BMC Medical Informatics and Decision Making

SYSTEMATIC REVIEW

Open Access

Semantics-driven improvements in electronic health records data quality: a systematic review

Yirong Wu[1], Mudan Ren[2], Na Chen[2] and Liu Yang[2*]

Abstract

Background Data quality (DQ) of electronic health record (EHR) is crucial for the advancement of health informatization, yet it remains a significant challenge. Scholars are showing a growing interest in leveraging semantic technologies to enhance EHR data quality. However, previous studies have focused predominantly on specific semantic technologies, scenarios, or objectives—such as interoperability—often overlooking the potential of a various semantic technologies across different scenarios.

Objective This systematic review aimed to explore the potential of employing a range of semantic technologies to improve EHR data quality in a broader spectrum of application scenarios.

Methods Our systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Three databases were searched, including PubMed, IEEE Xplore, and Web of Science Core Collection. The search terms used included "Semantic*", "Quality", "Electronic Health Record*", "EHR*", "Electronic Medical Record*", and "EMR*". These terms were combined via various Boolean operators to formulate multiple search queries.

Results Thirty-seven papers that met the inclusion criteria between 2008 and 2024 were analyzed. Six semantic techniques were identified as instrumental in improving EHR DQ: EHR standardization, controlled vocabulary, ontology, semantic web, knowledge graph, and natural language processing (NLP). These technologies were further mapped to 16 core data quality indicators and the FAIR principles (Findable, Accessible, Interoperable, and Reusable), highlighting their contributions across both technical and governance dimensions.

Conclusions The six identified semantic technologies can be categorized into three levels: foundational, general, and advanced. These technologies show significant potential in enhancing EHR DQ, particularly in the areas of conformance, portability, usability, and applicability, and they are suitable for a variety of contexs beyond interoperability, aligning with FAIR-aligned best practices in data management and reuse.

Keywords Semantic, Ontology, Data quality (DQ), Electronic health record (EHR)

*Correspondence:
Liu Yang
irisy826@gmail.com
[1]Institute of Advanced Studies in Humanities and Social Sciences, Beijing Normal University, Zhuhai 519087, China
[2]School of Government, Beijing Normal University, Beijing 100875, China

Introduction

The advent of electronic health record (EHR) systems has presented a significant opportunity to enhance healthcare quality, which is contingent upon the data's attributes: consistency, accuracy, completeness, and timeliness [1]. Despite this potential, the enhancement of EHR data quality frequently encounters substantial challenges. In light of recent advancements in artificial intelligence, semantic-based techniques are emerging as promising avenues for overcoming these challenges and improving EHR data quality.

Semantics refers to the meaning of data elements—such as words, phrases, sentences, or texts—and involves their interpretation and logical representation within specific domains. Semantic technologies, which enable the explicit articulation and processing of such meanings, are widely recognized for their potential in integrating heterogeneous data sources and improving data quality through automated reasoning and inference. Although a growing body of research has explored the role of semantic technologies in health informatics, existing reviews often suffer from limited scope. Many focus on a single type of semantic technology, such as ontologies [1] or semantic web technologies [2], without considering the broader technological landscape. Additionally, numerous literature reviews have concentrated specifically on the issue of semantic interoperability, which plays a pivotal role in data sharing and traceability, mitigating data silos and enhancing consistency [3–5]. However, these previous reviews have not comprehensively examined the breadth of semantic technologies or their contributions to multiple dimensions of electronic health record (EHR) data quality (DQ), such as accuracy, completeness, and usability. To address this gap, our systematic review takes a broader perspective beyond interoperability and isolated technologies. We provide a structured classification and analysis of six major types of semantic technologies and examine their potential roles across various dimensions of EHR DQ. Furthermore, to support systematic evaluation, we map these technologies to a pre-established DQ indicator framework from our prior research [8].

First, this study explores multiple semantic technologies. EHR standardization and controlled vocabulary facilitate data exchange by providing standardized terminologies, akin to semantic dictionaries [6], that map EHR vocabularies to standardized terms. Ontology, a central semantic technology [7], defines data concepts and relationships, forming the basis for ontology-based semantic web and knowledge graph technologies. Given the prevalence of unstructured text in EHRs, NLP is also included as a key technology for tasks such as entity extraction, information retrieval, question-answering, and knowledge graph development. Thus, the review encompasses six semantic technologies: EHR standardization, controlled vocabulary, ontologies, semantic web, knowledge graph, and NLP.

Second, this review considers broader application scenarios. In EHR scenarios, semantic technologies enhance data consistency by establishing standardized terminology lists and automatically detecting and correcting errors through analyzing term relationships, which in turn improves data accuracy. We adopt a 16-item EHR data quality (DQ) evaluation framework developed in our earlier study [8], which includes dimensions such as accuracy, completeness, timeliness, consistency, precision, conformance, uniqueness, credibility, plausibility, traceability, portability, usability, accessibility, relevance, applicability, and understandability. This framework was validated by four types of EHR stakeholders (doctors, nurses, hospital supervisors, and clinical researchers) and is suitable for general-purpose evaluation beyond specific diseases or departments. In this review, we use this DQ indicator system as an analytical lens to evaluate how the six representative semantic technologies relate to and potentially improve different aspects of EHR data quality in real-world settings. The mapping between technologies and indicators is presented in Table 4 to aid structured analysis.

Third, this review aligns semantic technologies with the FAIR data principles—Findable, Accessible, Interoperable, and Reusable—to emphasize their value in enhancing data discoverability, accessibility, and reusability. This not only situates semantic technologies within a widely endorsed framework for data stewardship but also reinforces their practical significance in supporting the long-term utility of EHR data across clinical, research, and administrative contexts.

Accordingly, our contributions are threefold:

(1) We provide a comprehensive classification of six major semantic technologies, covering their characteristics and levels of semantic processing;
(2) We systematically evaluate the potential of these technologies to improve EHR data quality using a validated 16-indicator DQ framework from our prior research;
(3) We highlight the alignment of semantic technologies with FAIR principles, revealing how these tools support both technical interoperability and broader goals of data governance and reusability.

Methods

Search strategy and selection process

Our review follows the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [9].

Table 1 Eligibility criteria

Inclusion criteria	Exclusion criteria
1. The study must engage the use of semantic technologies to improve EHR data quality.	1. Studies that do not focus on the use of semantic technologies to improve EHR data quality will be excluded.
2. The study must be peer-reviewed. We include the peer-reviewed conference papers and journal articles.	2. Books, book chapters, and dissertations, as well as other forms of literature, are excluded because they are often not peer-reviewed/inexperienced research [10].
3. The study must be published in English.	3. Studies not published in English will be excluded. Some evidence suggests that including only English studies does not bias meta-analysis, at least in the field of medicine [11].
4. Only original research studies are included.	4. Review articles are excluded. The purpose is to ensure a comprehensive analysis of original research and avoid interference from secondary analysis or subjective judgments of researchers.

Table 2 Initial search queries performed for the identification of studies

Database	Search Terms
PubMed: Medline (National Library of Medicine)	((semantic*) AND (quality) AND ((electronic health record*) OR (electronic medical record*) OR (EHR) OR (EMR))) Filters: English
WOS: Web of Science Core Collection (Clarivate Analytics)	(semantic* AND quality AND (electronic health record* OR EHR OR EMR OR electronic medical record*)) Languages: English
IEEE Xplore (Institute of Electrical and Electronics Engineers Xplore Digital Library)	(semantic* AND quality AND ("electronic health record*" OR "electronic medical record*" OR EHR OR EMR)) Conferences & Journals

Information sources

We selected three databases: PubMed/Medline, a database for biomedical literature; the Institute of Electrical and Electronics Engineers (IEEE) Xplore Digital Library, a database for literature in computer science and electronic technology; and the Web of Science Core Collection, a comprehensive citation index database for core journals.

Eligibility criteria

This review strictly followed the PRISMA principles and conducted a systematic search for peer-reviewed publications that examined the use of semantic technologies to improve EHR DQ. Table 1 summarized the general principles for the inclusion and exclusion of studies.

Search strategy

Our search terms included "Semantic*", "Quality", "Electronic Health Record*", "EHR*", "Electronic Medical Record*", and "EMR*". The search terms are linked by Booleans. We have placed the initial search terms in Table 2 performed for the identification of studies. Original search queries were performed on July 20, 2024, to identify the studies conducted in each. Since 2008 marked a turning point in global EHR adoption and policy reform, we selected it as the baseline to capture the evolution of semantic technologies for improving EHR data quality through 2024 [51, 52].

To ensure relevance and focus in the screening process, we applied a pragmatic and widely used strategy by limiting inclusion to articles with the term "semantic" (or its variants) explicitly mentioned in the title. This decision was based on the specificity of our research question, which centers on the application of semantic technologies in improving EHR data quality. Titles serve as the most immediate and indicative representation of a study's primary focus. Articles that do not mention "semantic" in the title were considered less likely to have semantic methods as their core contribution. This strategy was also essential for maintaining screening feasibility and methodological transparency given the large volume of retrieved records. We acknowledge that some potentially relevant papers might mention "semantic" only in the abstract or body text; however, our approach prioritized specificity over sensitivity to ensure consistency in relevance.

Study selection

Following the initial screening based on the presence of the term 'semantic' in the title, two reviewers independently conducted full-text screening to assess each study's relevance to the research objective: exploring how semantic technologies can improve electronic health record (EHR) data quality. To ensure consistency and objectivity during manual screening, we established a set of inclusion and exclusion criteria as follows:

Inclusion criteria (all must be met):

- The study involves semantic technologies (e.g., controlled vocabulary, ontology, Semantic Web, knowledge graph, or NLP), and these technologies must play a core methodological role—not merely be mentioned or used in unrelated contexts (e.g., image segmentation).
- The application context is healthcare, with explicit focus on structured EHR data.
- The study contributes to the improvement of EHR data quality, such as in terms of accuracy, completeness, consistency, interoperability, understandability, or governance.

Exclusion criteria (if any apply):

- "Semantic" is used only in non-informatic or vague contexts, or does not refer to an actual semantic technology.
- The study is focused on medical imaging, consumer health, or non-clinical data without linkages to EHR or structured clinical data.
- No identifiable or substantial contribution to improving EHR data quality is made.

To support this process, we designed a structured manual review form to guide evaluation. Each entry included: article title, semantic technology used (Y/N), semantic technologies (if applicable), focus on EHR data (Y/N), contribution to EHR data quality (Y/N), DQ indicators (if applicable), justification / notes, initial decision, and reason for inclusion/exclusion. The evaluation of semantic technologies is based upon the *Classification Principles and Coding Logic of Semantic Technologies* provided in the Supplementary material [53].

Two doctoral researchers with expertise in semantic health informatics independently applied the form to all candidate studies. Prior to the full screening, a training and calibration exercise using 20 randomly sampled studies was conducted to align interpretations. Discrepancies were resolved through consensus. Ongoing discrepancies were addressed via adjudication by a third expert. The final screening outcomes are summarized in Appendix Table A1, while Appendix Table A2 presents the screening process results for all literature [53].

Study evaluation

To evaluate the impact of semantic technologies on EHR data quality, we applied a structured evaluation framework derived from our previously validated 16-item DQ indicator system [8]. These indicators include accuracy, completeness, timeliness, consistency, precision, conformance, uniqueness, credibility, plausibility, traceability, portability, usability, accessibility, relevance, applicability, and understandability. The framework enabled us to map each included study to specific data quality dimensions it addressed.

Definitions of semantic technologies

The semantic technologies discussed in this article include EHR standardization, controlled vocabulary, ontology, semantic web, knowledge graph, and NLP. To demonstrate the role of each technology in improving EHR DQ, we have further classified and delineated the definitions or scopes of these technologies. To ensure conceptual clarity and theoretical rigor, we delineated clear boundaries between these technologies based on two primary criteria: (1) Conceptual definition and origin, and (2) Functional role in enhancing EHR data quality (DQ).

EHR standardization

OpenEHR

Conceptually, the open electronic health record (openEHR) is an open standard for representing and communicating health information, originating in Europe and Australia. It was endorsed by the International Standards Organization in 2008 as the ISO 13,606 standard. Functionally, openEHR employs a two-level modeling approach to separate domain knowledge from specific clinical information: a reference model that defines generic data structures and an archetype model that defines clinical concepts. The majority of current clinical archetypes are represented using the archetype definition language (ADL). However, the capacity for reasoning over ADLs is currently very limited, and the availability of tools for using and managing ADL content is insufficient [12]. These limitations hamper the semantic expressiveness of openEHR, as ADL primarily supports the syntactic definition of clinical structures, resulting in limited support for rich semantic content [12, 13].

HL7

Health Level Seven (HL7), develop in the United States, is a family of international standards for the exchange, integration, sharing, and retrieval of electronic health information. Conceptually, HL7 provides a comprehensive framework encompassing conceptual models, documents, applications, and messaging standards. These include the HL7 reference information model (RIM), clinical document architecture (CDA), clinical context object workgroup (CCOW), and messaging protocols such as HL7 V2.5 and V3.0. Functionally, the HL7 RIM defines structural relationships among data elements without assigning specific semantic meaning to them, limiting its expressiveness. HL7 fast health care interoperability resources (FHIR), a more recent standard, aims to enhance interoperability and usability by integrating features of previous HL7 versions with modern web technologies such as XML and JSON. FHIR provides a modular and developer-friendly framework that facilitates efficient healthcare data exchange.

Controlled vocabulary

A controlled vocabulary refers to a curated set of standardized terms used to describe data elements consistently across systems. Its primary aim is to reduce ambiguity and redundancy in natural language by enforcing the use of predefined labels selected and maintained by experts. Functionally, controlled vocabularies facilitate data retrieval, improve search precision, and support basic semantic interoperability through lexical normalization. Conceptually, they offer a flat or hierarchical list of terms without necessarily modeling the relationships among them.

Ontology

Ontology, rooted in philosophy, refers to the formal and explicit specification of concepts and their relationships within a domain. In biomedical informatics, an ontology is defined as [14] "a formal set of entities and the relationships between them, which is machine-processable and human-interpretable within a specific application domain". Functionally, ontologies go beyond term standardization by enabling logical reasoning, subclass inference, and semantic integration. They conceptualize not only entities but also properties, constraints, and logical axioms, supporting advanced knowledge representation and automated inference capabilities.

Semantic web

On December 18, 2000, Tim Berners-Lee introduced the concept of the semantic web at the XML2000 conference. Conceptually, the semantic web is a web-based framework that enables machines to understand, interpret, and link data through formal semantics. It integrates unique addressing mechanisms—uniform resource identifiers (URIs)—with structured knowledge representation using the resource description framework (RDF) and web ontology language (OWL), and employs simple protocol and RDF query language (SPARQL) [15]. Functionally, the semantic web facilitates semantic-level information sharing and interoperability across heterogeneous systems by encoding and connecting data in a machine-readable manner. Ontologies are a foundational component of the semantic web architecture, but the focus of semantic web technologies is not ontology development itself. In this review, studies focusing on ontology construction are classified under the ontology category, while those utilizing OWL and SPARQL in web-based implementations fall within the semantic web category.

Table 3 Summary of the distinctions among 6 semantic technologies

Semantic Technology	Conceptual Definition	Primary Function in EHR DQ
EHR Standardization	Structured frameworks and protocols (e.g., HL7, openEHR) for defining, exchanging, and storing healthcare data.	Ensures structural conformance, facilitates data exchange, and improves interoperability across systems.
Controlled Vocabulary	Curated lists of standardized terms used to enforce linguistic consistency (e.g., SNOMED CT, LOINC).	Reduces lexical ambiguity, enhances terminological consistency, and enables basic semantic interoperability.
Ontology	Formal specifications of domain concepts and logical relationships enabling semantic reasoning.	Provides semantic inference, supports high-level integration, and formalizes medical knowledge structures.
Semantic Web	A web-based data architecture using RDF, OWL, SPARQL to make structured data machine-readable and queryable.	Enables semantic-level information sharing and federated data access via standardized web technologies.
Knowledge Graph	Graph-structured knowledge representations integrating ontologies and instance data.	Encodes and links real-world entities and relationships; supports semantic search and reasoning.
NLP	Computational techniques for extracting structured data from unstructured clinical text.	Transforms narrative EHR data into structured formats; supports concept extraction, classification, and downstream inference.

Knowledge graph

In the architecture of the semantic web, the descriptions of instance data and ontological knowledge are separated [16]. In 2012, Google introduced the term "knowledge graph" to describe the use of graph-based structures for representing real-world entities and their relationships. Conceptually, a knowledge graph integrates ontological schemas and factual instance data into a unified graph model. Functionally, it enables semantic reasoning, entity linking, and contextual knowledge retrieval by connecting data through nodes (entities) and edges (relationships). A knowledge graph typically combines structured data, ontologies, and supporting standards and tools into a cohesive system that enables richer interpretation and intelligent data services.

NLP

Natural language processing (NLP) technology enables the extraction of variables from unstructured text reports to create structured databases, thereby enhancing the accuracy and standardization of data. Conceptually, NLP focuses on computational techniques for analyzing and understanding human language. Functionally, NLP plays a dual role in relation to semantic technologies: it is often used as a preprocessing step for constructing controlled vocabularies, ontologies, semantic web structures, and knowledge graphs; conversely, existing controlled vocabularies and ontologies can enhance NLP performance by supporting disambiguation and enhancing semantic understanding, such as optimizing word embeddings or entity linking.

Although some of these technologies are interdependent (e.g., ontologies are foundational to the semantic web and knowledge graphs), we classify them as separate categories due to their distinct implementation emphases, data processing goals, and usage patterns observed in the reviewed literature. The following table (Table 3) summarizes the conceptual and functional distinctions among the six semantic technologies, which guided our classification:

Results

Search results

Following the PRISMA process (Fig. 1), we retrieved a total of 1057 articles from three databases and ultimately identified 37 original research studies (excluding review papers) that met the inclusion criteria for this review (Table A2). We have detailed in Table 4 the basic information of each article and the relevant information related to our research, including the article number (Ref.), article title, author, database, article type, publication year, method of processing / verification, Semantic techniques involved, and the indicators involved in the quality of EHR data [8]. In terms of the semantic technologies applied, EHR standardization appeared the most (in 19 articles), followed by ontology and controlled vocabulary, which appeared 15 and 14 times, respectively, whereas semantic web and NLP appeared 10 and 6 times, respectively. Knowledge graph technology appeared the least common, only twice.

Semantic technique distribution of the included studies

Figure 2 illustrates the involvement of semantic technologies in journal publications across different years within the included literature. Six semantic techniques are depicted by bubbles of varying colors, with the size of the bubbles correlating to the frequency of each technique in the journals' annual publications. For example, *J Biomed Inform* has a substantial body of related publications, notably in 2010, 2014, 2016, 2017, 2018, and 2020. Additionally, the *Stud Health Technol Inform* published a considerable number of articles on the semantic web in 2009.

Dynamic flow of semantic technology

Figure 3 utilizes Sankey diagrams to depict the fluctuating prominence of six semantic techniques across different years and journals. In 2013, a significant number of papers addressed all semantic technologies except knowledge graph. By 2024, the focus had shifted predominantly toward controlled vocabulary and natural language processing (NLP) technologies. EHR standardization has the highest number of articles among all the technologies, appearing in the majority of journals. In contrast, articles on knowledge graph are limited to just two journals.

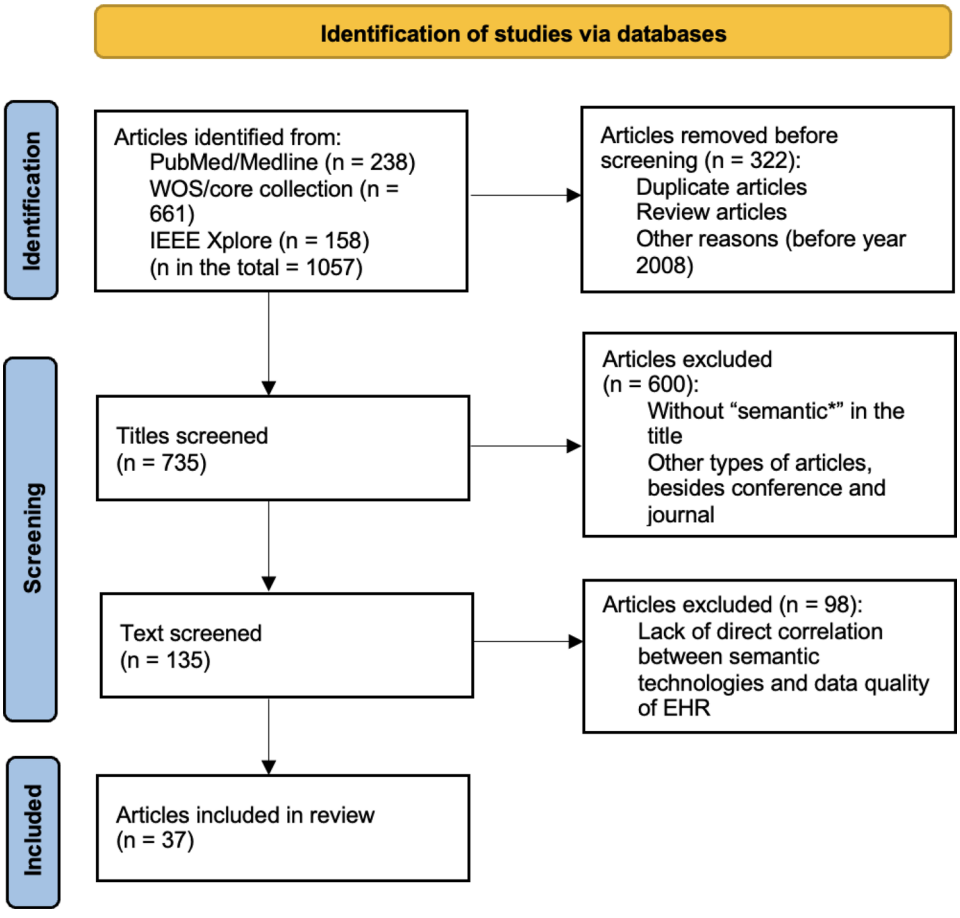

Fig. 1 PRISMA flowchart diagram for the systematic review

Table 4 Summary of all included studies ($n=37$)

Ref.		Author	Database	Article type	Year	Method	Semantic techniques involved	Indicators involved
[38]	Providing semantic interoperability between clinical care and clinical research domains	Laleci et al.	PubMed	J	2013	Quantitative	SD, CV, OT	portability, conformance, relevance, applicability
[50]	Mismatches between major subhierarchies and semantic tags in SNOMED CT	Bona et al.	PubMed	J	2018	Quantitative	CV	consistency, credibility
[33]	Design and development of a sharable clinical decision support system based on a semantic web service framework	Zhang et al.	PubMed	J	2016	Quantitative	SD, OT, SW	accuracy, completeness, understandability, conformance, portability
[18]	Contribution of clinical archetypes, and the challenges, towards achieving semantic interoperability for EHRs	Tapuria et al.	PubMed	J	2013	Qualitative	SD	accuracy, consistency, completeness, uniqueness, precision
[13]	An approach for the semantic interoperability of ISO EN 13,606 and OpenEHR archetypes	Martínez-Costa et al.	PubMed	J	2010	Qualitative	SD, OT, SW	portability, applicability
[39]	Achieving clinical statement interoperability using R-MIM and archetype-based semantic transformations	Kilic et al.	PubMed	J	2009	Qualitative	SD	conformance, portability, traceability
[29]	Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching	Kiourtis et al.	PubMed	J	2019	Quantitative	SD, OT	completeness, portability, conformance, precision, credibility, consistency
[21]	Development of an EHR system for sharing - a semantic perspective	Liu et al.	PubMed	J	2009	Qualitative	SD, CV, SW	accessibility, applicability
[22]	Construction of the cervical cancer common terminology for promoting semantic interoperability and utilization of Chinese clinical data	Hong et al.	PubMed	J	2021	Quantitative	NLP, CV	understandability, usability
[36]	Multi-ontology refined embeddings (more): a hybrid multi-ontology and corpus-based semantic representation model for biomedical concepts	Jiang et al.	PubMed	J	2020	Quantitative	NLP, CV	accuracy, relevance, uniqueness, accessibility
[31]	Creating personalised clinical pathways by semantic interoperability with electronic health records	Wang et al.	PubMed	J	2013	Qualitative	SD, CV, OT, SW, NLP	conformance, accessibility, applicability, portability
[45]	Creating hospital-specific customized clinical pathways by applying semantic reasoning to clinical data	Wang et al.	PubMed	J	2014	Quantitative	OT, SW	applicability, consistency, completeness, uniqueness
[32]	Ontology-based clinical pathways with semantic rules	Hu et al.	PubMed	J	2012	Qualitative	OT, SW	applicability, portability, accuracy, usability
[25]	Conducting research using the electronic health record across multi-hospital systems: semantic harmonization implications for administrators	Bowles et al.	PubMed	J	2013	Qualitative	CV	portability, accessibility, completeness, conformance
[28]	Detection of medical text semantic similarity based on convolutional neural network	Zheng et al.	PubMed	J	2019	Quantitative	CV, NLP	understandability, accessibility, applicability
[30]	Modeling and validating HL7 FHIR profiles using semantic web Shape Expressions (ShEx)	Solbrig et al.	PubMed	J	2017	Quantitative	SD	applicability, portability, credibility, conformance, understandability
[20]	HL7 FHIR with SNOMED-CT to achieve semantic and structural interoperability in personal health data: a proof-of-concept study	Chatterjee et al.	PubMed	J	2022	Qualitative	SD, CV	portability, completeness, conformance, accessibility, precision
[12]	Integrating semantic dimension into openEHR archetypes for the management of cerebral palsy electronic medical records	Ellouze et al.	PubMed	J	2016	Quantitative	SD, OT, SW	portability, credibility, uniqueness

Table 4 (continued)

Ref.		Author	Database	Article type	Year	Method	Semantic techniques involved	Indicators involved
[17]	An openEHR based approach to improve the semantic interoperability of clinical data registry	Min et al.	PubMed	J	2018	Qualitative	SD	conformance, uniqueness, accuracy, applicability, consistency
[27]	An ontology for healthcare quality indicators: challenges for semantic interoperability	White et al.	PubMed	J	2015	Qualitative	OT	conformance, accessibility, usability
[37]	Archetype-based knowledge management for semantic interoperability of electronic health records	Garde et al.	PubMed	J	2009	Qualitative	SD	portability, credibility, accuracy, consistency
[43]	A Flexible Semantic Integration Framework for Fully-integrated EHR based on FHIR Standard	Dridi et al.	WOS	C	2020	Quantitative	NLP, SD, OT, SW	conformance, consistency, relevance
[49]	Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings	Antunes et al.	WOS	C	2020	Quantitative	NLP	understandability, uniqueness, precision, credibility
[48]	A novel methodology for clinical semantic annotations assessment	Moreno-Fernandez-de-Leceta et al.	WOS	J	2018	Quantitative	NLP	completeness, accuracy, uniqueness, consistency, precision
[44]	Towards a secure semantic knowledge of healthcare data through structural ontological transformations	Kiourtis et al.	WOS	C	2019	Quantitative	SD, OT	conformance, portability, applicability
[47]	Research on the medical knowledge deduction based on the semantic relevance of electronic medical record	Qiao et al.	WOS	J	2023	Quantitative	NLP, KG	consistency, completeness, accuracy, understandability
[46]	Learning medical concept representation based on semantic information in medical textural data	Im et al.	WOS	J	2024	Quantitative	NLP, CV	precision, understandability, applicability
[42]	SAPHIRE: intelligent healthcare monitoring based on semantic interoperability platform: pilot applications	Nee et al.	WOS	J	2008	Quantitative	SD	portability, applicability
[41]	Gaining the semantic knowledge of healthcare data through syntactic models transformations	Kiourtis et al.	IEEE	C	2018	Quantitative	OT	conformance, understandability, usability, applicability
[19]	A multidimensional framework for semantic electronic health records in oncology domain	de Figueiredo et al.	IEEE	C	2022	Qualitative	SD, CV	conformance, applicability, accessibility, portability, relevance
[40]	Semantic constraints specification and Schematron-based validation for internet of medical things' data	Koren et al.	IEEE	J	2022	Quantitative	SD	relevance, conformance, applicability, portability, consistency, usability, precision
[15]	Ontology based semantic representation for Public Health data integration	Rao et al.	IEEE	C	2015	Quantitative	SW, OT, CV	accessibility, portability, applicability
[26]	Using ontologies to improve semantic interoperability in health data	Liyanage et al.	Pubmed	J	2015	Qualitative	OT	understandability, applicability, conformance
[23]	Semantic processing of EHR data for clinical research	Sun et al.	Pubmed	J	2015	Quantitative	CV	portability, conformance, applicability, usability, understandability
[35]	Capturing semantic relationships in electronic health records using knowledge graphs: an implementation using MIMIC III dataset and GRAPHDB	Aldughayfiq et al.	Pubmed	J	2023	Qualitative	OT, KG	accessibility, accuracy, conformance, portability, applicability

Table 4 (continued)

Ref.		Author	Database	Article type	Year	Method	Semantic techniques involved	Indicators involved
[24]	SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research	Wu et al.	Pubmed	J	2018	Quantitative	CV, SD, NLP	accessibility, accuracy, applicability
[34]	Data quality principles in the semantic web	Assaf et al.	IEEE	C	2012	Qualitative	SW	usability, conformance, applicability

[a] Document type: J = Journal paper; C = Conference paper

[b] Semantic techniques involved: SD = EHR standardization; CV = Controlled vocabulary; OT = Ontology; SW = Semantic web; KG = Knowledge graph; NLP = Natural language processing

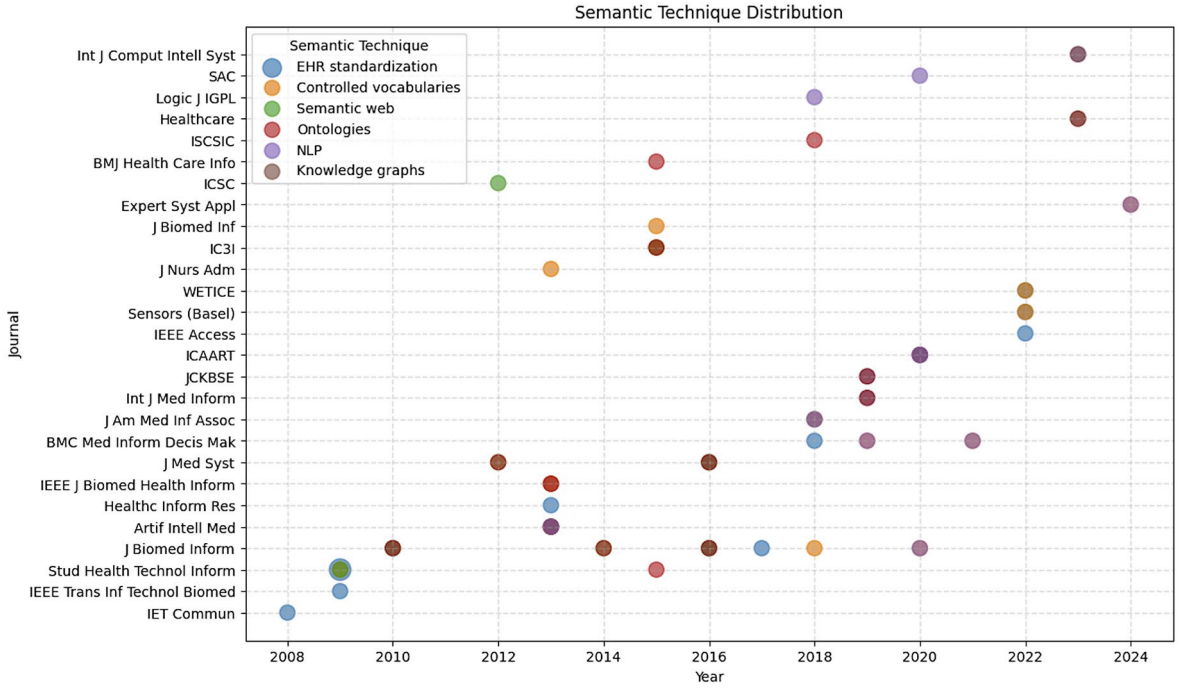

Fig. 2 Semantic technique distribution of the included studies

Classification based on the utilization of semantic technology

EHR standardization

OpenEHR In most openEHR-related studies, different strategies and modules have been developed to introduce semantic dimensions on the basis of openEHR archetypes, including Domain-involved Archetype Editor (DiAE) [17], clinical archetypes [18], and BioFrame [19]. Ellouze et al. [12] transformed EHRs of cerebral palsy patients into prototypes expressed according to ADLs, paving the way for the introduction of semantic dimensions. Specifically, they involved semantic dimensions in archetypes via semantic web technologies, with ontologies constructed via UML class diagrams.

For other openEHR-related studies, Garde et al. [37] presented the advancements and challenges related to openEHR technology aimed at improving semantic interoperability. Martínez-Costa et al. [13] discussed the semantic interoperability between two EHR standards: openEHR and ISO EN 13,606. They found that, using ontology-based technologies, it is possible to convert archetypes from openEHR to ISO EN 13,606, and vice versa.

HL7

Semantic technologies based on HL7 have also been widely used to achieve data interoperability, in which different strategies and modules have been developed, including SNOMED-CT [20], the R-MIM (Refined Message Information Model) [39], and ontology [31, 38]. The realization of data interoperability enables the development of multimodal data-based application systems,

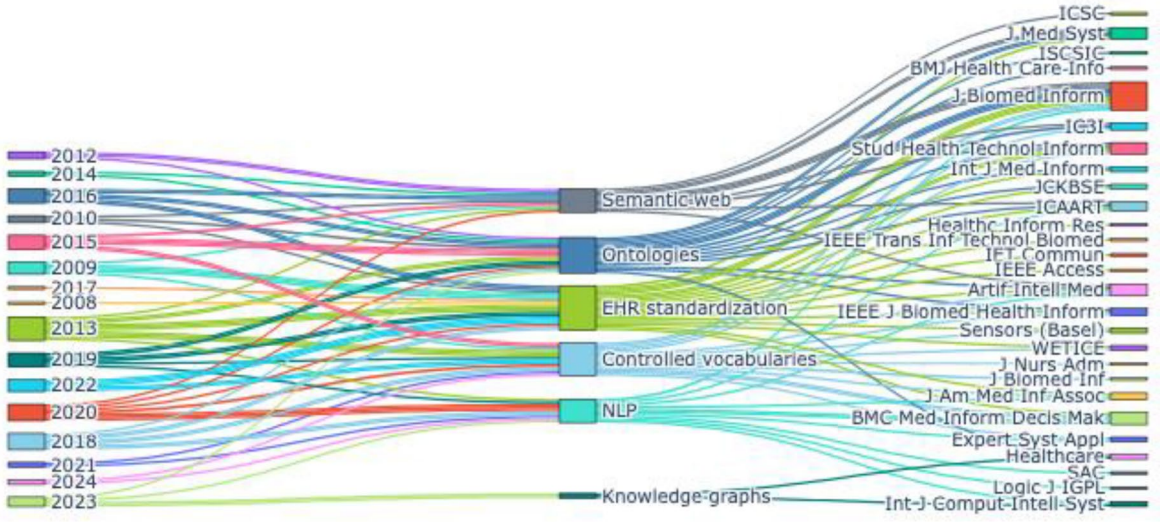

Fig. 3 Dynamic flow of semantic technology

including data integration systems [40, 42, 43] and retrieval systems [21, 24].

In particular, FHIR, as the latest HL7 framework, effectively facilitates the interoperability of healthcare data to address the challenges of data exchange in healthcare systems and improve the connectivity of healthcare information, thereby enhancing the quality and efficiency of healthcare services. Kiourtis et al. [29, 44] explored the use of ontologies to transform clinical data into the HL7 FHIR format. Solbrig et al. [30] reported that shape expressions (ShEx) could be used to describe and validate RDF-based representations of HL7 FHIR instances.

Furthermore, Zhang et al. [33] designed and developed a shared clinical decision support system based on a semantic web service framework, using the HL7 Virtual Medical Record (vMR) standard data model to describe clinical data.

Controlled vocabulary

Controlled vocabulary plays a crucial role in the process of data semantization, with different approaches used in its applications. One widely used approach is to combine controlled vocabularies with standardization technology to achieve semantic interoperability of data [19, 20, 24]. Another common approach involves converting controlled vocabularies into RDF format and then leveraging ontology technology to enable data semantization [15, 31, 38]. Some studies also directly combine controlled vocabularies in RDF format with other RDF-formatted data to build application systems without constructing ontologies [21, 23]. In other applications, Jiang et al. [36] developed a semantic similarity measurement method based on the Medical Subject Headings (MeSH) vocabulary. Im et al. [46] used controlled vocabularies, such as the ICD-9, to predict the likelihood of readmission or future diseases. Zheng et al. [28] employed Chinese version of MeSH (CMeSH) to train feature vectors aimed at finding correlations between imaging and pathology reports.

As one of the technologies for achieving data semantization, the development of controlled vocabularies themselves has also garnered significant attention from researchers and developers. Hong et al. [22] established a Cervical Cancer Common Terminology (CCCT), a vocabulary intended to facilitate clinical data exchange, ensure DQ, and support large-scale data analysis. Bowles et al. [25] created their standard terminology framework, which maps different names to standard terms to develop a clinical decision system. Bona and Ceusters [50] discussed the issue of semantic tags within SNOMED CT.

Ontology

This review categorizes studies related to ontology technology into two types. (1) Construct ontologies as a target. This type involves studies that focus on constructing ontologies to achieve data semantic interoperability or improve data quality. Most of these studies utilize various conversion tools or platforms to transform EHR data into XML format, which is then further converted into RDF format to build ontologies for achieving semantic interoperability of data [29, 38, 41, 44]. Moreover, some studies have directly constructed ontologies on the Protégé platform [26, 27]. (2) Construct ontologies as an intermediate step. This type involves studies that build ontologies as a foundation and use semantic web or knowledge graph technologies to achieve data semantic interoperability or enhance the quality of clinical care. In these studies, since ontology construction is an intermediate step rather than the final objective, most research

directly utilizes the Protégé platform for ontology construction [12, 15, 31–33, 35, 43, 45].

Semantic web

In semantic web-based studies (Fig. 4), the Jena semantic web framework (Jena) is widely used to build application systems. To enhance the use of clinical pathways (CPs) by leveraging the semantic interoperability between knowledge-based CPs and semantic EHRs, Protégé was utilized as the ontology editor tool, OWL was adopted for ontology description, and Jena was used for semantic transformation and reasoning [31, 32, 45]. Zhang et al. [33] designed and developed a shared clinical decision support system based on Jena. Rao et al. [15] designed a public health ontology based on Jena to describe domain knowledge and achieve semantic interoperability for data integration.

The combination of semantic web and model-driven engineering (MDE) technologies is also widely used to develop semantic-based systems. On the basis of semantic web and MDE technologies, Martínez Costa et al. [13] developed a tool called the poseacle converter in the Eclipse platform to achieve interoperability. Ellouze et al. [12] involved a semantic dimension in the prototype model to achieve semantic interoperability, in which the ADL2OWL translator tool was used to transform ADL archetypes into OWL ones.

In other applications of semantic web technology, Dridi et al. [43] utilized the MetaMap tool to recognize medical entities from unstructured data and generate a JSON-based structured output. The results are then transformed into FHIR-based document formats, which are mapped to some ontologies. Liu et al. [21] developed an EHR system prototype to implement the semantic interoperability of EHR systems, in which Semplore is used to retrieve semantic web data.

Additionally, Assaf and Senart [34] discussed the importance of DQ in the semantic web. They first identified five principles that affect the quality of semantic web data, including the data source, raw data, semantic conversion, linking process, and global quality. Then, for each principle, they listed the attributes followed in the data management process.

Knowledge graph

Aldughayfiq et al. [35] explored the use of knowledge graph to obtain and represent complex relationships in EHRs. The results indicate that a knowledge graph is an effective tool for capturing complex semantic relationships in EHRs, enabling more efficient and accurate data analysis. Qiao et al. [47] proposed a TCM-KR method of knowledge reasoning to efficiently mine first-order association rules from unstructured text data, which are used to infer a large amount of high-quality knowledge and predict the correlation characteristics of some diseases in a knowledge graph.

NLP

We classified the application of NLP technology into two types: existing technologies and self-developed tools (Fig. 5). The available NLP tools used to extract entities from free text reports include the Natural Language Toolkit (NLTK) [36, 49], the Medical Probabilistic Language Understanding System (M+) [31], MetaMap [43], and the Bio-YODIE NLP pipeline [24]. In other studies, researchers developed their own NLP models to extract entities and relations from clinical notes. Hong et al. [22] developed a combination method of a conditional random field (CRF) and a rule-based information extraction technique to recognize named entities and relations. Qiao et al. [47] proposed a TCM-KR method of knowledge reasoning to extract information from unstructured text data in the field of medicine. After removing all the punctuation, number, and stop words from raw report texts, Zheng et al. [28] developed a convolutional neural network model to detect medical text semantic similarity. Moreno-Fernandez-De-Leceta et al. [48] developed a mechanism to obtain accuracy indicators of semantic annotations generated by a language annotating system (LAS) from EHRs instead of human annotations. Im et al.

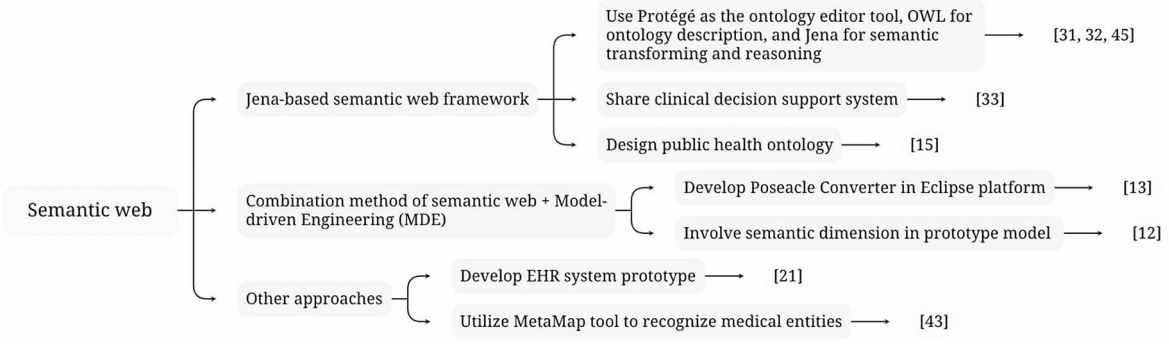

Fig. 4 Utilization of semantic web

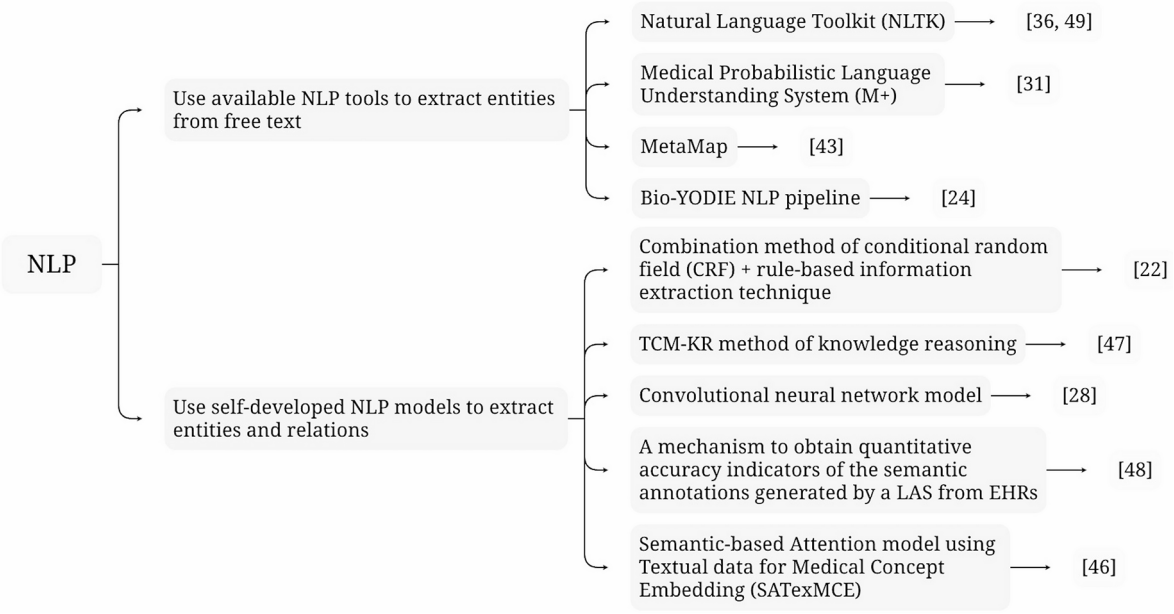

Fig. 5 Utilization of NLP

Table 5 Studies using semantic technologies to achieve semantic interoperability

Semantic technology	Ref.
EHR standardization	[17, 18, 30, 37, 39, 40, 42]
EHR standardization + controlled vocabulary	[19–21, 31, 38]
EHR standardization + ontology	[12, 13, 29, 33, 43, 44]
Controlled vocabulary	[15, 22, 23, 25]
Ontology	[26, 27, 41]
Semantic web / knowledge graph	[34, 35, 47]

[46] developed a semantic-based attention model using textual data for medical concept embedding (SATexMCE) to learn medical concept representations.

Utilization of semantic technologies for enhancing EHR data quality

Semantic technologies have demonstrated strong potential in improving the quality of EHR data. Technologies such as EHR standardization, controlled vocabulary, ontology, semantic web, and knowledge graph can improve data consistency through the construction of standardized terms and structures. Additionally, technologies like NLP can extract structured variables from unstructured text reports, thereby improving data accuracy. The utilization of these semantic technologies to enhance EHR data quality is illustrated across the 37 included studies, as described below.

First, most studies focus on improving data interoperability using semantic technologies (Table 5). Specifically, EHR standardization provides standardized structures and terminologies for health data transmission, facilitating data exchange between EHR systems and other health data sources [17, 18, 30, 37, 39, 40, 42]. Moreover, some studies explicitly propose integrating controlled vocabularies with standardization technologies to achieve semantic interoperability [19–21, 31, 38], while others introduce ontologies into standardization processes to enhance interoperability [12, 13, 29, 33, 43, 44]. Controlled vocabularies provide standardized terminologies for EHR data, improving data consistency and facilitating seamless exchange between systems [15, 22, 23, 25]. Ontologies, which formally represent medical concepts and their interrelationships, support standardized data to representation and sharing, thus enhancing data standardization, interoperability, and secondary use [26, 27, 41]. Furthermore, semantic web and knowledge graph technologies —both built upon ontologies— provide standardized frameworks for data representation and integration, thereby facilitating semantic interoperability [34, 35, 47]. In terms of data quality improvement, semantic interoperability directly aligns with conformance and consistency [1], and indirectly supports other indicators such as completeness, portability, usability, applicability, and understandability.

Second, several studies explore how different semantic technologies can improve data specific data quality indicators. For example, NLP is applied in multiple stages, including word annotation [48], word embedding [28, 36, 46, 49], variable extraction model development [24], and model evaluation. Improvements in each of these stages contribute to the overall enhancement of data accuracy. In another case of controlled vocabulary technologies, Bona and Ceusters addressed semantic label alignment issues in SNOMED CT and proposed a disambiguation

solution to improve terminology quality [50]. Moreover, Qiao et al. employed knowledge graph [47] to enhance data completeness through semantic reasoning based on the contextual relationships in EHR data.

Third, beyond studies targeting specific DQ indicators, some research proposes more comprehensive frameworks based on specific semantic technologies to improve data quality. For example, Assaf and Senart [34] proposed a multi-dimensional framework for enhancing data quality throughout the semantic web construction process. Their framework includes data sources quality, raw data quality, semantic conversion quality, linking process quality, and global quality. This study demonstrates that constructing a high-standard semantic web can systematically improve the quality of data.

Discussion

This study analyzed 37 papers to explore the potential of semantic technologies to improve the data quality of EHRs. The findings reveal that with significant advancements in semantic technology over recent years, a variety of semantic technologies have influenced EHR data quality, including NLP, EHR standardization technology, controlled vocabulary, ontology, semantic web technology, and knowledge graph. Most papers concentrate on a single technology [1, 2] or scenario [3–5], such as leveraging semantic techniques to improve data interoperability. This article encompasses a wider range of six semantic techniques that have the potential to enhance EHR data quality. To facilitate a more nuanced understanding of how different semantic technologies support EHR data quality improvement, we introduce a tiered classification scheme. This scheme, briefly mentioned in the abstract, is now further elaborated to clarify both its rationale and relevance.

Diverse semantic technologies at multiple levels

To better understand how these technologies vary in complexity and application scope, we classify them into three hierarchical levels: foundational, general, and advanced. This classification is based on two criteria: (1) the depth of semantic abstraction and reasoning supported by the technology, and (2) its ability to build upon or integrate with other semantic components. Foundational technologies (e.g., NLP) focus primarily on extracting and structuring information from unstructured data, often serving as the initial layer of semantic processing. General technologies (e.g., standardization systems and controlled vocabularies) introduce codified terminologies and schema mappings to facilitate data normalization and syntactic or semantic alignment. Advanced technologies (e.g., ontologies, knowledge graph, and semantic web) enable formal semantic modeling, rule-based inference, and cross-domain integration through high-level semantic constructs. This tiered classification helps illuminate the role of each technology in the EHR data quality enhancement pipeline.

First, foundational technology. NLP technologies [22, 24, 28, 31, 36, 43, 46–49] can extract semantic entities and relationships from unstructured clinical text. This enables following semantic processes such as ontology construction, knowledge graph development, and semantic enrichment of EHRs. Given its versatility and enabling role across other semantic technologies, NLP is categorized as foundational.

Second, general technologies include EHR standardization technology [12, 13, 17–21, 24, 29–31, 33, 37–40, 42–44] and controlled vocabulary [15, 19–25, 28, 31, 36, 38, 46, 50]. These approaches provide structured schemas and controlled terminologies, enabling syntactic and semantic alignment across heterogeneous EHR systems. They are frequently used to support interoperability, ensure documentation consistency, and address issues such as missing values or erroneous terms. As such, they are positioned as general-purpose tools for routine semantic enhancement.

Third, advanced technologies consist of ontologies [12, 13, 15, 26, 27, 29, 31–33, 35, 38, 41, 43–45], semantic web [12, 13, 15, 21, 31–34, 43, 45], and knowledge graph [35, 47]. These technologies offer higher-order semantic representations and logical inference capabilities. Ontologies define formal relationships among clinical concepts, which form the backbone of semantic web frameworks and knowledge graph architectures. Due to their capacity to enable machine reasoning, cross-domain integration, and dynamic semantic services, we categorize these as advanced technologies.

Multiple application scenarios

Semantic technologies can improve EHR data quality and optimize diagnostic and treatment processes, with widespread applications in clinical and related fields. First, numerous studies suggest that semantic technologies can enhance EHR data quality, including completeness, consistency, and conformance. High-quality data can facilitate clinical diagnosis, patient management, and clinical research. Second, semantic interoperability is a key goal of data semantic research. It integrates and exchanges data between EHR systems and other health data sources via different semantic technologies, enabling population health monitoring, health policy research, and policy evaluation. Third, some studies use semantic technologies to develop clinical decision support systems on the basis of EHR data. Finally, semantic web and knowledge graph make large datasets more accessible and understandable, benefiting clinical education and training. In summary, semantic technologies have the potential to improve EHR data quality, which is beneficial for better

application in clinical diagnosis and treatment, research and decision-making, population health monitoring, public health policy research, and evaluation, as well as clinical education and training, thus laying a solid foundation for the construction of high-performance health systems in the future.

Relationships between semantic technologies and EHR data quality assessment indicators

To provide a structured understanding of how different semantic technologies contribute to specific dimensions of EHR data quality, we utilized a previously developed and validated DQ indicator system as an analytical reference [8]. Building on this foundation, this study engaged a team of semantic and clinical experts to assess the relationships between semantic technologies and these indicators comprehensively across the 37 papers reviewed (Table 4). Within our research framework, these semantic technologies exhibit a clear correlation with most indicators, whereas a theoretical link is suggested with a few others. For example, a variety of semantic technologies contribute significantly to enhancing data conformance and are intricately connected to portability, usability, and applicability. In contrast, the associations between semantic technologies and data timeliness, precision, credibility, traceability, and accessibility appear less pronounced.

In alignment with current best practices in data stewardship, particularly the FAIR principles (Findable, Accessible, Interoperable, Reusable), our review further underscores the value of semantic technologies in enabling high-quality, FAIR-aligned EHR data. Semantic technologies inherently support FAIR goals by enriching data with machine-readable meaning, standard vocabularies, and structural consistency. Recognizing this synergy, we mapped the 16 data quality indicators identified in our prior work to the four dimensions of FAIR, highlighting how each indicator contributes to making EHR data more discoverable, accessible, interoperable, and reusable (see Table 6).

As shown in Table 7, each principle aligns with a distinct set of data quality dimensions, highlighting how semantic technologies serve as an enabling infrastructure for FAIR-aligned data governance. For instance, semantic identifiers and metadata improve the findability and traceability of health records, while ontological mappings and semantic standards (e.g., SNOMED CT, HL7 FHIR) drive interoperability and reusability across diverse systems. This conceptual alignment not only reinforces the relevance of our indicator framework but also positions semantic technologies as foundational tools for achieving both high data quality and FAIR-compliant data ecosystems in healthcare.

Advantages and challenges

Semantic technologies show great potential in improving the quality and practicability of EHR data. First, semantic technologies significantly enhance the completeness, consistency, conformance, and accessibility of EHR data, which is beneficial for addressing issues such as missing values, incorrect entries, variability in documentation, and lack of standardization. In alignment with the FAIR principles, semantic technologies help make EHR data more findable and accessible by introducing standardized terminologies and persistent identifiers, which also enhance traceability and metadata quality.

Second, the use of standardized terminologies and formats allows EHR data to be exchanged between different EHR systems, achieving semantic interoperability. This enables the integration of data from various EHR systems, facilitating population health monitoring, improving care quality, conducting clinical research, and evaluating public health policies. This interoperability aligns closely with the 'I' in FAIR and contributes to the reusability of data across institutional and geographical boundaries.

Third, semantic technologies can automate some manual data processing tasks. For example, NLP technologies can automatically extract variables from text to build structured databases, saving considerable time and cost and allowing healthcare providers to focus on higher-priority tasks. This also supports the 'Reusable' dimension of FAIR by transforming unstructured content into structured, semantically rich data that can be queried, interpreted, and applied in multiple contexts.

Finally, semantic technologies can provide visualized results. For example, the graph structure of knowledge graphs might offer interpretable results for clinical decision support. By enriching both content and context, such technologies contribute to improved data understandability and usability—core aspects of reusable and high-quality health data systems.

However, despite the great potential of semantic technologies in improving EHR data quality and practicability, several critical challenges and gaps remain to be addressed. First, while semantic technologies can improve EHR data quality, they are powerless against certain factors that cause low-quality data, such as data generated by low-performance devices and errors caused by data entry personnel's negligence. This indicates that enhancing EHR data quality is a systematic project that requires the effective combination of semantic technologies with other techniques. Second, implementing semantic technologies requires significant investment in time, money, and technical personnel. This is a challenge for many resource-limited healthcare institutions, especially because of the shortage of personnel with semantic knowledge and skills, which hinders the development

Table 6 Correlation analysis of semantic technologies with EHR data quality assessment indicators

Indicators	Definition	EHR standardization	Controlled vocabularies	Ontologies	Semantic web	Knowledge graph	NLP
accuracy	The patient information recorded in the EHR system is consistent with the actual situation of the patient.	⊗	⊗	√	√	√	⊗
completeness	The patient information recorded in the EHR system is detailed and complete.	⊗	⊗	√	√	√	⊗
timeliness	The patient status recorded in the EHR system is timely and effective.	⊗	⊗	⊗	⊗	⊗	⊗
consistency	The degree of internal and external consistency of EHR data should meet the indices claimed by managers.	√	√	√	√	√	⊗
precision	The qualitative or quantitative precision of the EHR data should meet the level claimed by the dataset producer.	⊗	⊗	⊗	⊗	⊗	⊗
conformance	EHR data is stored, processed, and circulated in a standardized format.	√	√	√	√	√	√
uniqueness	There is no duplication of EHR data records; some data must remain unique.	⊗	⊗	√	√	√	⊗
credibility	EHR data is sourced from professional institutions; data is frequently reviewed.	⊗	⊗	⊗	⊗	⊗	⊗
plausibility	The value of EHR data is reasonable.	⊗	⊗	√	√	√	⊗
traceability	Ensure the auditability of the EHR data access trail and change trail.	⊗	⊗	⊗	⊗	⊗	⊗
portability	The extent to which EHR data can be stored, replaced, or transferred from one system to another while maintaining the existing quality claimed by the data producer	√	√	√	√	√	√
usability	EHR data should be available at the level claimed by the data administrator.	√	√	√	√	√	√
accessibility	EHR data can be easily accessed and extracted with a user-friendly interface.	⊗	⊗	⊗	⊗	⊗	⊗
relevance	There is some desired correlation between EHR data.	⊗	⊗	√	√	√	⊗
applicability	The content recorded in the EHR is suitable for the patient's health management and disease diagnosis and treatment; the extracted data is suitable for the research or diagnosis carried out.	√	√	√	√	√	√
understandability	The preview and interpretation levels of EHR data should be at the level claimed by the data collector.	⊗	√	√	√	√	√

√: the indicator is related to the semantic technology; ⊗: the indicator is not directly related to the semantic technology

and widespread application of these technologies. Third, at the application level, while the use of semantic technologies to enhance data interoperability is widely accepted, data interoperability raises concerns about the secondary use of patient data at the management level, such as privacy and ethical issues. These concerns touch upon the 'Accessible' and 'Reusable' dimensions of FAIR, where access control, consent, and ethical reuse must be addressed alongside technical implementation.

In summary, although semantic technologies for EHR data offer significant benefits for data quality, interoperability, and clinical applications, there are also key challenges in cost and ethics that need to be addressed. The FAIR-aligned application of semantic technologies offers a structured path forward but also highlights the need for balanced governance, cross-disciplinary expertise, and resource equity.

Limitations

This study has the following limitations. First, although the inclusion criteria were designed to ensure methodological rigor and conceptual clarity—specifically focusing on studies in which semantic technologies play a central role in improving EHR data quality—this

Table 7 Mapping FAIR principles to EHR data quality indicators through semantic technologies

FAIR Principle	Mapped Data Quality Indicators	How Semantic Technologies Enable This Alignment
Findable	traceability, uniqueness, relevance, conformance	Semantic metadata, ontologies, and identifiers enhance traceability and ensure unique, relevant representations of health data that align with standard terminologies.
Accessible	accessibility, timeliness, usability	Semantic frameworks support controlled access, temporal validation, and user-friendly retrieval through standardized APIs (e.g., FHIR).
Interoperable	consistency, portability, conformance, applicability	Ontologies and semantic mappings (e.g., SNOMED CT, openEHR, HL7 FHIR) ensure syntactic and semantic interoperability across systems and domains.
Reusable	accuracy, completeness, credibility, plausibility, understandability, precision, applicability	Semantic annotation and validation rules ensure data quality dimensions that support reliable reuse for clinical decision-making and secondary research.

approach may have inadvertently excluded some relevant studies, particularly those where such technologies are discussed but not explicitly labeled in the title. This trade-off between specificity and recall may limit the generalizability of the findings. Second, since most of the studies were conducted in high-income countries and large healthcare institutions, the results may not reflect the challenges and benefits of implementing semantic technologies in low-income countries or smaller healthcare institutions. Third, while some studies have evaluated the impact of semantic technologies on improving data quality, there is a lack of comparative empirical evidence on the effectiveness of different semantic technologies. Future research should design studies to obtain empirical evidence to compare the performance of various semantic technologies.

Conclusion

This review summarizes six semantic technologies used to enhance the quality of EHR data, including NLP, EHR standardization, controlled vocabulary, ontology, semantic web, and knowledge graph. These semantic technologies have great potential for improving the conformance, portability, usability, applicability, and interoperability of EHR data. They also contribute to making health data more Findable, Accessible, Interoperable, and Reusable (FAIR), aligning with current best practices in data stewardship. Importantly, semantic technologies can be applied beyond traditional interoperability scenarios, supporting secondary uses such as clinical research, policy evaluation, and personalized decision-making. Nonetheless, while these technologies offer substantial benefits, they are not without challenges. These include issues related to source data quality, high implementation costs, technical resource limitations, and ethical considerations such as patient privacy and consent. By mapping these technologies to a comprehensive set of 16 data quality indicators, this review provides a structured understanding of how semantics-driven approaches can systematically enhance EHR data quality.

Abbreviations

ADL	Archetype Definition Language
CCCT	Cervical Cancer Common Terminology
CCOW	Clinical Context Object Workgroup
CDA	Clinical Document Architecture
CMeSH	Chinese version of MeSH
CPs	Clinical Pathways
CRF	Conditional Random Field
DiAE	Domain-involved Archetype Editor
DQ	Data Quality
EHREMR	Electronic Health RecordElectronic Medical Record
FHIRs	Fast Health Care Interoperability Resources
HL7	Health Level 7
LAS	Language Annotating System
MeSH	Medical Subject Headings
MDE	Model-Driven Engineering
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OWL	Web Ontology Language
RDF	Resource Description Framework
RIM	Reference Information Model
R-MIM	Refined Message Information Model
SATexMCE	Semantic-Based Attention Model Using Textual Data for Medical Concept Embedding
ShEx	Shape Expressions
SPARQL	Simple Protocol and RDF Query Language
URIs	Uniform Resource Identifiers
vMR	Virtual Medical Record

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03146-w.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements

Not applicable.

Author contributions

The authors confirm their contribution to the paper as follows: Study conception and design: YW, NC and LY. Data collection and Analysis: MR, NC and LY. Writing: original draft or/and review & editing: YW and LY. All authors read and approved the final manuscript.

Funding

This work was supported by the project "Research on data quality of electronic medical records based on data semantics" of the National Social Science Foundation of China, grant number 20BTQ066.

Data availability

Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 31 March 2025 / Accepted: 6 August 2025
Published online: 11 August 2025

References

1. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, Jalaludin B, Yeo AE, Talaei-Khoei A. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inf. 2013;82(1):10–24.
2. Hammad R, Barhoush M, Abed-Alguni BH. A semantic-based approach for managing healthcare big data: a survey. J Healthc Eng. 2020;2020(1):8865808.
3. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, Schunke LC. Semantic interoperability in health records standards: a systematic literature review. Health Technol. 2022;12(2):255–72.
4. Moreno-Conde A, Moner D, Cruz WD, Santos MR, Maldonado JA, Robles M, Kalra D. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inf Assoc. 2015;22(4):925–34.
5. Amar F, April A, Abran A. Electronic health record and semantic issues using fast healthcare interoperability resources: systematic mapping review. J Med Internet Res. 2024;26:e45209.
6. Lin Y, Song BH, Duan HB, Huang FL. Overview of semantic technology and applications. App Res Comp. 2005;6:130–2. 135.
7. Cregan AM. Overview of semantic technologies. In: Rittgen P, editor. Handbook of ontologies for business interaction. Hershey: IGI Global; 2008. p. 1–20.
8. Yang L, Ren MD, Sun SF, Lu J, Wu YR. Investigation on the preferences for data quality assessment indicators of electronic health records: user-oriented perspective. JAMIA Open. 2024;7(4):ooae142.
9. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467–73.
10. Dekker R, Bekkers V. The contingency of governments' responsiveness to the virtual public sphere: a systematic literature review and meta-synthesis. Gov Inf Q. 2015;32(4):496–505.
11. Morrison A, Polisena J, Husereau D, Moulton K, Clark M, Fiander M, et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. Int J Technol Assess Health Care. 2012;28(2):138–44.
12. Ellouze AS, Bouaziz R, Ghorbel H. Integrating semantic dimension into OpenEHR archetypes for the management of cerebral palsy electronic medical records. J Biomed Inf. 2016; 63:307–24.
13. Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT. An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. J Biomed Inf. 2010;43(5):736–46.
14. Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, et al. National center for biomedical ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS. 2006;10(2):185–98.
15. Rao RR, Makkithaya K, Gupta N. Ontology based semantic representation for Public Health data integration. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I). IEEE. 2014: 357–362.
16. Yuan YL, Cao H. Semantic web, ontology and knowledge graph: an introduction and their linguistic foundation. Chin Linguistics. 2021;2021(1):8–19.
17. Min L, Tian Q, Lu X, An J, Duan H. An OpenEHR based approach to improve the semantic interoperability of clinical data registry. BMC Med Inf Decis Mak. 2018;18:15.
18. Tapuria A, Kalra D, Kobayashi S. Contribution of clinical archetypes, and the challenges, towards achieving semantic interoperability for EHRs. Healthc Inf Res. 2013;19(4):286–92.
19. de Figueiredo EB, Dametto M, de Franco Rosa F, Bonacin R. A multidimensional framework for semantic electronic health records in oncology domain. In: 2021 IEEE 30th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). IEEE. 2021:165–170.
20. Chatterjee A, Pahari N, Prinz A. HL7 FHIR with SNOMED-CT to achieve semantic and structural interoperability in personal health data: a proof-of-concept study. Sensors. 2022;22(10):3756.
21. Liu H, Hou XQ, Hu G, Li J, Ding YQ. Development of an EHR system for sharing - a semantic perspective. Stud Health Technol Inf. 2009;150:113–7.
22. Hong N, Chang F, Ou Z, Wang Y, Yang Y, Guo Q et al. Construction of the cervical cancer common terminology for promoting semantic interoperability and utilization of Chinese clinical data. BMC Med Inf Decis Mak. 2021;21(Suppl 9):309.
23. Sun H, Depraetere K, De Roo J, Mels G, De Vloed B, Twagirumukiza M, Colaert D. Semantic processing of EHR data for clinical research. J Biomed Inf. 2015;58:247–59.
24. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inf Assoc. 2018;25(5):530–7.
25. Bowles KH, Potashnik S, Ratcliffe SJ, Rosenberg M, Shih NW, Topaz M, et al. Conducting research using the electronic health record across multi-hospital systems: semantic harmonization implications for administrators. J Nurs Adm. 2013;43(6):355–60.
26. Liyanage H, Krause P, De Lusignan S. Using ontologies to improve semantic interoperability in health data. J Innov Health Inf. 2015;22(2):309–15.
27. White P, Roudsari A. An ontology for healthcare quality indicators: challenges for semantic interoperability. Stud Health Technol Inf. 2015;210:414–8.
28. Zheng T, Gao Y, Wang F, Fan C, Fu X, Li M, et al. Detection of medical text semantic similarity based on convolutional neural network. BMC Med Inf Decis Mak. 2019;19(1):1–11.
29. Kiourtis A, Nifakos S, Mavrogiorgou A, Kyriazis D. Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching. Int J Med Inf. 2019;132:104002.
30. Solbrig HR, Prud'hommeaux E, Grieve G, McKenzie L, Mandel JC, Sharma DK, Jiang G. Modeling and validating HL7 FHIR profiles using semantic web shape expressions (ShEx). J Biomed Inf. 2017;67:90–100.
31. Wang HQ, Li JS, Zhang YF, Suzuki M, Araki K. Creating personalised clinical pathways by semantic interoperability with electronic health records. Artif Intell Med. 2013;58(2):81–9.
32. Hu Z, Li JS, Zhou TS, Yu HY, Suzuki M, Araki K. Ontology-based clinical pathways with semantic rules. J Med Syst. 2012;36(4):2203–12.
33. Zhang YF, Gou L, Tian Y, Li TC, Zhang M, Li JS. Design and development of a sharable clinical decision support system based on a semantic web service framework. J Med Syst. 2016;40:118.
34. Assaf A, Senart A. Data quality principles in the semantic web. In: 2012 IEEE Sixth International Conference on Semantic Computing. IEEE. 2012:226–229.
35. Aldughayfiq B, Ashfaq F, Jhanjhi NZ, Humayun M. Capturing semantic relationships in electronic health records using knowledge graphs: an implementation using MIMIC III dataset and GRAPHDB. Healthcare. 2023;11(12):1762.
36. Jiang S, Wu W, Tomita N, Ganoe C, Hassanpour S. Multi-ontology refined embeddings (more): a hybrid multi-ontology and corpus-based semantic representation model for biomedical concepts. J Biomed Inf. 2020(111):103581.
37. Garde S, Chen R, Leslie H, Beale T, McNicoll I, Heard S. Archetype-based knowledge management for semantic interoperability of electronic health records. Stud Health Technol Inf. 2009(150):1007–11.
38. Laleci GB, Yuksel M, Dogac A. Providing semantic interoperability between clinical care and clinical research domains. IEEE J Biomed Health Inf. 2013;17(2):356–69.
39. Kilic O, Dogac A. Achieving clinical statement interoperability using R-MIM and archetype-based semantic transformations. IEEE Trans Inf Technol Biomed. 2009;13(4):467–77.
40. Koren A, Jurčević M, Prasad R. Semantic constraints specification and Schematron-based validation for internet of medical things' data. IEEE Access. 2022;10:65658–70.
41. Kiourtis A, Mavrogiorgou A, Kyriazis D. Gaining the semantic knowledge of healthcare data through syntactic models transformations. In: 2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC). IEEE. 2017:102–107.
42. Nee O, Hein A, Gorath T, Hülsmann N, Laleci GB, Yuksel M, et al. Fruntelata, SAPHIRE: intelligent healthcare monitoring based on semantic interoperability platform: pilot applications. IET Commun. 2008;2(2):192–201.

43. Dridi A, Sassi S, Chbeir R, Faiz S. A Flexible Semantic Integration Framework for Fully-integrated EHR based on FHIR Standard. In: 12th International Conference on Agents and Artificial Intelligence (ICAART). SCITEPRESS. 2020:684–691.
44. Kiourtis A, Mavrogiorgou A, Kyriazis D. Towards a secure semantic knowledge of healthcare data through structural ontological transformations. In: Knowledge-Based Software Engineering: 2018: Proceedings of the 12th Joint Conference on Knowledge-Based Software Engineering (JCKBSE 2018), Springer. 2019:178–188.
45. Wang HQ, Zhou TS, Tian LL, Qian YM, Li JS. Creating hospital-specific customized clinical pathways by applying semantic reasoning to clinical data. J Biomed Inf. 2014;52:354–63.
46. Im SJ, Xu Y, Watson J. Learning medical concept representation based on semantic information in medical textural data. Expert Syst Appl. 2024(238):122123.
47. Qiao Z, Zhang F, Lu H, Xu Y, Zhang G. Research on the medical knowledge deduction based on the semantic relevance of electronic medical record. Int J Comput Intell Syst. 2023;16(1):38.
48. Moreno-Fernandez-de-Leceta A, Lopez-Guede JM, Ezquerro Insagurbe L, Ruiz de Arbulo N, Graña M. A novel methodology for clinical semantic annotations assessment. Log J IGPL. 2018;26(6):569–80.
49. Antunes R, Silva JF, Matos S. Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings. In: 35th Annual ACM/SIGAPP Symposium on Applied Computing (SAC '20). ACM. 2020:662–669.
50. Bona JP, Ceusters W. Mismatches between major subhierarchies and semantic tags in SNOMED CT. J Biomed Inf. 2018;81:1–15.
51. Everson J, Rubin JC, Friedman CP. Reconsidering hospital EHR adoption at the dawn of HITECH: implications of the reported 9% adoption of a basic EHR. JAMIA. 2020;27(8):1198–205.
52. CHINANEWS. The Chinese government has released a draft for soliciting opinions on deepening the reform of the medical and health system. 2008. Available from: https://www.chinanews.com.cn/jk/kong/news/2008/10-14/1410793.shtml
53. Yang L. Semantics-driven improvements in electronic health records da-ta quality: a systematic review. OSF. 2025. Available from: osf.io/s4qmb

Publisher's note