ORIGINAL RESEARCH **Open access**

Probabilistic Record Linkage and Machine Learning Model for Detecting Duplicate Patient Records Using Demographic Similarity, Address Variation, Encounter History, Insurance Identifiers, and Clinical Pattern Matching

Juan Perez[1], Ana Gutierrez[1*], Carlos Lopez[2]

Abstract

Duplicate patient records are a pervasive problem in electronic health records, endangering patient safety and inflating healthcare costs. In EHR-driven health systems, identity fragmentation can separate medications, allergies, diagnoses, laboratory results, and prior encounters across more than one record. Traditional probabilistic linkage relies on manual tuning of weights for demographic attributes and cannot fully exploit temporal address changes, insurance identifiers, or clinical patterns. These limitations become especially important when names are misspelled, addresses change, identifiers are missing, or patients receive care across multiple facilities. This article develops a conceptual model that combines probabilistic record linkage with machine learning to detect duplicate patient records. The model uses demographic similarity, address variation, encounter history, insurance identifiers, and clinical pattern matching as complementary evidence streams. The proposed model first applies probabilistic blocking to generate candidate record pairs, then uses a deep learning classifier, such as a Siamese network, to score pairs based on static and dynamic features. The output is a duplicate probability that can support automated ranking, human review, and master patient index maintenance. Conceptually, the model would be expected to improve duplicate detection compared with purely probabilistic linkage when demographic data are incomplete or unstable. Its main advantage is that it can anchor linkage decisions in more stable encounter patterns, longitudinal clinical trajectories, and repeated institutional contact signals. Such a hybrid model could enable a more accurate and self-maintaining master patient index while reducing the burden of manual record merging. It would also support safer registration workflows by identifying probable duplicates before identity fragmentation affects care delivery.

Keywords Machine learning, Patient identity management, Duplicate patient records, Probabilistic record linkage, Master patient index, Siamese networks

*Correspondence:
Ana Gutierrez
ana.gutierrez@gmail.com
[1] Department of Intelligent Healthcare Systems, Faculty of Medicine, National Autonomous University of Mexico, Mexico City, Mexico
[2] Department of Health Data Engineering, Faculty of Engineering, Monterrey Institute of Technology, Monterrey, Mexico

Introduction

Duplicate patient records create clinical, administrative, and financial risk because they divide one patient's information across separate identifiers and can obscure prior diagnoses, medications, allergies, procedures, and care gaps. Nelson and colleagues [1] frame master patient index optimization as an operational patient-safety problem, while Grannis and

colleagues [2] show that patient matching strategy must account for real-world identification uncertainty. Lippi and colleagues [3] further describe patient identification errors as a persistent healthcare and laboratory medicine crisis rather than a narrow technical problem. In this context, duplicate detection should be understood as essential infrastructure for safe EHR-driven care.

Probabilistic record linkage provides a practical framework for comparing uncertain identifiers, but it depends heavily on the completeness and stability of demographic fields. Li and colleagues [4] emphasize that missing data and field selection affect probabilistic linkage behavior, while Ferguson, Hannigan, and Stack [5] address the additional problem of field dependency and missing-data imputation. Scalable linkage systems such as CIDACS-RL, described by Barbosa and colleagues [6], demonstrate the importance of indexing and scoring design when data are large and imperfect. These limitations become especially important when patients change addresses, names are entered inconsistently, or identifiers are absent.

Machine learning offers a complementary approach because it can learn complex relationships among record-pair features from labeled duplicate and non-duplicate examples. Bian and colleagues [7] show that patient linkage can be implemented within a privacy-preserving health research network, while Tachinardi and colleagues [8] demonstrate the importance of privacy-preserving linkage across disparate institutions. Pathak and colleagues [9] position privacy-preserving record linkage as a public health requirement, and Nguyen and colleagues [10] show that deidentified public health records can be linked under privacy-preserving constraints. Therefore, machine learning for duplicate detection must combine predictive flexibility with secure handling of sensitive identity and clinical variables.

This article proposes a hybrid probabilistic record linkage and machine learning model for detecting duplicate patient records in a master patient index environment. The model retains probabilistic blocking, as supported by Li and colleagues [4], while extending the scoring layer through deep entity-matching concepts described by Mudgal and colleagues [11]. Röchner and Rothlauf [12] show that machine learning can reduce manual effort in health record linkage, making such a hybrid design operationally relevant. The proposed model fuses demographic, address, encounter, insurance, and clinical-pattern features into a duplicate probability that supports reviewable MPI maintenance rather than autonomous record merging.

Background

Duplicate patient records and the master patient index

Duplicate patient records arise when one patient is represented by more than one identity record within or across health information systems. In master patient index settings, Nelson and colleagues [1] describe machine learning as a way to improve patient record linkage, while Grannis and colleagues [2] compare real-world referential and probabilistic approaches to patient matching. Lippi and colleagues [3] emphasize that patient identification failures can affect clinical and laboratory safety, reinforcing the need for systematic identity management. Hospitals therefore require duplicate detection systems that are accurate enough for operational use but transparent enough for health information management review.

Probabilistic record linkage: fundamentals and limitations

Probabilistic record linkage compares pairs of records across multiple fields and assigns evidence weights reflecting agreement, disagreement, and uncertainty. The data-adaptive Fellegi-Sunter approach developed by Li and colleagues [4] illustrates how probabilistic linkage can incorporate missing data and field selection, while Ferguson, Hannigan, and Stack [5] address dependency among linkage fields. Almeida and colleagues [13] show that linkage quality must be examined carefully when administrative health databases are joined at scale. These studies support the use of probabilistic linkage as a foundation, but they also show why fixed weights and unstable identifiers can limit performance in complex clinical environments.

Machine learning and deep learning for record linkage

Machine learning reframes duplicate detection as a record-pair classification task in which a model learns how feature combinations distinguish true matches from nonmatches. Röchner and Rothlauf [12] demonstrate how machine learning can be applied to health record linkage while explicitly considering manual effort, and Christen, Häntschel, Christen, and Rahm [14] show how autoencoders can support privacy-preserving record linkage. Ranbaduge, Vatsalan, and Ding [15] extend this direction through privacy-preserving deep learning, while Mudgal and colleagues [11] describe the broader deep entity-matching design space. These approaches suggest that neural matching can improve conceptual flexibility, provided that clinical deployment remains conservative and auditable.

Using address variation and geographic data

Address variation is a central challenge in patient matching because patients relocate, addresses are reformatted, historical addresses remain in records, and household-level details may be incomplete. Randall and colleagues [16] propose privacy-preserving linkage using multiple dynamic match keys, which is directly relevant to changing identifiers over time. Vidanage and colleagues [17] also show that dynamic match-key approaches require careful privacy analysis, while Randall and colleagues [18] evaluate Bloom filter linkage under blinded conditions. Ranbaduge and Christen [19] further emphasize temporal linkage, supporting the idea that address history should be represented as time-dependent evidence rather than a single static field.

Encounter history and clinical pattern matching as stable identifiers

Encounter history and clinical pattern matching can provide linkage evidence when demographic identifiers are weak, because care-seeking behavior, facility use, diagnosis patterns, procedure sequences, and medication histories may remain coherent across duplicate records. Lim and colleagues [20] show that privacy-preserving linkage can unlock health-system data for chronic kidney disease outcome modeling, while Brown and Randall [21] evaluate secure linkage of large health datasets using a hybrid cloud model. Brown, Borgs, Randall, and Schnell [22] demonstrate cryptographic linkage on large medical datasets, suggesting that clinical evidence can be incorporated under secure computation. Suo and colleagues [23] also show that deep patient similarity learning can represent longitudinal clinical similarity, which supports the conceptual use of clinical trajectories as linkage evidence.

Model Development Overview

High-level duplicate detection pipeline

The proposed pipeline begins by standardizing incoming patient identity fields and applying probabilistic blocking rules to generate candidate record pairs from the MPI. Blocking is justified by Ferguson, Hannigan, and Stack [5], who address computational efficiency in record linkage, and by Barbosa and colleagues [6], whose CIDACS-RL system emphasizes scalable indexing and scoring. Li and colleagues [4] support the use of adaptive probabilistic methods before more complex classification, while Almeida and colleagues [13] show that linkage workflows require explicit quality examination. Candidate pairs would then be converted into rich feature vectors and scored by a deep classifier before being consolidated into duplicate clusters for steward review.

Figure 1 presents the proposed end-to-end workflow for hybrid probabilistic blocking, deep pairwise duplicate scoring, governance review, and operational master patient index action.

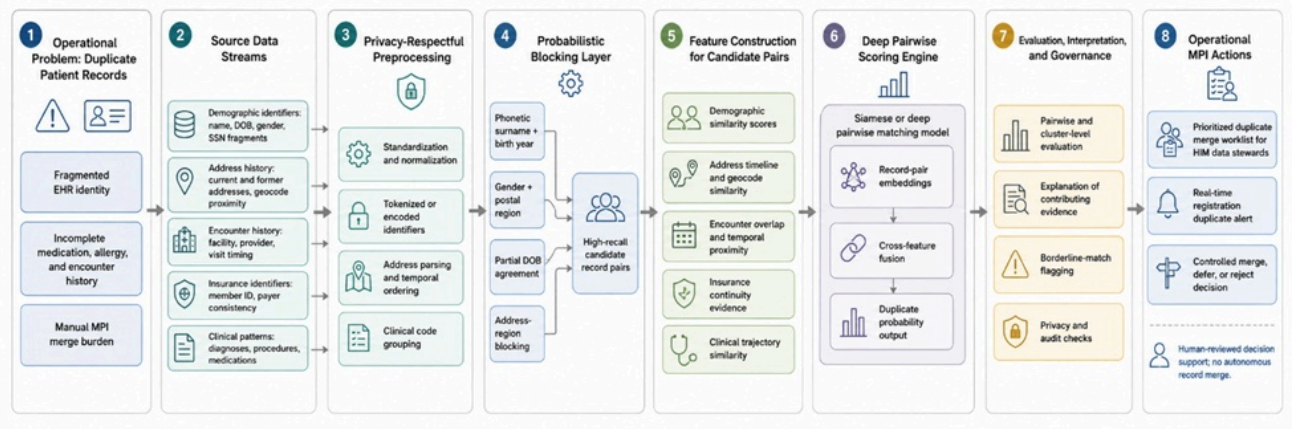


Figure 1. End-to-End Hybrid Probabilistic Record Linkage and Deep Learning Workflow for Duplicate Patient Record Detection in the Enterprise Master Patient Index

Core input features

The model's input feature space would include demographic similarity scores, address variation features, encounter-history overlap, insurance evidence, and clinical-pattern similarity. Grannis and colleagues [2] demonstrate that patient matching benefits from real-world comparison strategies, while Röchner and Rothlauf [12] show that health record linkage can incorporate machine learning to reduce manual review burden. Coeli and colleagues [24] highlight the difficulty of record linkage under suboptimal conditions, and de Paula and colleagues [25] show that linkage algorithms must be assessed for both feasibility and accuracy in public health databases. Suo and colleagues [23] support the inclusion of clinical-pattern similarity by showing how deep learning can model patient similarity for personalized healthcare.

Design principles

The model is designed around scalability, privacy-respectful feature construction, auditability, and adaptation to evolving patient identities. Bian and colleagues [7] show that hash-based privacy-preserving linkage can be implemented in a clinical research network, while Tachinardi and colleagues [8] describe privacy-preserving linkage across institutions in a learning health system. Pathak and colleagues [9] emphasize the public health importance of privacy-preserving record linkage, and Nguyen and colleagues [10] demonstrate privacy-preserving linkage of deidentified surveillance records. These studies support a model architecture that avoids unnecessary exposure of raw identifiers while preserving enough evidence for reviewable duplicate detection.

Data Sources and Feature Engineering

Demographic similarity features

Demographic similarity features would compare names, dates of birth, gender, and identifier fragments using approximate and rule-informed comparisons rather than strict equality alone. Li and colleagues [4] show that probabilistic linkage must account for missingness and field selection, while Ferguson, Hannigan, and Stack [5] demonstrate the importance of handling field dependency. Kasthurirathne and Grannis [26] specifically address machine learning approaches for identifying nicknames in health information exchange data, which supports flexible name comparison. Leshin and colleagues [27] further show that name transformation affects match rates, reinforcing the need for phonetic, token-level, and edit-distance features in patient matching.

Address variation and insurance features

Address features would represent normalized street elements, postal codes, geocode proximity, address recency, number of recorded address changes, and consistency between current and historical residences. Randall and colleagues [16] support the use of multiple dynamic match keys, while Vidanage and colleagues [17] caution that such dynamic linkage evidence can introduce privacy risks. Randall and colleagues [18] show how Bloom filter methods can support blinded linkage, and Ranbaduge and Christen [19] provide a framework for temporal record linkage. Insurance features would be treated similarly as time-varying administrative evidence, with stable member identifiers strengthening a match while payer transitions remain compatible with non-duplicate or duplicate interpretations.

Encounter and clinical pattern features

Encounter features would compare facility use, provider overlap, visit timing clusters, care setting transitions, and recurring service locations across two candidate records. Lim and colleagues [20] demonstrate the value of privacy-preserving linkage for chronic disease outcome modeling, while Brown and Randall [21] show that secure linkage can be applied to large health datasets. Brown, Borgs, Randall, and Schnell [22] further support secure linkage using cryptographic keys, which is relevant when clinical features are included in identity resolution. Suo and colleagues [23] provide a conceptual basis for representing clinical-pattern similarity through deep patient similarity learning.

Table 1 summarizes the patient identity, administrative, encounter, insurance, and clinical data structures used to construct representation-learning features for duplicate record detection.

Table 1. Input Structure and Representation Learning Strategy for Hybrid Duplicate Patient Record Detection

Input domain	Example source fields	Data structure used by the model	Representation or feature strategy	Linkage value	Practical risk addressed
Demographic identifiers	Name, date of birth, gender, partial SSN, contact information	Static structured fields with possible missingness, typographical variation, and field dependency	Approximate string matching, phonetic encoding, token-level name comparison, date-tolerance logic, identifier-fragment agreement	Provides conventional identity evidence for candidate generation and pairwise scoring	Reduces missed matches caused by spelling errors, nickname use, transposed digits, or incomplete registration fields
Address history	Current address, previous addresses, postal code, geocoded location, address effective dates	Longitudinal and time-stamped administrative sequence	Address normalization, geocode-distance comparison, temporal decay weighting, address-change frequency, regional stability features	Captures residential continuity even when the current address differs across records	Prevents false exclusion of true duplicates who moved, while limiting inappropriate matches based only on historical proximity
Encounter history	Facility, clinic, provider, visit type, encounter dates, registration location	Time-ordered utilization sequence	Encounter-overlap scores, visit-date clustering, facility-use similarity, provider continuity features	Adds operational evidence from repeated patterns of care access	Helps distinguish demographically similar patients by comparing where, when, and how they interact with the health system

Insurance identifiers	Member ID, payer name, plan type, coverage dates, subscriber relationship	Time-varying administrative identifiers	Member ID similarity, payer-continuity indicators, coverage-transition features, longitudinal consistency checks	Provides strong supporting evidence when stable, but remains flexible when coverage changes	Avoids overreliance on insurance fields that may shift because of employment, payer transition, or plan redesign
Clinical pattern data	Diagnosis codes, procedure codes, medication classes, laboratory pattern summaries, chronic disease indicators	Sparse longitudinal clinical sequence or bag-of-codes representation	Code co-occurrence, chronic disease trajectory similarity, medication-class overlap, temporal sequence alignment	Provides a stable clinical "story" when demographic identifiers are incomplete or unstable	Supports matching when identity fields are weak, while requiring privacy safeguards because clinical data are sensitive
Candidate pair metadata	Blocking rule that generated the pair, number of shared blocks, prior review status	Pair-level operational metadata	Blocking-source indicators, candidate-generation confidence, prior steward action flags	Adds context for why a pair entered the model pipeline	Supports auditability and helps identify systematic blocking or registration weaknesses
Review and adjudication data	HIM merge decision, reject decision, defer decision, reviewer notes, dispute history	Human-labeled decision record	Supervised labels, review-state features, adjudication categories, reviewer uncertainty flags	Provides ground-truth evidence for future model calibration and governance	Prevents the model from being treated as autonomous and preserves expert oversight in identity resolution

Probabilistic Blocking and Deep Learning Scoring

Blocking strategy

Blocking is necessary because comparing every record against every other record is operationally impractical in an enterprise MPI. Ferguson, Hannigan, and Stack [5] emphasize computational efficiency in record linkage, while Barbosa and colleagues [6] demonstrate scalable indexing and scoring for very large datasets. Almeida and colleagues [13] show that the quality of the linkage process must be examined after candidate generation, and Ranbaduge and Christen [19] show that temporal linkage adds further computational complexity. Therefore, the proposed system would use multiple indexing passes, such as phonetic surname plus birth year, gender plus postal region, partial date-of-birth agreement, or normalized address components, to create high-recall candidate sets.

Deep learning scorer architecture

The deep learning scorer could be implemented as a Siamese network that embeds each record and compares the embeddings, or as a feedforward classifier that receives engineered pairwise similarity features. Christen, Häntschel, Christen, and Rahm [14] show that autoencoders can support privacy-preserving record linkage, while Ranbaduge,

Vatsalan, and Ding [15] extend privacy-preserving linkage through deep learning. Mudgal and colleagues [11] describe the design space for deep entity matching, and Li, Li, Suhara, Doan, and Tan [28] show how pretrained language models can support deep entity matching. Govind and colleagues [29] further position entity matching as a data science workflow, which supports a modular architecture in which the neural scorer remains one reviewable component.

Output and clustering

The model would output a duplicate probability for each candidate pair and then aggregate pairwise evidence into patient-level clusters. Nelson and colleagues [1] show that patient record linkage can be optimized within a master patient index, while Grannis and colleagues [2] emphasize that real-world matching strategies must support patient identification strategy rather than isolated pairwise classification. Röchner and Rothlauf [12] highlight the operational tradeoff between linkage quality and manual effort, making threshold design and worklist creation central to deployment. Tachinardi and colleagues [8] further support the need for governed linkage across disparate institutions, so final clustering should remain configurable, auditable, and reviewable.

Integrating Address Variation, Insurance, and Clinical Patterns

Modelling address changes and relocations

Address changes should be modeled as longitudinal evidence rather than as a single static comparison field. Randall and colleagues [16] support this logic through dynamic match keys, while Ranbaduge and Christen [19] show that temporal record linkage can explicitly represent changes across time. In the proposed model, an address timeline embedding or recurrent address module could learn that short-distance relocations, repeated historical addresses, and stable regional patterns support possible linkage, whereas abrupt or inconsistent geographic jumps should weaken but not automatically exclude a candidate pair. Because Vidanage and colleagues [17] identify privacy risks in dynamic match-key linkage, address sequences should be encoded as comparison features rather than exposed as raw residential histories.

Insurance identifiers as linkage evidence

Insurance identifiers can provide strong administrative evidence, but they must be treated as time-varying rather than permanent identity anchors. Grannis and colleagues [2] show that real-world patient matching requires multiple evidence sources, while Li and colleagues [4] demonstrate why field selection and missingness must be handled adaptively in probabilistic linkage. A consistent member identifier across repeated encounters could strengthen a duplicate probability, whereas payer changes, plan transitions, or gaps in coverage should be interpreted in relation to the patient's broader demographic, address, and encounter evidence. In this model, insurance features would therefore contribute graded linkage evidence instead of functioning as absolute deterministic rules.

Clinical pattern matching and chronic disease trajectories

Clinical pattern matching could help recognize the same patient when demographic or administrative identifiers are incomplete, because chronic disease trajectories, medication histories, procedure sequences, and repeated specialty encounters may remain coherent across duplicate records. Suo and colleagues [23] provide a conceptual basis for deep patient similarity learning, while Lim and colleagues [20] show that linked health data can support chronic disease outcome modeling under privacy-preserving conditions. The proposed model could compare diagnosis and medication trajectories through sequence similarity, code co-occurrence, or temporal alignment methods, but it should treat clinical similarity as probabilistic support rather than proof of identity. Brown and Randall [21] and Brown, Borgs, Randall, and Schnell [22] further indicate that secure linkage methods are necessary when sensitive clinical patterns become part of identity-resolution logic.

Model Interpretability and Audit of Linkage Decisions

Explaining duplicate rates to HIM professionals

A duplicate-detection model should provide interpretable evidence to health information management professionals rather than only a probability score. Nelson and colleagues [1] describe patient record linkage optimization within an MPI environment, and Röchner and Rothlauf [12] emphasize the operational tradeoff between linkage quality and manual effort, both of which support review-facing explanations. For a specific candidate pair, explanation methods could indicate that recurring outpatient procedures, overlapping facility use, stable historical addresses, or compatible insurance evidence increased the duplicate probability despite a misspelled name. These explanations would make the model more suitable for steward review, threshold adjustment, and institutional governance.

Transparent override and dispute resolution

Transparent override mechanisms are necessary because duplicate detection directly affects patient identity and record integrity. Tachinardi and colleagues [8] show that linkage across institutions requires governance, while Pathak and colleagues [9] frame privacy-preserving linkage as a public health action that must manage both opportunity and risk. The proposed model would route borderline cases to human review, preserve an audit trail of demographic, address, insurance, encounter, and clinical evidence, and allow reviewers to accept, reject, or defer candidate merges. Reviewer feedback could then be captured as adjudication evidence for future model refinement without implying uncontrolled self-correction or autonomous merging.

Operational Deployment in the MPI Environment

Integration with enterprise master patient index

In an enterprise master patient index, the model could run as an offline batch process that scans recent registrations, newly imported records, or unresolved identity queues. Nelson and colleagues [1] directly support the operational relevance of MPI-focused machine learning linkage, while Grannis and colleagues [2] show that patient identification strategy benefits from evaluating both referential and probabilistic approaches. The output would be a prioritized worklist of candidate duplicate clusters, with each cluster accompanied by field-level and trajectory-level evidence for data stewards. This deployment mode would reduce unnecessary real-time burden while supporting regular MPI hygiene and controlled merge workflows.

Real-time duplicate checking at registration

The same trained scoring logic could also be deployed as a real-time service during patient registration, where it would flag potential duplicates before a new record is finalized. Bian and colleagues [7] demonstrate that linkage tools can be embedded within health data network infrastructure, and Nguyen and colleagues [10] show that privacy-preserving linkage can operate on deidentified records in public health surveillance. In a registration workflow, the model should return ranked candidate matches and explanatory evidence rather than forcing automatic merges. Such a design would support registrar decision-making, reduce creation of new duplicates, and preserve human accountability for identity confirmation.

Evaluation Strategy

Record linkage performance metrics

The proposed model should be evaluated conceptually through standard record linkage metrics, including precision, recall, and F-measure for pairwise duplicate classification, as well as cluster-level measures after global resolution. Almeida and colleagues [13] show that linkage quality must be examined when constructing large health cohorts, while de

Paula and colleagues [25] compare linkage algorithms in terms of accuracy and computational feasibility. Coeli and colleagues [24] further demonstrate that linkage under suboptimal conditions requires careful evaluation because real-world data imperfections can distort apparent performance. Therefore, evaluation should focus not only on pairwise scoring but also on whether the resulting clusters support safe and coherent patient identity management.

Comparison with probabilistic and deterministic baselines

The hybrid model should be compared against deterministic rules, standard probabilistic linkage, and data-adaptive Fellegi-Sunter baselines under the same governance and review assumptions. Li and colleagues [4] provide an appropriate probabilistic comparator through their data-adaptive Fellegi-Sunter model, while Ferguson, Hannigan, and Stack [5] offer a computationally efficient probabilistic linkage approach that accounts for field dependency and missing data. Barbosa and colleagues [6] also provide a scalable indexing and scoring reference point for large administrative datasets. Such comparisons should be framed as conceptual evaluation requirements rather than as reported experimental results.

Operational impact and data quality improvement

Operational evaluation should examine whether the model would be expected to improve MPI worklist prioritization, reduce manual review burden, and improve the timeliness of duplicate detection. Röchner and Rothlauf [12] explicitly connect machine learning linkage with the tradeoff between linkage quality and manual effort, while Nelson and colleagues [1] focus on master patient index optimization using machine learning. Lippi and colleagues [3] situate patient identification as a healthcare and laboratory medicine safety concern, suggesting that operational impact should include safer access to complete patient histories and reduced fragmentation of care information. The model should also be evaluated for governance outcomes, including reviewer trust, auditability, and appropriate escalation of uncertain matches.

Table 2 outlines the evaluation, interpretability, governance, privacy, and deployment safeguards required for practical implementation of the proposed MPI duplicate detection model.

Table 2. Evaluation, Interpretability, Governance, and Deployment Safeguards for the Proposed MPI Duplicate Detection Model

Implementation component	What should be evaluated	Practical evaluation approach	Governance or safety safeguard	Intended operational action
Probabilistic blocking	Whether candidate generation captures plausible duplicates without overwhelming the scoring engine	Assess blocking coverage, candidate volume, and missed-pair review using adjudicated samples	Maintain multiple blocking passes rather than relying on one deterministic rule	Generate a manageable set of high-recall candidate pairs for deeper scoring
Deep pairwise scoring	Whether the model appropriately ranks likely duplicate record pairs	Compare duplicate probability ranking against expert-reviewed pairs without using autonomous merge decisions	Require threshold review, calibration checks, and monitoring for unstable feature behavior	Prioritize candidate pairs for HIM review and MPI cleanup

Cluster resolution	Whether pairwise predictions form coherent patient-level duplicate clusters	Examine cluster consistency, many-to-one merge risks, and conflicting linkage chains	Prevent automatic many-record merges unless reviewed and approved	Present duplicate clusters as reviewable worklist items
Interpretability layer	Whether reviewers can understand why a pair was flagged	Provide field-level and trajectory-level evidence summaries for each candidate pair	Display contributing evidence such as name similarity, address continuity, encounter overlap, insurance consistency, and clinical trajectory similarity	Help HIM professionals accept, reject, or defer candidate merges
Privacy protection	Whether sensitive identity and clinical data are protected during linkage	Review tokenization, hashing, encryption, role-based access, and retention controls	Avoid unnecessary exposure of raw identifiers or full clinical histories during model scoring	Enable privacy-respectful duplicate detection across operational environments
Human review workflow	Whether the system supports accountable identity adjudication	Track reviewer decisions, override reasons, uncertainty categories, and escalation patterns	Require human confirmation before merge decisions, especially for borderline or clinically risky cases	Preserve human accountability for patient identity decisions
Registration integration	Whether real-time alerts reduce new duplicate creation without disrupting intake workflow	Evaluate alert relevance, registrar usability, and frequency of deferred or dismissed alerts	Present ranked candidates and evidence rather than forcing a registration decision	Support duplicate prevention at the point of patient registration
MPI batch deployment	Whether periodic model runs improve identity hygiene	Review nightly or scheduled worklists, unresolved queues, and stewardship workload patterns	Separate model recommendation from merge authorization	Support routine MPI maintenance and prioritization of high-risk duplicates
Failure mode monitoring	Whether the model creates unsafe false merges or misses subtle duplicates	Review false-positive patterns, false-negative patterns, demographic instability, and data-source gaps	Apply conservative thresholds for merge recommendations and escalate uncertain cases	Reduce unsafe merges and identify weaknesses in registration or data quality processes
Continuous governance	Whether the model remains trustworthy as data systems and patient populations change	Conduct periodic audits of model behavior, feature drift, reviewer feedback, and policy compliance	Use governance committee oversight involving HIM, clinical informatics, privacy, and patient safety stakeholders	Maintain safe, auditable, and institutionally accountable patient identity management

Limitations

Labeling and ground truth challenges

Ground truth for duplicate patient records is difficult to establish because expert adjudication is labor-intensive and may itself depend on incomplete or ambiguous evidence. Röchner and Rothlauf [12] show that linkage quality must be balanced against manual effort, while Coeli and colleagues [24] demonstrate that suboptimal data conditions complicate linkage evaluation. Training data may overrepresent obvious duplicates that have already been reviewed, leaving subtle identity splits underrepresented in labeled examples. As a result, the proposed model should be viewed as a decision-support system for candidate prioritization, not as a substitute for governed identity adjudication.

Privacy and computability

The model's most informative features may include sensitive demographic, insurance, address, encounter, and clinical-pattern data, which creates privacy and computability constraints. Bian and colleagues [7], Tachinardi and colleagues [8], and Pathak and colleagues [9] all emphasize that privacy-preserving linkage is essential when records cross organizational or public health boundaries. Christen, Häntschel, Christen, and Rahm [14] and Ranbaduge, Vatsalan, and Ding [15] further show that privacy-preserving representation can be combined with machine learning and deep learning approaches. Therefore, the proposed model should use tokenized, hashed, encrypted, or otherwise encoded comparison features wherever possible, while preserving enough interpretable evidence for safe human review.

Conclusion

A probabilistic record linkage and machine learning model for duplicate patient record detection could strengthen patient identity management in EHR-driven healthcare systems. The proposed framework begins with probabilistic blocking, constructs rich candidate-pair features, and applies a deep scoring layer to estimate duplicate probability. This structure allows the model to combine traditional linkage logic with learned evidence weighting. Its primary purpose is to support safer and more consistent master patient index maintenance.

The key strength of the proposed model is its integration of diverse identity evidence streams. Demographic similarity, dynamic address history, insurance identifiers, encounter patterns, and clinical trajectories each provide partial evidence, and their combination could support more robust duplicate detection than any single feature group. The model is also designed to be interpretable, with review-facing explanations that show why a candidate pair was flagged. This interpretability is essential because patient identity decisions require clinical and administrative accountability.

Important challenges remain before such a model could be responsibly deployed. Reliable labels are difficult to obtain, subtle duplicates may be underrepresented in training data, and automated linkage recommendations may be mistrusted if the evidence is not transparent. Privacy is also central because identity resolution depends on sensitive demographic, administrative, and clinical information. The model should therefore be governed as a human-reviewed decision-support system rather than as an autonomous merge mechanism.

Future work should prioritize collaborative, multi-site validation of patient matching accuracy, governance processes, and operational usability in real MPI environments. Public or semi-public benchmark datasets for patient record linkage would also help standardize evaluation without exposing sensitive identifiers. Health systems, informatics researchers, privacy experts, and health information management professionals should jointly define acceptable evidence standards for automated duplicate detection. Such collaboration would make patient identity management more accurate, auditable, and clinically trustworthy.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Sep 2024 Revised: 03 Nov 2024 Accepted: 28 Nov 2024
Published online: 25 February 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

Nelson W, Khanna N, Ibrahim M, Fyfe J, Geiger M, Edwards K, et al. Optimizing patient record linkage in a master patient index using machine learning: algorithm development and validation. JMIR Form Res. 2023;7:e44331.

Grannis SJ, Williams JL, Kasthuri S, Murray M, Xu H. Evaluation of real-world referential and probabilistic patient matching to advance patient identification strategy. J Am Med Inform Assoc. 2022;29(8):1409-15.

Lippi G, Mattiuzzi C, Bovo C, Favaloro EJ. Managing the patient identification crisis in healthcare and laboratory medicine. Clin Biochem. 2017;50(10-11):562-7.

Li X, Xu H, Grannis S. The data-adaptive Fellegi-Sunter model for probabilistic record linkage: algorithm development and validation for incorporating missing data and field selection. J Med Internet Res. 2022;24(9):e33775.

Ferguson J, Hannigan A, Stack A. A new computationally efficient algorithm for record linkage with field dependency and missing data imputation. Int J Med Inform. 2018;109:70-5.

Barbosa GC, Ali MS, Araujo B, Reis S, Sena S, Ichihara MY, et al. CIDACS-RL: a novel indexing search and scoring-based record linkage system for huge datasets with high accuracy and scalability. BMC Med Inform Decis Mak. 2020;20(1):289.

Bian J, Loiacono A, Sura A, Mendoza Viramontes T, Lipori G, Guo Y, et al. Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. JAMIA Open. 2019;2(4):562-9.

Tachinardi U, Grannis SJ, Michael SG, Misquitta L, Dahlin J, Sheikh U, et al. Privacy preserving record linkage across disparate institutions and datasets to enable a learning health system: the National COVID Cohort Collaborative experience. Learn Health Syst. 2024;8(1):e10404.

Pathak A, Serrer L, Zapata D, King R, Mirel LB, Sukalac T, et al. Privacy preserving record linkage for public health action: opportunities and challenges. J Am Med Inform Assoc. 2024;31(11):2605-12.

Nguyen L, Stoové M, Boyle D, Callander D, McManus H, Asselin J, et al. Privacy-preserving record linkage of deidentified records within a public health surveillance system: evaluation study. J Med Internet Res. 2020;22(6):e16757.

Mudgal S, Li H, Rekatsinas T, Doan A, Park Y, Krishnan G, et al. Deep learning for entity matching: a design space exploration. In: Proc Int Conf Manag Data (SIGMOD); 2018. p. 19-34.

Röchner P, Rothlauf F. Using machine learning to link electronic health records in cancer registries: on the tradeoff between linkage quality and manual effort. Int J Med Inform. 2024;185:105387.

Almeida D, Gorender D, Ichihara MY, Sena S, Menezes L, Barbosa GC, et al. Examining the quality of record linkage process using nationwide Brazilian administrative databases to build a large birth cohort. BMC Med Inform Decis Mak. 2020;20(1):173.

Christen V, Häntschel T, Christen P, Rahm E. Privacy-preserving record linkage using autoencoders. Int J Data Sci Anal. 2023;15(4):347-57.

Ranbaduge T, Vatsalan D, Ding M. Privacy-preserving deep learning based record linkage. IEEE Trans Knowl Data Eng. 2023;36(11):6839-50.

Randall SM, Brown AP, Ferrante AM, Boyd JH. Privacy preserving linkage using multiple match-keys. Int J Popul Data Sci. 2019;4(1):1094.

Vidanage A, Ranbaduge T, Christen P, Randall S. A privacy attack on multiple dynamic match-key based privacy-preserving record linkage. Int J Popul Data Sci. 2020;5(1):1345.

Randall S, Wichmann H, Brown A, Boyd J, Eitelhuber T, Merchant A, et al. A blinded evaluation of privacy preserving record linkage with Bloom filters. BMC Med Res Methodol. 2022;22(1):22.

Ranbaduge T, Christen P. A scalable privacy-preserving framework for temporal record linkage. Knowl Inf Syst. 2020;62(1):45-78.

Lim D, Randall S, Robinson S, Thomas E, Williamson J, Chakera A, et al. Unlocking potential within health systems using privacy-preserving record linkage: exploring chronic kidney disease outcomes through linked data modelling. Appl Clin Inform. 2022;13(4):901-9.

Brown AP, Randall SM. Secure record linkage of large health data sets: evaluation of a hybrid cloud model. JMIR Med Inform. 2020;8(9):e18920.

Brown AP, Borgs C, Randall SM, Schnell R. Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. BMC Med Inform Decis Mak. 2017;17(1):83.

Suo Q, Ma F, Yuan Y, Huai M, Zhong W, Gao J, et al. Deep patient similarity learning for personalized healthcare. IEEE Trans Nanobiosci. 2018;17(3):219-27.

Coeli CM, Saraceni V, Medeiros PM Jr, da Silva Santos HP, Guillen LC, Alves LG, et al. Record linkage under suboptimal conditions for data-intensive evaluation of primary care in Rio de Janeiro, Brazil. BMC Med Inform Decis Mak. 2021;21(1):190.

de Paula AA, Pires DF, Alves Filho P, de Lemos KR, Barçante E, Pacheco AG. A comparison of accuracy and computational feasibility of two record linkage algorithms in retrieving vital status information from HIV/AIDS patients registered in Brazilian public databases. Int J Med Inform. 2018;114:45-51.

Kasthurirathne SN, Grannis SJ. Machine learning approaches to identify nicknames from a statewide health information exchange. AMIA Summits Transl Sci Proc. 2019;2019:639-48.

Leshin J, Sanghvi A, Ravuri K, Owen M, Kho A. The impact of name transformation on match rates within a large consumer database. AMIA Annu Symp Proc. 2022;2022:692-701.

Li Y, Li J, Suhara Y, Doan A, Tan WC. Deep entity matching with pre-trained language models. arXiv. 2020;arXiv:2004.00584.

Govind Y, Konda P, Suganthan GCP, Martinkus P, Nagarajan P, Li H, et al. Entity matching meets data science: a progress report from the Magellan project. In: Proc Int Conf Manag Data (SIGMOD); 2019. p. 389-403.