SBC Computing Reviews, 2026, 05:1,
ISSN: 2966-3938 • doi: 10.5753/reviews.2026.6737
This work is licensed under a Creative Commons Attribution 4.0 International License.

SCOPING REVIEW

Harnessing Big Data in Healthcare: A Scoping Review of Predictive Analytics, Interoperability, and Security Challenges

André E. Gonçalves [Universidade Politécnica de Castelo Branco | *andre.goncalves1@ipcbcampus.pt*]
Bruno A. M. G. J. de Sousa [Universidade Politécnica de Castelo Branco | *bruno.sousa1@ipcbcampus.pt*]
Francisco M. Gonçalves [Universidade Politécnica de Castelo Branco | *gfrancisco@ipcbcampus.pt*]
Eurico Lopes [Universidade Politécnica de Castelo Branco | *eurico@ipcb.pt*]
Luís Miguel Barata [Universidade Politécnica de Castelo Branco / Instituto de Telecomunicações / NOVA Laboratory for Computer Science and Informatics | *luis.barata@ipcb.pt*]

Escola Superior de Tecnologia, Universidade Politécnica de Castelo Branco, Castelo Branco, Portugal.

Abstract. The increasing availability of electronic health records, connected medical devices, and large-scale clinical datasets has expanded the potential use of Big Data in healthcare. This focused scoping review examines evidence across three interconnected domains: health information management and interoperability, predictive analytics, and data security and privacy. Following PRISMA-ScR, a title-only search was conducted in Scopus for the period from 2013 to 2024, retrieving 124 records. The formal selection process was subsequently restricted to publications issued between 2019 and 2024, resulting in twelve included studies. The findings indicate potential improvements in hospital interoperability, disease prediction, resource planning, and privacy-preserving data management. However, system heterogeneity, limited standardization, incomplete clinical metric reporting, infrastructure requirements, and algorithmic bias remain important barriers. Given the use of a single database, a restrictive title-only strategy, and a small heterogeneous corpus, the findings should be interpreted as a focused mapping rather than an exhaustive representation of the field.

Keywords: Big Data, Healthcare, Scoping Review, Predictive Analytics, Interoperability, Data Privacy

Received: 22 October 2025 • **Accepted:** 06 September 2026 • **Published:** 23 November 2026

1 Introduction

The healthcare sector is currently undergoing a radical digital transformation, fueled by the massive adoption of Electronic Health Records (EHRs), medical Internet of Things (mIoT) devices, and high-throughput genomic sequencing. This shift has transitioned medicine from a traditional, reactive model to a data-driven, proactive paradigm. Central to this evolution is Big Data—defined as the storage, processing, and management of datasets so vast and complex that conventional software is inadequate for their analysis [Ristevski and Chen, 2018].

While Douglas Laney's original framework characterized Big Data by the three Vs: *Volume* (the scale of data), *Velocity* (the speed of generation), and *Variety* (diversity of sources) [Laney *et al*., 2001], the healthcare context necessitates the inclusion of two additional dimensions: *Veracity* and *Value*. *Veracity* refers to the reliability and quality of clinical data, which is often unstructured or noisy, while *Value* represents the ultimate goal of turning raw data into actionable clinical insights that improve patient outcomes and reduce hospital costs.

Despite the growing body of literature, the implementation of Big Data in clinical settings faces multi-layered obstacles. Technical challenges such as high data dimensionality [Wang *et al*., 2022] and the lack of semantic interoperability between heterogeneous sources [Chen *et al*., 2024] hinder the seamless flow of information. Simultaneously, ethical and social challenges, such as algorithmic bias and the risk of perpetuating healthcare disparities, remain critical concerns, as highlighted by the landmark study by Obermeyer et al. [Obermeyer *et al*., 2019], which revealed how biased training data can underestimate the care needs of marginalized populations.

1.1 Related Work and Rationale for the Review

Previous reviews and surveys have examined the use of Big Data and Artificial Intelligence (AI) in healthcare from different perspectives. For example, Ristevski and Chen [Ristevski and Chen, 2018] discuss the main characteristics, opportunities, and challenges of Big Data analytics in medicine, whereas Chattu et al. [Chattu, 2021] and Kaur et al. [Kaur *et al*., 2021] examine applications involving AI, Big Data, and blockchain technologies. Other contributions focus more specifically on predictive analytics, healthcare data infrastructures, interoperability, privacy, or security.

Within the literature identified during the preliminary examination of the topic, these dimensions are frequently discussed separately or with different levels of emphasis. Consequently, there remains value in examining how information management and interoperability, predictive analytics, and data security and privacy are represented and interconnected within a common analytical framework. This focused scoping review addresses that need by organizing and critically examining the evidence captured by the defined Scopus search strategy. Particular attention is given to the trade-offs represented in the included studies, including those between interoperability and privacy, data standardization and predictive performance, and technological capability and clinical applicability.

The review also examines the reporting of clinically relevant performance measures beyond overall accuracy, in-

cluding sensitivity and F1-score. However, because the search was restricted to the Scopus database, applied exclusively to publication titles, and produced a small and heterogeneous corpus, the study does not aim to provide an exhaustive assessment of the healthcare Big Data literature.

1.2 Objective and Contribution

This scoping review aims to provide a focused mapping of the Big Data applications in healthcare represented in the literature retrieved through the defined Scopus search strategy. The formal evidence-mapping process covers publications issued between 2019 and 2024, while records retrieved from 2013 to 2024 are used exclusively to contextualize publication activity.

The review examines three interconnected domains: health information management and interoperability, predictive analytics, and data security and privacy. Its contribution lies in organizing and critically examining the evidence represented in the retrieved corpus, with particular attention to the trade-offs between interoperability, predictive performance, privacy, infrastructure requirements, and algorithmic fairness. It also examines the reporting of clinically relevant performance measures beyond overall accuracy.

Given the use of a single database, a title-only search strategy, and a small heterogeneous corpus, the study is presented as a focused and exploratory mapping rather than as a comprehensive representation of the healthcare Big Data field.

2 Methodology

To ensure methodological transparency, this scoping review was reported in accordance with the PRISMA Extension for Scoping Reviews (PRISMA-ScR) guidelines [Tricco *et al.*, 2018]. The scoping-review approach was selected to support an exploratory mapping of the evidence represented within the retrieved corpus and to address the defined research questions.

2.1 Research Questions

The study is guided by the following research questions, designed to provide a focused overview of the evidence captured by the search strategy:

(RQ1) What are the main applications of Big Data in the healthcare sector?

(RQ2) In what ways has Big Data analysis contributed to clinical and administrative healthcare outcomes?

(RQ3) What are the primary challenges, ethical concerns, and technical limitations in the implementation of Big Data in healthcare?

2.2 Eligibility Criteria

Studies were selected according to the following inclusion criteria: (1) peer-reviewed publications written in English; (2) publication between 2019 and 2024; (3) full-text availability; and (4) original research or structured reviews focusing on Big Data applications in healthcare.

The following exclusion criteria were applied: (E1) studies not specifically related to the healthcare sector; (E2) publications in which Big Data was not a central component of the application or analysis; (E3) purely theoretical works without an empirical application, proof-of-concept, or structured review methodology; and (E4) editorials, conference abstracts, book reviews, or non-peer-reviewed white papers.

2.3 Literature Search and Strategy

The literature search was conducted in the Scopus database, selected for its multidisciplinary coverage of healthcare and computing-related research. The search combined the terms “Big Data”, “Health”, and “Application”, applied exclusively to the article-title field.

```
TITLE("Big Data" AND "Health" AND "Application") AND PUBYEAR > 2012 AND PUBYEAR < 2025
```

The initial search covered the period from 2013 to 2024 and retrieved 124 records. This broader retrieval set was used to describe the evolution of publication activity over time. For the formal scoping-review selection process, the eligibility period was subsequently restricted to publications issued between 2019 and 2024, resulting in 59 records. This restriction was adopted to focus the main evidence mapping on recent developments in artificial intelligence, blockchain, interoperability, and healthcare data infrastructures.

Although the title-only strategy increased thematic specificity, it also reduced search sensitivity. Relevant publications may not include all three concepts in their titles, even when Big Data applications in healthcare are discussed in their abstracts, keywords, or full texts. The use of the term “Application” may also have excluded studies using related terms such as “implementation”, “platform”, “framework”, “system”, or “deployment”. Therefore, the retrieved corpus should be interpreted as a focused sample rather than an exhaustive representation of the field.

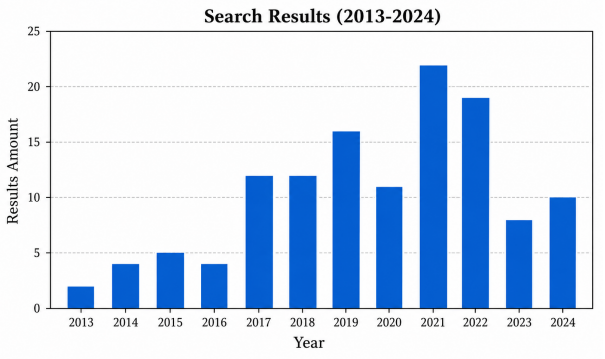

Figure 1. Annual distribution of the 124 Scopus records retrieved through the title-only search between 2013 and 2024. The formal scoping-review selection was subsequently restricted to publications issued between 2019 and 2024.

Figure 1 shows the annual distribution of the 124 records retrieved between 2013 and 2024. Publication activity increased after 2016, reached its highest level in 2021, and declined during the final years of the period. These values are presented for descriptive contextualization and should not be interpreted as evidence of a change in the research focus of the field.

2.4 Study Selection and Data Charting

The study selection process was conducted in sequential stages. First, the retrieved records were assessed based on title relevance and availability. The remaining records

were subsequently screened through abstract examination, followed by full-text assessment against the predefined eligibility criteria.

The first three authors independently conducted the screening process. Disagreements regarding study eligibility were resolved through discussion among them. The two supervising authors provided methodological guidance throughout the review process and supported the final revision of the manuscript.

A standardized data-charting form was developed before data extraction. For each included study, the following information was recorded: bibliographic reference, publication year, geographical context, study type, healthcare domain, Big Data application, technological approach, reported benefits, performance metrics, interoperability aspects, security and privacy mechanisms, ethical concerns, and implementation limitations.

The first three authors performed the data extraction and charting. Any discrepancies identified during this process were resolved through discussion among them.

The main data-charting process was restricted to the twelve studies directly included through the formal selection procedure. In addition, primary studies reported within included review articles were examined only when necessary to illustrate specific reporting practices. Such indirectly identified studies were not treated as formally included evidence and were clearly distinguished from the main review corpus.

Given the exploratory purpose of the scoping review and the heterogeneity of the included evidence, the findings were synthesized descriptively and organized into three analytical domains: health information management and interoperability, predictive analytics, and data security and privacy.

2.5 Study Characteristics

Data extraction focused on geographic location, core themes, and technological focus. Table 1 summarizes the twelve studies directly included through the formal scoping-review selection process. Review and survey articles are classified as secondary evidence. Primary studies cited within those reviews were not treated as independently included evidence and their performance values were therefore not incorporated into the direct comparative synthesis.

3 Results

The initial Scopus search identified 124 records published between 2013 and 2024. After restricting the formal eligibility period to 2019–2024, 59 records remained for the scoping-review selection process. Sixteen reports could not be retrieved, leaving 43 records for title screening. Ten records were excluded after title assessment and 13 after abstract screening. Twenty reports underwent full-text assessment, of which eight were excluded for not meeting the predefined eligibility criteria. The final scoping-review corpus comprised 12 studies.

Figure 2 presents the PRISMA-ScR flow diagram for the study-selection process.

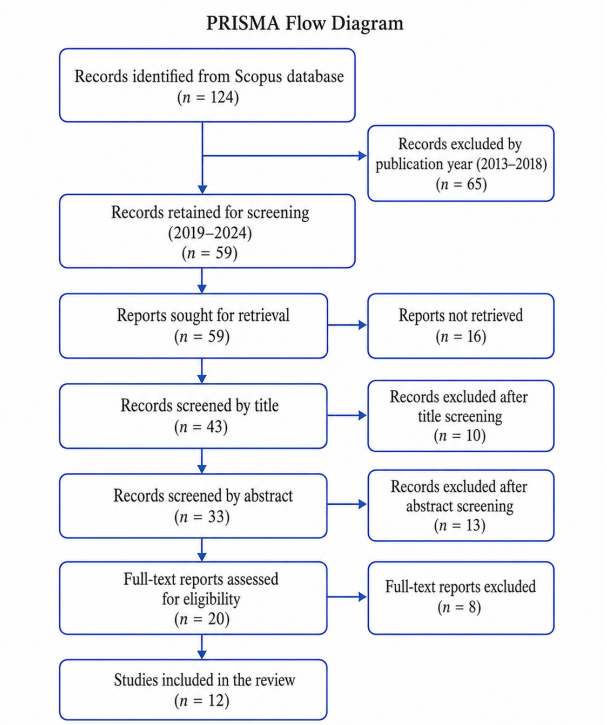

Figure 2. PRISMA-ScR flow diagram of the scoping selection process.

3.1 Mapping Health Information Management

For analytical purposes, this review introduces the author-defined terms “horizontal platform” and “vertical platform”. A horizontal platform refers to a general-purpose infrastructure integrating data across multiple clinical domains, departments, or institutional systems. A vertical platform refers to a disease-specific or specialty-specific infrastructure designed to provide greater clinical detail within a narrower domain. These terms are used as interpretive categories and do not represent an established taxonomy in the literature.

For the purpose of the present synthesis, the healthcare data platforms described in the included studies were grouped into two author-defined analytical categories: horizontal and vertical platforms. On one hand, studies such as Wang et al. [Wang *et al.*, 2022] and Sun et al. [Sun *et al.*, 2022] focus on general-purpose hospital platforms. Wang’s implementation at West China Hospital stands out for its scale, utilizing FHIR standards to govern over 1,400 variables across 13 health domains. This ”horizontal” approach prioritizes hospital-wide interoperability and high-performance computing (1,900 TFLOPS) to support diverse AI tools simultaneously.

Conversely, Chen et al. [Chen *et al.*, 2024] demonstrate a disease-specific approach by constructing a platform dedicated exclusively to nasopharyngeal carcinoma. While smaller in scope (15 EHR systems), this ”vertical” approach allows for deeper clinical granularity, such as patient timeline visualizations that are often lost in general management systems.

A critical comparison suggests a clear trade-off: general platforms (Wang) excel in operational efficiency and institutional scaling, whereas specialized platforms (Chen) provide superior support for longitudinal oncological research. Both,

Table 1. Summary of the publications considered: title and reference, country of origin, and main topics.

Title and Reference	Location	Topic
A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health [Chattu, 2021]	India	Predictive models and data privacy
Big data health care platform with multisource heterogeneous data integration and massive high-dimensional data governance for large hospitals [Wang *et al*., 2022]	China	Hospital information management
Risks and opportunities to ensure equity in the application of big data research in public health [Wesson *et al*., 2022]	United States	Predictive models and Ethics
Development of a predictive model for integrated medical and long-term care resource consumption based on circulatory diseases [Takura *et al*., 2021]	Japan	Predictive models
A comprehensive survey of AI, blockchain technology and big data applications in medical field and global health [Kaur *et al*., 2021]	India	Data privacy
A survey of data element perspective: Application of artificial intelligence in health big data [Xiong *et al*., 2022]	China	Predictive models
Analysis on the application of big data technology in medical and health industry [Qinghua and Weiguang, 2021]	China	Hospital information management
Construction and application of nasopharyngeal carcinoma-specific big data platform based on electronic health records [Chen *et al*., 2024]	China	Hospital information management
India's national family health survey datasets: exploration of big data applications toward evidence-informed public health [Talati *et al*., 2022]	India	Predictive models
Application of multimedia technology in health service management system under the background of big data fusion [Sun *et al*., 2022]	India	Hospital information management
Application of AI Technology Based on Medical Big Data in Disease Monitoring, Early Warning, and Prevention [Liu *et al*., 2024]	China	Predictive models
Research on the Application of Health Big Data based on Semi-Supervised Learning Algorithm [Ma, 2024]	India	Predictive models

however, face the persistent barrier of "data silos" and the lack of universal semantic standardization, as also noted by Sun et al. [Sun *et al*., 2022].

3.2 Predictive Models and Performance Metrics

Predictive modelling was addressed both by primary studies included in the formal scoping-review corpus and by secondary review articles. To illustrate the reporting of predictive performance metrics, Table 2 summarizes selected primary studies reported within the included reviews by Chattu et al. [Chattu, 2021] and Xiong et al. [Xiong *et al*., 2022].

These primary studies were not independently retrieved, screened, or included through the present scoping-review selection process. Therefore, they are presented exclusively as indirect evidence derived from included secondary sources and should not be interpreted as part of the twelve-study corpus.

The table is used to illustrate recurring metric-reporting patterns, particularly the frequent emphasis on Accuracy and the omission of Sensitivity and F1-score.

The ASHRO model by Takura et al. [Takura *et al*., 2021] provides a rare example of linking predictive accuracy to clinical outcomes, showing that high-adherence patients (correctly identified by the model) had significantly lower 36-month mortality (2% vs 7%). Similarly, Ma et al. [Ma, 2024] emphasize that for tumor detection, the F1-score (93.1%) is a more critical validator than accuracy alone, as it accounts for false negatives that could lead to missed diagnoses.

3.3 Data Security and Blockchain Integration

Security mapping identifies Blockchain as the primary emerging solution for the "privacy vs. accessibility" paradox. Supriya et al. [Chattu, 2021] and Kaur et al. [Kaur *et al*., 2021] both identify blockchain as a catalyst for decentralized data integrity. By utilizing asymmetric cryptography and purpose-based consent models (like Hyperledger Fabric), healthcare systems can move away from centralized vulnerabilities. However, the literature reveals a gap between theoretical security and practical deployment; high computational costs and the difficulty of integrating with legacy EHR systems remain the primary "real-world" implementation barriers.

4 Discussion

The comparative analysis of the selected studies provides a focused mapping of the healthcare Big Data applications represented within the retrieved corpus. This section synthesizes the evidence to address the proposed research questions through a critical lens.

4.1 Architectural Trade-offs in Management (RQ1)

The mapping of health information management reveals a fundamental tension between system universality and clinical specialization. Wang et al. [Wang *et al*., 2022] demonstrate the success of "horizontal" platforms based on the international FHIR (Fast Healthcare Interoperability Re-

Table 2. Predictive Models Discussed in the Review: Directly Included Studies and Illustrative Primary Studies Reported Within Included Secondary Reviews by Chattu et al. [Chattu, 2021] and Xiong et al. [Xiong *et al.*, 2022].

Ref.	Context	Technique Used	Performance Metrics (%)	Evidence source
[Amin *et al.*, 2019]	Heart disease	Voting (Naïve Bayes + LR)	Acc: 87.41 | Sens: NR | F1: NR	Indirect
[Mohan *et al.*, 2019]	Heart disease	Random Forest (Linear Model)	Acc: 88.4 | Sens: NR | F1: NR	Indirect
[Singh and Kumar, 2020]	Heart disease	K-Nearest Neighbor	Acc: 87.0 | Sens: NR | F1: NR	Indirect
[Haq *et al.*, 2018]	Heart disease	Logistic Regression	Acc: 89.0 | Sens: NR | F1: NR	Indirect
[Bashir *et al.*, 2019]	Heart disease	Logistic Regression (SVM)	Acc: 84.85 | Sens: NR | F1: NR	Indirect
[Jo *et al.*, 2019]	Alzheimer's	Auto-encoder (SAE)	Acc: 98.8 | Sens: NR | F1: NR	Indirect
[Lin *et al.*, 2018]	Alzheimer's	Conv. Neural Network (CNN)	Acc: 79.9 | Sens: NR | F1: NR	Indirect
[Kautzky *et al.*, 2018]	Alzheimer's (Early)	Random Forest and SVM	Acc: 77.0 | Sens: NR | F1: NR	Indirect
[Almubark *et al.*, 2019]	Alzheimer's (Early)	SVM	Acc: 91.08 | Sens: NR | F1: NR	Indirect
[wei Fei, 2010]	Cardiac arrhythmia	Particle Swarm Opt. SVM	Acc: 95.65 | Sens: NR | F1: NR	Indirect
[Huang *et al.*, 2008]	Breast cancer	SVM (Grid Search)	Acc: 86.0 | Sens: NR | F1: NR	Indirect
[Zheng *et al.*, 2014]	Breast cancer	K-SVM	Acc: 97.38 | Sens: NR | F1: NR	Indirect
[Gonzalez-Abril *et al.*, 2014]	Valvular heart disease	Genetic-SVM (GSVM)	Acc: 95.0 | Sens: NR | F1: NR	Indirect
[Khan *et al.*, 2008]	Breast cancer	Hybrid Fuzzy Decision Tree	Acc: 85.0 | Sens: NR | F1: NR	Indirect
[Levashenko *et al.*, 2016]	Breast cancer (DSS)	Fuzzy Decision Trees	Acc: 96.0 | Sens: NR | F1: NR	Indirect
[Hassan *et al.*, 2011]	Hepatitis C treatment	Decision Tree (CART)	Acc: 80.0 | Sens: NR | F1: NR	Indirect
[Moon *et al.*, 2012]	Smoking behavior	Decision Tree	Acc: 67.17 | Sens: NR | F1: NR	Indirect
[Chang and Chen, 2009]	Dermatosis classif.	Artificial Neural Network	Acc: 92.62 | Sens: NR | F1: NR	Indirect
[Er *et al.*, 2010]	Pulmonary disease	Probabilistic Neural Network	Acc: 92.16 | Sens: NR | F1: NR	Indirect
[Sharma and Parmar, 2020]	Heart disease	Deep Learning (Talos Opt.)	Acc: 90.76 | Sens: NR | F1: NR	Indirect
[Long *et al.*, 2015]	Heart disease	IT2FLS (Fuzzy Logic)	Acc: 88.3 | Sens: NR | F1: NR	Indirect
[Talati *et al.*, 2022]	Malnutrition (India)	Random Forest / XGBoost	Acc: NR | **Sens: 70–80%** | F1: NR	Directly
[Ma, 2024]	Tumor Detection	Semi-Supervised DCNN	**Acc: 94.2 | Sens: 92.5 | F1: 93.1**	Directly

Note: Most primary studies listed in this table were identified within the included secondary reviews by Chattu et al. [Chattu, 2021] and Xiong et al. [Xiong et al.*, 2022] and were not independently screened through the present review protocol. Talati et al. [Talati* et al.*, 2022] and Ma et al. [Ma, 2024] were directly included in the formal scoping-review corpus. The "Evidence source" column distinguishes direct from indirect evidence. Acc = Accuracy; Sens = Sensitivity; F1 = F1-score; NR = Not Reported.*

sources) standard to govern over 1,400 variables across large-scale hospital networks. This approach facilitates broad administrative oversight and multi-departmental research. In contrast, Chen et al. [Chen *et al.*, 2024] show that "vertical," disease-specific platforms—such as those dedicated to nasopharyngeal carcinoma—offer the longitudinal granularity and specialized search engines that are often sacrificed in general management systems.

However, a significant "infrastructure gap" emerges as a critical scientific lacuna. The platform described by Wang et al. requires a processing capacity of 1,900 TFLOPS and 20 PB of storage, an investment that is virtually inaccessible to rural hospitals or healthcare systems in developing regions. This disparity suggests that the current "one-size-fits-all" approach to Big Data may exacerbate the digital divide between tier-1 academic hospitals and community health centers. Future research must prioritize "Lightweight Interoperability Frameworks"—architectures that provide the semantic benefits of Big Data governance without the massive hardware overhead. Furthermore, as highlighted by Sun et al. [Sun *et al.*, 2022], the reliance on multimedia fusion introduces a severe "Veracity" challenge: as unstructured data like medical videos and high-resolution imaging are integrated, the complexity of maintaining data quality for clinical auditing

and real-time processing increases exponentially.

4.2 Beyond Accuracy: The Clinical Utility Gap (RQ2)

The illustrative models summarized in Table 2 suggest a recurring tendency to prioritize Accuracy while omitting metrics such as Sensitivity and F1-score. Because these examples were derived from included secondary reviews rather than independently screened primary studies, this pattern should be interpreted cautiously and cannot be generalized to the entire healthcare Big Data literature.

Most of the illustrative primary studies [Chattu, 2021; Xiong *et al*., 2022; Liu *et al*., 2024; Jo *et al*., 2019] reported within the two included secondary reviews prioritized Accuracy as the principal performance measure. However, because these studies were not independently screened through the present review protocol, this observation should be interpreted as an illustrative reporting pattern rather than as a prevalence estimate for the broader literature.

Accuracy may provide an incomplete or potentially misleading assessment when evaluated on substantially imbalanced clinical datasets. A model with 99% accuracy is clinically dangerous if its **Sensitivity** (Recall) is low, as it leads to high rates of false negatives for life-threatening conditions.

The incomplete reporting of metrics such as F1-score, Sensitivity, and Area Under the ROC Curve (AUC) in several studies represented in the reviewed corpus limits the comparative assessment of predictive models and their potential clinical applicability. Clinicians require models that minimize the risk of missing a diagnosis over models that simply maximize correct classifications. The ASHRO model developed by Takura et al. [Takura *et al*., 2021] provides a necessary methodological shift by correlating treatment adherence predictions directly with a reduction in mortality rates (from 7% to 2%). This demonstrates that the true value of Big Data lies not in raw predictive power, but in its translation into actionable, "evidence-informed" decisions that demonstrably improve patient survival and resource allocation.

4.3 Ethics, Equity, and the "Black Box" Problem (RQ3)

The human and social dimension of Big Data is often the most neglected link in its implementation. As demonstrated by the seminal work of Obermeyer et al. [Obermeyer *et al*., 2019], algorithms can unintentionally learn and perpetuate systemic socioeconomic or racial prejudices if the training data reflects historical inequities in healthcare access. For instance, using "healthcare costs" as a proxy for "health needs" can result in algorithms underestimating the care required by marginalized populations who historically spent less on health due to lack of access rather than lack of illness.

Our mapping found that while theoretical frameworks like "Equity by Design" are advocated by Wesson et al. [Wesson *et al*., 2022], the majority of the 12 analyzed studies rely on massive regional or national datasets (e.g., India's NFHS [Talati *et al*., 2022] or Japan's Kokyho [Takura *et al*., 2021]) without explicitly discussing demographic representativeness or algorithmic fairness. This creates a "Cross-Border Validation Gap": there is insufficient evidence that a predictive model trained in one specific demographic context remains accurate or ethically fair when applied to different ethnic or socioeconomic groups. Moreover, the shortage of specialized IT professionals mentioned by Sun et al. [Sun *et al*., 2022] exacerbates the "Black Box" problem, as institutional capacity to audit these models for bias remains critically low.

4.4 The Blockchain Paradox: Security vs. Performance

Finally, data security via blockchain presents a practical paradox. The inherent immutability and decentralized nature of blockchain—hailed by Kaur et al. [Kaur *et al*., 2021] and Chattu et al. [Chattu, 2021] for data integrity—are technically at odds with the "Right to be Forgotten" mandated by regulations like the GDPR. Furthermore, the computational overhead of consensus mechanisms often conflicts with the "Velocity" requirements of real-time Big Data processing.

The evidence represented in the included corpus suggests that hybrid architectures may provide a promising approach for reconciling privacy, scalability, interoperability, and regulatory requirements in real-world healthcare implementations. In these systems, sensitive clinical data is stored *off-chain* in secure, standardized hospital servers, while only the access logs and hashes of patient consent are recorded on the blockchain. This approach addresses scalability but demands new "Legal-Tech" frameworks that can bridge the gap between data protection laws and distributed ledger engineering—a frontier that this scoping review identifies as one of the most urgent for future interdisciplinary research.

4.5 Limitations

This review has several limitations. First, the search was conducted exclusively in Scopus. Although Scopus provides broad multidisciplinary coverage, relevant studies indexed only in PubMed/MEDLINE, Web of Science, IEEE Xplore, ACM Digital Library, Embase, or other specialized databases may not have been retrieved.

Second, the search terms were applied exclusively to publication titles. This choice increased thematic specificity but reduced sensitivity, since relevant studies may discuss Big Data applications in healthcare without including all selected concepts in their titles. The use of the term "Application" may have further excluded studies using alternative expressions such as "implementation", "platform", "framework", or "deployment".

Third, the final corpus contained only twelve heterogeneous publications, including original studies, reviews, surveys, and conceptual contributions. The findings should therefore be interpreted as a focused mapping of the included evidence and cannot be generalized to the entire healthcare Big Data literature.

Fourth, no formal methodological quality or risk-of-bias assessment was conducted. Although such assessment is not mandatory for all scoping reviews, its absence limits conclusions regarding the strength of evidence.

Finally, several included studies did not report clinically relevant metrics such as sensitivity, specificity, F1-score, calibration, external validation, or subgroup-level performance. This incomplete reporting restricted comparisons and prevented a definitive assessment of clinical safety and general-

izability. Table 2 includes primary studies identified within two secondary reviews included in the formal corpus. These studies were not independently retrieved or screened using the present eligibility criteria. Accordingly, the table provides an illustrative analysis of performance-metric reporting and should not be interpreted as a systematic synthesis of all predictive studies identified by the review.

5 Conclusions

This focused scoping review examined twelve studies published between 2019 and 2024 addressing Big Data applications in healthcare. A broader retrieval set covering 2013–2024 was used to contextualize publication activity. Within the formally included corpus, the studies indicate potential benefits in health information management, predictive analytics, interoperability, and data security, while also identifying persistent methodological, infrastructural, and ethical limitations.

The findings highlight the importance of reporting clinically meaningful performance measures beyond Accuracy, including Sensitivity, Specificity, F1-score, calibration, external validation, and subgroup-level evaluation. They also suggest that general-purpose and disease-specific healthcare platforms address different operational and clinical requirements.

The conclusions should be interpreted in light of the single-database, title-only search strategy and the small, heterogeneous formal corpus. The predictive models extracted from included secondary reviews were used only as illustrative indirect evidence and do not constitute an independently screened synthesis of primary studies.

Future research should employ broader multi-database search strategies, evaluate models across diverse demographic populations, strengthen interoperability and fairness assessments, and investigate scalable privacy-preserving architectures for healthcare data exchange.

Declarations

Authors' Contributions

AG, BS, and FG contributed to the investigation, study selection, data extraction, data charting, and preparation of the original draft. LB and EL provided methodological supervision and contributed to the critical revision of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Availability of research artifacts

The list of analyzed studies and data extraction sheets are available upon request.

Use of Artificial Intelligence

Generative AI (ChatGPT-4o) was used exclusively for linguistic refinement and grammar correction to ensure professional academic English. No data was synthesized or analyzed by AI tools.

Further relevant information

Ethical approval was not required as this study is based on a review of publicly available peer-reviewed literature.

References

Almubark, I., Chang, L.-C., Nguyen, T., Turner, R. S., and Jiang, X. (2019). Early detection of alzheimer's disease using patient neuropsychological and cognitive data and machine learning techniques. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5971–5973. DOI: 10.1109/BigData47090.2019.9006583.

Amin, M. S., Chiam, Y. K., and Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, 36:82–93. DOI: 10.1016/j.tele.2018.11.007.

Bashir, S., Khan, Z. S., Hassan Khan, F., Anjum, A., and Bashir, K. (2019). Improving heart disease prediction using feature selection approaches. In *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 619–623. DOI: 10.1109/IBCAST.2019.8667106.

Chang, C.-L. and Chen, C.-H. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications*, 36(2, Part 2):4035–4041. DOI: 10.1016/j.eswa.2008.03.007.

Chattu, V. K. (2021). A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health. *Big Data and Cognitive Computing*, 5(3):41. DOI: 10.3390/bdcc5030041.

Chen, Z.-K., Wang, X.-Q., Xiao, L.-L., Sun, J.-D., Mao, M.-Y., Zhang, H.-B., and Guan, J. (2024). Construction and application of nasopharyngeal carcinoma-specific big data platform based on electronic health records. *American Journal of Otolaryngology*, 45(3):104204. DOI: 10.1016/j.amjoto.2023.104204.

Er, O., Yumusak, N., and Temurtas, F. (2010). Chest diseases diagnosis using artificial neural networks. *Expert Systems with Applications*, 37(12):7648–7655. DOI: 10.1016/j.eswa.2010.04.078.

Gonzalez-Abril, L., Nuñez, H., Angulo, C., and Velasco, F. (2014). Gsvm: An svm for handling imbalanced accuracy between classes inbi-classification problems. *Applied Soft Computing*, 17:23–31. DOI: 10.1016/j.asoc.2013.12.013.

Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., and Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile information systems*, 2018(1):3860146. DOI: 10.1155/2018/3860146.

Hassan, M., Abdalla, M., Ahmed, S., Akil, W., Esmat, G., Khamis, S., and ElHefnaw, M. (2011). The decision tree mode for prediction the response to the treatment in patients with chronic hepatitis c. *New York Science Journal*, 4(7):69–79. Available at:`https://www.sciencepub.net/newyork/ny0407/13_6298ny0407_69_79.pdf`.

Huang, C.-L., Liao, H.-C., and Chen, M.-C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Systems with Applications*, 34(1):578–587. DOI: 10.1016/j.eswa.2006.09.041.

Jo, T., Nho, K., and Saykin, A. J. (2019). Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging

data. *Frontiers in aging neuroscience*, 11:220. DOI: 10.3389/fnagi.2019.00220.

Kaur, I., Kumar, Y., and Sandhu, A. K. (2021). A comprehensive survey of ai, blockchain technology and big data applications in medical field and global health. In *2021 international conference on technological advancements and innovations (ICTAI)*, pages 593–598. IEEE. DOI: 10.1109/ictai53825.2021.9673285.

Kautzky, A., Seiger, R., Hahn, A., Fischer, P., Krampla, W., Kasper, S., Kovacs, G. G., and Lanzenberger, R. (2018). Prediction of autopsy verified neuropathological change of alzheimer's disease using machine learning and mri. *Frontiers in aging neuroscience*, 10:406. DOI: 10.3389/fnagi.2018.00406.

Khan, M. U., Choi, J. P., Shin, H., and Kim, M. (2008). Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5148–5151. DOI: 10.1109/IEMBS.2008.4650373.

Laney, D. *et al*. (2001). 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1. Available at:`https://www.bibsonomy.org/bibtex/742811cb00b303261f79a98e9b80bf49`.

Levashenko, V., Zaitseva, E., Kvassay, M., and Deserno, T. M. (2016). Reliability estimation of healthcare systems using fuzzy decision trees. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 331–340. IEEE. DOI: 10.15439/2016f150.

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., Guo, G., Xiao, M., Du, M., Qu, X., *et al*. (2018). Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment. *Frontiers in neuroscience*, 12:777. DOI: 10.3389/fnins.2018.00777.

Liu, C., Ma, G., and Huang, D. (2024). Application of artificial intelligence technology based on health and medical big data in disease monitoring, early warning, and prevention and control. *Procedia Computer Science*, 247:70–77. DOI: 10.1016/j.procs.2024.10.009.

Long, N. C., Meesad, P., and Unger, H. (2015). A highly accurate firefly based algorithm for heart disease prediction. *Expert Systems with Applications*, 42(21):8221–8231. DOI: 10.1016/j.eswa.2015.06.024.

Ma, Q. (2024). Research on the application of health big data based on semi-supervised learning algorithm. In *2024 3rd International Conference for Innovation in Technology (INOCON)*, pages 1–4. IEEE. DOI: 10.1109/inocon60754.2024.10512045.

Mohan, S., Thirumalai, C., and Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542–81554. DOI: 10.1109/ACCESS.2019.2923707.

Moon, S. S., Kang, S.-Y., Jitpitaklert, W., and Kim, S. B. (2012). Decision tree models for characterizing smoking patterns of older adults. *Expert Systems with Applications*, 39(1):445–451. DOI: 10.1016/j.eswa.2011.07.035.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453. DOI: 10.1126/science.aax2342.

Qinghua, C. and Weiguang, W. (2021). Analysis on the application of big data technology in medical and health industry. In *Journal of Physics: Conference Series*, volume 1883, page 012136. IOP Publishing. DOI: 10.1088/1742-6596/1883/1/012136.

Ristevski, B. and Chen, M. (2018). Big data analytics in medicine and healthcare. *Journal of Integrative Bioinformatics*, 15(3):20170030. DOI: doi:10.1515/jib-2017-0030.

Sharma, S. and Parmar, M. (2020). Heart diseases prediction using deep learning neural network model. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(3):2244–2248. DOI: 10.35940/ijitee.c9009.019320.

Singh, A. and Kumar, R. (2020). Heart disease prediction using machine learning algorithms. In *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, pages 452–457. DOI: 10.1109/ICE348803.2020.9122958.

Sun, D., Shi, H., and Ma, H. (2022). Application of multimedia technology in health service management system under the background of big data fusion. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1134–1137. IEEE. DOI: 10.1109/icssit53264.2022.9716407.

Takura, T., Hirano Goto, K., and Honda, A. (2021). Development of a predictive model for integrated medical and long-term care resource consumption based on health behaviour: application of healthcare big data of patients with circulatory diseases. *BMC medicine*, 19:1–16. DOI: 10.1186/s12916-020-01874-6.

Talati, K. N., Madan-Patel, G. D., Gurjwar, R. K., and Yadav, A. R. (2022). A case for action: India's national family health survey datasets await exploration of big data applications toward evidence-informed public health decision-making to tackle malnutrition. *Indian Journal of Community Medicine*, 47(1):151–152. DOI: 10.4103/ijcm.ijcm$_6$98$_2$1.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Lanauze-Bruneau, K., Casthaldi, D., Vafaei, A., Wilson, K., *et al*. (2018). Prisma extension for scoping reviews (prisma-scr): checklist and explanation. *Annals of internal medicine*, 169(7):467–473. DOI: 10.7326/m18-0850.

Wang, M., Li, S., Zheng, T., Li, N., Shi, Q., Zhuo, X., Ding, R., Huang, Y., *et al*. (2022). Big data health care platform with multisource heterogeneous data integration and massive high-dimensional data governance for large hospitals: design, development, and application. *JMIR Medical Informatics*, 10(4):e36481. DOI: 10.2196/36481.

wei Fei, S. (2010). Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine. *Expert Systems with Applications*, 37(10):6748–6752. DOI: 10.1016/j.eswa.2010.02.126.

Wesson, P., Hswen, Y., Valdes, G., Stojanovski, K., and Handley, M. A. (2022). Risks and opportunities to ensure equity in the application of big data research in public

health. *Annual review of public health*, 43(1):59–78. DOI: 10.1146/annurev-publhealth-051920-110928.

Xiong, H., Chen, H., Xu, L., Liu, H., Fan, L., Tang, Q., and Cho, H. (2022). A survey of data element perspective: Application of artificial intelligence in health big data. *Frontiers in Neuroscience*, 16:1031732. DOI: 10.3389/fnins.2022.1031732.

Zheng, B., Yoon, S. W., and Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4, Part 1):1476–1482. DOI: 10.1016/j.eswa.2013.08.044.