

Review

Automation of Living Clinical Practice Guideline Development: Scoping Review

Abdullah Altammami^{1,2*}, MS, DMD; Olga Ares-Rufo^{3*}, MD; Anas Alzahrani^{4,5*}, MD, MPH, PhD; Abdullah Alkhotani^{4,5}, MD; Dinesh Mital², PhD; Jose F Florez-Arango³, MD, MsC, PhD

¹Health Informatics Department, College of Public Health and Health Informatics, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

²Department of Health Informatics, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States

³Center for Biomedical Informatics and Biostatistics (CB2), University of Arizona, Tucson, AZ, United States

⁴Department of Preventive Medicine and Public Health, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia

⁵King Abdulaziz University Hospital, King Abdulaziz University, Jeddah, Saudi Arabia

*these authors contributed equally

Corresponding Author:

Anas Alzahrani, MD, MPH, PhD

Department of Preventive Medicine and Public Health

Faculty of Medicine

King Abdulaziz University

80200 Abdullah Sulayman St

Jeddah, 21589

Saudi Arabia

Phone: 966 555669003

Email: ahalzhrani@kau.edu.sa

Abstract

Background: Clinical practice guidelines (CPGs) translate research evidence into actionable recommendations, but their development faces mounting challenges. With dozens of trials and systematic reviews published daily, the traditional 18- to 24-month development cycle struggles to incorporate emerging evidence, and manual screening, synthesis, and quality assessment consume thousands of person-hours per guideline. The resulting knowledge translation pipeline is slow, inconsistent, and quickly outdated.

Objective: This review aimed to chart the computational methods empirically tested for automating CPG development, the tasks they address, and their evaluation measures, and to map where that evidence is concentrated and where it is absent for living guidelines.

Methods: The review followed the Joanna Briggs Institute scoping review framework and PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidance. We searched PubMed and Scopus for studies published between January 2001 and December 2025, supplemented by backward and forward citation searching. Eligible studies used automated methods for a guideline development task, reported original research with empirical testing, and gave sufficient methodological detail. Two reviewers independently screened records and extracted data. Findings were charted across 4 pillars: retrieval, extraction, encoding, and evaluation.

Results: A total of 47 studies met eligibility criteria, spanning 2001 to 2025, with a marked increase after 2015 and a surge in large language model (LLM) work in 2024-2025. Pillar counts (nonexclusive) were as follows: retrieval, 6; extraction, 15; encoding, 26; and evaluation, 17. In retrieval pillar, query expansion achieved 80% recall, while semantic search tools traded recall for specificity (98%). In extraction, rule-based natural language processing achieved an F-measure of 81.96% for structured clinical actions and 71.75% for patient states, and hybrid ontology methods covered 86% of eligibility criteria; LLMs performed well on structured fields but degraded on interpretive content. In encoding, approaches progressed from formalisms such as Guideline Interchange Format version 3 (GLIF3) and guideline elements model toward Fast Healthcare Interoperability Resources (FHIR)-based representation; semiautomated tools assisted guideline encoding, and logic verification detected inconsistencies and coverage gaps missed by human reviewers. In evaluation, LLM appraisers assigned higher Appraisal of Guidelines for Research and Evaluation II (AGREE II) scores than human appraisers, by a mean difference of 12.5% in one study (95% limits

of agreement –30.6% to 55.5%; intraclass correlation coefficient 0.753) and by 22 to 60 percentage points across domains in another.

Conclusions: Automation shows promise for bounded, well-structured guideline development tasks, while robust evidence for end-to-end, production-scale living guideline pipelines remains limited. No included study demonstrated a validated pipeline connecting all 4 pillars in an operating guideline program, and evidence on clinical safety, workflow integration, and cost-effectiveness is largely absent. Realizing living guidelines will require interoperability standards such as FHIR, and evaluation in real guideline production rather than retrospective benchmarks.

(JMIR AI 2026;5:e103273) doi: [10.2196/103273](https://doi.org/10.2196/103273)

KEYWORDS

clinical practice guidelines; automation; artificial intelligence; machine learning; natural language processing; computer-interpretable guidelines; GRADE; AGREE II; evidence-based medicine; decision support systems; living guideline; Grading of Recommendations Assessment, Development and Evaluation; Appraisal of Guidelines for Research and Evaluation II

Introduction

Clinical practice guidelines (CPGs) translate research into actionable recommendations but face mounting challenges. Exponential literature growth—dozens of trials and systematic reviews published daily—creates information overload, making current, comprehensive guidelines increasingly difficult to maintain. Traditional manual processes are resource-intensive, time-consuming, and inconsistent across organizations [1].

Beyond initial development, translating narrative recommendations into clinical decision support systems requires manual encoding by knowledge engineers, a months-long process prone to errors. Quality assessment using AGREE II (Appraisal of Guidelines for Research and Evaluation II) and GRADE (Grading of Recommendations Assessment, Development and Evaluation) instruments is subjective and labor-intensive [2,3]. Keeping guidelines updated remains aspirational rather than a reality.

These challenges create a “knowledge pipeline” from evidence to practice that is slow, inconsistent, and rapidly outdated. Recent advances in AI, natural language processing (NLP), and knowledge representation offer promising solutions [4]. However, the landscape remains fragmented, with studies addressing isolated components using diverse methods, making it difficult to assess the state of the art.

Several reviews have examined subsets of this problem. Earlier work focused primarily on the encoding of guidelines into computer-interpretable formats, cataloging formalisms such as Asbru, Guideline Interchange Format (GLIF), and PROforma without evaluating their performance or scalability [5,6].

More recent reviews have addressed downstream applications—such as the methodology of living guidelines [7] or the integration of encoded guidelines into clinical decision support systems [8]—but none has examined the upstream process by which guidelines are extracted, encoded, and appraised in the first place. To our knowledge, no existing review maps all 4 stages of that process within a unified framework, nor covers the period of rapid large language model (LLM) adoption through 2025. This scoping review therefore aims to determine what kinds of evidence are available on the automation of CPG development, and to chart that evidence across 4 stages: retrieval, extraction, encoding, and evaluation.

Specifically, we ask: (1) what computational methods have been empirically tested for automating each stage; (2) what tasks, data sources, comparators, and outcome measures were used to evaluate them; (3) what degree of validation and human oversight each approach required; and (4) where evidence is concentrated or absent with respect to continuously updated, living guidelines.

Methods

Overview

Following the Joanna Briggs Institute (JBI) framework for scoping reviews [9], and reporting in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; [Multimedia Appendix 1](#)) [10], this review maps the available evidence on computational methods for automating CPG development across four pillars: (1) retrieval of relevant literature, (2) extraction of structured information from guideline documents, (3) encoding of guidelines into machine-interpretable formats, and (4) automated quality evaluation. The primary objective was to identify and chart the range of computational methods that have been empirically tested for automating guideline development, the tasks they address, and the outcome measures used to evaluate them, and to map the distribution of that evidence across the 4 pillars in order to expose knowledge gaps and areas of fragmentation. Consistent with scoping review methodology, the review charts what evidence exists and how it has been evaluated; it does not pool results, compare methods against one another, or critically appraise the quality of individual studies [9]. An OSF project for this review was created on March 16, 2026 [11]. The review was not prospectively registered. The project record described the review’s scope but did not specify databases, search strategies, or eligibility criteria in advance, and a registration documenting the methods as conducted was completed retrospectively in August 2026 [12]. The scope of the review evolved during conduct: it was initially framed around 3 pillars, and retrieval was subsequently added as a fourth.

Search Strategy and Eligibility

We searched PubMed and Scopus on August 10, 2026, for studies published between January 1, 2001, and December 31,

2025. The start date reflects the emergence of shareable computer-interpretable guideline (CIG) formalisms [13].

The search combined 2 sets of terms with OR. The first comprised phrases naming guideline formalisms, appraisal instruments, and automation tasks (eg, “computer-interpretable guideline,” “GLIF3,” “APPRAISE-RS,” and “AGREE II scores”). The second comprised paired terms, each combining a guideline concept with an automation, computational, or process concept (eg, “large language model” AND “clinical practice guideline”; “information extraction” AND “clinical guideline”). Terms were drawn from the vocabulary of studies known to the review team and refined iteratively; the search was not peer-reviewed. Full strategies for both databases are given in [Multimedia Appendix 1](#).

Reference lists of included studies were screened and forward citation searching performed, identifying 8 further records. Gray literature, preprints, and dissertations were not searched. Records were imported into Covidence (Future Evidence Foundation), where 898 duplicates were removed before screening.

Inclusion Criteria

Studies were eligible if the automated component acted on a guideline development task as defined below, reported original research with empirical testing, and provided sufficient methodological detail for assessment.

Exclusion Criteria

Studies were excluded if they addressed guideline implementation or adherence only, were opinion pieces without primary data, or were not published in English.

Operational Definitions and Pillar Assignment

For the purposes of this review, CPG development refers to the processes by which a guideline is produced or updated: identification and screening of evidence, extraction and structuring of evidence and recommendation content, representation of recommendation logic in a machine-interpretable form, and appraisal of the resulting evidence or recommendations. It excludes processes downstream of a completed guideline, namely dissemination, point-of-care execution of an existing guideline in a decision support system, and measurement of clinician adherence. A study was eligible only if the automated component acted on a guideline development task as defined above. Where a computable artifact was authored as part of guideline production, it was treated as development (encoding); where an existing guideline was executed against patient data at the point of care, it was treated as implementation and excluded. A living guideline was defined as a guideline maintained through continuous or scheduled surveillance of new evidence with incremental updating of affected recommendations, rather than through periodic full revision. Studies were assigned to a pillar according to the task the automated component performed: retrieval, where the component identified or screened candidate evidence; extraction, where it parsed narrative text into structured elements; encoding, where it produced or verified a formal, machine-interpretable representation; and evaluation, where it scored or appraised evidence or recommendation quality against an instrument such

as GRADE or AGREE II. Studies whose components spanned more than one task were assigned to every applicable pillar, so pillar counts are not mutually exclusive.

Data Extraction and Synthesis

Titles and abstracts were screened independently and in duplicate by 2 reviewers in Covidence systematic review software, followed by independent duplicate full-text screening against the same criteria. Before formal screening, both reviewers piloted the eligibility criteria on a sample of records and met to resolve interpretive differences; formal screening proceeded once agreement was judged acceptable. Disagreements at either stage were resolved by discussion, with a third reviewer available for arbitration where consensus could not be reached. Records retained after full-text screening were assessed against the operational definitions below. Two reports were excluded at this stage: one was a review and analysis of evaluation criteria rather than original research with empirical testing [14], and one met all other criteria but was published in Spanish [15].

- Pillar 1: Retrieval encompasses computational methods for identifying and retrieving relevant evidence from published literature, including automated search and screening tools [16-18].
- Pillar 2: Extraction encompasses computational methods for parsing narrative guideline text into structured data, including rule-based NLP, ontology-driven approaches, and LLM-based extraction systems [19].
- Pillar 3: Knowledge encoding refers to translating narrative clinical guidelines into formal, machine-interpretable representations that enable automated reasoning and decision support [13,20]. This encompasses representation formalisms (GLIF, guideline elements model [GEM], Arden Syntax, and PROForma) and semiautomated authoring tools. This pillar transforms guidelines from passive documents into active computational artifacts capable of logical verification [21,22].
- Pillar 4: Automated evaluation involves computational approaches to evaluating guideline quality and evidence strength using standardized instruments such as GRADE and AGREE II [23-26].

Ethical Considerations

This scoping review synthesized data from previously published, publicly available studies. It involved no human participants or animals and no access to individual-level or identifiable patient data. Institutional review board approval and informed consent were therefore not required, consistent with the policies of King Abdulaziz University. No primary data were collected. All included studies are cited in full, and no reproduction of copyrighted material beyond fair academic use was undertaken.

Results

Study Selection

The search identified 2577 records: 998 from PubMed, 1571 from Scopus, and 8 through backward and forward citation searching. After removal of 898 duplicate records, 1679 proceeded to title and abstract screening. Two reviewers

independently assessed records against the predefined eligibility criteria, with disagreements resolved as described in the “Methods” section. Of these, 1450 records were excluded.

The remaining 229 records were sought for full-text retrieval; 25 could not be obtained, leaving 204 reports for full-text eligibility assessment. Full-text exclusions totaled 157: 61 addressed evidence synthesis or not targeting CPG development automation (R1); 41 focused on clinical decision support or point-of-care use without automating any guideline development process (R2); 29 lacked original research on CPG automation and were reviews, opinions, commentaries, or protocols (R3); 20 did not present a transferable computational method for CPG development, including disease-specific tools or essentially manual approaches (R4); 5 provided insufficient methodological

detail to assess the computational approach (R5); and 1 met all other criteria but was not published in English (R6).

The final corpus of 47 studies [17-63] spans 2001-2025 and is characterized in Table 1 and detailed study by study in Table 2. Performance metrics by automation task are summarized in Table 3. The selection process is summarized in Figure 1. Publications showed a marked increase after 2015, with a specific surge in LLM-related work during 2024-2025. The relatively high attrition from 2577 records to 47 inclusions reflects the breadth of the search strategy alongside the specificity of the scope: studies empirically testing computational methods for automating the guideline development process itself—as opposed to using or implementing existing guidelines—represent a considerably smaller subset of the biomedical literature.

Table 1. Summary statistics of included studies (N=47 unique studies).

Characteristic	Studies, n (%)
Automation pillar^a	
Retrieval	6 (13)
Extraction	15 (32)
Knowledge encoding	26 (55)
Automated evaluation	17 (36)
Transparency	
Code or algorithm shared	12 (26)
Interrater reliability reported	18 (38)

^aPillar counts are nonexclusive: 47 unique studies contributed 64 pillar assignments, as several studies addressed more than one pillar.

Table 2. Characteristics of included studies categorized by pillar and approach type.

Study	Year	Country	Domain	Retrieval (n=6)	Extraction (n=15)	Encoding (n=26)	Evaluation (n=17)	Approach type
Duftschnid and Miksch [21]	2001	Austria	Neonatology			✓		Logic verification
Duftschnid et al [28]	2002	Austria	Neonatology			✓		Temporal reasoning
Georg et al [29]	2003	France	Hypertension			✓		GEM ^a encoding
Peleg et al [27]	2003	United States	Cross-domain			✓		CIG ^b formalism comparison
Boxwala et al [30]	2004	United States	Cross-domain			✓		GLIF3 ^c format
Shiffman et al [31]	2005	United States	Cross-domain			✓	✓	GLIA ^d instrument
Anselma et al [32]	2006	Italy	Cross-domain			✓		Temporal formalism
Hagerty et al [33]	2006	United States	Hypertension		✓	✓		ML ^e classifier
Kaiser et al [18]	2007	Austria	General		✓			NLP ^f (LASSIE ^g)
Hrabak et al [34]	2007	United States	Immunization domain			✓		Terminology standards
Serban et al [35]	2007	Netherlands and Spain	Oncology		✓			NLP pattern extraction
Peleg et al [36]	2008	Israel	Cross-domain			✓		CIG-EMR ^h mapping (KDOM ⁱ)
Fizman et al [37]	2008	United States	cardiovascular	✓	✓			Semantic (SemRep)
Bouaud et al [22]	2009	France	Hypertension			✓		Completeness check
Kaiser and Miksch [38]	2009	Austria	General		✓	✓		Living Guidelines (DELT/A ^j)
Peleg and Garber [39]	2010	Israel	Thyroid			✓	✓	Extended GLIA
De Backere et al [40]	2012	Belgium	ICU ^k			✓		CIG execution
Slaughter et al [41]	2013	United States	Cross-domain			✓		Ontology (C2PO ^l)
Zhu et al [19]	2013	China	Cross-domain		✓			Rule-based NLP
Farkash et al [42]	2013	Israel	Hypertension			✓		Model-driven engineering
Minard and Kaiser [43]	2013	Austria	Cross-domain		✓	✓	✓	ML action classification
Wenzina and Kaiser [63]	2013	Austria	Cross-domain		✓	✓	✓	Heuristic IE ^m + UMLS ⁿ semantic patterns
Wenzina and Kaiser [44]	2014	Austria	Cross-domain			✓		TimeML modeling
Galopin et al [45]	2014	France	Hypertension				✓	Ontology evaluation
Weng et al [46]	2014	United States	Cross-domain		✓			Hybrid NLP and ontology
Bui et al [17]	2015	United States	Multidomain	✓	✓			Query expansion
Stewart et al [24]	2015	United Kingdom	Multidomain				✓	Bayesian network
Llewellyn et al [47]	2015	United Kingdom	Multidomain				✓	Bayesian network

Study	Year	Country	Domain	Retrieval (n=6)	Extraction (n=15)	Encoding (n=26)	Evaluation (n=17)	Approach type
Hu et al [48]	2016	Netherlands	Cross-domain (breast cancer)	✓			✓	Semantic distance (topic-centric)
González-Ferrer et al [49]	2017	Spain	Hyponatremia			✓		PROForma encoder
Afzal et al [50]	2019	South Korea	Multidomain				✓	Context-aware GRADE ^o
Schlegel et al [51]	2019	United States	Cross-domain (diabetes)		✓	✓		NLU ^p (clinical tractor)
Iglesias et al [20]	2020	Spain	Infectious			✓		Formalism comparison
Fux et al [52]	2020	Israel	Asthma			✓		Multilayer PROForma
Shivers et al [53]	2021	United States	Family planning and STI ^q			✓		FHIR ^r encoding (WHO ^s DAK ^t)
Borchert et al [54]	2021	Germany	Oncology	✓	✓		✓	NLP disagreement detection
Šuster et al [25]	2023	Australia	Multidomain		✓		✓	Transformer ML
Lin et al [55]	2023	China	Musculoskeletal	✓				Deep learning screening
Lichtner et al [56]	2023	Germany	Infectious disease			✓		FHIR CPG-on-EBM ^u
López et al [57]	2023	Spain	Psychiatry (ADHD ^v)				✓	GRADE Automation (APPRAISE-RS)
Wan et al [26]	2025	China	Surgery				✓	LLM ^w (AGREE II ^x and RIGHT ^y)
Dos Santos et al [23]	2025	Brazil	Multidomain				✓	ML GRADE (URSE)
Wang et al [58]	2025	China	Imaging				✓	LLM evaluation
Hsu et al [59]	2025	Taiwan	Pharma		✓			LLM extraction
Boyle et al [60]	2025	Canada	Cross-domain	✓	✓			LLM evaluation
Wang et al [61]	2025	China	Drug Monitor				✓	LLM evaluation

Study	Year	Country	Domain	Retrieval (n=6)	Extraction (n=15)	Encoding (n=26)	Evaluation (n=17)	Approach type
von Dincklage et al [62]	2025	Germany	Intensive Care			✓	✓	FHIR (CPG-on-EBMonFHIR)

^aGEM: guideline elements model.

^bCIG: computer-interpretable guideline.

^cGLIF3: Guideline Interchange Format version 3.

^dGLIA: Guideline Implementability Appraisal.

^eML: machine learning.

^fNLP: natural language processing.

^gLASSIE: modeling treatment processes using information extraction.

^hEMR: electronic medical record.

ⁱKDOM: Knowledge-Data Ontological Mapper.

^jDELT/A: Document Exploration and Linking Tool/Addons.

^kICU: intensive care unit.

^lC2PO: Core Clinical Protocol Ontology.

^mIE: information extraction.

ⁿUMLS: Unified Medical Language System.

^oGRADE: Grading of Recommendations Assessment, Development and Evaluation.

^pNLU: natural language understanding.

^qSTI: sexually transmitted infection.

^rFHIR: Fast Healthcare Interoperability Resources.

^sWHO: World Health Organization.

^tDAK: Digital Adaptation Kit.

^uEBM: evidence-based medicine.

^vADHD: attention-deficit/hyperactivity disorder.

^wLLM: large language model.

^xAGREE II: Appraisal of Guidelines for Research and Evaluation II.

^yRIGHT: Reporting Items for Practice Guidelines in Healthcare.

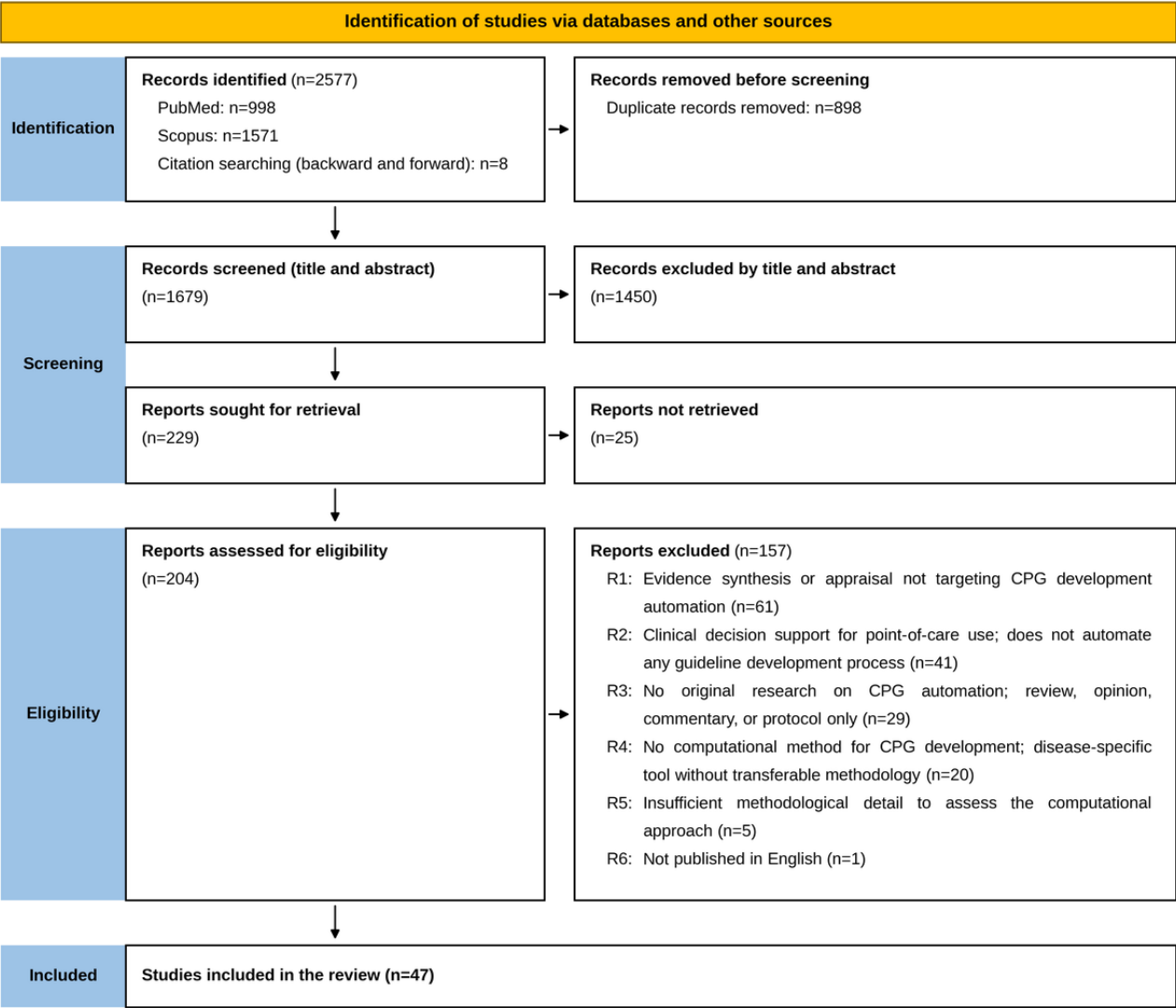
Table 3. Expanded performance metrics by automation task.

Task	Study	Approach	Performance metrics
Literature retrieval	Bui et al [17]	Query expansion	Recall 80%
Literature retrieval	Fiszman et al [37]	Semantic (SemRep)	Precision 88%; Recall 40%; Specificity 98%
Literature retrieval	Lin et al [55]	Deep learning (PAJO ^a [PubMed-BERT-based])	AUC ^b 91.84%; F_1 -score +1.91% vs baseline; dataset: 89,940 PubMed articles (neck pain)
Literature retrieval	Hu et al [48]	Topic-centric semantic-distance retrieval	Goal evidence identified for 12/16 guideline statements vs 5/16 with non–topic-centric retrieval; evidence-item coverage 46% vs 57%
Info extraction	Kaiser et al [18]	NLP ^c (LASSIE ^d)	Recall 0.84; Precision 0.87 (treatment steps)
Info extraction	Zhu et al [19]	Rule-based NLP	F-measure: 81.96% (actions), 71.75% (patient states)
Info extraction	Weng et al [46]	Hybrid ontology	86.2% coverage of eligibility criteria
Info extraction	Hsu et al [59]	LLM ^e (GPT-4o; Gemini-1.5-Pro)	Accuracy 93.5% / 92.7%; Cohen κ 0.90/0.89 (385 PharmGKB annotations, 4-class)
Info extraction	Šuster et al [25]	Transformer (EvidenceGRADER)	F_1 -score 0.74 (overall); F_1 -score 0.78 (RoB ^f); F_1 -score 0.31 (inconsistency)
Info extraction	Serban et al [35]	Ontology-driven linguistic pattern extraction (Asbru and MHB ^g)	Recall 72%-80%; Precision 20–35%; breast cancer guideline (Asbru and MHB formalization)
Info extraction	Wenzina and Kaiser [63]	Heuristic IE ^h + UMLS ⁱ semantic patterns	Recall 75%; Precision 88%; NPV ^j 85%; IF-patterns: Recall 82%, Precision 93%
Encoding	Hagerty et al [33]	Semiautomated	Precision 72%-79% (initial heuristics)
Encoding	Georg et al [29]	GEM ^k vs Manual	Broader coverage (104 vs 98 rules); 37% identical logic
Encoding	González-Ferrer et al [49]	PROForma	Agreement with experts: Cohen κ =0.86 (overall), Cohen κ =0.80 (top diagnosis)
Verification	Bouaud et al [22]	Formal logic	Completeness: 12.6% theoretical coverage; 55% patient coverage
Evaluation	Stewart et al [24]	Bayesian (SAQAT ^l)	Agreement with experts: Cohen κ =0.69
Evaluation	Dos Santos et al [23]	ML (URSE) for GRADE ^m	Accuracy: 97% (imprecision), 90% (heterogeneity), 73% (RoB)
Evaluation	Wan et al [26]	LLM (GPT-4o) for AGREE II ⁿ and RIGHT ^o	Domain gaps +22 to +60 pp (LLM > manual, all 6 domains); no ICC ^p or Cohen κ reported
Evaluation	Wang et al [61]	LLM (GPT-4o) for AGREE II	ICC 0.753 (95% CI 0.532-0.854); mean difference 12.5%, LLM higher (LoA ^q -30.6% to 55.5%)
GRADE evaluation	López et al [57]	Rule-based GRADE automation (APPRAISE-RS)	Avg 1.96 recommendations per patient (28 simulated); clinician satisfaction median 2.0/3.0; 348 RCTs ^r (ADHD ^s)

^aPAJO: Paper title, Abstract, and Journal.^bAUC: area under the receiver operating characteristic curve.^cNLP: natural language processing.^dLASSIE: modeling treatment processes using information extraction.^eLLM: large language model.^fRoB: risk of bias.^gMHB: Many-Headed Bridge.^hIE: information extraction.ⁱUMLS: Unified Medical Language System.^jNPV: negative predictive value.^kGEM: guideline elements model.^lSAQAT: Semi-Automated Quality Assessment Tool.^mGRADE: Grading of Recommendations Assessment, Development and EvaluationⁿAGREE II: Appraisal of Guidelines for Research and Evaluation II.

^oRIGHT: Reporting Items for Practice Guidelines in Healthcare.
^pICC: intraclass correlation coefficient.
^qLoA: limits of agreement.
^rRCT: randomized controlled trial.
^sADHD: attention-deficit/hyperactivity disorder.

Figure 1. Study selection flow diagram presented using the PRISMA 2020 flow diagram template as adapted for scoping reviews (PRISMA-ScR). Adapted from Page et al [64], which is published under the Creative Commons Attribution 4.0 International License. CPG: clinical practice guideline; R: reason for exclusion.



Pillar 1: Retrieval

Six studies investigated the automation of literature search and screening, revealing a trade-off between efficiency and comprehensive recall [17,37,48,54,55,60]. Query expansion methods improved literature retrieval, achieving 80% recall in identifying relevant citations [17]. Deep learning approaches demonstrated additional promise: Lin et al [55] proposed the Paper title, Abstract, and Journal (PAJO) model, combining PubMedBERT with journal-level features for automated literature screening in CPG development, achieving an area under the receiver operating characteristic curve (AUC) of 91.84% on a neck pain evidence corpus and outperforming text classification baselines by 1.91% in F_1 -score.

However, the transition to semantic and generative search tools highlighted limitations in reliability. While semantic processing tools (eg, SemRep [US National Library of Medicine]) achieved high specificity (98%) and precision (88%) in answering clinical questions, they suffered from low recall (40%), rendering them more suitable for filtering than for the comprehensive searching required in systematic reviews [37]. Furthermore, recent evaluations of generative AI indicated that while LLMs can successfully generate search syntax, they currently fail to independently execute retrieval or screening with the rigorous reliability necessary for systematic reviews [60].

Pillar 2: Extraction

Fifteen studies addressed the extraction of structured data from narrative text, showing a clear performance divide between rule-based systems and generative models [17-19,25,33,35,37,38,43,46,51,54,59,60,63]. Previously, ontology-driven approaches applied linguistic pattern extraction to map breast cancer guideline text to formal representations (Asbru/MHB [Many-Headed Bridge]), achieving recall of 72%-80% and demonstrating the feasibility of semiautomated formalization from narrative clinical text [35].

Early rule-based NLP systems demonstrated high reliability in specific domains; systems using Unstructured Information Management Architecture (UIMA) and MetaMap (US National Library of Medicine) achieved an F-measure of 81.96% for extracting clinical actions and >70% for fine-grained attributes such as dosage [19]. Hybrid approaches combining NLP with upper-level ontologies (eg, EliXR) successfully covered 86.2% of eligibility criteria in free-text guidelines [46].

In contrast, modern LLM approaches yielded mixed results. While capable of producing structurally coherent recommendations, LLMs demonstrated an accuracy gap between overall performance (77% accuracy) and complex evaluation items (29%-40% accuracy) [58]. Critical evaluations emphasized that LLMs struggle with extraction accuracy compared to human performance, necessitating human-in-the-loop verification to prevent data hallucinations [60].

Pillar 3: Knowledge Encoding

Overview

Twenty-six studies addressed the translation of narrative guidelines into computable formats, highlighting the progression from static formalisms to dynamic, living frameworks [20-22,27-34,36,38-44,49,51-53,56,62,63].

Formalisms and Standards

The foundational work in this pillar established flow-chart-based representations and rule-based models. Standards such as GLIF3 (GLIF version 3) were developed to enable shareable, nested, and iterative guideline representations [30], while the GEM was shown to produce rule bases with broader coverage of clinical situations than manually created alternatives [29].

The evolution of encoding standards was accompanied by advances in interoperability and integration. Peleg et al [27] systematically compared 6 major CIG formalisms (Asbru, EON, GLIF, GUIDE, PRODIGY [Prescribing Rationally with Decision Support in General Practice Study], and PROforma) using real clinical guidelines, identifying key modeling differences and paving the way toward consensus standards. Building on this foundation, the same group developed KDOM (Knowledge-Data Ontological Mapper), a framework that automatically generates SQL queries to map GLIF3-encoded guideline data items to institutional electronic medical record (EMR) schemas, enabling cross-institutional sharing of CIGs without manual reencoding [36].

More recently, Fast Healthcare Interoperability Resources (FHIR)-based encoding approaches have extended this

interoperability to internationally curated resources. Shivers et al [53] demonstrated automated generation of FHIR Implementation Guides from World Health Organization (WHO) Digital Adaptation Kit data dictionaries for family planning and sexually transmitted infection (STI) guidelines, mapping clinical concepts to standard terminologies (LOINC [Logical Observation Identifiers Names and Codes], SNOMED-CT [Systematized Nomenclature of Medicine Clinical Terms], ICD-10 [International Statistical Classification of Diseases, Tenth Revision], and RxNorm). Further advancing this line of work, an implementation guide directly linking structured evidence artifacts from evidence-based guideline development to computable FHIR CPG recommendations was validated with the COVID-19 dexamethasone recommendation, covering Population, Intervention, Comparison, and Outcome (PICO) relationships, metadata, and GRADE evidence-to-decision justification [56].

Semiautomation and Efficiency

Technological interventions have measurably impacted the efficiency of the encoding process. Semiautomated authoring tools have been shown to predict condition and recommendation terms with 72%-79% precision [33]. To address the “living” guideline challenge, integrated systems combining extraction tools with versioning systems (eg, DELT/A [Document Exploration and Linking Tool/Addons]; Vienna University of Technology) allowed for modular updates. These frameworks enabled new guideline versions to inherit unchanged logic from previous iterations, requiring reencoding only for modified text sections [38].

Logic and Verification

A distinct advantage of computational encoding identified across multiple studies is the capacity for automated logic verification. Formal encoding methods revealed significant coverage gaps in narrative text, with one study finding that only 12.6% of theoretical patient profiles were explicitly covered by traditional guidelines [22]. Computational logic verification successfully identified anomalies often missed by human reviewers, including contradictory recommendations, dead-end paths [21,28], and temporal conflicts such as overlapping medication schedules [32].

Different formalisms demonstrated specific utility based on the use case: ontologies were found superior for knowledge sharing, while production rules (eg, Drools) excelled at execution speed [20]. In complex clinical scenarios such as hyponatremia, the PROforma formalism achieved high agreement (Cohen κ 0.86) with expert decisions [49], while multilayered models were proposed to separate core guidelines from local adaptations to simplify the update process [52]. Farkash et al [42] used model-driven engineering with Health Level Seven (HL7) standards to produce semiautomated representations readable by physicians.

Pillar 4: Automated Evaluation

Seventeen studies explored automated quality assessment [23-26,31,39,43,45,47,48,50,54,57,58,61-63].

LLM-Based Evaluation

The application of LLMs to the AGREE II instrument was a dominant theme in the 2024-2025 literature. Both studies that compared LLM and human AGREE II appraisal found the same qualitative pattern: the LLM assigned higher absolute scores than human appraisers. The magnitude differed markedly and the two are not on a common scale. Wang et al [61] reported a mean difference of 12.5% across 28 therapeutic drug monitoring guidelines, with 95% limits of agreement of -30.6% to 55.5% and an intraclass correlation coefficient of 0.753. Wan et al [26] report domain-level gaps of 22 to 60 percentage points across all 6 AGREE II domains in 54 documents on day surgery, with no intraclass correlation coefficient or Cohen κ reported. This bias was particularly evident in domains such as “Stakeholder Involvement,” where models tended to hallucinate compliance when details were missing. Despite these accuracy issues, the efficiency gains were substantial, reducing appraisal time from 1.5 hours to under 3 minutes per guideline [61]. Furthermore, when applied to reporting checklists (RIGHT [Reporting Items for Practice Guidelines in Healthcare]), LLMs scored significantly higher than humans (gap ~25%), suggesting a sensitivity to form over substance [26].

Consistency and GRADE

Beyond generative AI, formal and machine learning methods demonstrated high utility for specific evaluation tasks. Ontological modeling was successfully used to map patient profiles across different guidelines, confirming logical consistency despite textual variations [45].

For evidence grading, transformer classifiers (EvidenceGRADER) achieved an overall F_1 -score of 0.74, with 0.78 for risk of bias but only 0.31 for inconsistency [25]. A separate machine learning tool for GRADE rating reached 97% accuracy for imprecision, 90% for heterogeneity, and 73% for risk of bias [23]. Performance therefore varies markedly by domain and the two tools are not directly comparable. Bayesian networks provided a structured alternative, achieving substantial agreement (Cohen κ =0.69) with expert judgment, thereby offering a replicable method for quality assessment [24].

Complementing these approaches, APPRAISE-RS is a methodology that automatically applies iterative GRADE criteria to a database of 348 randomized controlled trials (RCTs), generating personalized treatment recommendations. This system achieved good clinician satisfaction in 28 simulated patients and demonstrated the feasibility of automated, continuously updated evaluation at scale [57].

Discussion

Summary of Evidence

This scoping review identified 47 studies addressing the automation of clinical practice guideline development, spanning 2001-2025 across 4 pillars [17-63]. In the retrieval pillar (n=6) [17,37,48,54,55,60], the transition from keyword-based filtering to transformer-based classifiers represents the most consequential methodological shift, with deep learning approaches achieving AUC values exceeding 90% on structured evidence corpora [55]. In the extraction pillar (n=15)

[17-19,25,33,35,37,38,43,46,51,54,59,60,63], ontology-driven linguistic pattern systems established the baseline [35], while more recent NLP pipelines have extended coverage to complex conditional and temporal constructs; however, recall rates for free-text guideline sentences indicate that human oversight remains necessary [59,60]. In the encoding pillar (n=26) [20-22,27-34,36,38-44,49,51-53,56,62,63], 2 decades of fragmented formalism development have been followed by recent work oriented toward FHIR-based representation. Two studies in this corpus [56,62] demonstrate FHIR implementation guides that produce interoperable, incrementally updatable guideline artifacts compatible with existing health information systems. Whether this represents convergence across the field or an emerging direction cannot be determined from 2 studies [56,62] published within 2 years of each other. In the evaluation pillar (n=17) [23-26,31,39,43,45,47,48,50,54,57,58,61-63], LLM-based tools produced structurally complete AGREE II and GRADE appraisals but assigned higher absolute scores than human appraisers in both studies that made the comparison, with wide limits of agreement [26,61]. This supports their use for triage and prioritization rather than as standalone quality arbiters. Across all pillars, no study has yet demonstrated a fully integrated, validated end-to-end pipeline connecting automated retrieval, extraction, encoding, and evaluation in a production clinical environment. This fragmentation represents the defining knowledge gap of the field.

Principal Results

A caution applies to every performance figure reported in this review. The included studies used noncommensurable outcome measures—precision, recall, AUC, F-measure, accuracy, Cohen κ , intraclass correlation coefficient, logical coverage, time saved, and clinician satisfaction—assessed against different reference standards, on corpora ranging from 28 guidelines to 89,940 abstracts, and under evaluation conditions ranging from retrospective benchmarking to simulated cases. These values cannot be pooled, ranked, or read as a common scale of maturity. It is more informative to distinguish the dimensions on which evidence exists from those on which it does not. Across the corpus, technical feasibility is well established in all 4 pillars and accuracy in controlled settings is documented for many individual tasks. Reproducibility is reported in only a small number of studies. Interoperability is addressed by the recent FHIR work [56,62]. Clinical safety, workflow integration, scalability, cost-effectiveness, and real-world adoption are essentially unevidenced: no included study reported an evaluation of an automation component embedded in an operating guideline program. The most defensible reading of this literature is therefore that automation demonstrates clear promise for bounded, structured tasks within guideline development, while robust evidence for end-to-end, production-scale living guideline pipelines remains limited.

This scoping review establishes that automated technologies for guideline development have progressed from theoretical proof-of-concept to practical, augmentative tools. A clear dichotomy exists in the current state of the art: rule-based and ontology-driven methods offer the rigorous reliability required for clinical safety and logic verification[20,21,49], while emerging LLMs offer the processing speed needed to handle

vast literature volumes [59,61]. However, the latter is currently hindered by “hallucinations” and systematic scoring biases [60,61]. Consequently, a fully automated pipeline remains elusive, with the most successful models using automation for data-intensive retrieval and initial encoding [17,33], while retaining human oversight for contextual interpretation [60,65].

Comparison With Prior Work

The Evolution of Encoding

Encoding represents the highest-leverage investment in the guideline pipeline because it uniquely enables quality gains that narrative guidelines cannot deliver [32,33]. The progression from early formalisms such as GLIF3 and GEM to modern hybrid ontology approaches highlights a shift toward interoperability [29,30]. Crucially, the literature confirms the feasibility of “living” guidelines through modular updates. Systems that allow new guideline versions to inherit unchanged logic from previous versions address the core challenge of keeping pace with daily evidence publication [38,52], a capability that purely narrative updating methods lack.

The inclusion of recent FHIR-based studies strengthens this trajectory considerably. Recent work demonstrated that structured evidence artifacts from the guideline development workflow can now be directly translated into computable FHIR recommendations using WHO Digital Adaptation Kit data dictionaries, mapping clinical concepts to standard terminologies (LOINC, SNOMED-CT, ICD-10, and RxNorm) [53]. Building on this, an implementation guide directly linking evidence artifacts from guideline development to computable FHIR CPG recommendations was validated against a real clinical scenario—the COVID-19 dexamethasone recommendation—covering PICO relationships, metadata, and GRADE evidence-to-decision justification [56]. This represents a meaningful step beyond earlier encoding frameworks: where GLIF3 and GEM required significant manual knowledge engineering, modern FHIR-based approaches leverage internationally curated data structures to reduce that burden. Equally, the KDOM framework established the foundational principle of automatic schema mapping between CIG formalisms and institutional EMR data models—a precondition for any truly deployable living guideline system [36]. Together, these studies suggest the field is converging toward a common interoperability layer that was largely absent from earlier reviews.

The Role of Generative AI

The integration of recent findings positions LLMs as assistants rather than replacements. A comparison of retrieval and evaluation studies reveals a critical safety limitation: LLMs exhibit reliable structural outputs but struggle with independent factual recall and objective evaluation [59,60]. Both studies that compared LLM and human AGREE II appraisal found the LLM assigned higher scores on average, though the magnitude differed markedly (mean difference 12.5%, 95% limits of agreement –30.6% to 55.5%, in 28 therapeutic drug monitoring guidelines [61]; 22 to 60 percentage points across domains in 54 documents on day surgery [26]). Because the direction was not consistent across guidelines of differing quality [61], these

findings support the use of LLMs for summarization and triage rather than for definitive quality assurance. The consensus across the reviewed studies indicates that Generative AI must be constrained by formal verification methods (logic checks) and human review to be safely deployed in clinical guideline development [26,60].

Integration of Pillars

The four pillars—retrieval, extraction, encoding, and evaluation—are functionally interdependent. Extraction provides raw material for encoding [18,46]; encoding facilitates automated appraisal of logical consistency [22,28]; and appraisal determines which evidence warrants extraction [15,17]. However, the literature remains fragmented, with few studies bridging these pillars. The path forward identified by this review is the development of end-to-end pipelines where extracted text is automatically mapped to standards such as HL7 FHIR, then verified by logic engines [42].

One study in this corpus provides early empirical evidence that this loop is achievable: starting from a database of 348 RCTs, APPRAISE-RS automatically applied iterative GRADE methodology to generate personalized treatment recommendations for individual patients [57]. While limited to a single disease area (attention-deficit/hyperactivity disorder) and evaluated on simulated cases, it illustrates what a fully automated living guideline pipeline could look like in practice.

Practical Implications for Digital Health

The findings of this review have direct implications for digital health developers, guideline bodies, and health informaticians. In the retrieval and extraction pillars, transformer-based classifiers and rule-based NLP tools are sufficiently mature for use as screening and extraction assistants in literature surveillance workflows, provided human oversight is retained for borderline cases and interpretive content [19,46,55].

In the encoding pillar, FHIR-based authoring tools—specifically the WHO DAK-to-FHIR pipeline, developed by Shivers et al [53] in collaboration with the WHO, and CPG-on-EBM-on-FHIR, developed by Lichtner et al [56] as part of the COVID-19 Evidence Ecosystem (CEOsyst) project—are the outputs of this research program closest to practical deployment. They offer standardized, interoperable artifacts that can integrate with existing health information systems and support incremental updates. Neither has yet been evaluated in routine guideline production.

In the evaluation pillar, LLM-based tools show promise for ranking and flagging guideline quality but should not yet be used as stand-alone assessors given the documented upward scoring bias [26,61]; hybrid workflows combining LLM screening with human confirmation represent the most defensible current approach.

The most significant knowledge gap identified by this review is the absence of validated end-to-end systems connecting all 4 pillars: no existing study has demonstrated a complete pipeline from automated literature retrieval through extraction, encoding, and evaluation to recommendation generation at scale. This represents the primary development priority for the field.

Limitations

Several features of the search constrain what this review can claim. The search covered 2 databases: PubMed and Scopus. Scopus was included to reach the informatics and engineering literature that a biomedical database indexes incompletely. Even so, it is likely that IEEE Xplore, the ACM Digital Library, and Web of Science would surface further work, particularly in conference proceedings. The search strategy is a building-block design that pairs guideline terms with named formalisms, tasks, and instruments; it favors precision over recall and may underretrieve studies describing comparable methods in vocabulary outside the enumerated pairings, and it was not formally peer-reviewed. Gray literature, preprints, and dissertations were not searched. Eligibility was restricted to English-language publications; one otherwise eligible study was excluded on this basis, and the restriction may underrepresent work published in other languages, particularly from European, Latin American, and East Asian informatics groups active in this field. The search window opens in January 2001, and earlier work is therefore not represented.

The synthesis of evidence was significantly challenged by the substantial study heterogeneity across the included literature. This variability encompassed differences in study design, target clinical domains, specific automation technologies evaluated, and reported outcomes, which collectively precluded the execution of a formal meta-analysis. Consequently, the findings presented are primarily based on a narrative and descriptive synthesis.

Furthermore, it is essential to contextualize the findings within the rapidly evolving landscape of the field, particularly concerning advancements in sophisticated language models such as LLMs and other forms of generative AI [4,61]. Given the pace of innovation, some of the specific technological findings, or the general technological state described in the older included literature (eg, pre-2015 NLP methods), may be potentially dated. Future research will need to continuously update this evidence base to remain current with emerging technologies.

A significant limitation observed across the corpus was that most studies were proof-of-concept demonstrations or early-stage feasibility assessments (Table 2). They often focused on showing what was technologically possible in a controlled environment. As a result, there was a limited body of evidence pertaining to real-world implementation, practical utility, and sustained impact. Crucially, evidence on long-term effectiveness, scalability, cost-effectiveness, and integration into existing clinical workflows remains scarce. The lack of robust implementation evidence makes it difficult to draw definitive conclusions about the true clinical and operational value of these automation solutions.

Finally, the potential for publication bias must be acknowledged. Research dissemination practices may inherently favor the reporting of studies that demonstrate positive or novel findings. This bias could lead to an overrepresentation of successful proof-of-concept evaluations, potentially skewing the perception of efficacy and feasibility, while studies with neutral or negative results, or those focused purely on implementation challenges, may be underrepresented in the published literature.

Conclusions

Automation of clinical practice guideline development has established technical feasibility across retrieval, extraction, encoding, and evaluation, with the strongest evidence for bounded, well-structured tasks assessed under controlled conditions. Evidence for end-to-end, production-scale living guideline pipelines remains limited: no study in this corpus demonstrated a validated pipeline connecting all 4 pillars in an operating guideline program, and evidence on clinical safety, workflow integration, and cost-effectiveness is largely absent. The pattern across the literature supports a hybrid, human-in-the-loop model in which automation handles data-intensive retrieval and initial encoding while experts provide contextual judgment and validation. Realizing continuously updated living guidelines will require both interoperability standards that connect extraction to computable representation, and evaluation of these components in real guideline production rather than in retrospective benchmarks.

Acknowledgments

We thank the support staff at the Department of Community Medicine at King Abdulaziz University.

Funding

This work was supported by the Saudi National Institute of Health (Saudi NIH) under grant number PRI01-2401-KAU19-46641119.

Data Availability

The datasets generated or analyzed during this study (the data extraction sheet, study-level pillar classifications, and summary tables) are available in the Open Science Framework repository [11].

Generative AI Use

The authors declare that generative AI was not used in the preparation of this manuscript.

Authors' Contributions

A Altammami and A Alzahrani contributed to conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing the original draft, and writing (review and editing). DM contributed to methodology, supervision, and writing (review and editing). JFF-A contributed to conceptualization, methodology, resources, supervision, and writing (review and editing). OA-R contributed to data curation, search strategy development, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram construction, and writing (review and editing). A Alkhotani contributed to investigation, validation, and writing (review and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Completed PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist and search strategies for PubMed/MEDLINE and Scopus.

[[PDF File \(Adobe PDF File\), 304 KB-Multimedia Appendix 1](#)]

References

1. Tso GJ, Tu SW, Oshiro C, Martins S, Ashcraft M, Yuen KW, et al. Automating guidelines for clinical decision support: knowledge engineering and implementation. *AMIA Annu Symp Proc.* 2016;2016:1189-1198. [[FREE Full text](#)] [Medline: [28269916](#)]
2. Hoffmann-Eßer W, Siering U, Neugebauer EAM, Brockhaus AC, Lampert U, Eikermann M. Guideline appraisal with AGREE II: Systematic review of the current evidence on how users handle the 2 overall assessments. *PLoS One.* 2017;12(3):e0174831. [[FREE Full text](#)] [doi: [10.1371/journal.pone.0174831](#)] [Medline: [28358870](#)]
3. Hoffmann-Eßer W, Siering U, Neugebauer EAM, Brockhaus AC, McGauran N, Eikermann M. Guideline appraisal with AGREE II: online survey of the potential influence of AGREE II items on overall assessment of guideline quality and recommendation for use. *BMC Health Serv Res.* Feb 27, 2018;18(1):143. [[FREE Full text](#)] [doi: [10.1186/s12913-018-2954-8](#)] [Medline: [29482555](#)]
4. Luo X, Lyu H, Song Z, Liu H, Wang Z, Li H, et al. The impact of generative artificial intelligence on the development, evaluation, and application of clinical practice guidelines. *Med J Peking Union Med Coll Hosp.* 2024;15(5):1173-1181. [doi: [10.12290/xhyxzz.2024-0602](#)]
5. Mulyar N, van der Aalst WMP, Peleg M. A pattern-based analysis of clinical computer-interpretable guideline modeling languages. *J Am Med Inform Assoc.* 2007;14(6):781-787. [[FREE Full text](#)] [doi: [10.1197/jamia.M2389](#)] [Medline: [17712087](#)]
6. De Clercq P, Kaiser K, Hasman A. Computer-interpretable guideline formalisms. *Stud Health Technol Inform.* 2008;139:22-43. [[FREE Full text](#)] [Medline: [18806319](#)]
7. Fraile Navarro D, Cheyne S, Hill K, McFarlane E, Morgan RL, Murad MH, et al. Methods for living guidelines: early guidance based on practical experience. Paper 5: decisions on methods for evidence synthesis and recommendation development for living guidelines. *J Clin Epidemiol.* Mar 2023;155:118-128. [[FREE Full text](#)] [doi: [10.1016/j.jclinepi.2022.12.022](#)] [Medline: [36608720](#)]
8. Loftus TJ, Balch JA, Marquard JL, Ray JM, Alper BS, Ojha N, et al. Longitudinal clinical decision support for assessing decisions over time: State-of-the-art and future directions. *Digit Health.* 2024;10:20552076241249925. [[FREE Full text](#)] [doi: [10.1177/20552076241249925](#)] [Medline: [38708184](#)]
9. Peters MD, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. *JBIManual for Evidence Synthesis.* Adelaide, South Australia. Joanna Briggs Institute; 2024.
10. 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):467-473. [[FREE Full text](#)] [doi: [10.7326/M18-0850](#)] [Medline: [30178033](#)]
11. Altammami A, Ares-Rufo O, Alkhotani AM, Mital D, Alzahrani AH, Florez-Arango JF. Automation of living clinical practice guideline development: a scoping review. *Open Science Framework.* 2026. URL: <https://osf.io/v945f> [accessed 2026-10-03]
12. Altammami A, Ares-Rufo O, Alkhotani AM, Mital D, Alzahrani AH, Florez-Arango JF. Automation of living clinical practice guideline development: a scoping review. *OSF Registries.* Aug 12, 2026. URL: <https://osf.io/4c9zj> [accessed 2026-10-03]
13. Greenes RA, Peleg M, Boxwala A, Tu S, Patel V, Shortliffe EH. Sharable computer-based clinical practice guidelines: rationale, obstacles, approaches, and prospects. *Stud Health Technol Inform.* 2001;84(Pt 1):201-205. [Medline: [11604733](#)]
14. Khodambashi S, Slaughter L, Nytrø Ø. Computer-interpretable clinical guidelines: a review and analysis of evaluation criteria for authoring methods. *Stud Health Technol Inform.* 2015;216:954. [Medline: [26262256](#)]

15. Araya V, Ávila C, Verdugo-Paiva F, Bravo-Soto G, Rada G, Neumann I. [Comparison between a semi-automated method and a traditional search for identifying studies to inform GES clinical practice guideline recommendations: a meta-epidemiological study]. *Rev Med Chil*. Sep 2025;153(9):641-645. [[FREE Full text](#)] [doi: [10.4067/s0034-988720250009000641](https://doi.org/10.4067/s0034-988720250009000641)] [Medline: [41021846](#)]
16. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. Jul 11, 2019;8(1):163. [[FREE Full text](#)] [doi: [10.1186/s13643-019-1074-9](https://doi.org/10.1186/s13643-019-1074-9)] [Medline: [31296265](#)]
17. Bui DDA, Jonnalagadda S, Del Fiore G. Automatically finding relevant citations for clinical guideline development. *J Biomed Inform*. Oct 2015;57:436-445. [[FREE Full text](#)] [doi: [10.1016/j.jbi.2015.09.003](https://doi.org/10.1016/j.jbi.2015.09.003)] [Medline: [26363352](#)]
18. Kaiser K, Akkaya C, Miksch S. How can information extraction ease formalizing treatment processes in clinical practice guidelines? A method and its evaluation. *Artif Intell Med*. Feb 2007;39(2):151-163. [[FREE Full text](#)] [doi: [10.1016/j.artmed.2006.07.011](https://doi.org/10.1016/j.artmed.2006.07.011)] [Medline: [16962747](#)]
19. Zhu H, Ni Y, Cai P, Cao F. Automatic information extraction for computerized clinical guideline. *Stud Health Technol Inform*. 2013;192:1023. [Medline: [23920797](#)]
20. Iglesias N, Juárez JM, Campos M. Comprehensive analysis of rule formalisms to represent clinical guidelines: Selection criteria and case study on antibiotic clinical guidelines. *Artif Intell Med*. Mar 2020;103:101741. [doi: [10.1016/j.artmed.2019.101741](https://doi.org/10.1016/j.artmed.2019.101741)] [Medline: [31928849](#)]
21. Duftschmid G, Miksch S. Knowledge-based verification of clinical guidelines by detection of anomalies. *Artif Intell Med*. Apr 2001;22(1):23-41. [doi: [10.1016/s0933-3657\(00\)00098-1](https://doi.org/10.1016/s0933-3657(00)00098-1)] [Medline: [11259882](#)]
22. Bouaud J, Séroussi B, Falcoff H, Julien J, Simon C, Denke DL. Consequences of the verification of completeness in clinical practice guideline modeling: a theoretical and empirical study with hypertension. *AMIA Annu Symp Proc*. 2009;2009:60-64. [[FREE Full text](#)] [Medline: [20351823](#)]
23. Oliveira Dos Santos A, Belo VS, Mota Machado T, Silva ESD. Toward automating GRADE classification: a proof-of-concept evaluation of an artificial intelligence-based tool for semiautomated evidence quality rating in systematic reviews. *BMJ Evid Based Med*. 2026;31(1):36-45. [doi: [10.1136/bmjebm-2024-113123](https://doi.org/10.1136/bmjebm-2024-113123)] [Medline: [40194821](#)]
24. Stewart GB, Higgins JPT, Schünemann H, Meader N. The use of Bayesian networks to assess the quality of evidence from research synthesis: 1. *PLoS One*. 2015;10(3):e0114497. [[FREE Full text](#)] [doi: [10.1371/journal.pone.0114497](https://doi.org/10.1371/journal.pone.0114497)] [Medline: [25837450](#)]
25. Šuster S, Baldwin T, Lau JH, Jimeno Yepes A, Martínez Iraola D, Otmakhova Y, et al. Automating quality assessment of medical evidence in systematic reviews: model development and validation study. *J Med Internet Res*. Mar 13, 2023;25:e35568. [[FREE Full text](#)] [doi: [10.2196/35568](https://doi.org/10.2196/35568)] [Medline: [36722350](#)]
26. Wan X, Wang R, Zhao J, Liang T, Wang B, Zhang J, et al. From manual to machine: revolutionizing day surgery guideline and consensus quality assessment with large language models. *J Evid Based Med*. Mar 2025;18(1):e70017. [doi: [10.1111/jebm.70017](https://doi.org/10.1111/jebm.70017)] [Medline: [40123109](#)]
27. Peleg M, Tu S, Bury J, Ciccarese P, Fox J, Greenes RA, et al. Comparing computer-interpretable guideline models: a case-study approach. *J Am Med Inform Assoc*. 2003;10(1):52-68. [[FREE Full text](#)] [doi: [10.1197/jamia.m1135](https://doi.org/10.1197/jamia.m1135)] [Medline: [12509357](#)]
28. Duftschmid G, Miksch S, Gall W. Verification of temporal scheduling constraints in clinical practice guidelines. *Artif Intell Med*. Jun 2002;25(2):93-121. [[FREE Full text](#)] [doi: [10.1016/s0933-3657\(02\)00011-8](https://doi.org/10.1016/s0933-3657(02)00011-8)] [Medline: [12031602](#)]
29. Georg G, Séroussi B, Bouaud J. Does GEM-encoding clinical practice guidelines improve the quality of knowledge bases? A study with the rule-based formalism. *AMIA Annu Symp Proc*. 2003;2003:254-258. [[FREE Full text](#)] [Medline: [14728173](#)]
30. Boxwala AA, Peleg M, Tu S, Ogunyemi O, Zeng QT, Wang D, et al. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. *J Biomed Inform*. Jun 2004;37(3):147-161. [[FREE Full text](#)] [doi: [10.1016/j.jbi.2004.04.002](https://doi.org/10.1016/j.jbi.2004.04.002)] [Medline: [15196480](#)]
31. Shiffman RN, Dixon J, Brandt C, Essaihi A, Hsiao A, Michel G, et al. The GuideLine Implementability Appraisal (GLIA): development of an instrument to identify obstacles to guideline implementation. *BMC Med Inform Decis Mak*. Jul 27, 2005;5:23. [[FREE Full text](#)] [doi: [10.1186/1472-6947-5-23](https://doi.org/10.1186/1472-6947-5-23)] [Medline: [16048653](#)]
32. Anselma L, Terenziani P, Montani S, Bottrighi A. Towards a comprehensive treatment of repetitions, periodicity and temporal constraints in clinical guidelines. *Artif Intell Med*. Oct 2006;38(2):171-195. [doi: [10.1016/j.artmed.2006.03.007](https://doi.org/10.1016/j.artmed.2006.03.007)] [Medline: [16766167](#)]
33. Hagerty CG, Pickens DS, Chang J, Kulikowski CA, Sonnenberg FA. Prediction in annotation based guideline encoding. *AMIA Annu Symp Proc*. 2006;2006:314-318. [[FREE Full text](#)] [Medline: [17238354](#)]
34. Hrabak KM, Campbell JR, Tu SW, McClure R, Weida RT. Creating interoperable guidelines: requirements of vocabulary standards in immunization decision support. *Stud Health Technol Inform*. 2007;129(Pt 2):930-934. [Medline: [17911852](#)]
35. Serban R, ten Teije A, van Harmelen F, Marcos M, Polo-Conde C. Extraction and use of linguistic patterns for modelling medical guidelines. *Artif Intell Med*. Feb 2007;39(2):137-149. [doi: [10.1016/j.artmed.2006.07.012](https://doi.org/10.1016/j.artmed.2006.07.012)] [Medline: [16963241](#)]
36. Peleg M, Keren S, Denekamp Y. Mapping computerized clinical guidelines to electronic medical records: knowledge-data ontological mapper (KDOM). *J Biomed Inform*. Feb 2008;41(1):180-201. [[FREE Full text](#)] [doi: [10.1016/j.jbi.2007.05.003](https://doi.org/10.1016/j.jbi.2007.05.003)] [Medline: [17574928](#)]

37. Fisman M, Ortiz E, Bray BE, Rindfleisch TC. Semantic processing to support clinical guideline development. *AMIA Annu Symp Proc.* Nov 06, 2008;2008:187-191. [[FREE Full text](#)] [Medline: [18999127](#)]
38. Kaiser K, Miksch S. Versioning computer-interpretable guidelines: semi-automatic modeling of 'Living Guidelines' using an information extraction method. *Artif Intell Med.* May 2009;46(1):55-66. [[FREE Full text](#)] [doi: [10.1016/j.artmed.2008.08.009](#)] [Medline: [18950994](#)]
39. Peleg M, Garber JR. Extending the GuideLine Implementability Appraisal (GLIA) instrument to identify problems in control flow. *AMIA Annu Symp Proc.* Nov 13, 2010;2010:627-631. [[FREE Full text](#)] [Medline: [21347054](#)]
40. De Backere F, Moens H, Steurbaut K, Colpaert K, Decruyenaere J, De Turck F. Towards automated generation and execution of clinical guidelines: engine design and implementation through the ICU Modified Schofield use case. *Comput Biol Med.* Aug 2012;42(8):793-805. [doi: [10.1016/j.combiomed.2012.06.003](#)] [Medline: [22770522](#)]
41. Slaughter L, Brochhausen M, Hogan W, Nytrø Ø. The core clinical protocol ontology (C2PO): A realist ontology for representing the recommendations within clinical guidelines. *Stud Health Technol Inform.* 2013;192:997. [Medline: [23920771](#)]
42. Farkash A, Timm JTE, Waks Z. A model-driven approach to clinical practice guidelines representation and evaluation using standards. *Stud Health Technol Inform.* 2013;192:200-204. [Medline: [23920544](#)]
43. Minard AL, Kaiser K. Supporting computer-interpretable guidelines' modeling by automatically classifying clinical actions. Cham: Springer; 2013. Presented at: Process Support and Knowledge Representation in Health Care AIME 2013 Joint Workshop, KR4HC 2013/ProHealth 2013; June 1:39-52; Murcia, Spain. [doi: [10.1007/978-3-319-03916-9_4](#)]
44. Wenzina R, Kaiser K. Using TimeML to support the modeling of computerized clinical guidelines. *Stud Health Technol Inform.* 2014;205:8-12. [Medline: [25160135](#)]
45. Galopin A, Bouaud J, Pereira S, Séroussi B. Using an ontological modeling to evaluate the consistency of clinical practice guidelines: application to the comparison of three guidelines on the management of adult hypertension. *Stud Health Technol Inform.* 2014;205:38-42. [Medline: [25160141](#)]
46. Weng C, Payne PRO, Velez M, Johnson SB, Bakken S. Towards symbiosis in knowledge representation and natural language processing for structuring clinical practice guidelines. *Stud Health Technol Inform.* 2014;201:461-469. [[FREE Full text](#)] [Medline: [24943582](#)]
47. Llewellyn A, Whittington C, Stewart G, Higgins JP, Meader N. The use of Bayesian networks to assess the quality of evidence from research synthesis: 2. inter-rater reliability and comparison with standard GRADE assessment. *PLoS One.* 2015;10(12):e0123511. [[FREE Full text](#)] [doi: [10.1371/journal.pone.0123511](#)] [Medline: [26716874](#)]
48. Hu Q, Huang Z, ten Teije A, van Harmelen F, Marshall M, Dekker A. A topic-centric approach to detecting new evidences for evidence-based medical guidelines. SCITEPRESS - Science and Technology Publications; 2016. Presented at: 9th International Conference on Health Informatics; 2016 February 21 to February 23:282-289; Rome, Italy. [doi: [10.5220/0005698902820289](#)]
49. González-Ferrer A, Valcárcel MÁ, Cuesta M, Cháfer J, Runkle I. Development of a computer-interpretable clinical guideline model for decision support in the differential diagnosis of hyponatremia. *Int J Med Inform.* 2017;103:55-64. [doi: [10.1016/j.ijmedinf.2017.04.014](#)] [Medline: [28551002](#)]
50. Afzal M, Hussain M, Haynes RB, Lee S. Context-aware grading of quality evidences for evidence-based decision-making. *Health Informatics J.* Jun 2019;25(2):429-445. [[FREE Full text](#)] [doi: [10.1177/1460458217719560](#)] [Medline: [28766402](#)]
51. Schlegel DR, Gordon K, Gaudioso C, Peleg M. Clinical tractor: a framework for automatic natural language understanding of clinical practice guidelines. *AMIA Annu Symp Proc.* 2019;2019:784-793. [[FREE Full text](#)] [Medline: [32308874](#)]
52. Fux A, Soffer P, Peleg M. A layered computer-interpretable guideline model for easing the update of locally adapted clinical guidelines. *Health Informatics J.* Mar 2020;26(1):156-171. [[FREE Full text](#)] [doi: [10.1177/1460458218813705](#)] [Medline: [30518264](#)]
53. Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: Authoring FHIR implementation guides based on WHO recommendations. *J Biomed Inform.* Oct 2021;122:103891. [[FREE Full text](#)] [doi: [10.1016/j.jbi.2021.103891](#)] [Medline: [34450285](#)]
54. Borchert F, Meister L, Langer T, Follmann M, Arnrich B, Schapranow M-P. Controversial trials first: identifying disagreement between clinical guidelines and new evidence. *AMIA Annu Symp Proc.* 2021;2021:237-246. [[FREE Full text](#)] [Medline: [35308948](#)]
55. Lin Y, Li J, Xiao H, Zheng L, Xiao Y, Song H, et al. Automatic literature screening using the PAJO deep-learning model for clinical practice guidelines. *BMC Med Inform Decis Mak.* Nov 03, 2023;23(1):247. [[FREE Full text](#)] [doi: [10.1186/s12911-023-02328-8](#)] [Medline: [37924054](#)]
56. Lichtner G, Alper BS, Jurth C, Spies C, Boeker M, Meerpohl JJ, et al. Representation of evidence-based clinical practice guideline recommendations on FHIR. *J Biomed Inform.* Mar 2023;139:104305. [[FREE Full text](#)] [doi: [10.1016/j.jbi.2023.104305](#)] [Medline: [36738871](#)]
57. López B, Raya O, Baykova E, Saez M, Rigau D, Cunill R, et al. APPRAISE-RS: Automated, updated, participatory, and personalized treatment recommender systems based on GRADE methodology. *Heliyon.* Feb 2023;9(2):e13074. [[FREE Full text](#)] [doi: [10.1016/j.heliyon.2023.e13074](#)] [Medline: [36798764](#)]

58. Wang Z, Sun J, Liu H, Luo X, Li J, He W, et al. Development and performance of a large language model for the quality evaluation of multi-language medical imaging guidelines and consensus. *J Evid Based Med*. 2025;18(2):e70020. [doi: [10.1111/jebm.70020](https://doi.org/10.1111/jebm.70020)] [Medline: [40181523](#)]
59. Hsu HY, Chen LW, Hsu WT, Hsieh YW, Chang SS. Extracting clinical guideline information using two large language models: evaluation study. *J Med Internet Res*. Sep 05, 2025;27:e73486. [FREE Full text] [doi: [10.2196/73486](https://doi.org/10.2196/73486)] [Medline: [40911841](#)]
60. Boyle A, Huo B, Sylla P, Calabrese E, Kumar S, Slater BJ, et al. Large language model-generated clinical practice guideline for appendicitis. *Surg Endosc*. 2025;39(6):3539-3551. [doi: [10.1007/s00464-025-11723-3](https://doi.org/10.1007/s00464-025-11723-3)] [Medline: [40251310](#)]
61. Wang B, Luo X, Zhang J, Shi Q, Lai H, Liu H, et al. Evaluating the quality of guidelines using the AGREE II tool by a large language model vs human appraisers. *JAMA Netw Open*. 2025;8(5):e2512621. [FREE Full text] [doi: [10.1001/jamanetworkopen.2025.12621](https://doi.org/10.1001/jamanetworkopen.2025.12621)] [Medline: [40434777](#)]
62. von Dincklage F, Bublitx VK, Kumpf O, Jurth C, Riessen R, Deja M, et al. Computer-interpretable quality indicators for intensive care medicine: development and validation study. *J Med Internet Res*. 2025;27:e77077. [FREE Full text] [doi: [10.2196/77077](https://doi.org/10.2196/77077)] [Medline: [41004802](#)]
63. Wenzina R, Kaiser K. Identifying condition-action sentences using a heuristic-based information extraction method. *Cham. Springer*; 2013. Presented at: Process Support and Knowledge Representation in Health Care. *Lecture Notes in Computer Science*; 2013 June 1:26-38; Murcia, Spain. [doi: [10.1007/978-3-319-03916-9_3](https://doi.org/10.1007/978-3-319-03916-9_3)]
64. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. Mar 29, 2021;372:n71. [doi: [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71)] [Medline: [33782057](#)]
65. Arno A, Elliott J, Wallace B, Turner T, Thomas J. The views of health guideline developers on the use of automation in health evidence synthesis. *Syst Rev*. 2021;10(1):16. [FREE Full text] [doi: [10.1186/s13643-020-01569-2](https://doi.org/10.1186/s13643-020-01569-2)] [Medline: [33419479](#)]

Abbreviations

AUC: area under the receiver operating characteristic curve
AGREE II: Appraisal of Guidelines for Research and Evaluation II
CEOsyst: COVID-19 Evidence Ecosystem
CIG: computer-interpretable guideline
CPG: clinical practice guideline
DELT/A: Document Exploration and Linking Tool/Addons
EMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
GEM: guideline elements model
GLIA: Guideline Implementability Appraisal
GLIF: Guideline Interchange Format
GLIF3: Guideline Interchange Format version 3
GRADE: Grading of Recommendations Assessment, Development and Evaluation
HL7: Health Level Seven
ICD-10: International Statistical Classification of Diseases, Tenth Revision
JB: Joanna Briggs Institute
KDOM: Knowledge-Data Ontological Mapper
LLM: large language model
LOINC: Logical Observation Identifiers Names and Codes
MHB: Many-Headed Bridge
NLP: natural language processing
PAJO: Paper title, Abstract, and Journal
PICO: Population, Intervention, Comparison, and Outcome
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PRODIGY: Prescribing Rationally with Decision Support in General Practice Study
RCT: randomized controlled trial
RIGHT: Reporting Items for Practice Guidelines in Healthcare
SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms
STI: sexually transmitted infection
UIMA: Unstructured Information Management Architecture.
WHO: World Health Organization

Edited by I Steenstra; submitted 02.Jun.2026; peer-reviewed by P Katyal, T Yusuff; comments to author 16.Jul.2026; revised version received 07.Sep.2026; accepted 23.Sep.2026; published 28.Oct.2026

Please cite as:

Altammami A, Ares-Rufo O, Alzahrani A, Alkhotani A, Mital D, Florez-Arango JF
Automation of Living Clinical Practice Guideline Development: Scoping Review
JMIR AI 2026;5:e103273

URL: <https://ai.jmir.org/2026/1/e103273>

doi: [10.2196/103273](https://doi.org/10.2196/103273)

PMID:

©Abdullah Altammami, Olga Ares-Rufo, Anas Alzahrani, Abdullah Alkhotani, Dinesh Mital, Jose F Florez-Arango. Originally published in JMIR AI (<https://ai.jmir.org>), 28.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.