Health and Technology (2025) 15:1011–1020
https://doi.org/10.1007/s12553-025-01012-3

REVIEW PAPER

Tokenization of electronic health records and healthcare data: enhancing security and privacy while enabling usability

João Carlos Ferreira[1,2] · Luis B. Elvas[1,2] · Ricardo Correia[3] · Miguel Mascarenhas[3,4]

Received: 9 October 2024 / Accepted: 11 August 2025 / Published online: 17 September 2025
© The Author(s) 2025

Abstract

Purpose The purpose of this paper is to examine the role of tokenization in protecting Electronic Health Records (EHRs) and healthcare data, with a focus on ensuring data security and privacy compliance. Tokenization, which replaces sensitive patient data with unique identifiers (tokens), is explored as a means to prevent unauthorized access while maintaining data utility for healthcare operations, research, and analytics. Additionally, the paper aims to assess how tokenization helps healthcare organizations comply with regulations like HIPAA, and how it facilitates secure data sharing and analytics.

Methods This state-of-the-art review analyzes the latest advancements in tokenization techniques, particularly their applications in safeguarding patient data during storage, transmission, and analysis. The review includes an exploration of tokenization's role in secure data exchange, system interoperability, and the integration of natural language processing (NLP) for analyzing de-identified healthcare data. Various methodologies and real-world implementations of tokenization in healthcare are comprehensively assessed.

Results The analysis reveals that tokenization effectively enhances data security and privacy, supporting compliance with stringent regulatory frameworks like HIPAA. Furthermore, tokenization enables the secure exchange of healthcare data across systems while maintaining interoperability. The integration of NLP for de-identified data analysis further expands its application in research and analytics without compromising patient confidentiality.

Conclusions The review highlights the potential of tokenization to balance data security and usability within healthcare systems. It improves patient care by facilitating secure data sharing and supporting advanced analytics, while simultaneously ensuring operational efficiency and regulatory compliance. Tokenization emerges as a crucial tool in enhancing the protection and utility of healthcare data.

Keywords Tokenization · Electronic health records (EHRs) · Healthcare data security · Data privacy · HIPAA compliance · Data sharing · Interoperability · Natural language processing (NLP) · Healthcare analytics · Patient data protection · Health information security

✉ João Carlos Ferreira
joao.c.amaro.ferreira@himolde.no; jcaf2001@gmail.com

✉ Miguel Mascarenhas
miguelmascarenhassaraiva@gmail.com

Luis B. Elvas
luiselvaas@gmail.com

Ricardo Correia
ricardo@bioghp.com

[1] Logistics, Molde University College, Molde NO-6410, Norway

[2] ISTAR, Instituto Universitario de Lisboa ISCTE-IUL, Lisboa 1649-026, Portugal

[3] BioGHP, Lisboa 1000-260, Portugal

[4] Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine, of University of Porto, Porto 4200-319, Portugal

1 Introduction

The tokenization of Electronic Health Records (EHRs) and healthcare data has emerged as a vital technique for enhancing data security, privacy, and usability in the healthcare industry. Tokenization involves replacing sensitive patient information with unique tokens, ensuring that data remains protected during storage, transmission, and analysis. This approach not only helps healthcare organizations comply with stringent regulations like the Health Insurance Portability and Accountability Act (HIPAA) but also facilitates secure data sharing and analytics, essential for improving healthcare outcomes.

Tokenization is a security procedure used to replace sensitive health information with unique identifiers, or tokens, that preserve the data's usefulness while disclosing the true secret information [1]. This is an effective strategy for preventing unauthorized access and potential security and privacy breaches to patient data in the healthcare sector [2].

Once the tokens are formed, the original data cannot be recovered from them without access to the tokenization system. This strategy is especially important in light of healthcare laws like the Health Insurance Portability and Accountability Act (HIPAA), which impose strict requirements on the security of patient data [3]. Tokenization helps healthcare firms comply with these regulations by ensuring that sensitive data is never exposed in its unprocessed form. Tokens are used to store actual data in a central, secure location as opposed to being used for everyday operations, analytics, and data sharing.

Tokenization lowers the possibility of data breaches, which may be extremely costly and damage the reputations of healthcare providers [4, 5]. Data breaches are prevented by limiting the exposure of private health information. Furthermore, it increases patient trust as people feel more certain that the greatest degree of security is being applied to their personal data [6]. The efficient provision of healthcare services and the motivation of patients to contribute their data, which might enhance health outcomes through more informed clinician judgments and research, depend on this trust.

It can also facilitate data exchange and interoperability among various healthcare organizations and systems. Tokenization makes data sharing between entities safe and de-identified in situations like population health management or research partnerships [7]. This allows for collaboration without compromising privacy. Additionally, it gives medical professionals the ability to do sophisticated analytics on patient data without disclosing private information, allowing them to use data-driven insights to enhance both patient care and operational effectiveness [8].

The aim of this state-of-the-art review is to explore the latest advancements, methodologies, and real-world implementations of tokenization in EHRs and health data. Specifically, this review highlights technological innovations, including applications in natural language processing (NLP), and addresses the challenges faced by the healthcare industry in maintaining a balance between data security and utility. By synthesizing recent literature, this report wishes to provide an in-depth analysis of how tokenization, in conjunction with NLP techniques, can help safeguard sensitive health data while enabling its meaningful use in healthcare operations, research, and population health management.

2 Methodology

This section explains the overall research methodology we followed during the paper selection process. We adopted a Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) methodology to ensure the accuracy and reliability of our findings. The PRISMA statement is a widely recognized guideline for conducting and reporting systematic reviews and meta-analyses. It provides a checklist of preferred reporting items to help authors improve the transparency and completeness of systematic review publications [9]. The methodology comprised several key stages, which are explained in the following sections.

We began with a comprehensive search strategy across multiple databases to identify relevant studies, followed by screening based on predefined inclusion and exclusion criteria. Next, the full texts were assessed for eligibility, ensuring only high-quality studies were included. Finally, the data was extracted, synthesized, and analyzed to draw the study's conclusions. This systematic approach helps minimize bias and ensures transparency, as recommended by the PRISMA 2020 guidelines [10].

2.1 Search strategy

This research was conducted in August 2024 with the aim of carrying out a systematic review on the tokenization of EHR and healthcare data. The review sought to identify and analyze the most current and relevant publications on this topic.

The review targeted English-language journal articles published between 2017 and 2024, ensuring the inclusion of the latest research in this rapidly evolving field. Three reputable databases were utilized for the literature search: Web of Science Core Collection [11], IEEEXplore [12], and Scopus [13].

The search terms were carefully crafted to capture articles addressing tokenization in the context of EHR, healthcare records, and related technologies. Keywords such as "Tokenization," "EHR," "EMR," "Healthcare data," "Clinical Data," "Patient Records," "NLP," "Natural Language Processing,"

"Text Mining," "Blockchain," "Machine Learning," and "AI" were incorporated into the final search query. Additionally, the search was limited to journal articles and review papers to ensure the inclusion of high-quality, peer-reviewed publications.

Figure 1 of the PRISMA flow diagram provides a comprehensive overview of the study selection process used in this research. We employed the two main approaches to find relevant studies - identification through databases and registers and identification through other methods.

The database and register search pathway initially yielded 40 results, all from the Scopus website, with no duplicates present. These papers underwent a preliminary selection process where their abstracts were analyzed, resulting in the exclusion of 21 articles that did not fully align with the search objectives. In the subsequent round, one article was excluded due to retrieval issues, leaving 19 papers for further consideration. The remaining articles were then screened in their entirety, leading to the exclusion of 5 papers that did not sufficiently focus on tokenization, the primary theme of this review.

The second pathway, identification of studies via other methods, produced 9 results, all of which were successfully retrieved. These articles were subjected to a comprehensive analysis to determine their relevance to tokenization and potential contribution to the review. This process led to the exclusion of 4 articles. The remaining 5 articles were then combined with the 13 papers from the first pathway, resulting in a total of 18 articles for in-depth review.

By employing these two complementary approaches, we significantly enhanced our ability to identify all relevant literature on the tokenization of electronic health records and healthcare data. This dual-pronged methodology strengthens the validity and comprehensiveness of the systematic review, ensuring that the final set of included studies provides a robust foundation for subsequent analysis and synthesis. The rigorous selection process across both pathways ensures that only the most pertinent and high-quality studies are included, thereby enhancing the overall quality and relevance of our review.

2.2 Results

This chapter presents a comprehensive analysis of the findings from our literature review on tokenization in healthcare data analysis. Our examination of the current research landscape reveals a multifaceted picture of tokenization's role, applications, and challenges in the healthcare domain. The results are organized into three main sections to provide a structured overview of the field: Data Analysis, Technological Advancements and Implementations and Challenges and Limitations of the studies.

Through this structured presentation, we aim to provide a clear and comprehensive picture of the state of tokenization in healthcare data analysis, setting the stage for the discussion that will follow.

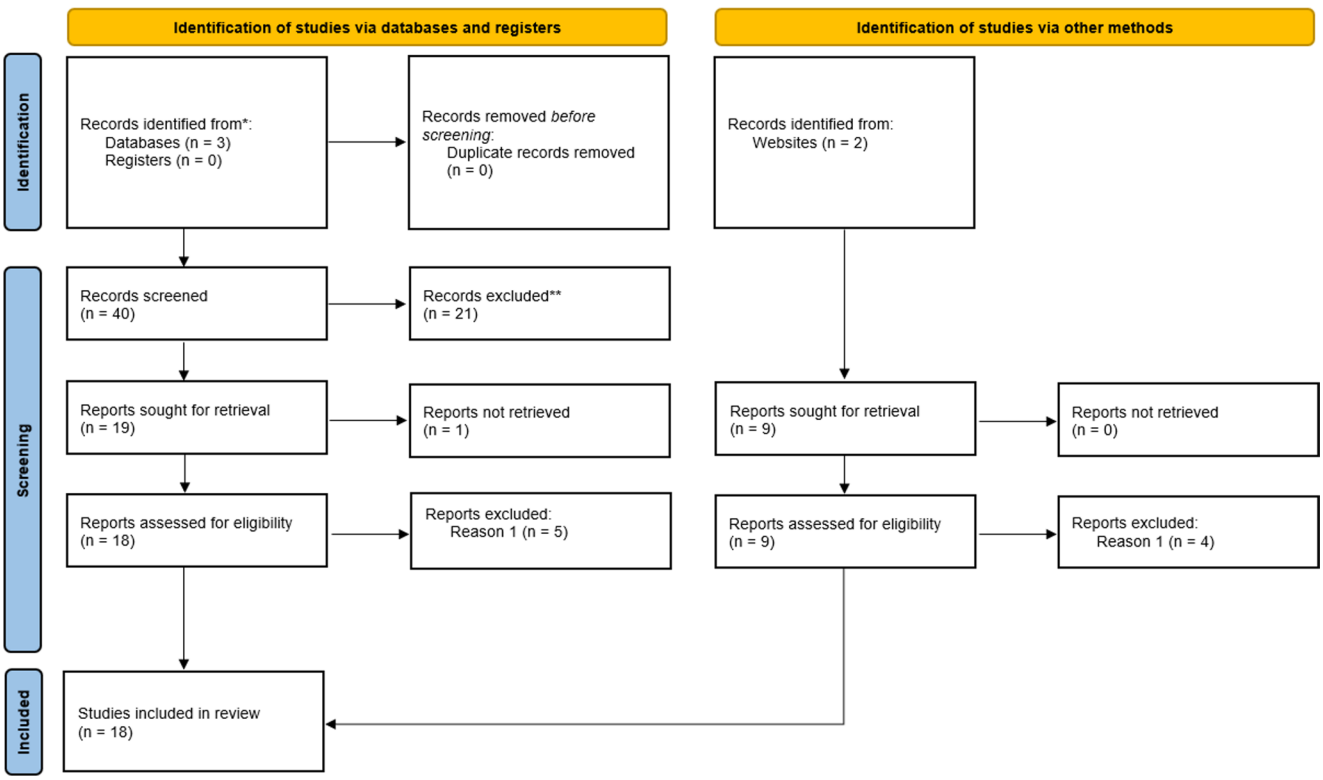


Fig. 1 PRIMSA workflow

2.3 Data analysis

Tokenization plays an important role across the reviewed articles, particularly in preprocessing EHR and clinical text for various NLP tasks. In many studies, tokenization is applied to transform unstructured text into structured data that can be analyzed using machine learning models. For example [14] and [15], use tokenization as an initial step in segmenting clinical text into tokens, facilitating further tasks like named entity recognition (NER) and predictive modeling. Both papers demonstrate how tokenization helps break down medical narratives and clinical notes into manageable units, allowing deep learning models to capture relevant patterns and improve the accuracy of the extraction and classification processes.

Tokenization also appears as a core preprocessing step in studies that focus on improving data quality and text normalization [16] and [17]. emphasize tokenization's role in handling noisy healthcare data, such as EHRs and clinical notes. Both studies focus on text normalization techniques, including tokenization, stop word removal, and punctuation filtering, to enhance data quality and improve machine learning model performance. The use of tokenization to standardize text, particularly in multilingual or complex medical terminologies, highlights its importance in reducing variability and improving NLP outputs across diverse healthcare settings.

Articles [18, 19] demonstrate the use of tokenization in conjunction with advanced NLP models, especially transformers and BERT-based models. Both utilize tokenization within transformer models, specifically through WordPiece and subword tokenization techniques. These techniques break down complex medical terminology into subwords, enabling the models to handle rare or out-of-vocabulary terms more effectively. Tokenization in these models ensures that medical terms are processed accurately, leading to improved outcomes in entity recognition and patient outcome predictions. Similarly, in [20], tokenization is crucial in transforming texts into manageable tokens that are used in large language models (LLMs) and multimodal LLMs for clinical decision-making.

In [21] it is presented a novel hierarchical framework for medical NLP, where tokenization is used as an initial step in the processing pipeline. The framework breaks down medical text into individual tokens that represent surface-level words, forming the basis for further semantic analysis. The paper proposes a hierarchical processing model inspired by human cognition, where tokens are transformed into logical representations of meaning. This hierarchical tokenization approach allows the system to progressively build higher-level interpretations from simple word tokens, ultimately enabling more complex understanding tasks like semantic memory and predictive coding. By organizing tokenization within a cognitive-inspired architecture, the framework facilitates more transparent and explainable NLP models, particularly in the medical domain.

In studies focused on patient data privacy and secure data sharing, tokenization is utilized to anonymize and protect sensitive information [22] and [23]. both highlight the use of tokenization to convert patient data into secure tokens, ensuring privacy during data exchanges across healthcare providers. In these cases, tokenization is used not just for data structuring, but also for generating encrypted tokens that safeguard patient privacy while enabling linkage across diverse datasets. This approach facilitates secure, large-scale analysis of healthcare data while ensuring compliance with data protection regulations.

Some articles [24, 25], use tokenization to deal with complex or irregular clinical data. The two of them discuss innovative methods of tokenizing continuous and discrete EHR data to improve the performance of machine learning models in predicting patient outcomes. These studies use tokenization to handle time-series data, converting it into sequences of tokens that can be fed into transformers or autoregressive models. By tokenizing both static and dynamic clinical measurements, these models can process data more effectively, improving their predictive power and interpretability in critical care settings.

Moreover, several studies use tokenization to enhance sentiment analysis and emotion recognition in healthcare contexts [26] and [27]. both employ tokenization to preprocess unstructured healthcare data, converting text into word tokens or vectors that can be analyzed for patterns related to emotions or behaviors. These tokenization strategies are essential for creating structured datasets from free-text medical notes, enabling accurate predictions in areas like mental health and smoking behavior classification.

Tokenization is also employed in domain-specific tasks such as medical entity extraction and disease diagnosis. In [28] and [29], tokenization is used to segment clinical texts and extract key medical entities such as disease diagnoses, treatment details, and biomarkers. These studies rely on tokenization to ensure that relevant medical terms are identified and processed efficiently, improving the accuracy of clinical entity recognition and information extraction tasks. Likewise, the study [30] emphasizes tokenization as a foundational step in preparing clinical notes for tasks like named entity recognition and relation extraction, further underlining its critical role in healthcare NLP pipelines.

In article [31], tokenization plays a crucial role in analyzing clinical discharge summaries for patient phenotyping, which involves predicting ten patient disorders. The study uses tokenization to segment text into word-level and sentence-level tokens, which are then processed through a convolutional

neural network (CNN) model. The authors highlight that only a small fraction of the tokens contributes significantly to the classification tasks, with token selection mechanisms employed to improve the model's interpretability and performance. This selective use of tokens enhances the explainability of the CNN's predictions, particularly for rare disorders like chronic pain, demonstrating the importance of token quality in patient classification from EHRs. Tokenization in this context enables a focused analysis of relevant linguistic features, which leads to moderate improvements in classification accuracy over traditional models. Table 1, resume this major findings.

Table 1 This table summarizes the key findings related to the role of tokenization in healthcare data analysis and natural Language processing (NLP)

Major Findings	Description	References
Tokenization for NLP Preprocessing	Tokenization is applied to convert unstructured EHR and clinical text into structured data for machine learning tasks like named entity recognition (NER) and predictive modeling.	[14, 15]
Text Normalization	Tokenization aids in improving data quality by handling noisy data, reducing variability, and enhancing machine learning model performance.	[16, 17]
Advanced Tokenization Techniques	Use of WordPiece and subword tokenization techniques in transformer and BERT-based models improves the handling of rare or out-of-vocabulary medical terms.	[18–20]
Hierarchical Tokenization Framework	Tokenization used in a hierarchical NLP model inspired by human cognition, enhancing explainability and semantic understanding.	[21]
Patient Data Privacy and Security	Tokenization helps anonymize sensitive patient data for secure sharing, enabling privacy while maintaining data utility across datasets.	[22, 23]
Tokenizing Time-Series and Irregular Data	Tokenization applied to continuous and discrete EHR data improves the performance of machine learning models in predicting patient outcomes, particularly in time-series analysis.	[24, 25]
Sentiment Analysis and Emotion Recognition	Tokenization converts unstructured healthcare data into word tokens or vectors for sentiment analysis and behavior prediction, aiding in mental health and smoking behavior classification.	[26, 27]
Medical Entity Extraction	Tokenization segments clinical text and extracts key medical entities (e.g., diagnoses, treatments), improving entity recognition and information extraction.	[28–30]
Tokenization for Patient Phenotyping	Tokenization in CNN models for clinical discharge summaries aids in patient classification, particularly for rare disorders, improving model interpretability.	[31]

2.4 Technological advancements and implementations

A prominent development in this field is the integration of tokenization with advanced deep learning models, particularly transformers and BERT-based architectures. For example [18] and [19], demonstrate how subword tokenization techniques, like WordPiece, are essential for breaking down complex medical terminology. These advancements enable models to handle rare or out-of-vocabulary terms, improving their performance in clinical entity recognition and predictive modeling. This shift towards transformers and BERT highlights the growing reliance on sophisticated tokenization methods to manage specialized healthcare texts.

Tokenization has also advanced in its role in generating secure, privacy-preserving datasets, in a way to secure data but also to anonymize it, enabling secure data exchanges across healthcare systems [22, 23]. These advancements ensure that sensitive patient information is protected while still allowing comprehensive data analysis across multiple platforms, showcasing the growing intersection between tokenization and blockchain technologies for secure healthcare applications.

Another key advancement is the application of tokenization in dynamic healthcare data, particularly time-series data. In [24] and [25], tokenization is used to handle continuous and discrete EHR data in real-time, facilitating accurate patient outcome predictions. These innovations in handling time-series data show the expanding utility of tokenization beyond static textual data, enhancing its role in real-time clinical decision support systems.

This is the major findings but on Table 2 we provide also criticisms of this approach.

This table provides a concise overview of the key criticisms of the findings from the literature review.

2.5 Challenges and limitations of the studies

Despite these advancements, several challenges and limitations related to tokenization persist across the reviewed literature. One of the major challenges is scalability. Studies such as [31] reveal that token selection strategies are crucial for improving performance in patient phenotyping models. However, the limited number of informative tokens and the need for larger sample sizes restrict the scalability of such models, especially in handling complex, unstructured clinical texts. Similarly, the need for more advanced architectures to fully exploit tokenization remains an issue, as demonstrated by the moderate improvements in deep learning models in this context.

Table 2 This table provides a concise overview of the key criticisms of the findings from the literature review

Criticism	Description
Limited Focus on Tokenization's Drawbacks	Tokenization can fragment or misinterpret complex medical terminology, especially multi-word expressions, leading to a loss of context, which is not adequately addressed in the literature.
Overreliance on Tokenization	Many studies emphasize tokenization as the core preprocessing step but overlook more advanced techniques (e.g., domain-specific embeddings) that could handle nuances in healthcare texts more effectively.
Scalability and Performance Issues	Tokenization's efficiency in processing large-scale, unstructured datasets like EHRs is a concern. Strategies for improving scalability, especially in real-time applications, are not sufficiently explored.
Lack of Token Interpretability	While tokenization enhances accuracy, the challenge of interpreting tokens in clinical settings is not thoroughly discussed. Subword tokenization techniques may obscure meaning, reducing transparency in healthcare applications.
Privacy Risks and Token Re-identification	Although tokenization helps anonymize data, the risks of re-identification in tokenized datasets are not adequately covered. More robust anonymization techniques, such as differential privacy, should be explored.
Tokenization's Role in Emerging Areas	The review does not explore tokenization's limitations in emerging areas like multimodal data analysis (combining text with images or clinical signals), where its effectiveness may vary significantly.

Integration issues with legacy systems also present significant challenges. Articles like [23] report difficulties in integrating tokenization frameworks within existing healthcare infrastructures, particularly when using blockchain for secure data exchanges. The complexity of implementing tokenization in a way that supports legacy systems while maintaining efficiency and security is an ongoing limitation that requires further research and development.

Another key limitation is the risk of re-identification, particularly when dealing with sensitive healthcare data. Although tokenization enhances privacy by anonymizing patient information, as seen in [22], there are still concerns about the robustness of these techniques in protecting against data re-identification attacks. Moreover, tokenization in time-series data models, such as those discussed in [25], faces challenges in maintaining patient privacy while ensuring that predictive models remain accurate and effective. Addressing these privacy concerns requires more advanced tokenization strategies, such as integrating differential privacy or other cryptographic methods.

Table 3 This table provides a concise overview of the main challenges and limitations found in the literature on tokenization in healthcare data analysis

Challenge/Limitations	Description
Scalability	Token selection strategies are crucial for patient phenotyping, but limited informative tokens and small sample sizes restrict scalability in handling complex clinical data.
Integration with Legacy Systems	Integrating tokenization frameworks into existing healthcare infrastructures, especially with blockchain, poses difficulties in maintaining efficiency and security.
Re-identification Risk	While tokenization enhances privacy, concerns remain about re-identification attacks, especially in time-series data models. Further privacy-preserving techniques are needed.
Performance Bottlenecks	Tokenization improves accuracy in tasks like emotion recognition but leads to computational delays and inefficiencies, especially with large datasets in resource-limited environments.

Finally, performance bottlenecks remain a challenge in tokenization workflows. For instance, in [26], although tokenization significantly improved accuracy in emotion recognition, the computational demands associated with processing large healthcare datasets led to delays and inefficiencies. These bottlenecks suggest that while tokenization improves accuracy, its computational complexity can limit its real-world application in resource-constrained environments. The Table 3, summary table of the findings from the literature review:

3 Discussion

The rapid advancement of digital technologies in healthcare has ushered in an era of unprecedented data generation and analysis. Within this landscape, tokenization has emerged as a critical tool for processing and securing healthcare data. This discussion chapter aims to synthesize the findings of our comprehensive literature review, offering a critical examination of the current state, challenges, and future prospects of tokenization in healthcare data analysis.

Our analysis reveals a complex picture of tokenization's role in healthcare informatics. From its foundational use in natural language processing to its innovative applications in privacy preservation and real-time data analysis, tokenization has demonstrated remarkable versatility. However, this review also uncovers persistent challenges that hinder its full potential. In the following sections, we will delve into the principal findings, explore future directions for research and development, acknowledge the limitations of our study, and draw conclusions about the trajectory of tokenization in healthcare.

By critically examining these aspects, we aim to provide researchers, healthcare professionals, and policymakers with a nuanced understanding of tokenization's impact on healthcare data analysis. This discussion not only highlights the transformative potential of tokenization but also underscores the need for continued innovation to address existing limitations and harness its full capabilities in improving patient care and advancing medical research.

3.1 Principal findings

Tokenization has emerged as a fundamental preprocessing step in healthcare NLP, serving as a bridge between unstructured clinical text and structured data suitable for machine learning analysis. Its widespread application across various studies underscores its importance in transforming complex medical narratives into manageable units. This transformation is crucial for subsequent tasks such as named entity recognition and predictive modeling, ultimately enhancing the accuracy of information extraction and classification in healthcare contexts.

The integration of tokenization with advanced NLP models, particularly transformers and BERT-based architectures, represents a significant leap forward in handling specialized medical terminology. The adoption of subword tokenization techniques, such as WordPiece, has markedly improved the processing of rare or out-of-vocabulary terms, addressing a longstanding challenge in medical NLP. This advancement has far-reaching implications for improving the accuracy of clinical entity recognition and patient outcome predictions, potentially revolutionizing clinical decision support systems.

A noteworthy finding is the dual role of tokenization in data privacy and secure sharing within healthcare systems. By converting sensitive patient information into secure tokens, tokenization facilitates anonymization while maintaining data utility. This approach enables large-scale analysis of healthcare data across institutions while ensuring compliance with stringent data protection regulations. The synergy between tokenization and blockchain technologies for secure healthcare applications represents a promising avenue for balancing data accessibility with privacy concerns in collaborative research efforts.

The application of tokenization to time-series and multimodal healthcare data marks a significant expansion of its utility beyond traditional text analysis. Innovative methods for tokenizing continuous and discrete EHR data have enhanced the performance of machine learning models in predicting patient outcomes, particularly in critical care settings. This evolution demonstrates tokenization's adaptability to diverse data types, paving the way for more comprehensive and real-time patient monitoring systems.

Performance limitations in processing large healthcare datasets present another significant challenge. While tokenization improves accuracy in tasks such as emotion recognition and sentiment analysis, the computational demands can lead to inefficiencies in resource-constrained environments. This constraint underscores the need for optimized tokenization algorithms and enhanced computational resources to facilitate real-world applications in healthcare settings.

3.2 Future directions

The future directions of tokenization in healthcare data present a broad range of opportunities for enhancing both the security and efficiency of data processing models. Many studies highlight the need for larger datasets to improve model performance, particularly in patient phenotyping and disease prediction. Increasing the size and diversity of EHR datasets would allow tokenization strategies to capture a broader array of clinical nuances, especially for rare phenotypes and underrepresented medical conditions. This expansion could lead to more robust models that are better suited for personalized medicine and precision healthcare applications.

In terms of technological advancement, there is a growing interest in integrating tokenization with cutting-edge machine learning models. The application of neuro-fuzzy theory, transformer-based architectures, and advanced deep learning techniques is a key area for future exploration. Researchers plan to combine tokenization with neuro-fuzzy systems and state-of-the-art language models to reduce dependency on expert knowledge and enhance model flexibility. Additionally, pre-training models with larger, more diverse datasets and employing self-supervised learning could further improve tokenization's effectiveness in handling diverse clinical data formats.

Privacy and security concerns remain a top priority for future research. To address the risks of data re-identification and improve the protection of sensitive patient information, future work will focus on integrating more advanced privacy-preserving techniques, such as differential privacy and cryptographic methods, into tokenization frameworks. This is particularly relevant for large-scale data exchanges across healthcare providers, where maintaining privacy while enabling meaningful data analysis is critical. Further, enhancing the security of blockchain-based tokenization methods to support dynamic and adaptive data exchanges is another focus area. This includes the development of more flexible public and private chain integrations to optimize security in real-time data exchanges.

Tokenization's role in time-series data and real-time healthcare applications will also evolve. Future research will seek to optimize tokenization processes for continuous and dynamic healthcare data, such as monitoring patient vitals in intensive care units. Enhancing the scalability and adaptability of tokenization in these settings will ensure models can

handle fluctuating data with minimal preprocessing, improving real-time clinical decision-making. Additionally, efforts will focus on reducing computational bottlenecks in tokenization methods to ensure their practical applicability in resource-limited environments, particularly in mobile and remote healthcare solutions.

Another promising area of future exploration is the integration of tokenization with domain-specific knowledge and ontologies. Researchers aim to incorporate more specialized medical knowledge into tokenization processes, which could improve the accuracy of models in understanding complex medical language and context. By leveraging medical ontologies and refining existing genomic signatures, tokenization could be adapted to provide more precise insights into disease progression and treatment outcomes.

There is also a push towards enhancing the transparency and explainability of tokenization models. Future work will aim to make tokenization-based models more interpretable for clinicians, ensuring that healthcare professionals can understand and trust the outputs of machine learning models. This will involve refining token selection mechanisms and developing frameworks that provide clear justifications for model decisions, making tokenization a more effective tool for clinical practice. In Table 4, we have a summary table of the future directions for tokenization in healthcare data.

Table 4 This table outlines the promising future directions for research and application of tokenization in the healthcare sector

Future Direction	Description
Larger and Diverse Datasets	Expanding EHR datasets to improve model performance, particularly for rare conditions and personalized medicine, enhancing tokenization's ability to capture clinical nuances.
Technological Integration	Exploring the combination of tokenization with neuro-fuzzy systems, transformer-based architectures, and advanced deep learning models to enhance flexibility and reduce expert dependency.
Privacy and Security Enhancements	Integrating advanced privacy-preserving techniques like differential privacy and cryptographic methods to mitigate re-identification risks and secure large-scale data exchanges.
Optimization for Time-Series Data	Adapting tokenization processes for continuous healthcare data (e.g., patient vitals) to improve scalability and real-time decision-making in clinical settings.
Domain-Specific Knowledge Integration	Incorporating specialized medical knowledge and ontologies into tokenization processes to improve understanding of complex medical language and context.
Transparency and Explainability	Enhancing interpretability of tokenization models for clinicians, refining token selection mechanisms, and developing frameworks for clearer model decision justifications.

3.3 Limitations

The search for relevant articles on tokenization in healthcare presented several challenges, primarily due to the niche nature of this specific application. Finding a broad and diverse range of studies, especially those published before 2020, proved difficult. This limitation resulted in a relatively small pool of directly relevant resources, necessitating a more focused search strategy and a greater reliance on recent publications. The scarcity of studies specifically addressing tokenization in healthcare often required expanding the search beyond the initial scope to include related fields or more general applications of tokenization. While this limitation highlights a potential gap in the literature, it also presents an opportunity for future research to explore this understudied area more thoroughly.

The interdisciplinary nature of tokenization in healthcare added another layer of complexity to the search process. This topic intersects healthcare, data security, and natural language processing (NLP), each with its own distinct body of literature, methodologies, and terminologies. Synthesizing insights from these diverse fields required a comprehensive understanding of technical concepts across multiple disciplines. Many articles that employed tokenization techniques did not focus explicitly on the process itself, making it necessary to carefully examine a wider range of studies to extract relevant information. This interdisciplinary complexity, while challenging in the search and synthesis process, also underscores the potential for combining ideas and techniques across these fields in future research.

4 Conclusion

The literature review clearly underscores the critical and evolving role of tokenization in healthcare data analysis. This technique has demonstrated substantial potential in enhancing data security, improving natural language processing (NLP) capabilities, and facilitating more accurate and efficient analysis of electronic health records (EHR). Tokenization has proven effective in addressing privacy concerns by converting sensitive patient information into secure tokens, which allows for anonymization without sacrificing data utility. As a result, it enables large-scale healthcare data analysis across institutions while ensuring compliance with stringent data protection regulations.

However, challenges remain. The scalability of tokenization in handling large, unstructured clinical datasets presents an ongoing issue, highlighting the need for more efficient token selection strategies. Integrating tokenization frameworks with legacy healthcare systems continues to be a significant obstacle, emphasizing the necessity for flexible,

interoperable solutions. Additionally, privacy concerns persist, particularly the risk of re-identification in tokenized datasets. This underscores the need for more advanced anonymization techniques that can withstand sophisticated attacks while preserving the integrity and utility of the data.

Looking ahead, the future of tokenization in healthcare data analysis is promising. Research efforts should focus on developing more robust privacy-preserving techniques, optimizing tokenization for real-time clinical applications, and incorporating domain-specific medical knowledge into tokenization processes. Enhancing the interpretability and explainability of tokenization-based models is also crucial to increasing their acceptance and utility in clinical practice.

In conclusion, while tokenization has significantly advanced healthcare data security and analysis, its full potential has yet to be realized. Addressing current challenges and exploring future research directions will be essential for widespread adoption and effectiveness in real-world healthcare applications. As the healthcare sector continues to digitize and generate ever-increasing volumes of data, tokenization will play an increasingly vital role in unlocking the value of this information, ultimately contributing to improved patient care and advancing medical research.

Author contributions Conceptualization, M.M.; methodology, L.B.E.; investigation, J.C.F.; writing—original draft preparation, J.C.F. and L.B.E.; writing—review and editing, R.C. and M.M.; funding acquisition, J.C.F. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by Molde University College - Specialized University in Logistics. This work is supported by the Blockchain project. PT (PRR–RE-C05-i01.02: AGENDAS/ALIANÇAS VERDES PARA A INOVAÇÃO EMPRESARIAL).

Data availability No data is used to share.

Code availability No Code is available.

Declarations

Conflicts of interest No Conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. What is Tokenization?. Security. Accessed: May 30, 2024. [Online]. Available: https://www.techtarget.com/searchsecurity/definition/tokenization
2. Vagadia B. Data integrity, control and tokenization, in *Digital disruption*. Future of business and finance. Cham: Springer International Publishing; 2020. pp. 107–76. https://doi.org/10.1007/978-3-030-54494-2_5.
3. Mudumbai SC, et al. Public health informatics and the perioperative physician: looking to the future. Anesth Analg. 2024;138(2):253–72.
4. Cost of a data. breach| IBM. 2024. Accessed: Sep. 09, 2024. [Online]. Available: https://www.ibm.com/reports/data-breach
5. Taitsman JK, Grimm CM, Agrawal S. Protecting patient privacy and data security. N Engl J Med. 2013;368(11):977–9. https://doi.org/10.1056/NEJMp1215258.
6. Linda Dimitropoulos P, Patel V, Scott M, Scheffler A, Steve Posnack MHS. Public attitudes toward health information exchange: Perceived Benefits and Concerns. 2011;17. Accessed: Sep. 09, 2024. [Online]. Available: https://www.ajmc.com/view/public-attitudes-toward-health-information-exchange-perceived-benefits-and-concerns
7. HealthManagement.org. Radiology Management. ICU management, healthcare IT, cardiology management, executive management. Accessed: Sep. 09, 2024. [Online]. Available: https://healthmanagement.org/c/hospital/issuearticle/blockchain-solves-healthcare-data-obstacles
8. Lordick T, Hoch A, Fransen B. Anonymization of electronic health care records: The EHR anonymizer. In Computational Life Sciences. Dörpinghaus J, Weil V, Schaaf S, Apke A, Eds. In Studies in Big Data, Cham: Springer International Publishing, 2022; vol. 112, pp. 485–499. https://doi.org/10.1007/978-3-031-08411-9_18
9. Liberati A, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700. https://doi.org/10.1136/bmj.b2700.
10. Page MJ, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021;10(1). https://doi.org/10.1186/s13643-021-01626-4.
11. Web of Science Core Collection, Clarivate. Accessed. Sep. 09, 2024. [Online]. Available: https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/web-of-science-core-collection/
12. IEEE Xplore. Accessed: Sep. 09, 2024. [Online]. Available: https://ieeexplore.ieee.org/Xplore/home.jsp
13. Scopus . Advanced search | Signed in. Accessed: Sep. 09, 2024. [Online]. Available: https://www.scopus.com/search/form.uri?display=advanced
14. Chng SY, Tern PJW, Kan MRX, Cheng LTE. Automated labelling of radiology reports using natural Language processing: comparison of traditional and newer methods. Health Care Sci. 2023;2(2):120–8. https://doi.org/10.1002/hcs2.40.
15. Pabón OS, Torrente M, Provencio M, Rodríguez-Gonzalez A, Menasalvas E. Integrating speculation detection and deep learning to extract lung cancer diagnosis from clinical notes. Appl Sci Switz. 2021;11(2):1–21. https://doi.org/10.3390/app11020865.
16. Nesca M, Katz A, Leung CK, Lix LM. A scoping review of preprocessing methods for unstructured text data to assess data quality. Int J Popul Data Sci. 2022;7(1). https://doi.org/10.23889/ijpds.v7i1.1757.
17. Akhtyamova L, Martínez P, Verspoor K, Cardiff J. Testing contextualized word embeddings to improve NER in Spanish clinical case narratives. IEEE Access. 2020;8:164717–26. https://doi.org/10.1109/ACCESS.2020.3018688.

18. Kim Y-M, Lee T-H. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inf Decis Mak. 2020;20. https://doi.org/10.1186/s12911-020-01241-8.
19. Mugisha C, Paik I. Bridging the gap between medical tabular data and NLP predictive models: a Fuzzy-Logic-Based textualization approach. Electron Switz. 2023;12(8). https://doi.org/10.3390/electronics12081848.
20. Guevara M, et al. Large Language models to identify social determinants of health in electronic health records. Npj Digit Med. 2024;7(1). https://doi.org/10.1038/s41746-023-00970-0.
21. Taira RK, Garlid AO, Speier W. Design considerations for a hierarchical semantic compositional framework for medical natural language understanding. PLoS ONE. 2023;18(3). https://doi.org/10.1371/journal.pone.0282882
22. Leapman MS, et al. Development of a longitudinal prostate cancer transcriptomic and clinical data linkage. JAMA Netw Open. 2024;7(6). https://doi.org/10.1001/jamanetworkopen.2024.17274.
23. Zhou T, Li X, Zhao H. Med-PPPHIS: Blockchain-Based personal healthcare information system for National physique monitoring and scientific exercise guiding. J Med Syst. 2019;43(9). https://doi.org/10.1007/s10916-019-1430-2.
24. Shickel B, et al. Multi-dimensional patient acuity estimation with longitudinal EHR tokenization and flexible transformer networks. Front Digit Health. 2022;4. https://doi.org/10.3389/fdgth.2022.1029191.
25. Lee Y, Chae Y, Jung K. Leveraging VQ-VAE tokenization for autoregressive modeling of medical time series. Artif Intell Med. 2024;154. https://doi.org/10.1016/j.artmed.2024.102925.
26. Lal M, Neduncheliyan S. Enhanced V-Net approach for the emotion recognition and sentiment analysis in the healthcare data. Multimed Tools Appl. 2024. https://doi.org/10.1007/s11042-024-18364-z.
27. Hossain E, et al. Natural Language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med. 2023;155:106649.
28. Mora S, et al. Towards the automatic calculation of the EQUAL Candida score: extraction of CVC-related information from EMRs of critically ill patients with candidemia in intensive care units. J Biomed Inf. 2024;156. https://doi.org/10.1016/j.jbi.2024.104667.
29. Solarte-Pabón O, et al. Transformers for extracting breast cancer information from Spanish clinical narratives. Artif Intell Med. 2023;143. https://doi.org/10.1016/j.artmed.2023.102625.
30. Kondra S, Xu W, Raghavan V. Unlocking the power of clinical notes: natural Language processing in healthcare. Acta Sci Med Sci. May 2024;8:37–44. https://doi.org/10.31080/ASMS.2024.08.1821.
31. Yang Z, Dehmer M, Yli-Harja O, Emmert-Streib F. Combining deep learning with token selection for patient phenotyping from electronic health records. Sci Rep. 2020;10(1):1432. https://doi.org/10.1038/s41598-020-58178-1.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.