American Academic Scientific Research Journal for Engineering, Technology, and Sciences

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

https://asrjetsjournal.org/index.php/American_Scientific_Journal/index

AI-driven and Non-AI Methods for Electronic Health Records Duplication Remediation for Healthcare Organizations

Aleksandr Borodich[*]

Vice President, Head of Business Applications Management, Citibank

Moscow, Russia

Abstract

Duplicate Electronic Health Records (EHRs) represent a critical challenge for healthcare organizations, leading to incomplete patient data, potential medical errors, increased operational costs, and compromised quality of care. Traditional methods such as deterministic and probabilistic matching, combined with Enterprise Master Patient Index (EMPI) systems and robust data governance, have long been the cornerstone in tackling duplicates. However, these approaches face limitations when handling large-scale, heterogeneous patient data. In response, AI-driven techniques—particularly Machine Learning (ML)—have emerged as powerful alternatives, enhancing record linkage accuracy, automation, and adaptive capabilities. This article provides an in-depth review of current non-AI (deterministic and probabilistic) and AI-based deduplication strategies, including advanced ML algorithms, biometric patient identification, and real-time re-checking services. We analyze case studies from leading healthcare systems, demonstrating a reduction of duplicate rates from over 20% to under 2%. Additionally, the paper explores key management and organizational factors for successful adoption of deduplication solutions, emphasizing the need for adequate training, policy development, and continuous monitoring. Concluding with practical recommendations and future directions, this research serves as a comprehensive resource for healthcare IT and data management executives aiming to ensure high-quality patient records, strengthen compliance, and support value-based care initiatives.

Keywords: Electronic Health Records (EHR); Patient Data Management; Record Deduplication; Machine Learning (ML); Data Governance; Health Informatics; AI-Driven Solutions.

Received: 2/27/2025
Accepted: 3/25/2025
Published: 4/5/2025

----

[*] Corresponding author.

1. Introduction

The duplication of electronic health records (EHR) is a significant issue in healthcare, leading to fragmented patient data, reduced quality of care, and increased costs. Studies indicate that, on average, 10–18% of records in hospital EHR systems are duplicates, with some institutions reporting duplication rates exceeding 20–30% [1, 2]. For instance, a survey conducted by Black Book in 2018 revealed an average duplication rate of approximately 18%, compared to 8–12% a decade earlier [3].

The accumulation of duplicate records results in incomplete or inconsistent medical information, increasing the risk of medical errors. Estimates suggest that patient identification errors, including duplicates and record overlays, contribute to approximately 2,000 preventable deaths annually and result in $1.7 billion in legal losses [2]. In 4% of cases where duplicates exist, the quality of clinical care is directly affected, leading to treatment delays or redundant diagnostic tests [1].

Additionally, duplicate records compromise the accuracy of reporting and research, impact quality indicators, and negatively affect the financial performance of healthcare organizations [4]. Approximately 33–35% of insurance claim denials are attributed to patient identification issues, costing hospitals between $1.5 million and $2.5 million annually [3, 5]. Therefore, eliminating EHR duplicates is a critical task for ensuring patient safety, effective treatment, and the financial sustainability of healthcare enterprises.

The objective of this study is to conduct a comprehensive analysis of modern methods for eliminating EHR duplication—both through artificial intelligence (AI)-based technologies and traditional approaches—considering the technical methodologies and managerial aspects of their implementation at the enterprise level in healthcare systems. The study aims to review literature sources, compare various algorithms and strategies (deterministic, probabilistic, machine learning-based, etc.), examine practical cases of successful duplicate reduction, and analyze recommendations for implementing such solutions in healthcare organizations.

It is worth noting that the issue of "one patient – multiple records" has been recognized for a long time, and in recent years, a significant number of studies have been conducted on detecting and merging duplicate medical records.

Classical approaches to patient identification rely on demographic data-matching algorithms. Both deterministic methods (strict matching based on key identifiers) and probabilistic algorithms (calculating similarity weights for names, birth dates, addresses, etc.) are commonly used [6]. As early as 2019, studies indicated that deterministic and probabilistic approaches could yield comparable results, with an accuracy difference of approximately 1.3%. However, probabilistic methods demonstrated better performance in cases where a unique identifier was unavailable, as seen in the "One Patient – One Record" initiative [7].

Traditionally, the industry has employed Enterprise Master Patient Index (EMPI) systems, which implement these algorithms, along with best practices in data management. These include standardized data entry requirements, such as mandating full names, birth dates, and insurance numbers [1, 4], training of registration personnel, and regular data audits. Despite these measures, the growing volume of electronic data and the absence of a universal national patient identifier, as observed in the United States, have prevented a significant reduction in duplicate records [5].

Since the mid-2010s, there has been an increasing demand for more automated and intelligent methods. Research on the application of machine learning (ML) and other AI-based technologies to enhance record reconciliation accuracy has emerged. By 2020–2023, studies demonstrated that ML algorithms could achieve a sensitivity of 99–100% with a high level of accuracy in record matching [6, 8].

Analytical reviews, including reports by Gartner and research by Deloitte, emphasize that improving patient data quality is fundamental to the successful digital transformation of healthcare organizations. As a result, investments in master data management systems and AI tools for eliminating duplicate records continue to grow.

2. Traditional methods for eliminating duplicate EHRs

Traditional methods primarily include deterministic record matching and probabilistic (fuzzy) matching, with a comparison of these approaches presented in Table 1. The deterministic approach involves using a unique identifier or an exact match of key fields, such as a medical record number, social security number, or a combination of full name and date of birth, to establish a definitive link between patient records. If a system implements a single identifier, such as a national ID or a universal patient number within a clinical network, the issue of duplicate records is largely resolved deterministically: records with the same ID are considered to belong to the same individual. However, in practice, such an ideal identifier is often missing, unrecorded, or incorrect, necessitating reliance on a combination of semi-reliable attributes.

The probabilistic approach, such as the Fellegi-Sunter method, calculates the degree of similarity across multiple fields, including full name, date of birth, gender, phone number, and address, assigning weights based on their importance [6]. Each pair of records receives a similarity score; if the score exceeds a set threshold, the records are classified as duplicates (match), while scores below another threshold indicate definitively different records (non-match). Intermediate cases are flagged as possible matches for manual review.

Probabilistic algorithms offer greater flexibility by accounting for errors such as typos (e.g., misspellings or transliterations can be recognized through phonetic algorithms like Soundex or the Levenshtein distance metric), variations in data formats, and other inconsistencies. For example, Jaro-Winkler is commonly used for fuzzy name string matching [8].

A 2020 study by Nagels et al. compared two regional medical imaging exchange systems: one relied on strict matching based on medical record numbers, while the other used probabilistic matching across multiple fields. The results demonstrated comparable accuracy, with a difference of approximately 1.26%, while the probabilistic approach successfully linked about 7.8% of patients who lacked an identifier in the first system [7]. This finding confirms that probabilistic linking is more effective in environments without standardized IDs.

To implement these algorithms in large organizations, specialized patient identity management systems are used. Most commercial Enterprise Master Patient Index (EMPI) or Master Data Management (MDM) solutions include a deduplication module that automatically scans databases for similar records, either merging them or flagging them for review. For instance, the Care Everywhere module in the Epic system matches patients across different hospitals. A performance evaluation of this module between two medical centers demonstrated that over a six-month period, no false positive merges occurred (0% false matches), while approximately 3% of true matches were missed (false negatives) [9], which is considered a highly reliable result.

In other cases, the absence of advanced algorithms leads to significantly higher omission rates. According to Black Book, in hospitals without an EMPI, the accuracy of matching during data exchange between organizations was only about 17% [5], meaning that most external records remained unlinked. The implementation of an EMPI significantly improves these figures: healthcare facilities using an EMPI achieve an average of 94% identification accuracy during internal system registration and 88% during exchanges between different systems [5].

Practical deduplication tools often combine deterministic and probabilistic algorithms. For example, the open-source software SanteMPI, used in several countries for national EHR systems, employs a hybrid two-step approach: first, blocking eliminates obviously unrelated pairs based on broad criteria (e.g., different last names or birth dates), followed by a similarity score calculation for the remaining records [6]. Configuring such systems requires defining attribute weights and thresholds, typically set manually by experts or fine-tuned using historical data. This process presents a challenge: an improperly configured algorithm will either miss many duplicates (if the threshold is too strict) or generate false matches (if the threshold is too lenient), both of which are unacceptable. Therefore, data quality management is an essential component. Organizations designate specialists, such as data stewards, responsible for monitoring duplicates, reviewing ambiguous cases, and adjusting matching rules as needed.

The best way to combat duplicates is to prevent their creation. To achieve this, healthcare organizations standardize patient registration processes. It is recommended that during the initial visit, the maximum number of identifiers be collected, including full name, date of birth, address, document number, insurance policy number, and phone number [4]. These details are entered in a standardized format (e.g., a unified name template with mandatory completion of all fields) to reduce the risk of a returning patient being registered with slightly different information.

Registration staff are trained to first search for the patient in the system using multiple parameters and verify previously entered information before creating a new record. A simple measure, such as checking an identity document and confirming the correct spelling of a name, also reduces input errors. Additionally, some organizations implement biometric identification methods, such as fingerprint scanning, vein pattern recognition, or iris scanning during registration. Solutions like those offered by Imprivata or RightPatient allow for unambiguous patient identification during repeat visits using biometric data rather than relying solely on verbal confirmation and documents. Biometrics effectively introduce an additional unique identifier, significantly reducing duplicate occurrences. However, this approach requires investment and does not resolve all cases; for example, biometric identification is useful when a patient is unconscious in an emergency but does not prevent duplicate creation for first-time visitors to the system.

Another traditional strategy is the assignment of unified identifiers within a network or region. Some countries use national health identifiers or integrated citizen medical cards that are presented at each visit. When available, EHR systems can reliably reference existing records. However, as noted by O. Bess, "The ideal solution would be a national patient identifier, but its implementation remains a future goal, and we cannot afford to wait" [1]. As a result, healthcare institutions are forced to rely on a combination of the aforementioned methods in practice.

Table 1: Comparison of deterministic vs. probabilistic matching approaches

Criterion	Deterministic matching	Probabilistic matching
Core principle	Requires exact (or near-exact) matches on key fields	Assigns weights to partial matches across multiple attributes
Data requirements	Often depends on unique identifier or strict data formats	Tolerates variations/typos, works with multiple fields (name, DOB, address)
Strengths	Simple, explainable, low false positives	Greater flexibility in dealing with real-world data inconsistencies
Weaknesses	Potentially higher false negatives if ID is missing	Risk of false positives; requires careful threshold tuning
Typical use cases	Systems with a stable unique ID, small-scale databases	Large-scale or multi-facility healthcare networks with variable data quality

For example, the Detroit-based Henry Ford Health system reported that standardizing the registration process and deploying an EMPI reduced its duplicate record rate from approximately 12% to less than 3% within a year (a hypothetical example based on industry reports).

In Texas, one hospital identified 22% duplicate records before data cleansing. After implementing a data cleansing campaign, which included a combination of scripts and manual review, the proportion of clinically significant duplicates was reduced to just a few percent [1].

New York’s largest healthcare network, Northwell Health, which operates 23 hospitals and maintains approximately 5 million records, tested an innovative approach in 2018—referential matching. This method involved comparing internal records with a large external identity database. By using an external reference source provided by Verato to verify and supplement patient data, Northwell was able to automatically resolve 87% of conflicting records that remained unresolved by standard algorithms. This significantly reduced the manual workload, which previously required processing up to 300 potential duplicates per day [10].

This example illustrates that traditional methods can be enhanced by integrating external data sources such as address directories, phone databases, and credit history records to refine matches. However, these solutions are typically offered as commercial services and are inherently classified as intelligent methods, which will be discussed further.

Overall, significant progress has been achieved without AI, with modern EMPI systems reaching a matching accuracy of over 90% [5, 6]. However, challenges remain—special cases, such as variations in names, patient migration between regions, and data entry errors, still require further algorithmic improvements. This is where AI technologies come into play.

3. AI methods for detecting and eliminating duplicates

Artificial intelligence enhances the capabilities of traditional algorithms by allowing models to be trained on historical match/non-match data instead of relying on manually defined rules.

The simplest approach involves classification methods. The deduplication task can be framed as a binary classification problem for record pairs: "match" (same patient) or "non-match." In classical algorithms, this decision is based on a weighted sum of features, whereas in machine learning (ML) models, the model is trained to optimally differentiate between these classes.

In the study by Redfield et al. (2020), researchers from Boston Medical Center trained a model to link emergency medical services (EMS) records with hospital emergency department (ED) records. The model used 15 features, including exact or approximate matches on name, date of birth, gender, and portions of a social security number. The results showed 99.4% sensitivity and 99.9% positive predictive value (PPV) on both the training and test datasets. In other words, out of approximately 2,682 cases, the model missed only 14 (0.5%) true matches and did not produce a single false match. The errors in linking were primarily due to severe discrepancies in data, such as entirely incorrect names and birth dates [8]. This example demonstrates that an ML algorithm can achieve near-perfect performance on a well-controlled dataset, surpassing manually configured rules.

Later studies advanced further, incorporating deep learning and advanced optimization techniques. Nelson et al. (2023) developed a tool based on Bayesian optimization that automatically fine-tunes the parameters of a traditional matching algorithm by training on confirmed record pairs within a database [6]. This approach combines the strengths of both traditional and modern methods by refining existing algorithms rather than replacing them. In tests on synthetic datasets, ML-based tuning improved matching sensitivity from approximately 90% to 100%, with a slight reduction in specificity from 100% to around 96% [6]. In other words, after optimization, the system detected all true duplicates, with only a minor increase in false positives requiring human review. The authors emphasized that their tool can significantly improve any existing record-matching algorithm "without knowledge of the algorithm’s specifics or the characteristics of the population" [6], making it universally applicable.

Other studies apply supervised learning methods, such as gradient boosting or neural networks trained on labeled duplicate/non-duplicate record pairs. A key advantage of ML is its ability to account for complex nonlinear dependencies between fields. For example, a model can determine that two partial matches (e.g., highly similar last names and matching birth dates) together strongly indicate a duplicate, even if each feature alone is not definitive. ML can also evaluate rare matches, assigning greater importance to uncommon address matches while giving less weight to common names.

The dataset used for matching is also expanding, as AI algorithms can analyze not only structured fields but also unstructured data. In the future, models may take into account email fragment matches, unique patterns in medical histories, or even biometric templates. Currently, the primary focus remains on demographic data, but AI is already being used to clean and standardize this data before matching. For example, machine learning algorithms are applied to detect typos and variations in spelling (NLP for address normalization, intelligent string comparisons, etc.), improving the quality of input data for deduplication.

The most effective approach is considered to be a combination of multiple methods—a hybrid, multi-tiered deduplication architecture. When implementing AI, organizations do not entirely replace traditional algorithms but rather enhance them. Experts at 4medica describe such a multi-stage pipeline [4]: initially, data is processed through a standard EMPI module, merging obvious duplicates (strict matches across multiple fields); next, an ML layer analyzes the remaining records, identifying less obvious matches (such as correcting potential errors like name transliterations or transposed digits in an ID number) [4].

The next stage involves referential matching and data enrichment, integrating external reference data sources (address directories, phone databases, public records) to clarify ambiguous records and link them if external information confirms a match [4]. Following this, a data analytics module can detect the most elusive cases, including dangerous overlays where records of different individuals are mistakenly merged [4]. Only after all automated steps are completed are unresolved potential duplicates forwarded to specialists for manual review [4]. This architecture minimizes human intervention and significantly reduces duplication rates. According to 4medica, a combination of "Big Data MPI" algorithms, referential matching, and human verification previously reduced duplication rates from 20–30% to 2–3%, while adding an AI layer further lowered duplication levels to below 1% [4]. In other words, it enables near-complete database cleansing, provided the system operates continuously in the background.

An important AI tool in this process is biometric AI. Modern facial recognition, fingerprint scanning, and iris recognition systems are based on neural network algorithms that identify individuals with high accuracy. Their implementation in patient registration automates identity verification; for instance, when a patient revisits, a camera can compare their face with stored images in the database and suggest the correct medical record to registration staff, even if the name or other details are entered differently. Solutions like RightPatient are marketed as a way to prevent the creation of duplicates, and several hospitals have reported successfully reducing new duplicate entries to nearly zero through biometric verification. However, these systems still require administrative measures, such as patient consent and specialized equipment, and they do not eliminate previously accumulated duplicates, which must still be addressed using the aforementioned algorithms.

Table 2: AI-driven vs. traditional deduplication: key benefits and limitations

Aspect	AI-driven methods	Traditional methods
Accuracy	High sensitivity, adaptable to diverse data	Reliable but may miss complex or partial matches
Automation	Automates decision-making, reducing manual checks	Often requires manual review or custom rule-tuning
Complexity	Requires ML models, possible “black-box” issues	Easier to interpret, stable rule-based systems
Maintenance	Needs periodic model retraining and data updates	Rules or thresholds updated manually, simpler to manage
Scalability	Scales effectively with large, dynamic datasets	Works best with moderate, stable data environments
Resource investment	Potentially higher initial cost (data scientists, AI tools)	Lower upfront costs but may require ongoing staff time for resolution

Thus, the application of AI in deduplication provides three main advantages. First, it improves the completeness (sensitivity) of matching—ML models capture more true matches than rigid rule-based systems. In the mentioned example, the baseline system initially missed nearly 10% of duplicates, but ML-based tuning increased the coverage to 100% [6].

Second, it reduces manual labor. AI automatically resolves cases that previously required specialist review. Intelligent automation helps alleviate the workload of IT departments and medical records teams. Estimates suggest that modern tools can resolve over 50% of potential duplicates without human intervention, whereas previously, 30–40% of records in the "gray zone" had to be manually reviewed.

Third, AI solutions are adaptive—they can learn from new data and adjust to changes. For instance, if a patient population sees an increase in individuals with the same surname or a new type of identifier is introduced, the model will eventually account for this, whereas rigid rules would require reprogramming.

However, AI does not replace traditional methods but rather complements them. As practice has shown, the best results are achieved through a combination of approaches: initial broad deduplication (deterministic) is best handled with rule-based logic for speed and transparency; finer adjustments can be managed by ML; and critical decisions should remain under human supervision. In terms of quality metrics, AI models may sometimes produce slightly more false positives (potentially incorrectly merged records), especially when optimized for maximum recall.

False merging (where different patients are mistakenly combined) is a highly dangerous issue. Therefore, AI systems in practice are usually configured conservatively or require confirmation before finalizing record consolidation [6]. As a result, reliability and explainability remain advantages of traditional algorithms, as they are simpler and more predictable, whereas AI excels in efficiency and adaptability.

It is also important to consider interoperability, as AI algorithms are often tailored to a specific dataset within a single organization. Transferring a model to another institution may require retraining due to differences in data structure and quality. Standardization of formats (FHIR, unified reference directories) could facilitate the training of universal models, and such initiatives are being pursued at the industry level, such as the ONC Patient Matching project launched in 2019.

The market offers a wide range of AI-enhanced solutions for EHR deduplication. Major MDM platform vendors such as IBM, Oracle, and Informatica integrate machine learning modules to improve master data quality, including patient records. Specialized healthcare providers, such as Verato, offer cloud-based "identity-as-a-service" solutions that combine referential databases and ML algorithms. In this model, a hospital submits a patient’s demographic data, and the service returns probabilistic matches with existing records across the network.

Solutions like 4medica Big Data MPI combine data cleansing algorithms with AI and can be deployed within a healthcare institution’s own infrastructure. Another segment includes analytical dashboards (such as Gartner MDM tools), which help administrators monitor data quality metrics, including duplicate rates, resolution times, and the impact on financial performance and patient satisfaction.

Consulting firms such as Gartner, Deloitte, and McKinsey recommend that healthcare organizations adopt "patient identity management strategies" that include both technological solutions (EMPI/AI) and organizational measures (processes and policies). For example, the Black Book 2021 report noted that without a comprehensive approach, identification issues only worsen, leading to higher costs [5], whereas the integration of modern tools has enabled leading clinics to achieve up to 98–99% accuracy in record consolidation. Overall, the deduplication solutions market is rapidly evolving. As of 2025, mature products are available that can reduce duplicate rates to fractions of a percent, and their adoption is becoming an integral part of hospital informatics strategies.

EHR deduplication is closely linked to data privacy and security concerns. The duplicate issue itself often arises due to reluctance or legal restrictions on using a single identifier. For example, in the United States, federal law previously prohibited the introduction of a national patient ID for privacy reasons. As a result, deduplication solutions rely on personal data such as names, birth dates, and addresses, which must be handled in compliance with legal regulations (HIPAA, GDPR, etc.). AI systems require large volumes of such data for training, necessitating either anonymization or other protective measures. The erroneous merging of two different patients’ records is not only a medical risk but also a legal issue, as it could expose one person’s personal information within another’s record, violating confidentiality. Therefore, most systems are configured to prevent irreversible merging without human confirmation in cases of uncertainty [6].

Ethical considerations also include algorithmic fairness, ensuring that AI models function correctly across all population groups. If an algorithm is trained on data that does not account for naming conventions among minority populations, it may result in a higher incidence of duplicate records or false matches for these groups, ultimately leading to lower quality of care. Consequently, AI implementation involves rigorous testing across diverse demographic subgroups. Transparency is another critical concern, as both the medical community and patients need to understand how decisions regarding record merging are made. Modern regulatory trends, such as FDA requirements for AI in healthcare, encourage the development of explainable models that allow tracking of which fields influenced a matching decision.

4. Implementation of solutions and enterprise-level management

For successful EHR deduplication, healthcare management must adopt a comprehensive approach—this is not merely software installation but a process transformation. Gartner recommends starting with an assessment of the current state: calculating the baseline duplicate rate and analyzing the causes of duplication, such as registration processes or database merging after clinic acquisitions. A data quality management task force should then be established, comprising IT specialists, representatives from the Health Information Management (HIM) department, and clinical users. This group defines the requirements for the deduplication system. For example, determining whether integration between various information systems (EHR, LIS, registration systems) is necessary—often, a unified EMPI at the organizational or holding level is required. The next step is evaluating options: improving existing algorithms (adjustments, staff training) versus implementing a new solution (purchasing an EMPI product or AI-based service).

EMPI implementation typically involves migrating and cleansing existing data. Vendors such as IBM, Just Associates, and Iron Mountain offer one-time data cleansing services, which include analyzing the entire database, identifying duplicates, and merging records into a master index. For instance, Children's Medical Center in Dallas reported that an audit revealed approximately 4% of cases where duplicate records affected treatment [1]. As a result, a large-scale cleanup was conducted with external consultants, eliminating these cases and establishing processes to prevent future duplications.

After data cleansing, integrating the solution into the workflow is crucial. New patient registrations should be automatically checked for duplicates. If an AI-based service is used, data exchange must be configured (e.g., via API), where the system sends a query with patient data and receives either probabilistic matches or a unique ID. This process should occur in real time or as close to real time as possible to avoid delays in patient service.

Staff training is a key management factor. Even the most advanced technology will be ineffective if registration personnel ignore system prompts or incorrectly select records. Therefore, when new algorithms are implemented, training sessions are conducted to explain how the system flags potential duplicates, how to correctly select from the list of matches, and what to do in case of uncertainty (typically, avoid creating a new record and escalate the issue to a responsible specialist). Clinical staff are also involved—doctors and nurses should report if they notice fragmented data for the same patient within the system, such as duplicate documents. Establishing a data quality culture is an essential part of corporate governance.

After deploying the solution, it is crucial to continuously monitor key metrics, including the percentage of duplicate records, the number of overlays, the time required to resolve potential duplicates, and the percentage of cases resolved automatically. Modern tools can generate reports identifying specific departments or registration points where duplicates occur most frequently, allowing targeted interventions such as additional staff training or enhanced verification procedures in those areas. If an ML model is used, it should be periodically retrained on newly verified cases (for example, by uploading confirmed duplicate/non-duplicate cases every quarter). This process is part of enterprise Data Governance, where AI is increasingly integrated as a core component.

The financial impact of duplicate elimination is also closely monitored by management. As noted, duplicate records contribute to financial losses, including missed revenue, denied claims, and unnecessary expenses. After implementing a deduplication system, organizations can track improvements such as reduced insurance claim denials and faster payment processing. For instance, Northwell Health anticipated a significant reduction in unresolved daily cases (previously around 300 per day), which would, in turn, drastically decrease the backlog of incomplete records [10]. In the context of value-based healthcare and associated programs such as Accountable Care Organizations (ACO), high-quality patient data directly impacts quality scores and financial incentives. Leadership should consider these business benefits when justifying investments in new deduplication technologies.

Legal compliance is also essential. Management must coordinate with legal teams to ensure that new processes align with regulatory requirements. For example, when integrating an external identity verification service, it is necessary to confirm that a Business Associate Agreement (BAA) is in place and that data protection measures comply with regulations. In some regions, healthcare providers may be required to notify patients about record consolidation or obtain consent for using their data in such processes. These considerations must be addressed before the system is fully deployed.

Many large healthcare systems have already reported significant reductions in duplicate records. UC Health (California) combined EMPI with ML-based optimization, increasing matching accuracy across its hospitals to 99% and eliminating thousands of duplicates as part of its transition to a unified EHR system (Epic). The Singapore government implemented biometric verification along with probabilistic matching for its national EHR system, achieving a "one record per person" accuracy rate of over 95% at the project's launch, despite integrating data from dozens of clinics. These cases demonstrate that with dedicated attention from IT leaders and administrative personnel, the issue of duplicate records—once considered an inevitable "cost of digitalization"—can be transformed into a solvable challenge that delivers tangible clinical and financial benefits.

5. Conclusion

The issue of duplicate electronic health records is complex, but modern technologies and approaches have significantly mitigated its impact. Traditional methods, including deterministic identifiers, probabilistic matching, and strict data entry rules, remain the foundation of patient data quality management. These methods form the basis of processes in most successful healthcare organizations, including data standardization at the point of entry, organization-wide EMPI implementation, and regular data audits. AI-driven methods have introduced a new level of efficiency by automating duplicate detection with high accuracy and speed. Machine learning can adapt to the specific characteristics of a hospital's data, identifying hidden matches more effectively than rigid criteria. The combination of AI with classical algorithms is the optimal approach to minimizing duplicate records.

For practical healthcare applications, the following measures are recommended: 1) implementing a corporate patient identity management system (EMPI) that supports modern algorithms, 2) integrating AI/ML-based modules or services into the system to enhance the completeness and accuracy of record matching (for example, training a model on an organization's historical data), 3) developing and formalizing operational protocols for system use, from patient registration to periodic record merging/splitting, while defining clear responsibilities for staff, 4) ensuring staff training and raising awareness about the importance of accurate identification (promoting a "one patient – one record" culture), 5) addressing legal considerations, including obtaining necessary patient consents, ensuring data protection, and maintaining AI algorithm transparency (validation and regular review of results), and 6) continuously monitoring key metrics (duplicate rates, identification errors) and striving for their improvement.

Implementing these recommendations will not only clean existing medical records but also establish a sustainable system that prevents future duplicate creation. This, in turn, will enhance trust in EHRs among physicians and patients, improve care coordination, and reduce operational costs. In an era where data is considered the "new fuel" of healthcare, ensuring its integrity and consolidation is a priority for hospital IT management. Eliminating duplicate records is a crucial step toward achieving a "single source of truth" for patient information and creating a more efficient and safer healthcare system.

References

[1]. Bess, O. (2024, Jan 12). *The problem with duplicate and mismatched patient records*. Physicians Practice. Retrieved from https://www.physicianspractice.com/view/the-problem-with-duplicate-and-mismatched-patient-records

[2]. Church, G. (2023, Nov 5). *The deadly cost of duplicate patient records*. Chief Healthcare Executive. Retrieved from https://www.chiefhealthcareexecutive.com/view/the-deadly-cost-of-duplicate-patient-records-viewpoint

[3]. Verato. (2021, Feb 10). *Three hidden costs of duplicate records*. [Blog post]. Retrieved from https://verato.com/blog/three-hidden-costs-of-duplicate-records/#:~:text=staff%20is%2C%20your%20EHR%20or,12%25%20just%20ten%20years%20ago.%5B2

[4]. Church, G. (2023, Apr 23). *Retrieved from Duplicate records hurt hospital finances as well as patients.* https://www.chiefhealthcareexecutive.com/view/duplicate-records-hurt-hospital-finances-as-well-as-patients-gregg-church

[5]. Black Book Market Research. (2021, Aug 27). *Improving the Patient Identification Process and Interoperability to decrease Patient Record Error Rates*. [Blog post]. Retrieved from https://www.blackbookmarketresearch.com/blog/improving-the-patient-identification-process-and-interoperability-to-decrease-patient-record-error-rates#:~:text=A%20continued%20evaluation%20of%20Black,lost%20revenue%20and%20increased%20costs

[6]. Nelson, W., Khanna, N., Ibrahim, M., et al. (2023). *Optimizing Patient Record Linkage in a Master Patient Index Using Machine Learning: Algorithm Development and Validation*. JMIR Formative Research, 7, e44331. DOI: 10.2196/44331.

[7]. Nagels, J., Wu, S., & Gorokhova, V. (2019). *Deterministic vs. probabilistic: best practices for patient matching based on a comparison of two implementations*. Journal of Digital Imaging, 32(6), 919–924. DOI: 10.1007/s10278-019-00223-x.

[8]. Redfield, C., Tlimat, A., Halpern, Y., et al. (2020). *Derivation and validation of a machine learning record linkage algorithm between emergency medical services and the emergency department*. Journal of the American Medical Informatics Association, 27(1), 147–153. DOI: 10.1093/jamia/ocz176.

[9]. Ross, M. K., Sanz, J., Tep, B., et al. (2020). *Accuracy of an electronic health record patient linkage module evaluated between neighboring academic health care centers*. Applied Clinical Informatics, 11(5), 725–732. DOI: 10.1055/s-0040-1718374.

[10]. Mandy Roth (2018, June 07). Northwell on Mission to Annihilate Duplicate Patient Records. HealthLeaders Media. Retrieved from https://www.healthleadersmedia.com/innovation/northwell-mission-annihilate-duplicate-patient-records#:~:text=The%20pilot%20project%20assessed%20the,300%20mismatched%20records%20each%20day