

REVIEW

Open access

Artificial Intelligence for Clinical Documentation Improvement from 2017 to 2023: A Systematic Review of Natural Language Processing Methods for Coding Accuracy, Note Quality, Billing Support, and Physician Workflow Efficiency

Ahmed Benali^{1*}, Karim Boudiaf², Samir Touati¹

Abstract

Clinical documentation is central to continuity of care, coding, billing, compliance, and quality measurement, yet it remains a major source of administrative burden for physicians. Artificial intelligence, especially natural language processing, has been proposed as a way to improve documentation quality, coding accuracy, billing support, and clinical workflow efficiency. This systematic review synthesised evidence from 2017 to 2023 on natural language processing methods applied to clinical documentation improvement. The review focused on automated clinical coding, note quality, billing support, computer-assisted physician documentation, and physician workflow efficiency. A PRISMA 2020-compliant search strategy was applied to PubMed, Scopus, IEEE Xplore, and Web of Science for publications from 1 January 2017 to 31 December 2023. Screening was performed in duplicate, and eligible studies were narratively synthesised by documentation domain, model type, deployment maturity, and evaluation approach. The evidence base expanded rapidly during the review period, especially in automated coding and clinical note generation. Several studies reported technically promising systems for ICD coding, documentation summarisation, and speech-derived notes, whereas billing outcomes and sustained workflow effects were less commonly evaluated. Natural language processing for clinical documentation improvement appears most mature for automated coding and note analysis. Evidence for billing support and physician workflow transformation remains less developed, particularly in prospective clinical environments.

Keywords Natural language processing, Artificial intelligence, Clinical documentation improvement, Automated coding, Physician workflow, Billing support

*Correspondence:

Ahmed Benali
ahmed.benali@gmail.com

¹ Department of Digital Healthcare Systems, Faculty of Medicine, University of Algiers, Algiers, Algeria

² Department of Health Data Analytics, Faculty of Engineering, University of Tunis El Manar, Tunis, Tunisia

Introduction

Clinical documentation has become both a clinical asset and an administrative burden, shaping continuity of care, reimbursement, compliance, quality reporting, and medico-legal accountability. Studies using electronic health record logs and time-motion observations reported that physicians spend substantial time on documentation and desktop medicine, contributing to perceived workload and after-hours work [1, 2]. Documentation burden also has implications for billing accuracy and regulatory compliance, because incomplete, redundant, or poorly structured notes may weaken the link

between clinical care, coded diagnoses, and medical necessity [3, 4]. These concerns created a strong rationale for clinical documentation improvement strategies that extend beyond conventional auditing and manual coding workflows [5].

Natural language processing has emerged as a major technical approach for clinical documentation improvement because much of the clinically and financially relevant evidence is embedded in free-text notes. Reviews of clinical natural language processing described a growing body of systems for extracting concepts, normalising terminology, identifying clinical events, and structuring narrative documentation [6, 7]. Since 2017, the field has increasingly shifted from rule-based extraction and classical machine learning toward neural architectures, attention mechanisms, and transformer-based approaches for coding and documentation-related tasks [8, 9]. This technical evolution has made it possible to address clinical note generation, automated coding, and documentation quality at a scale that would be difficult through manual review alone [10, 11].

The evidence base remains fragmented across several domains that are often evaluated separately. Automated coding studies commonly focus on ICD code prediction from discharge summaries or clinical notes, whereas note quality studies examine redundancy, copy-forward behaviour, documentation completeness, or speech recognition errors [12-15]. Workflow studies more often assess ambient speech, automated medical scribes, and physician time, while billing support is frequently inferred through coding alignment rather than directly measured through denials or revenue cycle outcomes [15, 16]. As a result, prior evidence has not always connected coding accuracy, note quality, billing compliance, and physician workflow into a single clinical documentation improvement framework [17, 18].

Figure 1 illustrates the integrated clinical documentation improvement framework linking natural language processing methods to coding accuracy, note quality, billing support, and physician workflow outcomes.

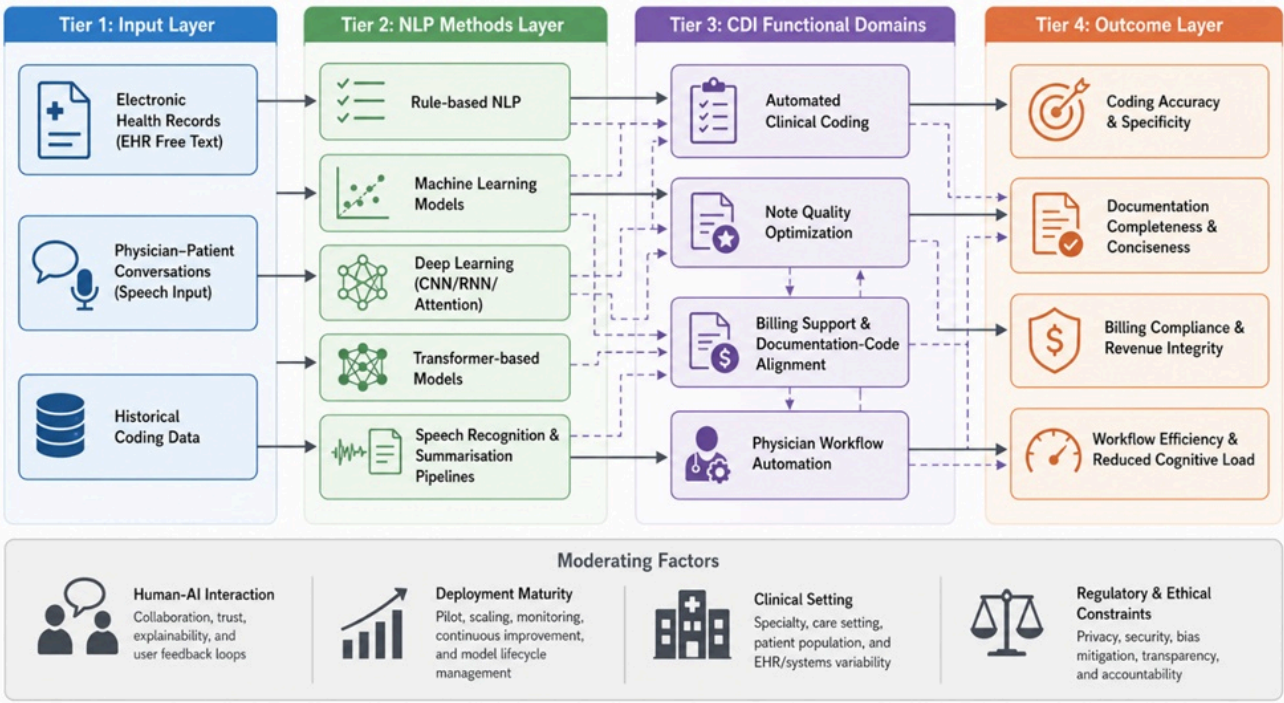


Figure 1. Integrated Clinical Documentation Improvement (CDI) Framework Linking NLP Capabilities to Clinical, Operational, and Financial Outcomes

The objective of this systematic review was to synthesise peer-reviewed evidence from 2017 to 2023 on artificial intelligence and natural language processing for clinical documentation improvement. The review followed PRISMA 2020

principles and examined four linked domains: automated clinical coding, note quality, billing support, and physician workflow efficiency. Because computer-assisted physician documentation and ambient documentation systems increasingly blend these domains, the review also considered human factors, deployment maturity, and evaluation methods. The emphasis was on systematic review-level synthesis rather than meta-analysis, because the included studies varied substantially in clinical setting, documentation type, model architecture, and outcome definition.

Materials and Methods

Search strategy

A structured search was conducted across PubMed, Scopus, IEEE Xplore, and Web of Science for studies published from 1 January 2017 to 31 December 2023. Search terms combined concepts for natural language processing, clinical documentation improvement, automated coding, note quality, billing support, computer-assisted physician documentation, ambient speech, physician workflow, ICD-10, CPT, SNOMED CT, and electronic health records [6, 7]. Example search concepts included natural language processing for clinical documentation improvement, automated coding from clinical notes, documentation completeness and accuracy, billing support using artificial intelligence, and ambient documentation for physician workflow.

Inclusion and exclusion criteria

Studies were eligible if they reported peer-reviewed research on artificial intelligence, machine learning, natural language processing, speech recognition, or automated summarisation applied to clinical documentation tasks. Eligible domains included automated coding, clinical note generation, note quality assessment, documentation burden, billing support, documentation-code alignment, and computer-assisted physician documentation. Studies were excluded if they were not focused on clinical documentation, did not involve clinical text or documentation workflow, were editorials or opinion pieces without empirical or systematic review content, or fell outside the publication window except where a prior review was used to contextualise the field. English-language studies were prioritised because the review focused on documentation systems and evaluation methods that could be consistently compared across included publications.

Screening and selection

Records were screened in duplicate using title and abstract review followed by full-text assessment for studies that appeared relevant to clinical documentation improvement. The search identified 1,904 records, of which 1,512 remained after deduplication; 250 full-text articles were assessed, and 31 publications were selected in the final synthesis after exclusions for non-documentation focus, insufficient NLP detail, non-clinical text, or lack of relevance to CDI outcomes.

Figure 2 presents the PRISMA 2020 flow diagram detailing the identification, screening, eligibility, and inclusion of studies in this review.

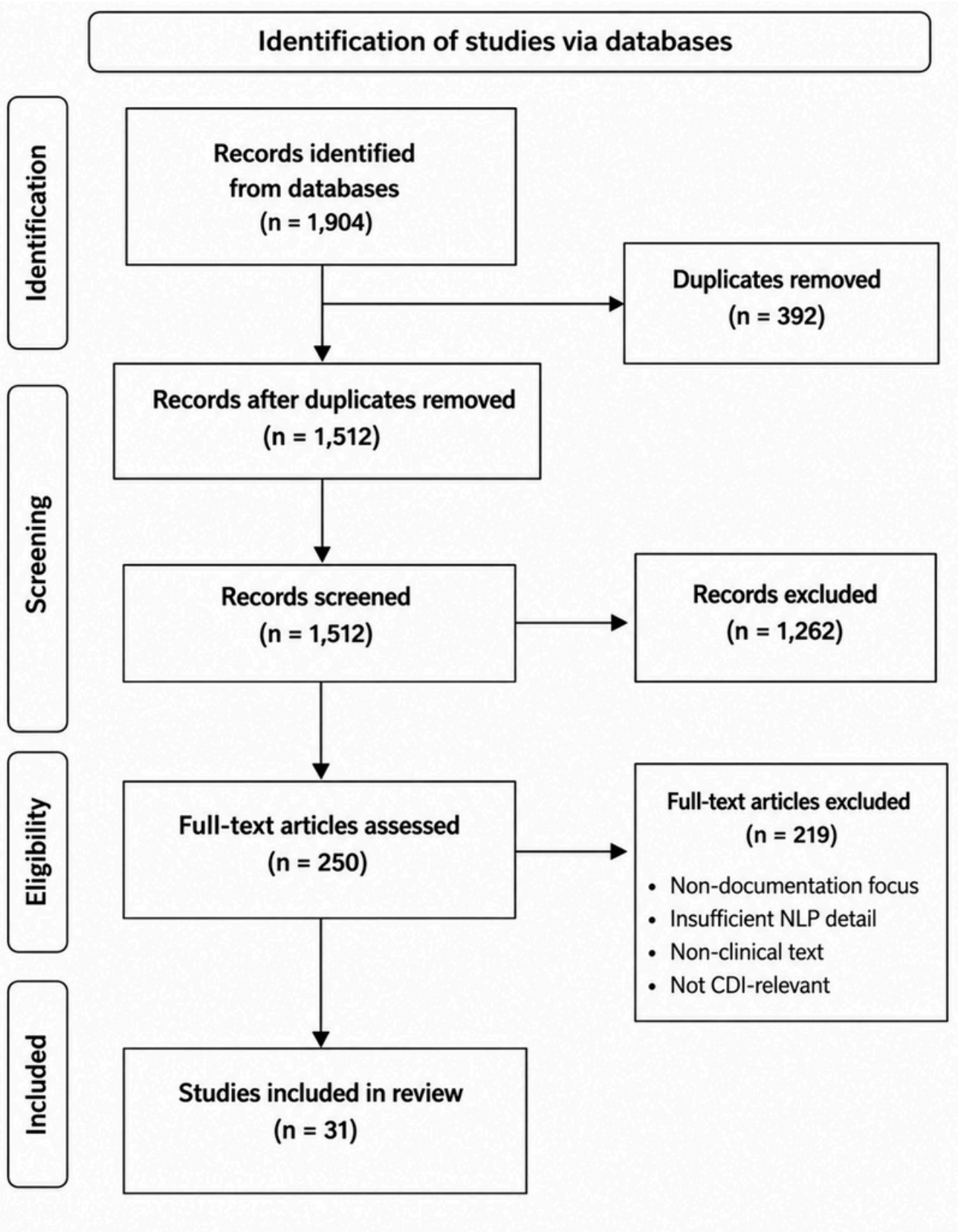


Figure 2. PRISMA 2020 Flow Diagram of Study Selection for AI-Based Clinical Documentation Improvement Review

Data extraction

Data extraction captured bibliographic information, clinical setting, documentation type, NLP method, task definition, evaluation metrics, outcome domain, and deployment status. For automated coding studies, extraction focused on code system, model type, input note type, label granularity, explainability, and whether the system was evaluated retrospectively or in a clinical workflow [8–10]. For note quality, billing, and workflow studies, extraction captured documentation completeness, redundancy, speech-derived error patterns, physician time, after-hours charting, user experience, and evidence of revenue-cycle relevance [3, 4, 14]. Each study was assigned to one or more CDI domains because several systems addressed multiple downstream objectives, such as coding support and documentation completeness [18, 19].

Risk of bias assessment

Risk of bias was assessed narratively using domains adapted from PROBAST-AI and from methodological concerns specific to clinical NLP evaluation. Relevant domains included representativeness of training data, label quality, clinical setting, missingness, local coding practices, external validation, model transparency, and appropriateness of outcome measures [8, 10]. For workflow and documentation burden studies, additional attention was given to whether the design relied on retrospective logs, observation, self-report, or prospective deployment, because each approach captures a different aspect of physician work [1, 2, 4]. Studies of speech recognition and automated note generation were also reviewed for human correction requirements and potential documentation safety risks [14, 16].

Synthesis methods

Because the studies varied substantially in tasks, datasets, clinical settings, outcome measures, and deployment maturity, a meta-analysis was not performed. Instead, findings were narratively synthesised across four domains: automated coding, note quality, billing support, and physician workflow efficiency [7, 10]. Vote-counting by conceptual category was used to describe which model families and evaluation outcomes were most common, including convolutional neural networks, recurrent neural networks, attention models, transformer-based models, speech-recognition pipelines, and sequence-to-sequence summarisation [8, 9, 18]. The synthesis emphasised whether systems were technically validated, clinically deployed, evaluated by users, or linked to measurable financial and workflow outcomes [4, 15].

Results and Discussion

Study selection

The final cited evidence set included 31 publications that collectively represented systematic reviews, automated coding models, clinical note quality studies, documentation burden analyses, speech-recognition evaluations, and automated note generation systems. The most common reasons for exclusion at full-text review were absence of a clinical documentation task, insufficient detail on natural language processing methods, focus on non-clinical text, or lack of relevance to coding, note quality, billing, or workflow outcomes [6, 7]. Several excluded studies addressed general EHR prediction or clinical decision support but did not evaluate documentation improvement as an explicit target. The included set therefore prioritised studies with direct implications for clinical documentation improvement rather than broader medical artificial intelligence [10, 15].

Study characteristics

The evidence base showed a clear increase in publications after 2017, with early work focusing heavily on documentation burden, speech recognition, and neural coding models, followed by greater interest in transformers and automated note generation. Studies were conducted across inpatient, outpatient, emergency, and primary care settings, although many coding models relied on retrospective hospital datasets and benchmark corpora rather than prospective clinical

deployments [8, 10, 12]. Geographically, the evidence included studies from North America, Europe, and Asia, with language and coding-system differences shaping the transferability of automated coding models [13, 20]. Clinical documentation improvement was therefore represented not as a single intervention type but as a collection of NLP-enabled tasks embedded in different health system contexts [6, 21].

NLP for automated coding: accuracy and methods

Automated coding was the most technically developed domain, with studies applying neural and attention-based methods to assign ICD or related codes from free-text clinical documentation. Early deep learning approaches used convolutional, recurrent, and attention mechanisms, while later systems increasingly examined transformer-based or hybrid architectures for mapping long clinical notes to diagnosis codes [8, 9, 22]. Several studies reported that label-aware attention, residual convolutional networks, semantic matching, and knowledge integration could improve code assignment relative to simpler baselines, although evaluation conditions varied across datasets and code systems [13, 23, 24]. Systematic synthesis suggests that automated coding is the CDI domain with the strongest technical literature but not necessarily the strongest evidence of clinical implementation [10].

NLP for automated coding: real-world adoption

Despite extensive modelling work, prospective adoption of automated coding systems appeared limited in the included literature. Several studies evaluated models retrospectively on historical notes, which is valuable for technical benchmarking but does not fully capture coder trust, physician documentation behaviour, local coding variation, or payer-specific requirements [11, 12, 25]. Studies that used explainable approaches and label attention were relevant to implementation because they attempted to make code predictions interpretable to coders and clinicians [8, 26]. However, the review found that real-world adoption remains constrained by differences between benchmark performance and the operational requirements of coding departments [10].

NLP for note quality: completeness and conciseness

Note quality studies addressed a different aspect of clinical documentation improvement by examining whether notes are complete, accurate, concise, and clinically meaningful. Evidence on note length and redundancy suggested that outpatient progress notes may become longer and more repetitive over time, creating risks of information overload and copy-forward bloat [3]. Studies of documentation concordance and speech-recognition-assisted notes also showed that clinical documents may diverge from observed physician behaviour or contain errors introduced during dictation and transcription workflows [14, 27]. These findings support the relevance of NLP tools that detect missing elements, reduce redundant text, and identify potential documentation inconsistencies [28].

NLP for note quality: impact on clinician satisfaction

Evidence linking NLP-driven note quality improvement to clinician satisfaction was less developed than evidence describing documentation burden itself. EHR log studies and workload assessments showed that documentation consumes substantial physician and nursing time, but fewer studies evaluated whether NLP interventions directly improved perceived note usefulness, satisfaction, or burnout-related outcomes [1, 4]. Automated summarisation and SOAP-note generation systems suggested a potential route to more concise and structured documentation, although many evaluations remained technical rather than user-centred [18, 19]. Overall, the literature indicates that reducing note length alone should not be treated as equivalent to improving clinical documentation quality unless completeness, accuracy, and clinician trust are also assessed [3, 28].

NLP for billing support: medical necessity and compliance

Billing support appeared in the evidence base primarily through automated coding, documentation-code alignment, and extraction of clinical evidence that could justify diagnoses or procedures. Automated ICD coding studies implicitly address

revenue-cycle needs because diagnosis codes influence reimbursement, risk adjustment, quality reporting, and audit exposure [10, 13]. Explainable coding models were particularly relevant to billing support because they attempted to connect predicted codes with textual evidence in the note, which may assist coders in verifying medical necessity and documentation sufficiency [8, 26]. However, the literature rarely evaluated denial prevention, payer audit outcomes, or charge capture as primary endpoints [10].

NLP for billing support: financial outcomes

Financial outcomes were among the least directly measured endpoints in the included literature. Although automated coding studies frequently evaluated accuracy-oriented metrics, few studies connected model outputs to actual reimbursement, denial reduction, coder productivity, or audit outcomes in deployed revenue-cycle workflows [9, 23, 29]. This gap may reflect commercial sensitivity, institutional variation in billing rules, and the difficulty of attributing financial outcomes to documentation tools alone. Consequently, the review found that billing support is a high-potential CDI domain but one with weaker independent evidence than automated coding or documentation burden measurement [4, 10].

NLP for physician workflow: ambient speech and note generation

Physician workflow studies focused on reducing documentation burden through speech recognition, automated medical scribes, and generation of structured clinical notes from patient-physician conversations. Automated scribe and medical conversation summarisation studies reported pipelines that transform dialogue into documentation, including systems for generating reports or SOAP-style notes [17-19]. These studies are important for CDI because generated notes can affect coding specificity, note completeness, billing defensibility, and physician time. However, evaluation methods varied substantially, with many systems assessed through technical or simulated documentation outputs rather than sustained clinical deployment [15].

NLP for physician workflow: after-hours charting and cognitive load

Documentation burden studies using EHR logs showed that physicians divide work between patient-facing care and desktop medicine, with after-hours documentation contributing to workload concerns [1, 2]. Clinical documentation burden measurement using event log metadata provided a foundation for evaluating whether AI documentation tools reduce time in notes, orders, messages, and other EHR activities [4]. Although some documentation automation systems aimed to reduce cognitive load by converting conversations into structured notes, fewer studies measured task switching, after-hours charting, or cognitive burden as primary outcomes [15,19]. The evidence therefore supports workflow efficiency as a key promise of AI-CDI but not yet as a consistently demonstrated outcome [5].

Model types and technical approaches

The technical trajectory moved from rule-based and classical NLP methods toward deep learning, attention, and transformer-based architectures. Automated coding studies included convolutional neural networks, recurrent convolutional models, label attention models, semantic matching methods, knowledge integration, and transformer-based representations [8, 9, 11, 22]. Speech and note generation studies used automated speech recognition, sequence-to-sequence modelling, modular summarisation, and conversation-to-note pipelines [16, 18, 19]. Across domains, the strongest methodological trend was the increasing use of models that encode long narrative text while attempting to preserve interpretability for clinical or coding users [23, 26].

Table 1 provides a cross-domain analytical mapping of NLP capabilities to clinical documentation improvement outcomes and their relative evidence maturity.

Table 1. Cross-Domain Analytical Mapping of NLP Capabilities to CDI Outcomes and Evidence Maturity

NLP Capability	Primary CDI Domain	Secondary Effects Across Domains	Typical Evaluation Metrics	Evidence Maturity	Key Implementation Gap
Automated ICD Coding	Coding Accuracy	Billing compliance, audit defensibility	F1 score, precision/recall, top-k accuracy	High (retrospective)	Limited prospective validation and coder trust integration
Clinical Note Summarisation	Note Quality	Coding specificity, workflow efficiency	ROUGE, BLEU, human evaluation	Moderate	Risk of information omission affecting billing and care continuity
Speech Recognition & Ambient Scribes	Workflow Efficiency	Note quality, coding downstream effects	Word error rate, time saved	Moderate	Error propagation and need for human correction
Documentation Completeness Detection	Note Quality	Billing justification, audit readiness	Recall of missing elements, rule-based checks	Low–Moderate	Lack of linkage to clinical and financial endpoints
Documentation-Code Alignment Models	Billing Support	Coding accuracy, compliance	Alignment accuracy, explainability scores	Low	Rare evaluation against real payer audit outcomes
Explainable AI for Coding	Coding + Billing	Trust, workflow integration	Attention visualization, interpretability metrics	Emerging	Limited validation in real-world coder workflows
Conversation-to-Note Generation	Workflow + Note Quality	Coding accuracy, clinician satisfaction	Human review, structural completeness	Emerging	Insufficient longitudinal workflow impact data

Evaluation metrics

Evaluation metrics varied by CDI domain and often reflected technical convenience more than operational relevance. Automated coding studies commonly relied on accuracy-oriented and classification metrics, while note generation studies used text-generation and human-review methods, and documentation burden studies used EHR logs, time-motion observations, or workload proxies [1, 8, 18]. Note quality studies examined length, redundancy, concordance, and error types, but these measures were not always connected to patient safety, billing defensibility, or clinician satisfaction [3, 14, 27]. Billing support evaluations were particularly limited because claim-level accuracy, denial prevention, and revenue-cycle return were rarely reported in a prospective and independently replicated way [10].

Deployment challenges and human factors

Deployment challenges were repeatedly linked to EHR integration, clinician trust, workflow fit, transparency, local documentation norms, and the risk of introducing new errors. Automated coding models may perform well on retrospective datasets but still require integration with coder workflows, local code policies, and audit processes [10, 13]. Speech and automated note systems must also manage transcription errors, summarisation omissions, inappropriate

phrasing, and physician accountability for AI-generated documentation [14, 15]. Human factors are therefore central to AI-CDI because clinicians and coders must decide when to accept, edit, override, or ignore machine-generated suggestions [8, 28].

Coding accuracy leads the evidence base

Automated clinical coding was the most mature area in the reviewed literature, supported by a dense technical literature using neural architectures, attention mechanisms, and transformer-based representations. Several studies reported that modern models can assign diagnosis codes from narrative clinical notes under retrospective evaluation conditions, but these results do not automatically translate into operational coding reliability [8, 11, 22]. Explainability and evidence highlighting are important because coders must understand why a code was suggested and whether the documentation supports it [8, 26]. The central gap is therefore not whether models can predict codes in principle, but whether they can be safely embedded into real-world coding and documentation workflows [10].

Note quality improvements often lack meaningful clinical endpoints

The note quality literature showed that documentation can become long, redundant, discordant, or vulnerable to speech-recognition errors. Studies on note length, observed behaviour, and dictated documents indicated that documentation quality cannot be reduced to the mere presence of more text [3, 14, 27]. NLP systems that reduce redundancy or generate summaries may improve usability, but they must preserve clinically relevant detail and avoid omitting information needed for billing, audit, or care coordination [18, 19]. Future studies should therefore treat conciseness, completeness, accuracy, and clinical usefulness as linked but distinct outcomes [28].

Billing support remains the least mature domain

Billing support was frequently implied through automated coding but less often evaluated as a primary clinical documentation improvement outcome. Coding studies are relevant to reimbursement because they connect clinical text to billable diagnoses, but most evaluations stopped at code prediction rather than measuring denial prevention, audit outcomes, or documentation sufficiency [13, 23, 29]. Explainable models could support billing workflows by showing the textual evidence behind suggested codes, yet this capability still requires validation against payer rules and medical necessity criteria [8, 26]. The review therefore suggests that billing support is an important but underdeveloped bridge between NLP research and revenue-cycle practice [10].

Physician workflow studies reflect a promise yet to be verified at scale

Physician workflow improvement is a major motivation for AI documentation systems, but the evidence remains uneven. EHR log studies documented the scale of documentation burden, while automated scribe and note-generation studies demonstrated technical feasibility for converting conversations into structured documentation [1, 4, 16]. However, few studies provided long-term independent evidence that these tools reduce after-hours charting, improve clinician satisfaction, preserve note quality, and maintain billing integrity in routine care [15, 19]. This limits confidence in broad claims that AI documentation systems have already transformed physician workflow at scale [5].

The integration problem: NLP as a feature, not a product

Many included studies evaluated standalone models rather than tools embedded in native EHR documentation, coding, and billing workflows. This distinction matters because an NLP model that performs well offline may fail if it interrupts clinicians, generates excessive suggestions, requires duplicative review, or does not align with local templates and coding rules [10, 15]. Documentation burden studies show that workflow context is essential when assessing time and efficiency, because EHR work is distributed across notes, orders, inboxes, and after-hours tasks [1, 4]. AI-CDI should therefore be evaluated as a workflow intervention rather than merely as a predictive model [5].

Table 2 presents a conceptual evaluation framework for assessing AI-driven clinical documentation improvement across clinical, operational, and financial dimensions.

Table 2. Conceptual Framework for Evaluating AI-Driven Clinical Documentation Improvement Across Clinical, Operational, and Financial Dimensions

Evaluation Dimension	Subcomponents	Current Measurement Approach	Limitation Identified in Review	Recommended Future Metric
Clinical Documentation Quality	Completeness, accuracy, conciseness	Note length, redundancy, concordance	Weak linkage to clinical outcomes	Composite clinical documentation quality index
Coding Reliability	Code accuracy, specificity, consistency	Classification metrics (F1, precision)	Lack of real-world coding validation	Agreement with expert coders + audit outcomes
Billing Integrity	Medical necessity, compliance, denial rates	Proxy via coding accuracy	Rare direct financial evaluation	Claim acceptance rate, denial reduction, revenue delta
Workflow Efficiency	Time in EHR, after-hours work, task switching	EHR logs, time-motion studies	Limited longitudinal evidence	Sustained time savings and cognitive load indices
Human-AI Interaction	Acceptance, override, editing burden	Rarely measured	Critical gap in usability evidence	Acceptance rate, override taxonomy, editing time
Deployment Maturity	Retrospective vs prospective use	Dataset-based validation	Lack of real-world implementation studies	Multi-site prospective trials
Equity & Bias	Performance across populations	Rare subgroup analysis	Underdeveloped domain	Stratified performance and fairness metrics

Human-AI collaboration and override behavior

Human-AI collaboration remains under-studied in clinical documentation improvement. Explainable coding models and automated note systems can provide suggestions, but the practical value depends on whether physicians and coders accept, revise, or override those outputs [8, 26]. Studies of documentation variation and concordance suggest that clinicians already differ in how they document care, which may influence how they respond to CDI prompts or generated text [27, 28]. Future evaluation should therefore examine acceptance rates, override reasons, editing burden, and the effect of suggestions on clinical reasoning and documentation ownership [15].

Regulatory and ethical landscape

AI-generated clinical documentation raises ethical and regulatory questions related to accountability, documentation integrity, privacy, bias, and auditability. Speech recognition and automated note generation can introduce errors or omissions that may affect patient safety, coding, billing, and medico-legal responsibility [14, 16]. Automated coding and documentation-code alignment systems may also reflect historical documentation and coding practices, including local biases and incomplete labels [10, 11]. These issues suggest that CDI systems require transparent audit trails, human oversight, monitoring for drift, and governance processes that extend beyond model development [6, 7].

Limitations

Review limitations

This review was limited by the heterogeneity of clinical documentation tasks, model architectures, datasets, clinical settings, and outcome measures. Because automated coding, note quality, billing support, and workflow efficiency were evaluated using different endpoints, a quantitative meta-analysis was not appropriate [7, 10]. The English-language focus may have underrepresented non-English documentation systems, although several studies demonstrated that language and coding-system context can materially influence model transferability [20]. The review also relied on published literature, which may overrepresent technically successful systems and underrepresent failed deployments or commercially sensitive billing outcomes [15].

Evidence base limitations

The evidence base was strongest for retrospective automated coding and weaker for prospective workflow, billing, and implementation outcomes. Many studies used historical datasets or controlled evaluations rather than real-world deployment, limiting conclusions about clinician adoption, coder trust, revenue-cycle impact, and long-term sustainability [8-10]. Documentation burden and note quality studies provided important context but did not always test AI-CDI interventions directly [1, 3, 4]. As a result, the current literature supports cautious optimism about AI for CDI while also showing the need for independent replication, prospective evaluation, and multidimensional outcome assessment [5, 28].

Comparison with prior reviews

Prior reviews helped establish the foundations of clinical natural language processing, but most addressed either broad clinical text processing or a narrower technical area rather than clinical documentation improvement as an integrated field. Reviews of clinical information extraction, chronic disease NLP, and clinical text machine learning described major progress in extracting structured information from notes, but they did not consistently connect extraction performance to coding, billing, documentation quality, and physician workflow outcomes [6, 7, 17, 21]. Similarly, reviews and methodological papers on automated coding emphasised the growth of ICD prediction models but gave less attention to note bloat, ambient documentation, or revenue-cycle implementation [10]. This review therefore extends prior work by organising the evidence around CDI functions rather than around NLP task type alone.

Some prior evidence focused on speech recognition, automated scribes, and medical conversation summarisation as solutions to physician documentation burden. These studies were important because they moved beyond retrospective note classification and examined documentation as a workflow problem involving conversation capture, report generation, and clinician review [15, 16, 18, 19]. However, this literature often remained separate from automated coding research, even though generated notes may directly influence coding specificity, medical necessity, billing compliance, and audit defensibility [13, 14]. By bringing these areas together, the present review highlights that documentation generation and coding support should be evaluated as connected parts of the same CDI ecosystem.

The novel contribution of this review is the emphasis on implementation gaps and multidimensional outcome evaluation. Automated coding studies often reported technical metrics, while documentation burden studies reported time or workload proxies, and note quality studies reported length, redundancy, concordance, or error patterns [1, 3, 8, 27]. Fewer studies connected these outcomes in a way that would allow health systems to determine whether an AI-CDI tool improves documentation quality, physician efficiency, coding reliability, and billing performance simultaneously [4, 26]. This synthesis therefore suggests that future CDI evidence should move from isolated model validation toward integrated clinical, operational, and financial evaluation [5, 29].

Research gaps

Long-term clinical and financial outcomes

A major gap is the absence of long-term studies showing whether AI-CDI tools improve patient outcomes, documentation reliability, physician well-being, or sustained revenue-cycle performance. Automated coding systems have been extensively evaluated on historical notes, but few studies followed their use through coder review, claim submission, payer response, audit processes, and institutional financial outcomes [13, 23, 29]. Workflow and note-generation studies similarly require longer follow-up to determine whether initial time savings persist after novelty effects, training periods, and workflow adaptations [15, 19]. Future studies should therefore track clinical, financial, and operational outcomes together rather than treating them as separate endpoints [1, 4].

Physician-AI interaction and trust

The literature provides limited evidence on how physicians and coders interact with AI-CDI suggestions in routine practice. Explainable coding models can highlight relevant note evidence, but the field still lacks detailed evaluation of when users accept, ignore, modify, or distrust suggestions [8, 24, 26]. This matters because poorly timed or poorly justified prompts may increase cognitive load, while well-designed suggestions could support more complete and specific documentation [27, 28]. Future research should examine trust calibration, override behaviour, editing burden, and the effect of AI suggestions on clinical reasoning and documentation habits [5, 15].

Equity and bias in automated documentation

Equity and bias remain underdeveloped topics in AI-CDI research. Models trained on historical notes and codes may reproduce existing differences in documentation intensity, access to care, diagnostic labelling, or coding practices across patient populations [10, 11]. Automated note generation may also amplify biases if speech recognition, summarisation, or template selection performs differently across accents, languages, specialties, or clinical contexts [14, 16, 18]. Future evaluations should therefore stratify performance across demographic groups, clinical settings, and language contexts while also examining whether AI-CDI changes documentation quality or billing outcomes unevenly [21, 20].

Implications

For research practice

The main implication for research practice is the need to shift from isolated accuracy studies to implementation-oriented evaluation across the four CDI domains. Automated coding models, note generation systems, note quality tools, and workflow interventions should be studied together because each can affect the others in clinical practice [8, 10, 18]. For example, a generated note may reduce physician typing time but still create coding ambiguity, billing risk, or downstream review burden if it lacks specificity or contains summarisation errors [14, 19]. Implementation science frameworks would help researchers examine adoption, fidelity, user adaptation, and sustainability in addition to model performance [4, 15].

For clinical practice

For clinical practice, AI-CDI tools should be treated as assistive systems rather than autonomous replacements for physicians, coders, or CDI specialists. The included literature supports the technical feasibility of code prediction, speech-derived documentation, and clinical text summarisation, but it also shows persistent challenges in error detection, interpretability, integration, and user trust [13, 16, 26]. Human oversight remains critical because documentation is both a clinical communication tool and a legal and billing record [14, 27]. Health systems adopting these tools should therefore define review responsibilities, escalation processes, and monitoring procedures before broad deployment [5, 10].

Conclusion

Natural language processing for clinical documentation improvement progressed rapidly between 2017 and 2023, especially in automated clinical coding and note analysis. The reviewed literature shows that AI methods can extract, classify, summarise, and generate documentation-relevant information across a variety of clinical contexts.

Billing support and physician workflow efficiency remain less mature as evidence domains. Although many systems have plausible value for revenue-cycle support and documentation burden reduction, fewer studies have evaluated real-world billing outcomes, sustained physician time savings, or long-term implementation effects.

A coordinated research agenda is needed to standardise evaluation, improve external validation, measure implementation outcomes, and assess equity. Future studies should connect coding accuracy, note quality, billing integrity, and workflow impact rather than evaluating each domain in isolation.

The future of clinical documentation improvement lies not in replacing physicians, coders, or CDI specialists, but in augmenting their work with trustworthy, context-aware, auditable AI. The most valuable systems will be those that improve documentation while preserving clinical judgment, professional accountability, and patient-centred care.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Jul 2023

Revised: 27 Sep 2023

Accepted: 22 Oct 2023

Published online: 25 February 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med*. 2017;15(5):419-26.
<https://doi.org/10.1370/afm.2121>.
- Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Aff (Millwood)*. 2017;36(4):655-62.

<https://doi.org/10.1377/hlthaff.2016.0811>.

Rule A, Bedrick S, Chiang MF, Hribar MR. Length and redundancy of outpatient progress notes across a decade at an academic medical center. *JAMA Netw Open*. 2021;4(7):e2115334.

<https://doi.org/10.1001/jamanetworkopen.2021.15334>.

Moy AJ, Schwartz JM, Chen R, Sadri S, Lucas E, Cato KD, et al. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *J Am Med Inform Assoc*. 2021;28(5):998-1008.

Holmgren AJ, Downing NL, Bates DW, Shanafelt TD, Milstein A, Sharp CD, et al. Assessment of electronic health record use between US and non-US health systems. *JAMA Intern Med*. 2021;181(2):251-9.

<https://doi.org/10.1001/jamainternmed.2020.7421>.

Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform*. 2017;73:14-29.

<https://doi.org/10.1016/j.jbi.2017.07.012>.

Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;77:34-49.

<https://doi.org/10.1016/j.jbi.2017.11.011>.

Mullenbach J, Wiegrefe S, Duke J, Sun J, Eisenstein J. Explainable prediction of medical codes from clinical text. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018;1:1101-11.

Li F, Yu H. ICD coding from clinical text using multi-filter residual convolutional neural network. *Proc AAAI Conf Artif Intell*. 2020;34(05):8180-7.

<https://doi.org/10.1609/aaai.v34i05.6406>.

Dong H, Falis M, Whiteley W, Alex B, Matterson J, Ji S, et al. Automated clinical coding: what, why, and where we are? *NPJ Digit Med*. 2022;5(1):159.

<https://doi.org/10.1038/s41746-022-00691-7>.

Ji S, Hölttä M, Marttinen P. Does the magic of BERT apply to medical code assignment? A quantitative study. *Comput Biol Med*. 2021;139:104998.

<https://doi.org/10.1016/j.compbiomed.2021.104998>.

Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes: case study on ICD code assignment. In: *AAAI Workshops*; 2018. p. 409-16.

Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Med Inform*. 2021;9(8):e23230.

<https://doi.org/10.2196/23230>.

Zhou L, Blackley SV, Kowalski L, Doan R, Acker WW, Landman AB, et al. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. *JAMA Netw Open*. 2018;1(3):e180530.

<https://doi.org/10.1001/jamanetworkopen.2018.0530>.

Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digit Med*. 2019;2(1):114.

<https://doi.org/10.1038/s41746-019-0190-1>.

Finley G, Edwards E, Robinson A, Brenndoerfer M, Sadoughi N, Fone J, et al. An automated medical scribe for documenting clinical encounters. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*; 2018. p. 11-5.

Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform.* 2020;8(3):e17984.
<https://doi.org/10.2196/17984>.

Enarvi S, Amoia M, Teba MD, Delaney B, Diehl F, Hahn S, et al. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In: *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*; 2020. p. 22-30.

Krishna K, Khosla S, Bigham JP, Lipton ZC. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In: *Proc ACL-IJCNLP*; 2021. p. 4958-72.

Coutinho I, Martins B. Transformer-based models for ICD-10 coding of death certificates with Portuguese text. *J Biomed Inform.* 2022;136:104232.
<https://doi.org/10.1016/j.jbi.2022.104232>.

Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform.* 2019;7(2):e12239.
<https://doi.org/10.2196/12239>.

Vu T, Nguyen DQ, Nguyen A. A label attention model for ICD coding from clinical text. *arXiv [Preprint]*. 2020 Jul 13:arXiv:2007.06351.

Chen Y, Chen H, Lu X, Duan H, He S, An J. Automatic ICD-10 coding: deep semantic matching based on analogical reasoning. *Heliyon.* 2023;9(4):e15178.
<https://doi.org/10.1016/j.heliyon.2023.e15178>.

Sonabend A, Cai W, Ahuja Y, Ananthakrishnan A, Xia Z, Yu S, et al. Automated ICD coding via unsupervised knowledge integration (UNITE). *Int J Med Inform.* 2020;139:104135.
<https://doi.org/10.1016/j.ijmedinf.2020.104135>.

Masud JH, Kuo CC, Yeh CY, Yang HC, Lin MC. Applying deep learning model to predict diagnosis code of medical records. *Diagnostics (Basel).* 2023;13(13):2297.
<https://doi.org/10.3390/diagnostics13132297>.

Hu S, Teng F, Huang L, Yan J, Zhang H. An explainable CNN approach for medical codes prediction from clinical text. *BMC Med Inform Decis Mak.* 2021;21(Suppl 9):256.
<https://doi.org/10.1186/s12911-021-01642-3>.

Berdahl CT, Moran GJ, McBride O, Santini AM, Verzhbinsky IA, Schriger DL. Concordance between electronic clinical documentation and physicians' observed behavior. *JAMA Netw Open.* 2019;2(9):e1911390.
<https://doi.org/10.1001/jamanetworkopen.2019.11390>.

Cohen GR, Friedman CP, Ryan AM, Richardson CR, Adler-Milstein J. Variation in physicians' electronic health record documentation and potential patient harm from that variation. *J Gen Intern Med.* 2019;34(11):2355-67.
<https://doi.org/10.1007/s11606-019-05215-5>.

Zhao S, Diao X, Xia Y, Huo Y, Cui M, Wang Y, et al. Automated ICD coding for coronary heart diseases by a deep learning method. *Heliyon.* 2023;9(3):e13998.
<https://doi.org/10.1016/j.heliyon.2023.e13998>.