Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Artificial Intelligence in Auditing the Quality of Electronic Nursing Records: A Critical Narrative Literature Review

João Gabriel Leal Sousa

Faculdade Presidente Antônio Carlos, Brazil

Corresponding author: joaogabriel202340@hotmail.com

Orcid ID: https://orcid.org/0009-0003-5187-8374

Abstract

Electronic nursing records contain assessments, interventions, and evaluations that can be examined for documentation quality, but extracting information from a record is different from verifying its clinical accuracy. This critical narrative and thematic review asks which applications, results, and limitations of artificial intelligence (AI) in auditing electronic nursing documentation are described in the scientific literature. Targeted scholarly web discovery, primary-source retrieval, and reference checking were completed on October 6, 2026, with a reproducible supplementary PubMed search on October 7, 2026. Twenty-six publications and official documents were retained and classified by their evidentiary role. The synthesis distinguishes direct documentation auditing from terminology mapping, AI-assisted documentation generation, contextual nursing research, and methodological or regulatory guidance. Only one of the 14 retained empirical publications directly evaluated AI-supported quality control of existing nursing documentation. That implementation reported fewer omissions and inconsistencies, although a nonrandomized design and inconsistent reporting limit causal interpretation. Terminology-mapping studies showed configuration-dependent performance and substantial residual errors; scenario-based instrument validation and documentation-generation studies support selected evaluation methods but do not establish comprehensive autonomous auditing. Across the corpus, documentation

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

quality, classifier performance, workload, and patient outcomes require separate endpoints. An author-developed, unvalidated governance workflow is proposed to connect criterion definition, authorized data access, evidence-linked flags, nurse adjudication, and prospective monitoring. The literature supports supervised assistance for bounded audit tasks; it does not establish that AI can independently certify record truth, judge nursing performance, or improve patient safety across institutions. Multicenter validation, clinically weighted error analysis, transparent reference standards, and evaluation of total verification effort remain priorities.

Keywords: artificial intelligence; nursing documentation; electronic health records; documentation quality; nursing audit

1. Introduction

Nursing documentation quality concerns more than whether a record is legible or electronically accessible. It includes whether the nursing process is represented coherently and whether the documented information is sufficiently complete and useful for its purpose. De Groot et al. (2019) synthesized criteria, instruments, and requirements for nursing documentation, providing an established basis for evaluating records before introducing an automated auditor. This review therefore treats documentation quality as the object of assessment and AI as a possible method of assistance, rather than assuming that digitalization or fluent text constitutes quality.

The technical opportunity is that nursing information exists in both structured fields and narrative text. AI-based classification can help organize narrative content, while language models can produce structured outputs from clinical material. However, such capabilities address different tasks: finding a documented intervention, checking whether a required assessment is present, and determining whether the record accurately represents an intervention actually performed are not interchangeable. Contemporary nursing studies separately investigate terminology classification,

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

generated care plans, and documentation quality control (Abudari et al., 2025; Li et al., 2026; Lv et al., 2025).

The distinction matters for information systems because an audit result can trigger a correction request, influence a dashboard, or become an input to professional review. This manuscript defines an AI-supported documentation audit as a criterion-based examination of an existing nursing record that produces an evidence-linked finding for review. It defines documentation generation as creation or transformation of a draft. These are operational definitions proposed for this synthesis; studies of drafting are included only where they clarify evaluation or verification requirements relevant to auditing.

The research question is: What applications, results, and limitations of AI in auditing the quality of electronic nursing records are described in the scientific literature? The objectives are to identify the functions being investigated, distinguish the outcomes actually measured, assess methodological boundaries, and propose a practical research agenda. The review does not evaluate diagnostic accuracy for disease prediction, financial billing audit, or autonomous clinical care.

2. Literature Review

2.1 Documentation quality, data quality, and care quality

Data quality provides a useful vocabulary but should not displace nursing-specific reasoning. Weiskopf and Weng (2013) identified completeness, correctness, concordance, plausibility, and currency in EHR assessment. Kahn et al. (2016) organized harmonized terminology around conformance, completeness, and plausibility, distinguishing verification from validation. These frameworks inform the operational dimensions in Table 1; their categories are not evidence that a nursing audit algorithm is effective.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

A field may conform to an allowed format while containing an incorrect observation. A note may be internally plausible while describing care that cannot be independently verified. Conversely, an absent value may mean that a criterion was inapplicable, that another document contains the information, or that documentation was omitted. The interpretation proposed here is that auditors should distinguish a technical defect, a possible clinical inconsistency, and an unresolved evidentiary question. The EHR data-quality literature supports assessing multiple dimensions rather than treating field presence as a universal proxy for quality (Lewis et al., 2023).

Nursing-sensitive information can support quality assessment, but the route from documentation to care outcomes requires separate validation. Von Gerich et al. (2022) manually examined structured and narrative material from 230 acute cardiac-care records and extracted 1,676 expressions related to selected indicators. This establishes the availability and interpretive complexity of relevant content, not the performance of a deployed AI auditor. In this review, better documentation is therefore a documentary outcome; better care remains a separate empirical claim.

2.2 Distinguishing automation, machine learning, and generative AI

A deterministic check can identify a missing signature or a timestamp outside a predefined interval. A learned classifier can map text to an intervention category. A generative model can produce a rationale or proposed revision. These functions may coexist in one application, but their contributions should be evaluated separately. Zhang et al. (2024), for example, evaluated an automated ICU documentation quality-control system based on integration and monitoring. It is retained as a relevant automation comparator, not counted as evidence of a specifically demonstrated LLM effect.

Mitha et al. (2023) identified 43 eligible studies of NLP in nursing notes, with heterogeneous methods and reporting. Their review demonstrates that nursing NLP

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

extends beyond recent conversational systems and that nursing-specific terminologies and complete performance reporting have not been consistently used. Accordingly, this manuscript does not use “AI,” “NLP,” and “LLM” as synonyms. An extraction result should identify the method, target label, and reference standard against which it was assessed.

2.3 Audit findings as claims requiring evidence

This review proposes three distinct output categories: confirmed technical nonconformance, suspected documentary inconsistency, and indeterminate finding. A required field that is absent can sometimes be established directly. A contradiction between two entries requires contextual interpretation. A statement about an undocumented bedside event may remain indeterminate without another source. These categories are a conceptual aid developed for this article, not a validated scale or an existing professional standard.

The distinction also prevents an automated system from treating its own suggestion as new clinical evidence. A proposed diagnosis can be relevant to a case without having been assessed or documented by the responsible nurse. A draft intervention can be appropriate without having been delivered. Under the proposed audit workflow, a recommendation remains a recommendation until an authorized professional verifies and records the corresponding clinical fact.

3. Review Methods

3.1 Review design and actual discovery route

A critical narrative and thematic review was conducted. Initial discovery and source verification were completed on October 6, 2026. A supplementary structured PubMed verification search was completed on October 7, 2026, extending the final publication cutoff to that date. The actual route consisted of targeted scholarly web searches,

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

retrieval of journal and repository pages, examination of PubMed/PMC records, and consultation of official regulatory and organizational documents. Searches were iteratively refined around nursing documentation, electronic nursing records, quality control, auditing, NLP, machine learning, LLMs, terminology mapping, and documentation evaluation. Representative queries and the selection rules are supplied in Appendix A.

PubMed and PMC initially served as retrieval and verification sources. On October 7, 2026, PubMed was also searched directly through the NCBI ESearch service using the exact query in Appendix A. The query returned 26 unique records; their titles and abstracts were examined and two additional primary studies were retained after full-text appraisal. Five records were already included; the other 19 were not added to the focused narrative synthesis for the reasons recorded in Appendix B. This supplementary ledger is distinct from the unavailable historical discovery ledger. No direct searches were performed in CINAHL, Embase, Scopus, Web of Science, ACM Digital Library, or IEEE Xplore. Subscription access to CINAHL was unavailable for this revision; Scopus and Web of Science verification was not performed. The search remains selective rather than exhaustive, and its title/abstract terms can miss publications using different terminology.

3.2 Eligibility and evidence roles

Eligible empirical sources concerned electronic nursing documentation quality control, classification or structuring of nursing narratives, nursing-specific evaluation instruments, or generation workflows with relevant quality or verification outcomes. Contextual studies addressed interpretation of nursing records or organizational sources of distortion. EHR quality frameworks, reporting guidelines, and official governance documents were eligible to support methodology and safeguards. Primary

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

published versions were preferred over earlier preprints when an accessible final record existed.

Disease-risk prediction from nursing notes alone was excluded from the focal evidence map because predictive discrimination is not documentation-audit accuracy. Commercial product claims, news items, Wikipedia, unsupported promotional material, and studies without a clear relationship to the research question were not used as scientific evidence. Incident-investigation reports and scenario-generated care plans were retained as adjacent documentation domains and explicitly labeled, rather than presented as routine bedside-record audits.

3.3 Corpus composition and source accounting

The original corpus contained 24 sources: 12 primary empirical publications, three review syntheses, four methodological/framework publications, and five official governance or legal documents. The supplementary PubMed search added Kim et al. (2026) and Leurs et al. (2026), bringing the final corpus to 26 sources: 14 primary empirical publications, three reviews, four methodological/framework publications, and five official documents. Only one empirical publication directly evaluates AI-supported nursing documentation quality control; ten provide classification, structuring, generation, instrument-evaluation, or adjacent event-extraction evidence; and three provide automation-comparator or contextual evidence. Table 2 maps all 14 empirical sources. Twenty-one of the 26 sources (80.8%) are dated 2021–2026; five older sources provide foundational or legal context.

Corpus counts describe retained sources. Appendix B separately records the 26 results of the new supplementary search; these are not retrospectively invented historical screening totals. A complete deduplicated ledger of all web results and exclusions was not maintained, so identification counts and a PRISMA flow diagram are not reported. The supplied example manuscripts and reviewer letters informed

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

organization and reporting discipline but were not treated as nursing-effectiveness evidence or included in the corpus denominator.

3.4 Extraction, appraisal, and synthesis

Extraction recorded task, clinical setting, study design, input data, reference standard, reported outcome, and principal transfer limitation. Table 2 functions as the focal evidence matrix. Quantitative results were checked against available primary article sections; discrepancies between prose and tables were retained as limitations instead of silently resolved. Comparable percentages were not pooled, because the sources measure different constructs, populations, units, and workflows.

Appraisal considered reference-standard independence, sampling, blinding, study design, external validation, clinical applicability, and reporting consistency. This is a structured narrative appraisal, not GRADE, a validated numerical quality score, or a formal risk-of-bias instrument. No duplicate independent screening or extraction was undertaken. Consequently, this synthesis reports neither kappa nor inter-rater reliability for its own selection process; reliability figures from included studies refer only to those studies. No independent second nursing-informatics reviewer was available for the revision; the supplementary appraisal therefore remains single-reviewer and does not establish independent reproducibility.

4. Results

4.1 Direct documentation quality control

Lv et al. (2025) evaluated a locally deployed closed-loop system using 278 records before and 278 after implementation. Missing assessment items declined from 20/278 to 5/278, logical errors from 26/278 to 2/278, and timeliness errors from 24/278 to 0/278. These are record-level observations within the evaluated workflow, not proof of zero future error or independently established patient-safety benefit.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

The paper inconsistently reports audit times (prose: 238–74 seconds; table medians: 337–91 seconds) and design (retrospective pre/post methods; randomized wording in discussion). This synthesis uses the table medians and the nonrandomized design, without isolating a model effect from the bundled workflow (Lv et al., 2025).

4.2 Terminology mapping and semantic structuring

Niyirora et al. (2025) compared machine-learning approaches with nurse coding of home-health narratives into NIC. The strongest model reported F1=0.31, with lower performance than the human raters and difficulties involving less common or context-dependent categories. This is evidence of a classification gap, not validation of comprehensive auditing. A missed category could reflect model error, difficult source text, or annotation uncertainty.

Li et al. (2026) evaluated CCC mapping on 100 ICU records. DeepSeek-R1 with retrieval and few-shot examples achieved F1=0.7836 for diagnoses and 0.8413 for interventions; independent nurse agreement before consensus was Fleiss κ=0.6449 and 0.6180, respectively. The results concern a particular configuration and reference standard. They do not imply that inferred codes establish bedside truth or that performance transfers to another taxonomy, language, or institution.

Moen et al. (2020) investigated sentence classification and coherent document organization under nursing subject headings. Such work supports assistance with locating and organizing recorded content. Its relevance to audit is indirect: structure can make information easier to inspect, but correct placement is not the same endpoint as identifying clinically important omissions or verifying care delivery.

Kim et al. (2026) mapped 107 selected ICU intervention notes from MIMIC-IV to CCC terminology using GPT-4o mini and expert references. In the most contextualized prompt phase, match rates were 61.7% for care components, 72.0% for interventions, and 72.9% for sub-interventions. These are hierarchy-specific mapping rates, not audit sensitivity. The

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

small, selected set, single model, and context-dependent reference limit transfer. Reasoning failures and fabricated content remain relevant risks, reinforcing source inspection rather than autonomous confirmation of care.

4.3 Instruments for evaluating generated documentation

Abudari et al. (2025) developed NPET around the nursing process and tested 64 generated care plans representing eight scenarios, producing 368 expert ratings. Content validity was S-CVI/Ave=0.88. Average-rater ICCs of 0.85–0.94 should be interpreted alongside lower single-rater values, approximately 0.49–0.75. The overall comparison among models was nonsignificant (p=.144). The study supports initial scenario-based instrument evidence, not autonomous assessment of real-world records or definitive superiority of paid models.

The practical implication is a distinction between an instrument's measurement properties and a model's audit performance. This review proposes using validated instruments where their scope fits the record being examined, followed by local adaptation and reliability testing when scope or language changes. A rubric should not become an AI-generated answer key against which that same AI system is evaluated. Clinical reviewers need access to the source material and a documented method for resolving disagreements.

4.4 Generation, voice input, and handover workflows

Ju et al. (2025) studied 40 nurses documenting virtual-patient scenarios. Mean time changed from 467.18 to 182.68 seconds; accuracy was rated 3.62/5 and fluency 4.50/5. The abstract describes an approximately 40% reduction, but the published means imply approximately 60.9%. This review reports the original times and flags the discrepancy. The finding concerns assisted drafting in a simulation, with quality ratings and a sequential design; it does not demonstrate audit sensitivity.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Chen et al. (2026) observed shorter handover-documentation times after implementation across three Taiwanese hospitals. Nurses verified generated content, but the study did not conduct formal text-quality evaluation against a gold-standard reference set. The result supports workflow feasibility and a time association; it cannot establish content fidelity, an independent causal model effect, or improved patient outcomes.

Su et al. (2026) combined speech recognition with schema-constrained DART generation. The study reported macro-F1=0.82 and a hallucination rate of 2.51%, with nurse verification before upload. These outcomes concern transcription and structuring under that evaluation protocol. A nonzero hallucination rate makes review of clinical facts essential; perceived completeness among users is not an objective estimate of audit correctness.

4.5 Adjacent evidence and organizational context

Teo et al. (2026) evaluated 150 patient-safety investigation reports in a voluntary pre/post NARRATE workflow. Cluster-adjusted differences favored assisted reports for completeness (1.95 points) and adapted PDQI mean (0.53 points). Self-selection, a different documentary purpose, and the bundled intervention limit transfer to routine nursing-record audits. The study explicitly does not establish downstream patient-safety effects.

Wang et al. (2025) identified nurse-, patient-, operational-, and organizational-level factors influencing information distortion. This contextual evidence cautions against interpreting every problematic record as an individual writing failure. An audit tool may detect a pattern while leaving its organizational cause unresolved. Workflow pressures, documentation habits, and interface constraints require attention alongside model performance.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

The automation comparator of Zhang et al. (2024) reported improvements following ICU system implementation in 600 patients, including QC times of 264 versus 62 seconds. Because the intervention emphasizes data integration and automated checks, its benefits should not be attributed to generative AI. It instead provides a reason to compare an AI auditor with an adequately implemented rules-based alternative rather than with an unnecessarily weak baseline.

Leurs et al. (2026) evaluated fall extraction from 30,615 nursing records in one Dutch hospital. Their confusion matrix contained 99 true positives, 505 false positives, 29,987 true negatives, and 25 false negatives: sensitivity was approximately 80%, specificity 98%, and positive predictive value 16%. High specificity therefore coexisted with substantial flag-review burden at low prevalence. This is adjacent event-extraction evidence, not validation of documentation-defect auditing or independent proof of undocumented falls. The reported 92.6% time saving used an estimated incident-report comparator and should not be generalized to total audit effort or improved patient outcomes.

5. Discussion

5.1 Answering the research question

The identified applications form four groups: checking documentary omissions or inconsistencies, extracting and standardizing nursing concepts, evaluating generated plans or reports, and assisting record production. Only the first directly matches comprehensive criterion-based audit assistance. The remaining groups provide components or adjacent evaluation evidence. The literature supports feasibility for selected tasks, but the evidence map does not establish an institution-independent automated auditor.

The central limitation is mismatch between task and conclusion. A terminology classifier can be accurate within its reference standard while missing contextual safety information. A generation workflow can reduce writing time without measuring defect

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

detection. A measurement instrument can be reliable for scenario plans without validating a classifier for authentic notes. The appropriate conclusion is therefore conditional: an application is supported for the task, population, configuration, and endpoint actually tested, with further claims requiring additional evidence.

5.2 Avoiding documentation–reality substitution

This review proposes an explicit rule: documentary absence is a finding about the available record, not automatic proof that care was omitted. Conversely, documentary presence is not automatic proof that an intervention occurred. An auditor should first identify what evidence it inspected, then distinguish the inconsistency it can demonstrate from the event it cannot verify. This interpretation follows the separation between EHR verification and validation in Kahn et al. (2016).

Consider a hypothetical record containing an elevated pain score followed by no visible reassessment. A bounded audit may flag that no reassessment was found in the authorized documents within the defined interval. It should not state that the nurse failed to reassess the patient unless another evidentiary process establishes that fact. If a reassessment appears in a different section, the issue may concern retrieval rather than bedside care. This example is conceptual and does not represent an implemented system or observed patient case.

The proposed consequence is that institutional dashboards retain both confirmed findings and unresolved flags. Merging them would conceal how much of an apparent defect rate results from uncertain model interpretation. Human adjudication should preserve reasons for rejection, alternative evidence, and unresolved cases. The rate at which reviewers dismiss plausible-looking flags is itself a useful endpoint for future research, rather than a failure to be hidden during reporting.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

5.3 Reference standards and clinically meaningful errors

A defensible reference standard requires more than two people agreeing with a generated answer. The proposed evaluation process starts with explicit criteria and an independent review of source records. It records whether disagreement concerned a true clinical ambiguity, a taxonomy boundary, an incomplete record, or a reviewer error. Agreement is reported before consensus, and consensus does not erase the original uncertainty. This is an author-proposed evaluation specification informed by the reference-standard limitations identified in Table 2. Reliability statistics should match the scale and rater design: nominal categories may use an appropriate kappa statistic, ordinal ratings weighted kappa, and continuous scores an explicitly specified ICC. A universal ≥0.75 cutoff is not imposed on this narrative review or retrospectively on all included studies; lower pre-consensus agreement represents reference-standard uncertainty.

Performance should be reported at the relevant unit: document, criterion, clinical event, or label. An application may flag multiple defects in one record, and one clinical fact may appear repeatedly. Counting every repeated mention as an independent success would exaggerate information. The proposed protocol therefore specifies a unit of analysis in advance and reports clustering by patient, nurse, or encounter where appropriate. Table 3 lists endpoints that preserve these differences. Defect-level sensitivity must not be presented as record-level performance. Repeated observations require cluster-robust confidence intervals, mixed-effects models, or generalized estimating equations when appropriate to the sampling design; the patient, encounter, nurse, and institutional clustering structure should be prespecified, with cross-classification addressed where relevant. These are proposed requirements for future empirical studies, not analyses performed on this heterogeneous review corpus.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Not all mistakes have the same consequence. Missing an essential reassessment and misplacing a heading are different audit failures. This review recommends reporting clinically prioritized categories separately, with reviewer-defined severity rules and uncertainty intervals. Overall accuracy should not obscure rare but consequential false negatives. Equally, apparent sensitivity obtained by flagging almost every record would impose excessive verification work. The intended balance must be defined locally rather than reduced to an unsupported universal threshold. Future protocols should define locally justified low-, moderate-, and high-consequence error categories before evaluation, ideally adjudicating severity blinded to model identity. No universal severity weighting is proposed. Sensitivity, flag precision, criterion-specific defect prevalence, and confidence intervals should be reported together. Specificity requires a meaningful true-negative opportunity denominator; when this cannot be defined, false flags per audited criterion or document and flags reviewed per confirmed defect provide interpretable alternatives (see Table 3). Leurs et al. (2026) illustrate why high specificity alone does not establish a manageable alert workload.

For the proposed protocol, abstention rate is the number of eligible cases left unresolved or explicitly abstained divided by all eligible cases, multiplied by 100%. System abstention and unresolved professional adjudication should be reported separately, including their overlap. Metrics restricted to resolved cases must state that denominator alongside coverage and an analysis of unresolved cases; difficult cases must not disappear from reporting. This is an author-proposed reporting specification rather than an observed result.

5.4 Human oversight and professional accountability

The proposed workflow places registered nurses with relevant clinical and audit expertise in charge of adjudication. Competence means being able to evaluate the underlying criterion without relying on the model's explanation. An editable rationale is useful only when the reviewer can inspect the supporting record, detect missing

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

context, and reject the suggestion. Model-generated confidence should not override an unresolved evidentiary gap. Model-generated rationale is explanatory output, not independent clinical evidence.

For Brazil, COFEN Resolution 736/2024 describes the nursing process and formal documentation responsibilities, including the nurse's role in nursing diagnoses and prescriptions. Resolution 754/2024 addresses electronic records, confidentiality, access, and identification/signature requirements. These documents support preserving professional responsibility within an assisted workflow; they do not authorize autonomous AI sign-off or certify a particular product (Conselho Federal de Enfermagem [COFEN], 2024a, 2024b).

The institutional policy proposed here prohibits automatic retrospective insertion of a clinical event merely because an auditor expects it. A correction can clarify or supplement documentation only through the authorized professional process, with provenance and timing preserved. Likewise, automated flags should not determine disciplinary outcomes, professional competence, or care adequacy without an appropriate human review. These are proposed governance boundaries, not empirical findings from an implemented intervention.

5.5 Privacy, security, and processing architecture

Health-related personal data receive specific protection under Brazil's LGPD. An audit implementation must establish a lawful purpose, an appropriate basis for processing, and security measures; pseudonymization should not be described as universal anonymization. This legal discussion is restricted to Brazil and is not a complete compliance analysis. International deployment requires jurisdiction-specific assessment (Brasil, 2018).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

This article proposes separating authorization to retrieve a record from permission to export its content. Audit access should be limited to the authorized episode and necessary fields, and logs should avoid unnecessary sensitive payloads. Local deployment is an architectural choice, not evidence by itself of complete security. Provider retention, administrative access, training use, deletion, and processing location should be specified before real records are transmitted. These controls are implementation recommendations; their effectiveness was not experimentally evaluated in this review.

WHO (2024) discusses risks and governance of large multimodal models in health, while Tabassi (2023) presents a voluntary, non-sector-specific risk-management framework. They provide governance context rather than empirical effect-size evidence. Applying their principles does not make the proposed nursing workflow an endorsed standard. For an audit application, accountability should identify who defines criteria, operates infrastructure, reviews flags, approves changes, and responds to incidents.

5.6 Net workload rather than isolated speed

The proposed productivity boundary includes preparation, retrieval, flag review, correction, escalation, monitoring, and maintenance. Time saved in initial screening can be offset by false positives or repeated checking. Therefore, this review recommends matched comparisons of total nurse-auditor effort under the same quality requirements. Faster screening with more missed critical defects should not be considered a productivity improvement. Comparable workload measures should include minutes per record, minutes per confirmed defect, flags reviewed per confirmed defect, and corrections per 100 records. Reports should state the number of confirmed defects and avoid undefined ratios when none are confirmed. Alert

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

burden, unnecessary correction, automation bias, and override or rejection reasons remain distinct process outcomes.

A future study could compare manual audit, rules-based audit, and hybrid rules-plus-AI review. Each group would use the same audit criteria and an independent reference standard. Outcomes would include verified defects detected, important defects missed, time to final adjudication, unnecessary corrections, and unresolved cases. The AI group would record model version, prompt, retrieved context, and configuration. This is a proposed experiment, not a claim that the comparison has already been conducted.

5.7 Proposed Governance Workflow and Research Agenda

5.7.1 Nurse-supervised audit workflow

Table 4 presents an author-developed workflow connecting criterion definition, authorized access, hybrid detection, evidence-linked review, controlled correction, and monitoring. It is a decision aid synthesized for this review, not a validated score, certification, or professionally endorsed standard. Its practical aim is to make every consequential transition reviewable: from text to suspected defect, from suspected defect to adjudicated finding, and from finding to record amendment.

Before deployment, the institution would specify the audit scope and distinguish administrative requirements from clinical interpretation. Deterministic checks would remain available for criteria that do not need probabilistic language interpretation. The AI component would examine a bounded snapshot and return the criterion, source location, proposed category, and uncertainty. A reviewer would confirm, reject, or defer the finding. No correction would be written automatically into the authoritative clinical record. Failure analysis must distinguish retrieval failure from interpretation failure: a missing retrieved entry, an access restriction, or information stored outside the

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

snapshot is different from misreading an inspected entry. Each flag should record the authorized source scope, source location, retrieval status, and the reason for uncertainty, without expanding access privileges.

The workflow would retain a minimal audit trail describing what was assessed and why the decision changed. A rejected flag would remain available for aggregate error analysis without becoming a defect attributed to a nurse. Repeated disagreement around the same criterion would trigger a review of both the model and the criterion definition. This protects against interpreting poor system design as poor professional practice.

5.7.2 Prospective validation and reporting

The recommended research sequence is retrospective development, external testing, prospective silent evaluation, and controlled assisted use. During silent evaluation, the model would generate findings without changing the workflow, allowing assessment of errors before staff respond to them. During assisted use, investigators would measure verification effort and unintended effects, including delayed documentation or excessive interruptions. The sequence is proposed here; it is not a mandatory regulatory pathway. The silent phase should quantify raw flag rate, confirmed-defect sensitivity, precision, abstention, and reviewer time before users adapt their documentation to alerts. This separates baseline algorithmic performance from intervention-induced changes in documentation behavior.

DECIDE-AI emphasizes early clinical evaluation and human factors, while TRIPOD-LLM addresses transparent reporting of LLM studies. Their relevance is methodological: future nursing audits should describe intended users, setting, data provenance, prompts, model access, changes to workflow, and failure analysis. These guidelines do not validate the clinical accuracy of an audit application merely because its report is complete (Gallifant et al., 2025; Vasey et al., 2022).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

External testing should include another institution, a later time period, and relevant document types. Splitting notes while allowing material from the same patient episode to appear in both training and test sets must be prevented in evaluations intended to measure transfer to unseen patients. For Portuguese deployment, expert annotation should address local terminology, abbreviations, negation, temporal statements, and mixed structured/narrative content. English-language success does not establish performance in Brazilian records. Patient-level partitions should also prevent encounter leakage across development and evaluation sets. Temporal validation should test later periods independently of initial external validation, because templates, abbreviations, criteria, and case mix can change.

A model update, altered retrieval pipeline, new documentation template, or revised professional criterion changes the evaluated system. The proposed monitoring plan consequently records versions and repeats criterion-specific checks after consequential changes. Researchers should disclose failures and indeterminate outputs, not only successful examples. Results should remain stratified by task and document type when combining them would conceal meaningful differences. Version records should include the model, prompt, retrieval configuration, system instructions where disclosure is permitted, document templates, and terminology versions. Material updates require prespecified regression testing and renewed validation before consequential assisted use; Portuguese-language validation remains independent of English-language performance.

5.7.3 Priority questions

The first priority is whether AI assistance improves confirmed defect detection against a strong reference standard. The second is whether improvements persist after all reviewer effort is included. The third is whether the workflow remains useful across institutions, languages, and patient groups. The fourth is whether documentation

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

improvements lead to better communication or care outcomes in separately designed studies. None of these transitions should be presumed from writing-time reduction alone.

Additional priorities include reviewer automation bias, meaningful abstention, taxonomy drift, and the effect of alerts on documentation behavior. A system could encourage genuinely useful correction or merely more text that satisfies a rubric. Future studies should therefore assess fidelity and usefulness alongside completeness. Incentives for producing longer, more standardized notes should be examined rather than accepted as a universal improvement.

5.8 Limitations

This is a selective narrative synthesis based on targeted web discovery and one supplementary structured PubMed query, not complete searches across nursing and computing databases. It may omit eligible studies, particularly publications using different terminology, non-English material, subscription-only literature, and local nursing research. CINAHL access was unavailable, and optional Scopus/Web of Science checks were not performed. A complete historical search ledger, independent duplicate appraisal, and formal certainty grading are unavailable. The supplementary search ledger does not repair those historical limitations or confer systematic-review status. Source-role assignments and the proposed governance workflow remain interpretive judgments.

The empirical corpus includes different documentary purposes, clinical settings, languages, and reference standards. Simulated care plans, incident reports, terminology mapping, and automated ICU checks are not equivalent to comprehensive audits of routine nursing records. The observed results were not pooled. Findings from one configuration cannot establish performance across a rapidly evolving class of models.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Several sources have reporting or design constraints, including inconsistent numerical summaries, nonrandomized comparisons, small or local evaluation sets, voluntary adoption, and limited external validation. These were considered in the interpretation rather than treated as reasons to infer results not reported. The proposed workflow has not been implemented, independently validated, or shown to reduce adverse events.

6. Conclusion

The literature describes AI applications for documentary quality control, nursing concept mapping, structured drafting, and evaluation of generated content. Their results support selected supervised functions under specific conditions, with substantial differences in performance and evidence maturity. They do not establish comprehensive autonomous auditing, clinical truth verification, or general patient-safety benefit. Only one of the 14 retained empirical publications directly evaluated AI-supported quality control of existing nursing documentation. The supplementary evidence strengthens the mapping and event-extraction components without changing that central limitation. Documentation and workflow outcomes remain much better supported than patient-safety or clinical benefit.

The most defensible role is to assist criterion-based review while preserving source evidence, uncertainty, and nurse accountability. Research should evaluate defect detection, total adjudication effort, external transfer, and downstream outcomes separately. A useful audit system must help reviewers reach better-supported decisions, not merely produce a more fluent assessment or a faster-looking workflow.

Data Availability

No patient-level dataset was created or accessed for this review. The cited publications are the evidence sources; Table 2 provides the extracted focal evidence matrix, and Appendix A records the discovery approach and source-role rules.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Declaration on Generative AI Use

Generative AI was used to assist with targeted literature discovery, bibliographic checking, organization, revision, and academic English editing. The author is responsible for reviewing the manuscript, verifying its claims and references against the cited sources, and approving the final submission, and takes full responsibility for its content and interpretation.

Appendix A. Search and Selection Transparency

Initial discovery date: October 6, 2026. Supplementary search and final publication cutoff: October 7, 2026. Initial discovery used iterative scholarly web queries and title-specific follow-up searches. PubMed/PMC and publisher pages initially served as retrieval and verification sources. The supplementary PubMed query was executed directly using NCBI ESearch; no subscription-database search is claimed.

Representative queries actually used included: “nursing documentation quality audit natural language processing artificial intelligence electronic nursing records study”; “nursing records quality machine learning”; “nursing documentation quality natural language processing”; “nursing documentation audit artificial intelligence”; “documentation quality nursing automated NLP”; and “nursing documentation quality control large language study.” Title-specific queries verified the focal publications and official COFEN documents.

Supplementary PubMed query executed on October 7, 2026: ("nursing documentation"[Title/Abstract] OR "nursing records"[Title/Abstract]) AND ("artificial intelligence"[Title/Abstract] OR "natural language processing"[Title/Abstract] OR "large language model"[Title/Abstract] OR "machine learning"[Title/Abstract]) AND (audit[Title/Abstract] OR quality[Title/Abstract] OR completeness[Title/Abstract]) AND ("1900/01/01"[Date - Publication] : "2026/10/07"[Date - Publication])

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

No language or article-type filter was applied. ESearch returned 26 unique PMIDs and the corresponding PubMed records were retrieved with EFetch. Five records were already retained; two new primary studies were added after PMC full-text appraisal. Nineteen were not added to this focused narrative update; Appendix B records each disposition. The corpus changed from 24 to 26 sources and the empirical map from 12 to 14 publications. No additional direct documentation-quality-control implementation was retained. These counts concern only the new query and do not reconstruct the initial search.

Retained source roles were assigned by their contribution to the question, not by the presence of “AI” in the title. Direct audit evidence required assessment of existing nursing documentation against quality criteria. Classification and generation studies were mapped as components or adjacent evidence. Comparator/context sources informed interpretation. Reviews and guidance informed constructs and governance without being treated as additional independent implementations.

A complete initial-result and exclusion ledger is unavailable. The updated retained corpus and its composition are reported in Section 3.3; only the new supplementary query has a complete result ledger. No retrospective PRISMA accounting or formal inter-rater statistics are asserted. A future systematic or scoping review should add structured nursing and computing database searches, deduplication, a documented screening ledger, and independent reviewer procedures.

Appendix B. Supplementary PubMed Result Ledger

This ledger documents only the October 7, 2026 supplementary query. It does not assert exhaustive eligibility screening of the field. Records not added may still be relevant to broader reviews. Abstract-only candidates were not promoted to fully appraised primary evidence.

PMID 42551937 — [Creation of Patient Summaries from Electronic Medical Records and Integration of Medical Knowledge via Large Language Models-Next-Generation Medical DX and Drug Discovery Support Opened Up by Natural Language

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Processing]. Not added: adjacent summarization/drug-discovery overview; not a direct documentation audit.

PMID 42421458 — [The Structured Information Collection as a data source for AI applications: a qualitative study of data quality]. Not added: potentially relevant German contextual study; only indexed abstract verified, no full-text appraisal.

PMID 42393704 — Mapping ICU nursing intervention notes to clinical care classification using a large language model: an evaluation study. Added: CCC mapping component; full text appraised.

PMID 42349996 — Using AI to identify in-hospital falls: a comparative analysis with incident reports and manual chart review. Added: adjacent fall-event extraction; full text appraised.

PMID 42247576 — Voice-Based Structured Nursing Documentation Using Automatic Speech Recognition and Large Language Models: Development and Evaluation Study. Already included: Su et al. (2026).

PMID 41819121 — Integrating a Large Language Model to Streamline Nursing Handover Documentation Across Multiple Hospitals in Taiwan: Development and Implementation Study. Already included: Chen et al. (2026).

PMID 41431886 — Use of Artificial Intelligence in Electronic Health Records With Nursing Data Across Multiple Care Settings: A Scoping Review. Not added: broader scoping review; indexed abstract verified, full text not appraised in this focused supplement.

PMID 41361970 — A Machine Learning Approach to Predict Health Care-acquired Urinary Tract Infections From Electronic Nursing Documentation. Not added: urinary-infection prediction, outside documentation-audit target.

PMID 41357433 — Transforming critical care: the digital revolution's impact on intensive care units. Not added: broad ICU digitalization overview rather than nursing-record audit evaluation.

PMID 41272536 — Development and validation of the Nursing Process Evaluation Tool (NPET): a multidimensional instrument for assessing the quality of AI-generated nursing documentation. Already included: Abudari et al. (2025).

PMID 40776240 — Analyzing Nursing Records in Wound Care Using a Large Language Model. Not added: potentially relevant wound-note summarization; indexed abstract verified, full text not appraised.

PMID 40668938 — A DeepSeek-powered locally deployed closed-loop system for enhancing quality control in electronic nursing documentation: development and clinical validation. Already included: Lv et al. (2025).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

PMID 40526735 — Application and Challenges of Large Language Models in Clinical Nursing: A Systematic Review. Not added: broad clinical-nursing LLM review; not an additional primary audit implementation.

PMID 40384067 — Generative AI-Based Nursing Diagnosis and Documentation Recommendation Using Virtual Patient Electronic Nursing Record Data. Already included: Ju et al. (2025).

PMID 40129115 — An AI-Enabled Nursing Future With no Documentation Burden: A Vision for a New Reality. Not added: prospective vision rather than evaluated audit performance.

PMID 39733385 — [Artificial intelligence in health education: blessing or curse?]. Not added: health-education focus, outside record-audit target.

PMID 39316418 — Automated System to Capture Patient Symptoms From Multitype Japanese Clinical Texts: Retrospective Study. Not added: adverse-event extraction from mixed clinical document types, outside focused nursing-only supplement.

PMID 39176926 — Enhancing Healthcare Efficiency: Integrating ChatGPT in Nursing Documentation. Not added: brief adjacent generation/workflow report; no additional appraised audit-defect endpoint.

PMID 38913994 — A Scalable and Extensible Logical Data Model of Electronic Health Record Audit Logs for Temporal Data Mining (RNteract): Model Conceptualization and Formulation. Not added: EHR interaction-log model for workload, not content-quality audit.

PMID 38711689 — Embracing Artificial Intelligence: Revolutionizing Nursing Documentation for a Better Future. Not added: editorial, not empirical audit evaluation.

PMID 37863902 — SONAR, a nursing activity dataset with inertial sensors. Not added: sensor-based activity dataset, not audit of existing record content.

PMID 37528252 — Development and Validation of a Machine Learning Algorithm Using Clinical Pages to Predict Imminent Clinical Deterioration. Not added: deterioration prediction from clinical pages, outside scope.

PMID 36925384 — Predicting Workplace Violence in the Emergency Department Based on Electronic Health Record Data. Not added: workplace-violence prediction, outside documentation-audit scope.

PMID 35612225 — Clustering Nursing Sentences - Comparing Three Sentence Embedding Methods. Not added: adjacent sentence-clustering methods; no audit-defect endpoint.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

PMID 27919387 — Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data. Not added: broad secondary-use structuring review; not a new primary nursing-audit evaluation.

PMID 17604685 — Applying language technology to nursing documents: pros and cons with a focus on ethics. Not added: historical ethics discussion; governance role already covered, not an audit-effect study.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Table 1

Operational Dimensions for Auditing Electronic Nursing Documentation

Dimension	Proposed audit question	Interpretive boundary
Completeness	Are applicable assessments, actions, and evaluations documented?	Absence in available material does not establish omitted care.
Conformance	Does the entry meet the defined format, vocabulary, and identification rules?	Format compliance does not verify clinical truth.
Correctness and fidelity	Does the entry agree with an appropriate independent source?	A reference source is necessary; plausibility is insufficient.
Consistency	Do relevant entries agree after timing and context are considered?	Clinical change can explain an apparent contradiction.
Timeliness	Are event and documentation times appropriate and distinguishable?	Entry time is not necessarily bedside-event time.
Nursing-process coherence	Are assessments, diagnoses, plans, interventions, and evaluations linked?	A plausible generated link must not be treated as observed care.
Traceability and usefulness	Can reviewers identify the source and responsible professional?	Readable summaries should not replace authoritative records.

Note. Author-developed operationalization informed by De Groot et al. (2019), Kahn et al. (2016), Lewis et al. (2023), and Weiskopf and Weng (2013). The questions are illustrative, not a validated audit instrument.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Table 2

Empirical Evidence Map: Applications, Outcomes, and Transfer Limits

Source and role	Design / input	Reported result or contribution	Principal limitation
Lv et al. (2025) Direct AI audit	Retrospective pre/post; 556 records; blended QC workflow.	Fewer documentary defects; table medians 337 versus 91 seconds per audit.	Nonrandomized; time summaries and design labels conflict within the paper.
Niyirora et al. (2025) Mapping component	Home-health notes; models compared with nurse NIC classification.	Best model F1=0.31; lower than nurse performance.	Context-dependent and less common classes remain difficult; not defect detection.
Li et al. (2026) Mapping component	100 ICU records; CCC reference labels from three nurses.	Best RAG/few-shot F1: diagnoses 0.7836; interventions 0.8413.	Small local reference set; mapping does not verify undocumented events.
Moen et al. (2020) Structuring component	Nursing sentence classification and document organization.	Supports coherent organization under subject headings.	Organizational quality is different from comprehensive audit accuracy.
Abudari et al. (2025) Instrument evaluation	64 scenario-generated plans; 368 expert ratings.	Initial NPET validity and reliability evidence.	Simulated plans; average-rater reliability is not single-rater reliability.
Ju et al. (2025) Adjacent drafting	40 nurses; sequential virtual-patient documentation task.	Mean writing time 467.18 versus 182.68 seconds.	Simulation, subjective ratings, and an inconsistent percentage summary.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Table 2 (continued)

Empirical Evidence Map: Applications, Outcomes, and Transfer Limits

Source and role	Design / input	Reported result or contribution	Principal limitation
Chen et al. (2026) Adjacent handover	Pre/post implementation logs across three hospitals.	Shorter documentation time with nurse verification.	No formal gold-standard text-fidelity evaluation; bundled workflow.
Su et al. (2026) Adjacent voice drafting	Adapted speech recognition plus LLM DART output.	Macro-F1=0.82; reported hallucination rate 2.51%.	Residual factual errors; human verification and protocol-specific metrics.
Teo et al. (2026) Adjacent incident reporting	150 reports; voluntary pre/post use; clustered analysis.	Higher completeness and adapted PDQI ratings.	Self-selection; investigation reports differ from bedside notes.
Zhang et al. (2024) Automation comparator	Prospective temporal comparison; 600 ICU patients.	Fewer documentary problems; QC time 264 versus 62 seconds.	Data integration and automated checks are not an isolated LLM intervention.
von Gerich et al. (2022) Contextual empirical	Descriptive manual analysis of 230 cardiac-care records.	Identified expressions relevant to nursing-sensitive indicators.	Not an evaluated automated auditor or proof of patient benefit.
Wang et al. (2025) Contextual empirical	Qualitative study of information distortion.	Identified interacting organizational and individual factors.	Explains context; provides no AI audit effect estimate.
Kim et al. (2026) Mapping component	107 selected MIMIC-IV ICU intervention notes; GPT-4o mini; expert CCC references.	Contextualized-prompt match rates: 61.7% care components; 72.0% interventions; 72.9% sub-interventions.	Single model and selected notes; mapping, not documentation-defect detection or care verification.
Leurs et al. (2026) Adjacent event extraction	Retrospective single-hospital study; 30,615 nursing records; manual chart reference.	99 TP; 505 FP; 29,987 TN; 25 FN. Sensitivity ≈80%; specificity ≈98%; PPV ≈16%.	Chart-based fall extraction is not defect auditing; low PPV and estimated time comparator limit interpretation.

Note. AI = artificial intelligence; LLM = large language model; ICU = intensive care unit; MIMIC-IV = Medical Information Mart for Intensive Care IV; TP/FP/TN/FN = true positive/false positive/true negative/false negative; PPV = positive predictive value. QC = quality control; NIC = Nursing Interventions Classification; CCC = Clinical Care Classification; RAG = retrieval-augmented generation; NPET = Nursing Process Evaluation Tool; DART = data, action, response, and teaching; PDQI = Physician Documentation Quality Instrument. Sources are mapped separately to avoid counting generation or automation as direct AI audit validation. Values are not pooled.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Table 3

Proposed Endpoints for Prospective Evaluation of an AI-Supported Audit

Endpoint	Operational definition	Required reporting
Defect sensitivity	TP / (TP + FN) for reference-standard defects at the prespecified criterion level.	State eligible opportunities, defect count, severity, false negatives, and confidence intervals; distinguish record-level detection.
Flag precision	TP / (TP + FP) among flags with resolved reference adjudication.	Also report all flags, unresolved flags, adjudication coverage, and prevalence; retain rejection reasons and difficult cases.
Critical-error analysis	Missed or incorrectly resolved findings in locally defined low-, moderate-, and high-consequence categories.	Prespecify severity before evaluation; ideally blind severity adjudicators to model identity; report high-consequence errors separately.
Net audit effort	Preparation + retrieval + screening + verification + correction + escalation + monitoring + maintenance time.	Report minutes per record and per confirmed defect, flags reviewed per confirmed defect, and corrections per 100 records.
Unnecessary correction	Record amendments initiated from unsupported flags.	Verify provenance and whether changes introduced inaccuracies.
Abstention and unresolved cases	Unresolved or abstained eligible cases / all eligible cases × 100%.	Separate system abstention from unresolved adjudication; report overlap and reasons. Present coverage with resolved-case metrics.
External and temporal transfer	Performance at another institution or language and in an independent later period.	Report template, terminology, population, and system versions; prevent patient/encounter leakage; retest after material changes.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Table 3 (continued)

Endpoint	Operational definition	Required reporting
Human factors	Alert burden, override reasons, reviewer reliance, and interruption costs.	Combine observable workflow measures with perceptions, reported separately.
Clinical outcomes	Communication or patient outcomes in a separately designed evaluation.	Do not infer from F1, writing time, or documentary completeness.
Specificity and false flags	Specificity = TN / (TN + FP); false-positive rate = FP / (FP + TN), when negative opportunities are meaningful.	Prespecify criterion/document denominator. Otherwise report false flags per criterion or record and flags reviewed per confirmed defect.
Defect prevalence and uncertainty	Reference-standard defects / all eligible audited opportunities at the stated unit.	Report numerator, denominator, prevalence, confidence intervals, unresolved cases, and clustering by patient/encounter/nurse.
Retrieval and interpretation failures	Failure to retrieve an authorized relevant entry versus incorrect interpretation of an entry actually inspected.	Record retrieval status, source scope/location, access restrictions, and uncertainty; report these failures separately.
Silent-phase baseline	Raw flag rate, sensitivity, precision, abstention, and reviewer time before alerts influence users.	State eligible units and reference adjudication; separate algorithmic baseline from subsequent behavior or workflow changes.

Note. Author-proposed evaluation protocol informed by Tables 1–2 and reporting principles in Gallifant et al. (2025) and Vasey et al. (2022); it is not a validated composite score. TP/FP/TN/FN = true positive/false positive/true negative/false negative. Prespecify units, denominators, severity, and cluster structure. Use cluster-robust intervals, mixed-effects models, or generalized estimating equations when repeated observations require them. Undefined ratios must be reported as undefined, with their counts.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Table 4

Author-Proposed Nurse-Supervised Governance Workflow

Stage	Required control	Decision responsibility
1. Define scope	Specify applicable criteria, exclusions, source records, and escalation rules.	Nursing audit leadership; clinical and information-system specialists.
2. Authorize access	Limit episode access; document lawful processing, retention, and export rules.	Institutional data controller, security team, and authorized users.
3. Detect bounded findings	Combine deterministic checks with task-specific AI; preserve source links. Record retrieval failure separately from interpretation failure.	Technical team operates detection; model output remains provisional.
4. Adjudicate	Confirm, reject, or defer each consequential flag using original evidence. Model-generated rationale is explanatory output, not independent clinical evidence.	Qualified nurse auditor with independent task competence.
5. Correct with provenance	Prevent automatic invention of events; retain amendment author and time.	Authorized professional follows institutional record-correction procedure.
6. Monitor and reevaluate	Track false flags, missed defects, burden, incidents, and version changes. Include temporal validation and regression tests after material updates.	Nursing, informatics, safety, privacy, and security governance.

Note. Conceptual synthesis proposed by this review, informed by WHO (2024), Tabassi (2023), and Brazil-specific documentation responsibilities in COFEN (2024a, 2024b). It has not been implemented or validated and is not an endorsed standard, certification, or scoring instrument. WHO = World Health Organization; COFEN = Conselho Federal de Enfermagem.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

References

Abudari, M. O., Abu-abbas, M., Al-Ma'ani, M., Alradaydeh, M. F., & Alduraidi, H. (2025). Development and validation of the nursing process evaluation tool (NPET): A multidimensional instrument for assessing the quality of AI-generated nursing documentation. *BMC Nursing, 24*, Article 1422. https://doi.org/10.1186/s12912-025-04068-8

Brasil. (2018). Lei nº 13.709, de 14 de agosto de 2018: Lei Geral de Proteção de Dados Pessoais (texto compilado) [General Personal Data Protection Law, consolidated text]. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709compilado.htm

Chen, R.-J., Wu, M.-S., Tsai, L.-W., Chang, S.-S., Shen Hsiao, S.-T., & Lo, Y.-S. (2026). Integrating a large language model to streamline nursing handover documentation across multiple hospitals in Taiwan: Development and implementation study. *Journal of Medical Internet Research, 28*, Article e81604. https://doi.org/10.2196/81604

Conselho Federal de Enfermagem. (2024a). Resolução Cofen nº 736, de 17 de janeiro de 2024: Dispõe sobre a implementação do Processo de Enfermagem em todo contexto socioambiental onde ocorre o cuidado de enfermagem. [Resolution No. 736: Implementation of the Nursing Process]. https://www.cofen.gov.br/resolucao-cofen-no-736-de-17-de-janeiro-de-2024/

Conselho Federal de Enfermagem. (2024b). Resolução Cofen nº 754, de 16 de maio de 2024: Normatiza o uso do prontuário eletrônico e plataformas digitais no âmbito da Enfermagem [Resolution No. 754: Electronic records and digital platforms in nursing]. https://www.cofen.gov.br/resolucao-cofen-no-754-de-16-de-maio-de-2024/

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

De Groot, K., Triemstra, M., Paans, W., & Francke, A. L. (2019). Quality criteria, instruments, and requirements for nursing documentation: A systematic review of systematic reviews. *Journal of Advanced Nursing, 75*(7), 1379–1393. https://doi.org/10.1111/jan.13919

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., Demner-Fushman, D., Dligach, D., Daneshjou, R., Fernandes, C., Hansen, L. H., Landman, A., Lehmann, L., McCoy, L. G., Miller, T., Moreno, A., Munch, N., Restrepo, D., Savova, G., … Bitterman, D. S. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine, 31*, 60–69. https://doi.org/10.1038/s41591-024-03425-5

Ju, H., Park, M., Jeong, H., Lee, Y., Kim, H., Seong, M., & Lee, D. (2025). Generative AI-based nursing diagnosis and documentation recommendation using virtual patient electronic nursing record data. *Healthcare Informatics Research, 31*(2), 156–165. https://doi.org/10.4258/hir.2025.31.2.156

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs, 4*(1), Article 1244. https://doi.org/10.13063/2327-9214.1244

Kim, Y., Kim, J., Cho, Y., & Choi, M. (2026). Mapping ICU nursing intervention notes to clinical care classification using a large language model: An evaluation study. *BMC Nursing, 25*, Article 880. https://doi.org/10.1186/s12912-026-04949-6

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Leurs, W. L. M., Groeneveld, M., Haan, B., Lammers, L. A. S., Bouwman, A. R. A., & van der Linden, C. M. J. (2026). Using AI to identify in-hospital falls: A comparative analysis with incident reports and manual chart review. *BMJ Open Quality, 15*(2), Article e003786. https://doi.org/10.1136/bmjoq-2025-003786

Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R. O., & Gupta, A. (2023). Electronic health record data quality assessment and tools: A systematic review. *Journal of the American Medical Informatics Association, 30*(10), 1730–1740. https://doi.org/10.1093/jamia/ocad120

Li, M., Chen, D., Xiao, Q., He, Z., Zhang, Y., Zhong, J., Luo, Y., & Ma, H. (2026). Automated identification of nursing diagnoses and interventions from nursing records using a retrieval-augmented large language model approach: Quantitative study. *Journal of Medical Internet Research, 28*, Article e89850. https://doi.org/10.2196/89850

Lv, J., Xu, Y., Jiang, M., Lv, Y., Sun, J., Lu, J., Wang, L., & Wang, H. (2025). A DeepSeek-powered locally deployed closed-loop system for enhancing quality control in electronic nursing documentation: Development and clinical validation. *Journal of the American Medical Informatics Association, 32*(10), 1526–1532. https://doi.org/10.1093/jamia/ocaf109

Mitha, S., Schwartz, J., Hobensack, M., Cato, K., Woo, K., Smaldone, A., & Topaz, M. (2023). Natural language processing of nursing notes: An integrative review. *CIN: Computers, Informatics, Nursing, 41*(6), 377–384. https://doi.org/10.1097/CIN.0000000000000967

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Moen, H., Hakala, K., Peltonen, L.-M., Matinolli, H.-M., Suhonen, H., Terho, K., Danielsson-Ojala, R., Valta, M., Ginter, F., Salakoski, T., & Salanterä, S. (2020). Assisting nurses in care documentation: From automated sentence classification to coherent document structures with subject headings. *Journal of Biomedical Semantics, 11*, Article 10. https://doi.org/10.1186/s13326-020-00229-7

Niyirora, J., Longtin, L., Grabski, C., Patrishkoff, D., & Semko, A. (2025). A comparative analysis of machine learning models and human expertise for nursing intervention classification. *JAMIA Open, 8*(3), Article ooaf057. https://doi.org/10.1093/jamiaopen/ooaf057

Su, M.-H., Wang, W.-C., Hsu, Y.-M., Hou, S.-Y., Chuang, S.-J., & Chang, S.-S. (2026). Voice-based structured nursing documentation using automatic speech recognition and large language models: Development and evaluation study. *JMIR Nursing, 9*, Article e88567. https://doi.org/10.2196/88567

Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Teo, K. Y., Huang, L., Woh, K. C. Y., Ang, S. Y., & Ng, J. G. N. (2026). Nurse-led ambient AI scribe for patient safety incident investigation reports (project NARRATE): Retrospective pre-post comparative document-quality study. *JMIR Nursing, 9*, Article e100775. https://doi.org/10.2196/100775

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., … McCulloch, P. (2022). Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ, 377*, Article e070904. https://doi.org/10.1136/bmj-2022-070904

von Gerich, H., Moen, H., & Peltonen, L.-M. (2022). Identifying nursing sensitive indicators from electronic health records in acute cardiac care—towards intelligent automated assessment of care quality. *Journal of Nursing Management, 30*(8), 3726–3735. https://doi.org/10.1111/jonm.13802

Wang, J., Xu, Y., Yang, Z., Zhang, J., Zhang, X., Li, W., Sun, Y., & Pan, H. (2025). Factors influencing information distortion in electronic nursing records: Qualitative study. *Journal of Medical Internet Research, 27*, Article e66959. https://doi.org/10.2196/66959

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association, 20*(1), 144–151. https://doi.org/10.1136/amiajnl-2011-000681

World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. https://iris.who.int/handle/10665/375579

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 05/09/2026 - **Accepted:** 08/10/2026
Vol: 01.05
DOI: 10.66104/rbsr1p60
Pages: 1–39

Zhang, S., Quan, Y. Y., & Chen, J. (2024). Construction and application of an ICU nursing electronic medical record quality control system in a Chinese tertiary hospital: A prospective controlled trial. *BMC Nursing, 23*, Article 493. https://doi.org/10.1186/s12912-024-02178-3

CC BY