth**	**mun. of resid.**	**mun. of birth**	**sex**	**race/ ethinicity**
overall							
N = 26,107,682							
non-missing - N	4,204,496	26,100,362	26,107,682	26,107,676	26,107,682	26,107,682	26,107,682
(%)	(16.1%)	(99.97%)	(100%)	(100%)	(100%)	(100%)	(100%)
missing - N	21,903,186	7,320	0	6	0	0	0
(%)	(83.9%)	(0.03%)	(0%)	(0%)	(0%)	(0%)	(0%)

Mortality Database (SIM) - Under-five years old - 2018-2020							
	name	**mothers name**	**date of birth**	**mun. of resid.**	**mun. of birth**	**sex**	**race/ ethinicity**
overall							
N = 640,331							
non-missing - N	632,349	639,064	630,626	640,331	640,331	640,331	640,331
(%)	(98.95%)	(99.80%)	(98.68%)	(100%)	(100%)	(100%)	(100%)
missing - N	7,982	2,567	9,705	0	32,467	0	0
(%)	(1.25%)	(0.4%)	(1.52%)	(0%)	(5.08%)	(0%)	(0%)
Paraná (IBGE code = 41)							
N = 234,285							
non-missing - N	232,570	233,636	230,619	234,285	225,207	234,285	234,285
(%)	(99,27%)	(99,72%)	(98,44%)	(100%)	(96,13%)	(100%)	(100%)
missing - N	1,715	649	3,666	0	9,078	0	0
(%)	(0.73%)	(0.28%)	(1.56%)	(0%)	(3,87%)	(0%)	(0%)
Santa Catarina (IBGE code = 42)							
N = 132,161							
non-missing - N	129,605	131,760	129,715	132,161	127,990	132,161	132,161
(%)	(98.07%)	(99.70%)	(98.15%)	(100%)	(96.84%)	(100%)	(100%)
missing - N	2,556	401	2,446	0	4,171	0	0
(%)	(1.93%)	(0.30%)	(1.85%)	(0%)	(3.16%)	(0%)	(0%)
Rio Grande do Sul (IBGE code = 43)							
N = 273,885							
non-missing - N	270,174	272,368	270,292	273,885	254,667	273,885	273,885
(%)	(98.65%)	(99.45%)	(98.69%)	(100%)	(92.98%)	(100%)	(100%)
missing - N	3,711	1,517	3,593	0	19,218	0	0
(%)	(1.35%)	(0.55%)	(1.31%)	(0%)	7.02	(0%)	(0%)

5.2. Manual review of accuracy

To comply with best practices for manual accuracy review processes [Joffe et al. 2014], we engaged nine highly trained data professionals to label the first auxiliary dataset. The demographic of reviewers was diverse, encompassing gender, race/ethnicity, and age groups. Such diversity is essential given the cultural, temporal, and regional variations present in Brazilian administrative databases.

Our manual review protocol establishes a structured and reproducible framework for assessing record linkage accuracy in Brazilian administrative databases. We implemented a set of hierarchical decisions designed to reduce subjectivity, mitigate systematic biases, and ensure consistent match classifications across reviewers. The protocol comprises four general rules and a set of attribute-specific rules addressing names, dates, and auxiliary variables. Table 2 summarizes these rules, which serve as guidelines for the reviewers and are iteratively enhanced after each discussion led by the third reviewer in the event of disagreement.

Nine reviewers participated in the manual validation process. In the first and second rounds, each reviewer independently evaluated the match status of approximately 280–356 record pairs, resulting in a balanced distribution of the review workload. A third round of review was required only for a small subset of cases (approximately 45 pairs)

requiring additional verification, involving at most nine pairs per reviewer. Overall, the process ensured a consistent review effort across participants while limiting the number of pairs requiring discussion.

Table 2. Summary of hierarchical manual review rules for linkage decisions.

Rule block	Guidelines (what to check)	Decision guidance (how to decide)
General R1 (core consistency)	High agreement on core identifiers: name, mother's name, and date of birth; allow minimal/plausible differences (abbreviations, typos).	Classify as *link* when core identifiers strongly agree and differences remain plausible.
General R2 (compensation)	Moderate discrepancy in a single attribute (often auxiliary) with strong agreement in remaining attributes.	Accept as *link* when multiple other attributes provide strong support that plausibly explains the discrepancy.
General R3 (critical inconsistency)	Structural incompatibilities (e.g., substantially different date of birth; distinct mother's name without plausible explanation).	Classify as *non-link* when a critical inconsistency undermines identity coherence.
General R4 (contextual support)	Context variables (race/ethnicity, education, year of registry) used only to support interpretation.	Use contextual variables for confirmation/disambiguation; do not use them as standalone criteria.

During manual review, attribute-level considerations complemented the general rules. Name discrepancies, such as abbreviations, phonetic variations, omissions, and surname changes, were tolerated when plausible, but required stronger agreement in the mother's name, date of birth, or other attributes as discrepancies increased. Date of birth was treated as a highly discriminative attribute with limited tolerance for inconsistencies, although plausible typographical errors were considered. Auxiliary variables such as father's name, place of birth, and municipality information were used to support disambiguation and to invalidate candidate pairs when incompatible contextual information suggested different individuals.

5.3. Experiment design

The optimized parameter configuration obtained through the genetic algorithm was compared against the original CIDACS-RL parameter settings. Performance was assessed using a labeled reference dataset constructed through manual review of candidate record pairs, allowing the estimation of confusion matrix counts and standard evaluation metrics. The genetic algorithm model used was trained using the PyGAD with the parameters and operators explicited in Table 3.

To assess the robustness of the optimized configuration, experiments were conducted both on the aggregated dataset and separately for each state in the southern region of Brazil (Paraná, Santa Catarina, and Rio Grande do Sul). For each configuration, the optimal decision threshold was determined using the Youden index derived from the ROC curve, enabling a consistent comparison of linkage performance across the different evaluation settings.

Table 3. Main genetic algorithm configuration parameters and operators

Component	Parameter	Value
GA Core	Population size (`sol_per_pop`)	3000
	Number of generations (`num_generations`)	60
	Parents per mating (`num_parents_mating`)	10
	Random seed	2026
Selection	Parent selection type	Tournament
	Tournament size (K)	3
	Elitism	5% best individuals preserved
Variation Operators	Crossover type	Single-point
	Mutation type	Random (20% genes)
Stopping Criterion	Fitness target	0.98
	Maximum generations	60
Surrogate Model	Method	kNN surrogate
	Number of neighbors (k)	15

The original CIDACS-RL configuration assigns equal weights to the main QIDs (name, mother's name, and date of birth), with $w_{\text{name}} = 1.0$, $w_{\text{mother's name}} = 1.0$, and $w_{\text{date of birth}} = 1.0$, while lower weights are given to auxiliary attributes such as sex ($w_{\text{sex}} = 0.8$) and municipality of residence ($w_{\text{mun. res.}} = 0.5$). The attributes municipality of birth and race/ethnicity are not considered in the similarity score ($w = 0$). Penalty parameters for missing values are uniformly set to $p = 0.05$ for most attributes, except for municipality of birth and race/ethnicity, not used in the original settings. These parameters serve as the baseline configuration against which the optimized solution is evaluated.

The linkage quality was evaluated using standard metrics derived from the confusion matrix. Precision ($\frac{TP}{TP+FP}$) measures the proportion of predicted matches that correspond to true links, while recall or sensitivity ($\frac{TP}{TP+FN}$) represents the proportion of true links correctly identified. Specificity ($\frac{TN}{TN+FP}$) quantifies the ability to correctly identify non-links, and accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$) summarizes the overall proportion of correctly classified pairs. All metrics were computed from the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

5.4. Experimental Results

The optimization process produced the following parameter configuration: $w_{\text{name}} = 0.15$, $w_{\text{mother's name}} = 2.85$, $w_{\text{date of birth}} = 2.80$, $w_{\text{sex}} = 0.55$, $w_{\text{mun.res}} = 0.10$, $w_{\text{mun.birth}} = 0.75$, $w_{\text{race}} = 0.20$, and penalties $p_{\text{name}} = 0.70$, $p_{\text{mother's name}} = 0.30$, $p_{\text{date of birth}} = 0.55$, $p_{\text{sex}} = 0.35$, $p_{\text{mun.res}} = 0.05$, $p_{\text{mun.birth}} = 0.30$, and $p_{\text{race}} = 0.50$. In the Table 4, we consolidate the linkage quality metrics derived from the confusion matrix counts and the impact of the optimized parameters on the number of linked records. Using the Youden index to determine the optimal decision threshold for each method, as demonstrated in Figure 2, the optimized CIDACS-RL configuration improved linkage quality across all evaluation settings. The most expressive gains resulted from the reduction of false positive matches.

In the general dataset, false positives decreased from 295 to 29 (approximately 90%), while recall remained high (0.98 vs. 0.998). Consequently, precision increased from 0.70 to 0.96 and overall accuracy from 0.89 to 0.99. Similar patterns were observed across all southern states analyzed, where the optimized configuration reduced false pos-

itives by 88-98% without compromising recall. Additionally, the highlighted cells in Table 4 show an approximate 5% increase in matches with higher accuracy.

		General (Southern region)		**Paraná** (code 41)		**Santa Catarina** (code 42)		**Rio Grande do Sul** (code 43)	
		Original	**GA**	**Original**	**GA**	**Original**	**GA**	**Original**	**GA**
Confusion Matrix	VP	695	708	301	304	152	154	225	228
	FP	295	29	116	14	80	3	92	2
	FN	14	1	4	1	2	0	4	1
	VN	1 892	2 158	775	877	397	474	658	748
Linkage Quality	precision	0.702	0.960	0.721	0.955	0.655	0.980	0.709	0.991
	recall	0.980	0.998	0.986	0.996	0.987	1.000	0.982	0.995
	specificity	0.865	0.986	0.869	0.984	0.832	0.993	0.877	0.997
	accuracy	0.893	**0.989**	0.899	0.987	0.870	0.995	0.901	0.996
Linked records	*N*	485 411	**519 549**	476 520	492 969	488 005	537 031	491 708	528 646
	%	80.34	**85.99**	78.87	81.59	80.77	88.88	81.38	87.49
Non-linked records	*N*	118 802	84 664	127 693	111 244	116 208	67 182	112 505	75 567
	%	19.66	14.01	21.13	18.41	19.23	11.12	18.62	12.51

Table 4. Experimental comparison of linkage quality and proportion of linked records between the original CIDACS-RL implementation and our extension with genetic algorithm-based parameter optimization

The Figure 2a illustrates the threshold estimation obtained using the original CIDACS-RL parameter configuration and the optimized parameters proposed in this study. The optimized configuration increases the area under the ROC curve (AUC) from 0.96 to 0.994, indicating improved discriminative performance. Figure 2b shows the linkage quality across different decision thresholds for both parameter settings. The optimized configuration consistently outperforms the original CIDACS-RL implementation, producing a more robust composite similarity score and higher overall accuracy.

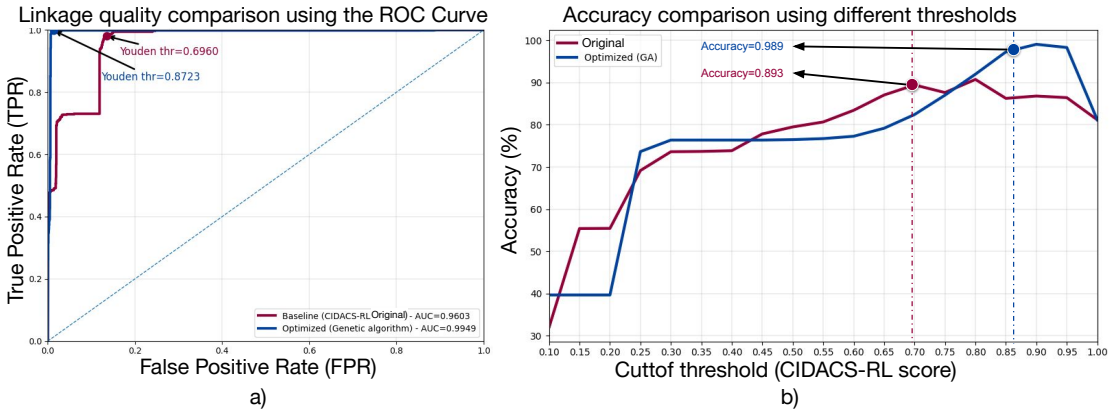

Figure 2. Comparison of Linkage quality using the ROC Curve and the accuracy over different score thresholds

Figure 2a illustrates the threshold estimation obtained using the original CIDACS-RL parameter configuration and the optimized parameters proposed in this study. The optimized configuration increases the area under the ROC curve (AUC) from 0.96 to 0.994,

indicating improved discriminative performance. Figure 2b shows the linkage quality across different decision thresholds for both parameter settings. The optimized configuration consistently outperforms the original CIDACS-RL implementation, producing a more robust composite similarity score and higher overall accuracy.

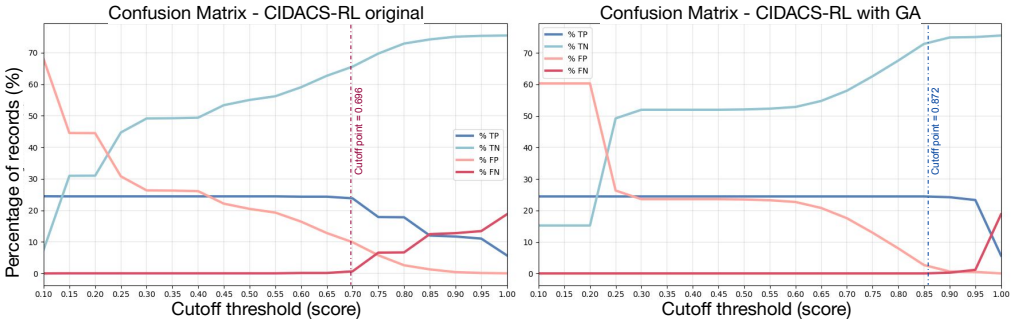

Figure 3. Comparison of the linkage quality using the Confusion Matrix

6. Discussion

The results demonstrate that the proposed optimization strategy improves the linkage quality of the CIDACS-RL workflow, primarily through the reduction of false positive matches while maintaining extremely high recall. These achievements may mitigate the information bias arising from false matches [Doidge and Harron 2019], preventing spurious associations between individuals and outcomes affecting the validity of population-level studies. By learning attribute weights and penalties directly from labeled data, the optimized configuration produces a more discriminative composite similarity score, improving precision and overall accuracy without compromising sensitivity. Our approach provide a practical solution for exploring a large parameter space in complex linkage pipelines, avoiding an infeasible search over parameter combinations. Additionally, combining a genetic algorithm with a k-nearest neighbors surrogate model allows the optimization process to approximate the linkage quality of candidate configurations without repeatedly executing the full linkage workflow.

The main limitation of our approach lies in its reliance on the construction of a labeled reference dataset, which requires manual review and may introduce subjectivity or reviewer variability. Although estimating linkage quality from a sample of reviewed pairs is part of the original CIDACS-RL workflow, future work may explore strategies to reduce this dependency. One possible direction is the use of machine learning models trained on a knowledge base to support the automated review of larger samples [Pita et al. 2017]. Additional alternatives may include active learning techniques to prioritize informative record pairs for labeling, as well as semi-supervised approaches capable of leveraging the large pool of unlabeled candidate pairs generated during the linkage process.

7. Conclusion

This study proposed a data-driven strategy to optimize CIDACS-RL linkage parameters using genetic algorithms combined with a k-nearest neighbors surrogate model. The optimized configuration substantially reduced false positive matches while maintaining extremely high recall, leading to consistent improvements in precision and overall accuracy

across all evaluated settings. These findings highlight the importance of systematic parameter optimization in record linkage workflows and suggest that evolutionary search strategies can improve the reliability of large-scale health data integration initiatives.

References

Ali, M. S., Ichihara, M. Y., Lopes, L. C., Barbosa, G. C., et al. (2019). Administrative data linkage in brazil: potentials for health technology assessment. *Frontiers in pharmacology*, 10:984.

Barreto, M. L., Ichihara, M. Y., Pescarini, et al. (2022). Cohort profile: the 100 million brazilian cohort. *International journal of epidemiology*, 51(2):e27–e38.

De Bruin, J. (2022). Record linkage toolkit documentation.

Doidge, J. C. and Harron, K. L. (2019). Reflections on modern methods: linkage error bias. *International journal of epidemiology*, 48(6):2050–2060.

Gkoulalas-Divanis, A., Vatsalan, et al. (2021). Modern privacy-preserving record linkage techniques: An overview. *IEEE Transactions on Information Forensics and Security*, 16:4966–4987.

Harron, K., Goldstein, H., and Dibben, C. (2016). *Methodological developments in data linkage*. Wiley Online Library.

Joffe, E., Byrne, M. J., et al. (2014). A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. *Journal of the American Medical Informatics Association*, 21(1):97–104.

Linacre, R., Lindsay, S., Manassis, et al. (2022). Applyisplink: free software for probabilistic record linkage at scale. *International Journal of Population Data Science*, 7(3):1794.

Nelson, W., Khanna, N., Ibrahim, et al. (2023). Optimizing patient record linkage in a master patient index using machine learning: Algorithm development and validation. *JMIR Formative Research*, 7:e44331.

Paixao, E. S., Cardim, L. L., Falcao, I. R., Ortelan, N., Silva, et al. (2021). Cohort profile: Cidacs birth cohort. *International journal of epidemiology*, 50(1):37–38.

Pita, R., Mendonça, E., Reis, S., Barreto, M., and Denaxas, S. (2017). A machine learning trainable model to assess the accuracy of probabilistic record linkage. In *DaWaK*, pages 214–227. Springer.

Rebouças, P., Paixão, E. S., et al. (2024). Ethno-racial inequalities on adverse birth and neonatal outcomes. *The Lancet Regional Health–Americas*, 37.

Shaikh, F. and Ragkhitwetsagul, C. (2008). Evaluating genetic algorithms for selection of similarity functions for record linkage. *Carnegie Mellon University*.

WAYKOLE, J. R. and SHINDE, S. (2014). An approach towards record linkage using genetic algorithm along with hash algorithm_2014. *International Journal of Current Engineering and Technology*, 4(3):2142–2146.

Yu, J., Nabaglo, J., Vatsalan, et al. (2020). Hyper-parameter optimization for privacy-preserving record linkage. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 281–296. Springer.