HTA Austria
Austrian Institute for
Health Technology Assessment
GmbH

Artificial Intelligence for Hospital Documentation Support

A Scoping Review of Current Use Cases

Final report

AIHTA Project Report No.: 171a | ISSN: 1993-0488 | ISSN-online: 1993-0496

HTA Austria
Austrian Institute for
Health Technology Assessment
GmbH

Artificial Intelligence for Hospital Documentation Support

A Scoping Review of Current Use Cases

Vienna, January 2026

Project Team

Project leader: Judit Erdös, MA

Authors: Judit Erdös, MA; Lena Grabenhofer, BA, MSc

Project Support

Systematic literature search: Tarquin Mittermayr, BA(Hons), MA

Internal review: Dr. PH, Gregor Goetz, MSSc MPH

External review: Miguel Ángel Armengol de la Hoz, PhD; Head of the Data Science Lab, Fundación Progreso y Salud, Ministry of Health and Consumer Affairs, Regional Government of Andalusia

Correspondence: Judit Erdös, Judit.erdos@aihta.at

Cover photo: @ LALAKA – stock.adobe.com

This report should be referenced as follows:

Erdos J, Grabenhofer L. Artificial Intelligence for Hospital Documentation Support. A Scoping Review of Current Use Cases. AIHTA Project Report No.: 171a; 2026. Vienna: HTA Austria – Austrian Institute for Health Technology Assessment GmbH.

Conflict of interest

All authors and the reviewers involved in the production of this report have declared they have no conflicts of interest in relation to the technology assessed according to the Uniform Requirements of Manuscripts Statement of Medical Journal Editors (www.icmje.org).

Disclaimer

The external reviewers did not co-author the scientific report and do not necessarily all agree with its content. Only the AIHTA is responsible for errors or omissions that could persist. The final version and the policy recommendations are under the full responsibility of the AIHTA.

During the preparation of this work, the authors used Claude.ai and ChatGPT.com to enhance the writing process. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

IMPRINT

Publisher:
HTA Austria – Austrian Institute for Health Technology Assessment GmbH
Josefstädter Straße 39 | 1080 Vienna – Austria
https://www.aihta.at/

Responsible for content:
Dr. rer. soc. oec. Ingrid Zechmeister-Koss, MA, managing director

AIHTA Project Reports do not appear on a regular basis and serve to publicise the research results of the Austrian Institute for Health Technology Assessment.

AIHTA Project Reports are only available to the public via the Internet at http://eprints.aihta.at/view/types/hta_report.html.

AIHTA Project Report No.: 171a

ISSN 1993-0488

ISSN online 1993-0496

Content

Executive Summary 9

Zusammenfassung 9

1 Introduction 16
1.1 Background and Rationale 16
1.2 Objectives and Scope 17

2 Methods 19
2.1 Literature search 19
2.1.1 Flow Diagram 19
2.2 Data Extraction and Analysis 21
2.3 Quality assessment 22
2.4 Piloting Guidance from AIHTA and ASSESS DHT 22

3 Results 23
3.1 Survey 23
3.2 Scoping Review 23
3.2.1 Outcomes 23
3.2.2 Characteristics of included reviews 24
3.2.3 Results by case vignette 25
3.2.4 Error detection, clinical note quality assessment 35
3.2.5 Billing codes 37
3.3 Piloting the AIHTA and ASSESS DHT guidance documents 38
3.4 Implementation and procurement considerations for hospital managers 39

4 Discussion 43
4.1 Main findings 43
4.2 Regulatory context and evidence expectations at EU level 44
4.3 Limitations 45
4.4 Implementation context in Austria 47

5 Conclusion 49

6 References 50

Appendix 53
Appendix A: Glossary of terms 53
Appendix B: Extraction tables 55
Appendix C: Procurement checklist for decision-makers 69
Appendix D: Search strategies 70

List of figures

Figure 2-1: Selection process (PRISMA Flow Diagram) ... 20
Figure 3-1: Workflow of AI scribe technology ... 25
Figure 3-2: Implementation enablers ... 39
Figure 3-3: Decision flow diagram for procurement decisions of AI-supported documentation tools ... 41

List of tables

Table 1-1: Inclusion and exclusion criteria for RQ2 ... 17
Table 3-1: Overview of documentation support uses cases addressed across included reviews ... 24
Table 3-2: Key characteristics, implementation costs, evaluation metrics of AI scribe ... 26
Table 3-3: Vignette 1 – AI scribe ... 28
Table 3-4: Key characteristics, implementation costs, evaluation metrics of structuring unstructured text use case ... 29
Table 3-5: Vignette 2 – Structuring unstructured text ... 30
Table 3-6: Key characteristics, implementation costs, evaluation metrics of AI-generated documentation ... 31
Table 3-7: Vignette 3 – AI-generated medical documentation without speech recognition ... 33
Table 3-8: Key characteristics, implementation costs, evaluation metrics of patient-friendly summaries ... 34
Table 3-9: Vignette 4 – Patient engagement and communication ... 35
Table 3-10: Key characteristics, implementation costs, evaluation metrics of error detection and note quality assessment ... 35
Table 3-11: Vignette 5 – Error detection in clinical notes, assessing note quality ... 36
Table 3-12: Key characteristics, implementation costs, evaluation metrics of assigning billing codes ... 37
Table 3-13: Vignette 6 – AI-generated billing codes ... 38
Table A-1: Characteristics of included reviews ... 55
Table A-2: Use case AI scribes (part 1) ... 60
Table A-2: Use case AI scribes (part 2) ... 62
Table A-3: Use case structuring free-text data ... 64
Table A-4: Use case AI-generated medical documentation ... 65
Table A-5: Use case AI-generated billing codes ... 67
Table A-6: Use case AI-generated patient-friendly summaries ... 67
Table A-7: Use case Error detection & note quality assessment ... 68

List of abbreviations

AI	Artificial Intelligence
AIHTA	Austrian Institute for Health Technology Assessment
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
BERTSUM	BERT Summarisation
CDA	Clinical Document Architecture
CE	Conformité Européenne
CNN	Convolutional Neural Network
CUR	Current Use
DHT	Digital Health Technology

DPIA	Data Protection Impact Assessment
ECO	Economic
ED	Emergency Department
EFF	Effectiveness
EHR	Electronic Health Record
EHDS	European Health Data Space
ELGA	Elektronische Gesundheitsakte
EMR	Electronic Medical Record
ETH	Ethical
EU	European Union
EUnetHTA	European Network for Health Technology Assessment
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
FLAN	Finetuned Language Net
GDPR	General Data Protection Regulation
GÖG	Gesundheit Österreich GmbH
GP	General Practitioner
GPT	Generative Pre-trained Transformer
HCP	Healthcare Professional
HIPAA	Health Insurance Portability and Accountability Act
HIV	Human Immunodeficiency Virus
HPI	History of Present Illness
HTA	Health Technology Assessment
ICD-10	International Classification of Diseases, 10th Revision
IQR	Interquartile Range
KIS	Krankenhaus Information System
LLM	Large Language Model
MDR	Medical Device Regulation
ML	Machine Learning
MMAT	Mixed Methods Appraisal Tool
NA	Not Applicable
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NLP	Natural Language Processing
NR	Not Reported
ORG	Organisational
PDQI-9	Physician Documentation Quality Instrument-9
PEGASUS	Pre-training with Extracted Gap-sentences for Abstractive Summarisation
PPV	Positive Predictive Value
PRISMA-ScR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RAG	Retrieval-Augmented Generation
RCT	Randomised Controlled Trial
REDCap	Research Electronic Data Capture
ROC	Receiver Operating Characteristic
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RQ	Research Question
RTR	Rundfunk und Telekom Regulierungs-GmbH
RVU	Relative Value Unit
SAF	Safety
ScR	Scoping Review
SOC	Social
SR	Systematic Review
TEC	Technical
TPR	True Positive Rate
U.S.	United States
UK	United Kingdom
USD	United States Dollar
VBC	Value-Based Care
WER	Word Error Rate
wRVU	ork Relative Value Unit

Executive Summary

Background

Clinical documentation has become one of the most time-consuming tasks for healthcare professionals. In hospitals in particular, electronic health record (EHR) systems generate substantial administrative burden, contributing to clinician stress, burnout, and reduced time for direct patient care. AI-enabled digital health technologies (DHTs) have emerged as promising tools to reduce documentation burden by supporting or automating workflow parts.

From a regulatory perspective, documentation support tools are typically classified as low-risk or as non-medical software, because they do not provide diagnostic, prognostic, or therapeutic clinical decision support. Even when regulatory requirements are limited, a structured evaluation of such tools supports procurement and implementation decisions in hospitals (e.g., usability, workflow fit, data protection, interoperability, and organisational impact). In Austria, national digital transformation initiatives – such as the AI Mission Austria 2030 and the eHealth Strategy 2024-2030 – further underline the need for a structured assessment to support hospital decision-making.

Because documentation support in hospitals comprises a broad range of functions and use cases rather than a single type of application, it is unclear whether evaluation criteria should be applied consistently across use cases or tailored to function and use-case-specific requirements. A mapping of documentation support functions to relevant evaluation criteria is therefore a prerequisite for a structured assessment and for determining whether criteria is transferable across use cases.

This report presents a scoping review that maps key AI-enabled documentation support functions, describes the evidence base for their performance and impact in hospitals, and pilots the applicability of existing guidance (AIHTA procurement checklist, ASSESS-DHT taxonomy and guidance) for their evaluation.

Research Questions

The project addresses two research questions (RQ):

- **RQ1**: Which AI-enabled documentation support functions are currently used or considered relevant in Austrian hospitals?
- **RQ2**: What is the current landscape of AI-enabled DHTs for documentation support in hospitals, and what evidence exists regarding their functions, implementation needs, performance, and outcomes?

Methods

To answer **RQ1**, a recent report by Gesundheit Österreich GmbH on pilot and routine AI applications in the Austrian healthcare system was used to identify documentation support applications. In addition, a short expert survey among Austrian healthcare experts was conducted.

To answer **RQ2**, a systematic literature search was conducted in four databases. Eligible sources included systematic reviews (SRs), scoping reviews (ScRs), and HTA/policy reports containing at least two primary studies evaluating documentation-related AI functions in hospital settings. Reports on non-AI tools and developmental studies without outcomes limited to describing the development of a specific AI-enabled DHT were excluded.

Overall, 64 full texts were assessed for eligibility, and seven reviews (three SRs, four ScRs) comprising 200 primary studies were included. Data were extracted on function, technology, setting, outcomes, implementation aspects, reported benefits and challenges.

Reported functions were grouped into six **"use cases":**

1. AI scribes: speech-based drafting of clinical notes.
2. Structuring unstructured text: extracting discrete data elements from narrative notes and embedding them into structured EHRs.
3. AI-generated documentation without speech recognition: large language model (LLM)-based auto-drafting tools generating clinical documents directly from digital inputs.
4. Patient-friendly summaries: converting clinical notes into lay-language summaries.
5. Error detection and quality assessment of clinical notes: flagging missing documentation domains, contradictions, redundant content, or unclear phrases.
6. Automated billing code assignment: standardised billing or classification codes are assigned directly from clinical documentation.

Results

For **RQ1**, the mapping of AI-enabled documentation support functions currently used or considered relevant in Austrian hospitals remained exploratory due to limited survey participation; Dragon Medical emerged as the only identified tool.

To address **RQ2**, the results below summarise the landscape of documentation-support DHTs and the evidence available across key functions and outcomes. AI scribes were most frequently studied. Their performance (accuracy and completeness) across studies varied; several studies reported omissions, meaning that the generated notes sometimes left out clinically relevant details. Clinician satisfaction with AI scribes was generally high, and documentation burden was perceived as lower, but time savings and productivity gains were inconsistent. Studies on structuring unstructured text showed improved accuracy but limited completeness for rare concepts and complex context; evidence on user- or organisational outcomes was scarce, with a few studies reporting reduced documentation time. Studies on AI-generated documentation reported improved completeness, however, omissions and hallucinations remained common, necessitating human oversight. Studies on automated billing code assignment as well as error detection and quality assessment of clinical notes mainly reported technical metrics, generally demonstrating high accuracy. AI-generated patient-friendly summaries showed improved patient comprehension and satisfaction. Across use cases, studies repeatedly identified information omissions, hallucinations, and oversight requirements as key risks, together with data protection concerns (especially for audio-recording systems), uncertainties regarding legal responsibility, and limited evidence on downstream clinical or organisational outcomes.

Discussion

Although the evidence base for AI documentation support is expanding, it remains heterogeneous in terms of the analysed datasets, evaluation metrics and outcome measures, and uneven across use cases, with most studies focusing on AI scribes. AI scribes and LLM-based auto-drafting tools without speech recognition show the most promising clinician-reported benefits, but quantitative improvements vary widely. Hallucinations, omissions, variable accuracy, and medico-legal risks underscore the need for human oversight and local validation.

Organisational impacts (productivity, workflow changes, cost savings) are documented mainly for AI scribes, while for other use cases, evidence is sparse. Technical evaluations dominate the literature, with few studies assessing implementation prerequisites such as integration requirements, training needs, or long-term performance.

These uncertainties intersect with an evolving regulatory landscape (European Medical Device Regulation, European Health Data Space, and AI Act). These frameworks introduce expectations around transparency, risk management, monitoring, and incident reporting, and support a shift toward structured, interoperable, provenance-tracked documentation – even for tools that are not classified as medical devices. Although many documentation support tools are not considered medical devices, their output may still affect the clinical record and therefore indirectly influence patient care.

In practice, deployment is a socio-technical process, rather than a technical upgrade. Implementation barriers include EHR integration, workflow fit, data quality limitations, governance gaps, privacy concerns, and limited institutional technical capacity. Safe implementation requires training and governance, integration with local hospital information systems, local validation and bias monitoring, and sustained human oversight supported by audit trails, traceable validation cycles and explainability checks. In Austria, the ELGA architecture and the upcoming Fast Healthcare Interoperability Resources-based expansions will further increase interoperability and provenance requirements, reinforcing the need for data protection impact assessments, and ongoing oversight of AI outputs.

Conclusion

AI-enabled documentation support tools offer potential to reduce administrative burden and improve documentation quality, with positive signals for clinician experience and workflow efficiency. However, evidence remains limited, inconsistent, and highly context dependent. A proportionate, risk-based approach – balancing potential benefit with safety – is essential. Hospitals should adopt structured validation, human oversight, fairness and bias monitoring, and governance mechanisms before deploying these tools at scale. Continued evaluation, methodological development, and stakeholder engagement will be necessary as technologies – and regulations – evolve.

Zusammenfassung

Hintergrund

Die klinische Dokumentation ist zu einer der zeitaufwändigsten Aufgaben für Ärzt:innen und andere Angehörige der Gesundheitsberufe geworden. Insbesondere in Krankenhäusern erzeugen elektronische Gesundheitsaktensysteme (EHR) erhebliche administrative Belastungen, die zu Stress, Burnout und weniger Zeit für die direkte Patient:innenversorgung beitragen. KI-gestützte digitale Gesundheitstechnologien (DHTs) bieten sich als vielversprechende Instrumente an, um diese Belastung zu verringern, indem sie Teile der Dokumentationsabläufe unterstützen oder automatisieren.

Aus regulatorischer Sicht werden Dokumentationsunterstützungstools in der Regel als risikoarm bzw. als nicht-medizinische Software eingestuft, da sie keine diagnostischen, prognostischen oder therapeutischen Empfehlungen liefern. Auch bei begrenzten regulatorischen Anforderungen unterstützt eine strukturierte Bewertung solcher Tools Beschaffungs- und Implementierungsentscheidungen in Krankenhäusern (z. B. hinsichtlich Benutzerfreundlichkeit, möglicher Verzerrungen, Datenschutz sowie Steuerung). In Österreich unterstreichen nationale Initiativen zur digitalen Transformation – wie die KI-Mission Austria 2030 und die eHealth-Strategie 2024-2030 – zusätzlich die Notwendigkeit einer strukturierten Bewertung dieser Technologien.

Da die Dokumentationsunterstützung im Krankenhaus ein breites Spektrum an Funktionen und Anwendungsfällen umfasst und nicht nur einen einzelnen Anwendungsfall darstellt, ist es unklar, ob Evaluierungskriterien einheitlich über alle Anwendungsfälle hinweg oder spezifisch nach Funktion und Anwendungsfallanforderungen angepasst werden sollten. Eine Zuordnung von Dokumentationsunterstützungsfunktionen zu relevanten Evaluierungskriterien ist daher eine Voraussetzung für eine strukturierte Bewertung und für die Beurteilung der Übertragbarkeit der Kriterien über verschiedene Anwendungsfälle hinweg.

Der Bericht führt einen Scoping-Review (ScR) zu KI-gestützten Dokumentationsunterstützungsfunktionen im Gesundheitswesen durch. Zentrale Analyseschwerpunkte bilden die Evidenzgrundlage zu Leistung und Auswirkungen der identifizierten KI-Dokumentationsunterstützungsfunktionen, insbesondere hinsichtlich ihrer klinischen und organisatorischen Outcomes. Ein weiteres Ziel war die Pilotisierung bestehender Evaluierungsinstrumente wie der AIHTA-Beschaffungscheckliste und der ASSESS-DHT-Taxonomie.

Forschungsfragen

Das Projekt adressiert zwei zentrale Forschungsfragen (FF):

- **FF1:** Welche KI-gestützten Dokumentationsunterstützungsfunktionen werden derzeit in österreichischen Krankenhäusern verwendet oder als relevant erachtet?
- **FF2**: Wie ist die aktuelle Landschaft der KI-gestützten DHTs zur Dokumentationsunterstützung in Krankenhäusern, und welche Evidenz gibt es hinsichtlich ihrer Funktionen, Implementierungsbedarfe, Leistung und Outcomes?

Methoden

FF1: Zur Identifikation von KI-Dokumentationsunterstützungsfunktionen in österreichischen Krankenhäusern wurde primär ein aktueller Bericht der Gesundheit Österreich GmbH über Pilot- und Routineanwendungen von KI im Gesundheitswesen herangezogen. Ergänzend wurde eine Expert:innenbefragung unter österreichischen Kliniker:innen und IT-Verantwortlichen durchgeführt.

FF2: Eine systematische Literaturrecherche wurde in vier Datenbanken durchgeführt. Inkludiert wurden systematische Reviews (SRs), Scoping Reviews (ScRs) und HTA-/Policy-Berichte, die mindestens zwei Primärstudien enthielten, welche dokumentationsbezogene KI-Funktionen im Krankenhauskontext evaluierten. Reviews zu technischen Systemen ohne künstliche Intelligenz, zur Primärversorgung, und Entwicklungsstudien ohne praktische Ergebnisse wurden ausgeschlossen.

Insgesamt wurden 755 Datensätze gescreent, 64 Volltexte überprüft und sieben Reviews (drei SRs, vier ScRs) mit 200 Primärstudien eingeschlossen. Es wurden Daten zu Art der KI-Funktion, Technologie, Settings, Ergebnissen, Implementierungsaspekten und berichteten Vorteilen oder Herausforderungen extrahiert.

Um die heterogene Evidenz zu strukturieren und den Vergleich zwischen Reviews zu erleichtern, wurden die berichteten Funktionen in sechs „Use Cases“ (dt. Anwendungsfälle) gruppiert:

1. KI-basierte Medical Scribes
2. Strukturierung unstrukturierter Texte
3. KI-generierte Dokumentation ohne Spracherkennung
4. Patient:innenfreundliche Zusammenfassungen
5. Fehlererkennung und Bewertung der Notizqualität
6. Automatisierte Zuweisung von Abrechnungscodes

Ergebnisse

Die Analyse umfasst sieben Reviews mit unterschiedlicher Evidenzbasis. Pro Anwendungsfall wurden zwischen einem und sieben Reviews identifiziert, wobei die meisten zu KI Scribes gehörten. Die Reviews umfassen 200 Primärstudien welche vorwiegend aus US-amerikanischen Krankenhaus- und Ambulanzkontexten stammen.

KI-basierte Medical Scribes

Die Ergebnisse für KI Scribes basieren auf allen sieben Reviews (vier SRs, drei ScRs) mit insgesamt 36 Primärstudien. Scribes (dt. Schreiber) dokumentieren Gespräche zwischen Ärzt:innen und Patient:innen: menschliche Scribes erledigen dies live, während AI-Scribes die Gespräche automatisch aufzeichnen und in klinische Notizen umwandeln. Die berichtete Genauigkeit variierte stark, je nach Aufgabe und Datensatz. Häufig ausgelassene Informationen betrafen insbesondere die Patient:innengeschichte, körperliche Untersuchungsergebnisse, relevante Nebendiagnosen und sozialmedizinische Angaben. Die Analyse identifizierte verschiedene potenzielle Vorteile, darunter verbesserte Lesbarkeit klinischer Notizen, erhöhte Vollständigkeit der Dokumentation und eine wahrgenommene Reduktion der Dokumentationslast für Kliniker:innen. Die Evidenz zu Zeitersparnissen ist jedoch inkonsistent: Einzelne Studien berichten von kürzeren Konsultationsdauern und moderaten Produktivitätssteigerungen, während andere keinen messbaren Vorteil identifizieren konnten.

Kritisch zu bewerten sind potenzielle Risiken, die in den Reviews hervorgehoben werden. Dazu gehören Unsicherheiten bezüglich der Genauigkeit der Notizen, die Notwendigkeit einer kontinuierlichen Überwachung und Qualitätskontrolle, Datenschutzrisiken durch Audioaufnahmen sowie ungeklärte medizinrechtliche Verantwortungsfragen. Diese Aspekte erfordern eine sorgfältige Abwägung bei der Implementierung von KI Scribe-Systemen.

Strukturierung unstrukturierter Texte

Dieses Anwendungsgebiet wird durch zwei Reviews (ein SR und ein ScR) mit über 90 Primärstudien abgedeckt. Diese KI-Tools extrahieren einzelne Datenelemente aus narrativen Notizen und fügen sie in strukturierte EHR-Felder ein, um die Dokumentation besser zu organisieren. Erwartet wird, dass dadurch Daten besser zugänglich und analysierbar werden und die Datenqualität für Forschung und klinische Entscheidungen steigt. Über alle Aufgaben hinweg erzielten die Modelle eine hohe technische Performance, die Vollständigkeit war jedoch bei seltenen Konzepten oder komplexen Kontextinformationen eingeschränkt. Evidenz zu klinischen oder organisatorischen Ergebnissen war begrenzt vorhanden, wobei mehrere Studien eine Reduktion der Dokumentationszeit von bis zu 56% berichteten, allerdings teilweise mit einer leichten Qualitätsminderung.

KI-generierte Dokumentation ohne Spracherkennung

Die Ergebnisse für KI-generierte Dokumentation basieren auf vier Reviews (ein SR, drei ScRs) mit insgesamt 24 Primärstudien. Bei dieser Anwendung erstellen Large Language Models (LLMs) und traditionelles maschinelles Lernen (ML) klinische Dokumente, wie Entlassungsbriefe, direkt aus digitalen Eingaben. Erwartet wird, dass die Dokumentation gleichwertig oder besser als manuell verfasste Notizen ist, die Prägnanz und Vollständigkeit gesteigert wird und übersehene Informationen erfasst werden.

Bei der Evaluierung mit validierten Instrumenten (z. B. PDQI-9) erzielten die generierten Notizen oft gleich hohe oder höhere Bewertungen als von Menschen verfasste Texte hinsichtlich Prägnanz und Vollständigkeit. Einige Modelle erfassten sogar Informationen, die Kliniker:innen übersehen hatten. Die Reviews zeigten jedoch, dass die Modelle teilweise Halluzinationen generierten – also Informationen, die nicht in den Originaldaten vorhanden waren und somit potenziell irreführend sein können. Gleichzeitig dokumentierten die SRs, dass trotz der generell hohen Bewertungen Informationsauslassungen und nicht verifizierbare Inhalte auftraten, die eine kontinuierliche klinische Überprüfung der KI-generierten Dokumente erforderlich machen. Zusätzlich wurden Herausforderungen wie Datenschutzbedenken, Workflow-Integration und rechtliche Implikationen identifiziert.

Patient:innenfreundliche Zusammenfassungen

Die Ergebnisse zu patient:innenfreundlichen Zusammenfassungen basieren auf vier Reviews (ein SR und drei ScRs) mit acht Primärstudien. Diese KI-Systeme übersetzen ärztliche Notizen in eine einfache Sprache. Die Evidenz deutet auf konsistent verbesserte Lesbarkeit, besseres Verständnis und höhere Patient:innenzufriedenheit. Patient:innen bewerteten die Zusammenfassungen als hilfreich und akzeptabel. Organisatorische Outcomes wurden nicht berichtet.

Fehlererkennung und Bewertung der Dokumentationsqualität

Dieses Anwendungsgebiet basiert auf einem SR, das 20 Primärstudien einschließt. Dabei werden KI-Tools eingesetzt, um fehlende Dokumentationsbereiche, Widersprüche, redundante Inhalte oder unklare Formulierungen zu erkennen. Erwartet wird, dass die technische Leistungsqualität und die Vollständigkeit der Dokumentation erhöht wird. Die technische Leistungsmetriken waren generell hoch. Die Evidenz konzentrierte sich auf technische Performance; Auswirkungen auf den Workflow, Arbeitsbelastung der Kliniker:innen oder Patient:innenergebnisse wurden nicht berichtet.

Automatisierte Zuweisung von Abrechnungscodes

Die Ergebnisse zu automatisierter Zuweisung von Abrechnungscodes basieren auf einem ScRS mit insgesamt zwei Primärstudien. KI-Tools analysieren Behandlungsdokumente und weisen diesen automatisch die passenden medizinischen Codes für die Abrechnung zu. Erwartet wird, dass Codierungsgenauigkeit, Vollständigkeit und Fehlerreduktion verbessert werden. Die in diesem ScR enthaltenen Primärstudien zeigten Potenzial für verbesserte Codierungsgenauigkeit, höhere Vollständigkeit und Fehlerreduktion im Vergleich zu manuellen Methoden. Die Genauigkeit reichte abhängig vom jeweiligen Modell von moderat bis hoch. Die Evidenz beschränkte sich auf technische Benchmarks ohne Bewertungen der Implementierung in der Praxis.

Diskussion

Die Evidenzlage zur KI-gestützten Dokumentationsunterstützung wächst, zeigt jedoch eine heterogene Verteilung über verschiedenen "Use Cases". KI-Scribes und LLM-basierte Auto-Drafting-Tools zeigen die vielversprechendsten von Kliniker:innen berichteten Vorteile, aber die quantitativen Verbesserungen variieren stark. Auslassungen, schwankende Genauigkeit, Halluzinationen und medizinrechtliche Risiken unterstreichen die Notwendigkeit menschlicher Überwachung und lokaler Validierung.

Die Evidenz zu organisatorischen Voraussetzungen – verstanden als Bereitstellung notwendiger Strukturen, Ressourcen, Verantwortlichkeiten und Abläufe zur sicheren, effektiven und nachhaltigen Implementierung einer KI-Lösung – sowie zu Integrationsanforderungen, Schulungsbedarf und langfristiger

Leistungsfähigkeit ist begrenzt. Organisatorische Auswirkungen wie Produktivitätssteigerungen, Workflow-Optimierungen oder Kosteneinsparungen sind hauptsächlich für KI-Scribes dokumentiert, während für andere Use Cases nur wenige Erkenntnisse vorliegen. In der Literatur dominieren technische Bewertungen.

Die Entwicklung von KI-gestützten Dokumentationsunterstützungssystemen vollzieht sich in einem komplexen regulatorischen Umfeld. Europäische Regulierungsrahmen wie die EU-Medizinprodukteverordnung (MDR), der Europäische Gesundheitsdatenraum (EHDS) und der KI-Regulierungsrahmen (AI Act) schaffen zunehmend einen Regulierungsrahmen und normativen Rahmen für diese Technologien. Im Anwendungsbereich der KI Scribes zeichnet sich bereits eine verstärkte Klassifizierung als Medizinprodukte ab, begründet durch die Möglichkeit, klinische Entscheidungsprozesse potenziell zu beeinflussen. Während KI Scribes eine erhöhte regulatorische Aufmerksamkeit erfahren, verbleiben alternative Anwendungsgebiete wie Textstrukturierung, patient:innenfreundliche Zusammenfassungen oder automatisierte Abrechnungscodes in einem weniger stringenten Regulierungskontext. Der KI-Regulierungsrahmen implementiert gleichwohl einheitliche Mindestanforderungen für KI-Systeme, die Transparenz, Risikomanagement, Monitoring und Vorfallmeldungen umfassen. Parallel zielt der Europäische Gesundheitsdatenraum auf eine systematische Transformation der Gesundheitsdokumentation in Richtung Interoperabilität und Nachverfolgbarkeit ab.

Die Einführung von Dokumentationsunterstützungssystemen erweist sich als komplexer sozio-technischer Prozess und nicht als reine technische Verbesserung. Wesentliche Herausforderungen umfassen die Integration in elektronische Patient:innenakten, Unstimmigkeiten in Arbeitsabläufen, Datenqualitätsprobleme, unzureichende Governance- und Steuerungsprozesse, Datenschutzbedenken (insbesondere bei Audioaufnahmen) und begrenzte technische Kapazitäten der Einrichtungen. Eine sichere Implementierung erfordert organisatorische Voraussetzungen – einschließlich Schulungen, klarer Governance-Strukturen (Rollen, Verantwortlichkeiten, Entscheidungs- und Eskalationswege) und technischer Integration in lokale Krankenhausinformationssysteme – sowie lokale Validierung, Änderungsmanagement, Überwachung von Verzerrungen und kontinuierliche menschliche Aufsicht, unterstützt durch nachvollziehbare und regelmäßige Validierungszyklen und, soweit möglich, Plausibilitätsprüfungen.

In Österreich werden diese Entwicklungen durch die ELGA-Architektur und bevorstehende Erweiterungen auf Basis von Fast Healthcare Interoperability Resources geprägt. Diese Vorgaben stellen hohe Anforderungen an die KI-generierte Dokumentation, insbesondere hinsichtlich der Interoperabilität und des Datenursprungs. Krankenhäuser müssen umfassende Risikobewertungen zum Datenschutz durchführen, Kontrollspuren implementieren und eine kontinuierliche fortlaufende Überwachung der KI-Ergebnisse sicherstellen.

Schlussfolgerung

KI-gestützte Dokumentationsunterstützungstools versprechen Effizienzgewinne im Gesundheitswesen, die Evidenzlage bleibt jedoch limitiert. Die bisherigen Studienergebnisse deuten auf mögliche Vorteile wie Reduktion administrativer Belastungen und Verbesserung der Dokumentationsqualität hin, sind aber inkonsistent und kontextabhängig.

Die Implementierung erfordert einen risikoadaptierten Ansatz mit zentralen Elementen: lokale Validierung, kontinuierliche menschliche Überwachung, Bias- und Datenschutzkontrollen sowie Interoperabilität mit Krankenhausinformationssystemen. Programmweite Einführungen müssen klare Ziele, ausreichende Ressourcen und Evaluationspläne mit definierten Metriken zu Prozessen, Qualität und Nutzererfahrung berücksichtigen.

Kontinuierliche Evaluierung und methodische Weiterentwicklung sind entscheidend, um die Potenziale vor dem Hintergrund sich wandelnder technologischer und regulatorischer Rahmenbedingungen zu bewerten.

1 Introduction

1.1 Background and Rationale

KI zur Reduktion der Dokumentationslast und Verbesserung von Effizienz und Versorgungsqualität

The increasing documentation burden for physicians, especially in hospital settings, has been identified as a major contributor to clinician burnout, inefficiency, and reduced time available for patient care. A substantial proportion of working hours is spent on electronic health records (EHRs), diverting attention from direct patient interaction. Artificial intelligence (AI) is becoming increasingly prevalent in healthcare, with applications across many specialties. One area where AI holds particular promise is in supporting clinicians with their documentation tasks. By automating or assisting with note-taking, discharge summaries, coding, or structuring unstructured text, **AI-enabled digital health technologies (DHTs)**- digital tools and systems used to in healthcare (see *Glossary* for the formal definition)- may have the potential to reduce administrative burden, improve workflow efficiency, and enhance documentation quality [1-6].

KI zur Unterstützung der Dokumentation im Krankenhaus

For the purpose of this report, *documentation support in hospitals,* is defined as the use of AI-enabled DHTs to reduce, assist, or enhance the administrative tasks of clinicians that are directly related to clinical documentation in hospital settings.

unklare Regulierung KI-basierter Dokumentationssysteme in der EU

Abgrenzung zwischen reiner Dokumentationshilfe und patient:innenenrelevanter KI schwierig

From a regulatory perspective, the classification of such tools remains somewhat ambiguous. Under the EU Artificial Intelligence Act[1] [7], AI-based systems used primarily for administration in healthcare are classified as minimal-to-no risk, meaning no specific regulatory requirements apply to system deployers. Similarly, under the EU Medical Device Regulation (MDR)[2] [8], software used purely for documentation support would typically not be classified as a medical device, as it does not directly influence medical decisions or patient-relevant outcomes. However, distinguishing whether such technologies are purely documentation support or also impacting patient or clinician care remains difficult, with exceptions such as AI-based scribing tools that may be classified as medical devices in some contexts or jurisdictions, reflecting ongoing regulatory considerations and evolving practice [6].

trotz geringem Risiko: KI beeinflusst Ressourcen, Personal und Versorgungsorganisation

Despite being often regarded as low-risk technologies, these systems raise critical questions, as their adoption is expected to affect multiple dimensions of care – resource allocation, staffing, patient outcomes, and the organisation of health services. In Austria, this relevance is reinforced by national strategies such as the Artificial Intelligence Mission Austria 2030 [9] and the eHealth Strategy 2024-2030 [10], which explicitly promote the integration of digital and AI solutions in healthcare to improve efficiency, quality, and innovation in hospital care.

[1] Regulation (EU) 2024/1689 is a cornerstone of the EU's regulatory framework for governing AI systems, addressing risks associated with their design, deployment, and use. The AI Act is conceived as safety legislation that complements existing sectoral measures (Medical Device Regulation/In Vitro Diagnostic Medical Devices Regulation), by specifically targeting hazards posed by AI systems. With its risk-based approach, the AI Act provides a foundation for ensuring the safety, transparency, and trustworthiness of AI technologies, particularly in critical sectors like healthcare.

[2] Regulation (EU) 2017/745 is the framework for the regulatory review and approval of medical devices for sale in all EU Member States.

1.2 Objectives and Scope

Überblick über KI-gestützte Unterstützung der klinischen Dokumentation – Funktionen, Nutzer, Ressourcenbedarf und Auswirkungen

The aim of this project is to provide an overview of AI-enabled DHTs in the field of documentation support with a focus on their core functions, target users, implementation requirements (including the types of resources required), anticipated clinical and organisational impact, and reported outcomes. To address this aim, this scoping review focuses on two research questions (RQ):

Forschungsfragen

RQ1: Which AI-enabled DHT functions in documentation support are considered most relevant in Austrian hospitals by Austrian healthcare experts?

RQ2: What is the current landscape of AI-enabled digital health technologies (DHTs) used for clinical documentation support in hospitals, and what evidence exists regarding their functions, implementation requirements, and reported outcomes?

In particular, RQ2 explores how AI documentation support may affect the time clinicians spend on administrative tasks, what potential benefits and challenges are perceived in terms of usability, accuracy, and satisfaction, and what types of resources are required for acquisition, setup, and ongoing integration with existing systems.

Table 1-1: Inclusion and exclusion criteria for RQ2

	Inclusion	Exclusion
Population	All healthcare providers engaged in clinical documentation.	-
Intervention	AI-enabled DHTs designed to support clinical documentation	Non-AI documentation support tools, AI-enabled DHTs not strictly used for documentation support, such as: ■ answering patient questions (medical chatbot), ■ processing of patient data/information extraction from EHR for research purposes, ■ creation of patient education materials, ■ risk prediction, predictive modelling, ■ diagnostic and clinical decision-making support (generating differential diagnosis, drug and treatment recommendations), ■ non-clinical use cases.
Comparator	No comparator, or Usual administrative practices without additional AI support	Sole focus on comparing two or more AI applications (e.g. GPT 3.5 vs GPT 4)
Outcomes	■ **Technical performance and documentation quality:** accuracy, completeness, relevance, reduction in manual revisions. ■ **Clinician-reported outcomes:** stress, burnout, administrative time reduction (during and after encounters). ■ **Patient-reported outcomes:** quality of care, satisfaction, comprehension, safety. ■ **Organisational outcomes:** workflow impacts, task redistribution, training requirements, implementation challenges, resource use, and cost implications, business efficiency (wait times, throughput).	Outcomes limited to describing the *development* of a specific AI-enabled DHT (e.g., algorithm training/performance testing without implementation or impact).

	Inclusion	Exclusion
Settings/ Context	All clinical settings in hospitals (secondary care and above), across specialties.	Sole focus on primary care. *(Reviews including both hospital and primary care settings were retained if at least part of the evidence concerned hospitals.)*
Study types	HTAs and policy documents, narrative reviews and systematic reviews including at least two primary studies analysing at least one outcome from the defined outcome categories.	Protocols, ongoing studies, qualitative-only studies, commentaries, and studies of non-AI tools.
Language	English language	Other languages

Abbreviations: HTAs … health technology assessments; GPT … Generative Pre-trained Transformer

The objectives, inclusion criteria and methods for this scoping review were specified in advance and documented in a protocol [11, 12]. Any deviations are documented in the discussion section.

Scoping-Review

Additionally, the Austrian Institute for Health Technology Assessment (AIHTA) procurement checklist [13] and ASSESS DHT guidance documents [14-16] are piloted to examine their applicability for the assessment of AI-supported documentation tools and to identify potential adaptations needed to improve their usability and relevance.

Anwendbarkeit der AIHTA-Beschaffungscheckliste und des ASSESS-Bewertungsleitfadens

2 Methods

2.1 Literature search

To address RQ1, we mapped the documentation support tools currently used or tested as pilot projects in Austria using the report from Gesundheit Österreich GmbH (GÖG) [17] as a starting point. To verify the information, we invited selected healthcare experts (providers, healthcare professionals, chief IT officers in selected Austrian hospitals) to complete an online survey with a free-text field for listing documentation support tools not captured in the GÖG report. Given the limited participation, validation was not possible, even with additional expert outreach.

**RQ1:
Übersicht über KI-Dokumentationstools in Österreich;
Expert:innenbefragung zur Ergänzung,
Validierung begrenzt**

To answer RQ2, a systematic literature search was conducted in four databases on 1st of July 2025. The full search strategies are provided in Appendix D. In addition, a targeted hand search of reference lists and relevant websites was performed using the search terms: "artificial intelligence", "AI", "documentation", "medical documentation", "clinical documentation", "clinical notes", "ChatGPT", "digital scribe", "machine learning", "natural language processing", "ambient scribe", and "automatic speech recognition".

**RQ2:
Systematische Literatur- und Handrecherche zu KI-Dokumentationsunterstützung**

2.1.1 Flow Diagram

The search resulted in 755 records after deduplication. The titles and abstracts were screened independently by two researchers. 64 publications were retrieved for full-text inspection, also by two independent researchers.

**Literaturauswahl:
755 Treffer,
64 Volltexte gescreent**

The selection process is depicted in Figure 2-1:

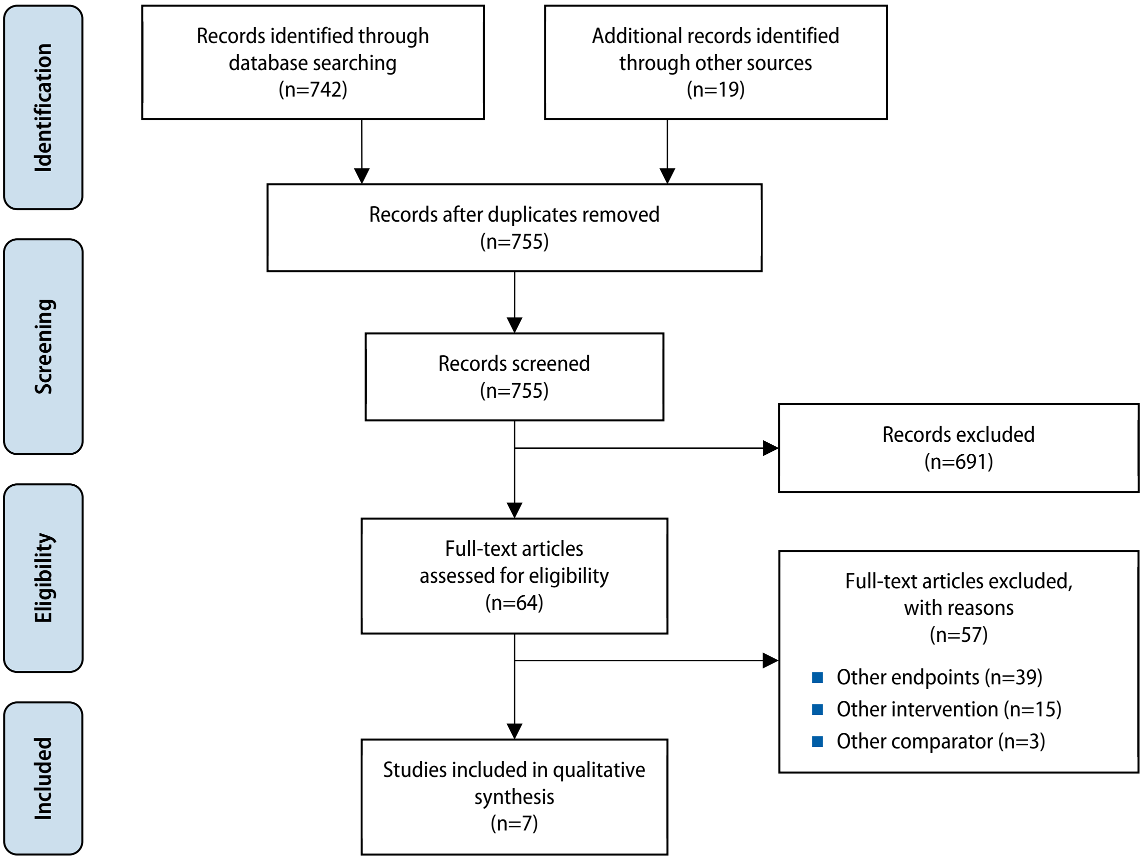
Identification
Screening
Eligibility
Included
Records identified through database searching (n=742)
Additional records identified through other sources (n=19)
Records after duplicates removed (n=755)
Records screened (n=755)
Records excluded (n=691)
Full-text articles assessed for eligibility (n=64)
Full-text articles excluded, with reasons (n=57)
■ Other endpoints (n=39)
■ Other intervention (n=15)
■ Other comparator (n=3)
Studies included in qualitative synthesis (n=7)

Figure 2-1: Selection process (PRISMA Flow Diagram)

2.2 Data Extraction and Analysis

Data were extracted into standardised tables, including author, year of publication, study design, number of included primary studies, outcomes analysed, type of AI technology, function of the AI, clinical setting, medical speciality, overall conclusions of the review, and any quality assessment of primary studies reported by review authors. Data extraction was performed by one assessor (JE) and checked for accuracy and completeness by a second assessor (LG). All evidence was narratively synthesised.

Datenerhebung: standardisierte Tabellen, Doppelkontrolle, narrative Synthese

Because *documentation support in hospitals* encompasses a broad range of AI applications, each supporting different processes and tasks [1], we applied a two-step approach to map and summarise the main use cases.

2-Stufen-Ansatz zur Abbildung und Zusammenfassung

Identification of **"use cases"**: Functions (clinical applications) of AI technologies, as described by review authors, were extracted and labelled as use cases.

Thematic analysis into **"case vignettes"**: Similar use cases were standardised, categorised, and clustered into broader case vignettes. This process yielded six distinct vignette categories representing AI-enabled documentation support:

Use-Cases identifizieren, thematisch zu Vignetten zusammenfassen

6 Vignetten:

a. **AI scribes** (ambient or dictation-based systems that transcribe spoken encounters into clinical notes. Earlier versions relied on speech recognition or dictation, whereas newer systems integrate generative AI and large language models to produce structured notes.)

KI-basierte Medical Scribes,

b. **Structuring unstructured text** (extracting data from free text into structured formats or coded fields)

Textstrukturierung,

c. **AI-generated documentation without speech recognition** (clinical documents generated directly from existing digital inputs, e.g. discharge summaries, operation notes, referral letters, without relying on real-time transcription of speech)

KI-generierte Dokumentation ohne Spracherkennung,

d. **Patient engagement through patient-friendly summaries** (conversion of medical text into plain-language summaries)

patient:innenfreundliche Zusammenfassungen,

e. **Error detection and note quality assessment** (AI tools that automatically identify errors, omissions or inconsistencies)

Fehlererkennung und Bewertung der Notizqualität,

f. **Automated billing codes** (systems that assign standardised billing or classification codes (e.g., ICD-10 codes) directly from clinical documentation, reducing manual coding work).

automatisierte Zuweisung von Abrechnungscodes

To provide additional structure for the mapping, we referred to the EUnetHTA Core Model® (v3.0) [18], particularly the domains *Description and technical characteristics (TEC)*[3], *Safety (SAF)*[4], and *Organisational aspects (ORG)*[5], the ASSESS-DHT taxonomy [15] and the Glossary of Terms for AI Validation in Healthcare [19], which extend Core Model considerations to AI-enabled DHTs.

EUnetHTA Core Model, ASSESS-DHT, Glossar KI-Validierung

[3] B0001, B0002, B0003, B0007

[4] C0008

[5] G0001, G0100, G0003, G0012, G0006, G0007, G0008, G0010

2.3 Quality assessment

This scoping review followed PRISMA-ScR guidance [20]. Consistent with PRISMA-ScR, we did not conduct risk-of-bias or critical appraisal of included reviews or primary studies; where available, we recorded quality-related information reported by the included reviews without independent verification or synthesis.

PRISMA-ScR: keine Bias-Bewertung, nur berichtete Qualitätsinformationen

2.4 Piloting Guidance from AIHTA and ASSESS DHT

The AIHTA procurement checklist [13] was examined to determine whether it is fit for purpose in the context of the defined use cases, while the ASSESS DHT guidance documents [14-16] were piloted to evaluate their applicability. First, ASSESS DHT taxonomy was piloted for use cases. Second, the guiding questions from the AIHTA guidance were systematically mapped to each use case to determine relevance (e.g., GDPR-related requirements presumed applicable across cases; additional items aligned with selected EUnetHTA Core Model domains). Third, the ASSESS DHT guidance was examined to identify components pertinent to documentation-focused applications – particularly AI scribes – including appropriate metrics and recommended assessment frequency. Evidence from the scientific literature was integrated to complement and substantiate these determinations.

AIHTA-Beschaffungscheckliste und ASSESS-DHT-Guidance: Relevanzprüfung

3 Results

3.1 Survey

Based on the GÖG report [17], we identified *Dragon Medical One* by Nuance as a clinical documentation tool used in some of the participating hospitals. The survey we conducted did not provide the names of other tools or specify particular use cases or functionalities where AI was implemented in hospital documentation. Therefore, we did not limit our analysis in the scoping review to predefined use cases but instead considered broad implementation cases.

keine Erhebung konkreter Tools/spezifischer Use Cases; Analyse auf übergeordnete Implementierungsfälle erweitert

3.2 Scoping Review

3.2.1 Outcomes

We grouped the outcomes and evaluation metrics reported in the included evidence into four broad categories:

Endpunkte in 4 Untergruppen unterteilt

Technical performance and **documentation quality** reflect how well the AI system performs in its intended task and how this translates into the quality of resulting documentation

1. technische Leistung und Dokumentationsqualität:

a. *Technical performance* was measured with established metrics from computer science, such as accuracy, recall, precision, specificity, F1-score, area under the ROC curve (AUC-ROC), and word-error-rate (WER for speech-to-text systems)[6]. These metrics reflect how reliably the AI system captures and reproduces information.

Genauigkeit, Präzision, F1-Score, AUC-ROC, WER;

b. *Documentation quality* was judged using different approaches:
   - Validated instruments, e.g., the Physician Documentation Quality Instrument (*PDQI*-9)[7], which evaluates attributes like accuracy, completeness, and comprehensibility.
   - Custom quality scores: some developers apply bespoke strategies, e.g., combining indicators such as significant error rates, relevance and precision of captured information, user acceptance, and transcription quality control [21].
   - By counting *errors*, *omissions*, or *hallucinations*. Hallucinations refer to fabricated information that is absent in the source, while omissions capture failures to include relevant details. Both are linked to training data quality, model design, and prompting strategies. Although safety and risks are not standard quality metrics, they are closely tied to hallucinations and omissions; for this reason, reported safety concerns were included within this category.

validierte Qualitätsbewertungsinstrumente, Einzel-Scores sowie die Erfassung von Fehlern und Auslassungen einschließlich Halluzinationen

[6] Definitions of the listed metrics are presented in the Glossary in Appendix A.

[7] Definition is presented in the Glossary in Appendix A.

Clinician-reported outcomes: capture how the AI tool affects the clinician's experience, workload, and performance. These outcomes directly impact clinicians and other healthcare professionals (e.g., nurses, administrative staff) who use AI tools for documentation support. They include changes in workflow, documentation burden, and documentation time, which in turn may influence levels of stress and burnout. Measures of user satisfaction, trust, and acceptance of the tool also belong here, as does provider engagement or disengagement – that is, the extent to which clinicians actively adopt and integrate the AI tool into their practice versus resisting or abandoning its use.

2. kliniker:innenrelevante Endpunkte: Arbeitsbelastung, Workflow, Stress, Zufriedenheit, Vertrauen und Akzeptanz der KI-Dokumentationstools

Patient-reported outcomes: reflect how documentation support tools impact patients directly or indirectly. Satisfaction with the encounter, perceived quality of patient-clinician visit, understanding of plain-language summaries (readability and comprehensibility), trust and acceptance of the tool, safety and risks (errors introduced or prevented by AI).

3. patient:innenrelevante Endpunkte: Zufriedenheit, Verständnis, Vertrauen, Sicherheit

Organisational outcomes: address hospital-level or meso-level implications. Examples include workflow or business efficiency, patient throughput, administrative staffing and resource use, costs, and feasibility of implementation, including integration with EHR systems. Some studies also reported billing outcomes, such as the accuracy and timeliness with which billing-related information could be created and submitted.

4. organisatorische Ergebnisse: Effizienz, Personal, Ressourcen, Kosten, EHR-Integration, Abrechnung

3.2.2 Characteristics of included reviews

In total, seven reviews were included in this report – three systematic (SR) and four scoping reviews (ScR). The unit of analysis for this report was the use case level, derived from the included reviews. Across the reviews, a total of 211 primary studies were reported, of which 11 were identified as duplicates across reviews, resulting in 200 individual studies. An overview of which reviews addressed each use case is shown in Table 3-1.

7 Reviews analysiert, 200 Primärstudien

The included primary studies were published between 2003 and 2024, with the majority appearing after 2020. Most studies were conducted in the United States (US), with additional evidence from the United Kingdom (UK), Netherlands, China, South Korea, Japan, and Israel. Clinical settings included hospitals, outpatient clinics, and specialised care units, and the medical specialties most frequently represented were internal medicine, oncology, paediatrics, dermatology, orthopaedics, and emergency medicine. Some reviews also included studies conducted in primary care setting to a minor extent.

Studien 2003-2024, v. a. ab 2020; internationale Settings, v. a. Krankenhäuser, Ambulanzen, Spezialbereiche

Detailed information on the characteristics and results of each review is presented in the Appendix B Table A-1.

Table 3-1: Overview of documentation support uses cases addressed across included reviews

Use case	Perkins 2024	Bracken 2025	Sasseville 2025	Hassan 2025	Lee 2024	Lumbiganon 2025	Vrdoljak 2024
Ambient AI scribe	✓	✓	✓	✓	✓	✓	✓
Structuring free text	✓						✓
AI-generated documentation		✓				✓	✓
Error detection	✓						
Patient-friendly summaries	✓				✓		✓
Assigning billing codes						✓	

3.2.3 Results by case vignette

The following chapters describe, for each use case, the technologies' core function, target group, claimed benefits, implementation-related cost considerations, deployment strategies and suggested valuation metrics. They also provide an evidence snapshot for each use case – summarizing the number of reviews and primary studies, the clinical settings examined and reported benefits and risks – presented according to the predefined outcome categories for each case vignette.

pro Use-Case: Funktionen, Zielgruppe, Nutzen, Kosten, Implementierung, Metriken, Evidenzübersicht, Nutzen und Risiken

AI scribes

Technology

AI scribes use advanced speech recognition technologies to automatically convert spoken clinician–patient interactions into draft clinical notes with minimal user intervention [22, 23]. Earlier versions relied on speech recognition and digital dictation, while the latest systems, available since 2021, integrate generative AI and large language models (LLMs) [22]. The typical workflow is shown in Figure 3-1. AI scribes, often referred to as "ambient scribes", are scalable, potentially cost-saving solutions that claim to enhance documentation speed and accuracy, optimize workflow, reduce clinician burnout, and improve patient care [23].

KI Scribes: Sprach- und KI-gestützte automatische Erstellung klinischer Notizen

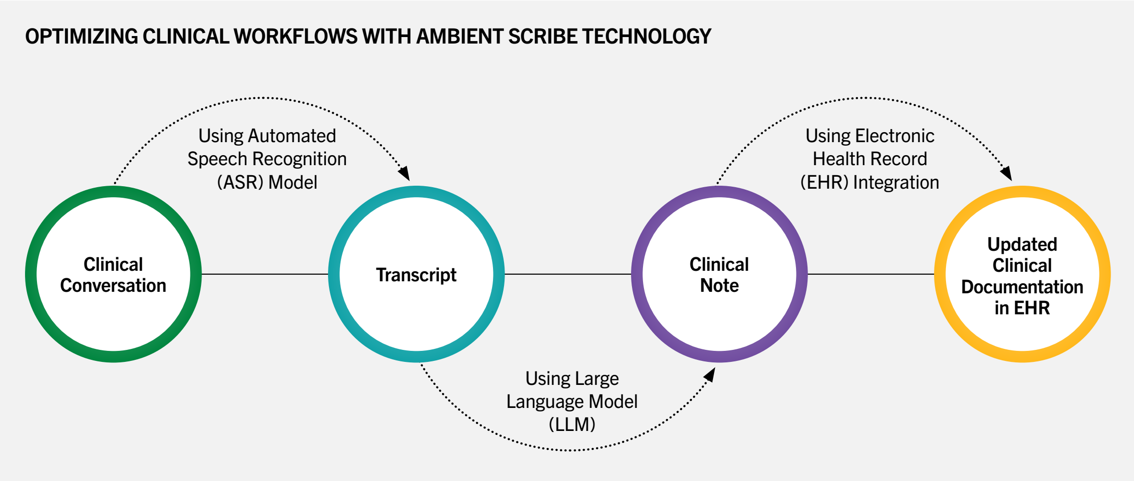
OPTIMIZING CLINICAL WORKFLOWS WITH AMBIENT SCRIBE TECHNOLOGY
Clinical Conversation
Using Automated Speech Recognition (ASR) Model
Transcript
Using Large Language Model (LLM)
Clinical Note
Using Electronic Health Record (EHR) Integration
Updated Clinical Documentation in EHR

Figure 3-1: Workflow of AI scribe technology (Source: [21])

The purchase and implementation of AI scribes entail several types of costs. These include licensing, investments in the integration with existing EHR systems and clinical workflows, technical infrastructure (servers, cloud storage, secure data pipelines), change management, staff training, and ongoing governance and oversight to safeguard performance and safety (review processes, evaluation mechanisms) [21].

Implementierungskosten: Lizenzen, EHR-Integration, Infrastruktur, Change Management, Schulung, Governance

Ambient scribing is increasingly commoditised, with limited differentiation between products and low switching costs. This is reflected in the crowded landscape – spanning established vendors, older dictation software and other documentation tools alongside applications of general-purpose LLMs.

viele Anbieter, geringe Differenzierung, niedrige Wechselkosten

Key characteristics, available products, anticipated benefits, implementation costs, recommended evaluation metrics and deployment category and strategies are summarized in Table 3-2.

Übersicht zu Produkten, Nutzen, Kosten, Evaluierung und Implementierung

Table 3-2: Key characteristics, implementation costs, evaluation metrics of AI scribe

Core function [22, 23]	■ Speech recognition automatically converting spoken interaction into draft text ■ Integrated generative AI and LLMs.
Target group	■ Physicians and other healthcare professionals
AI products[8]	■ Established commercial applications: *Dragon Ambient eXperience, Abridge, Nabla Copilot, Sunoh.ai, Amazon HealthScribe* ■ Legacy dictation tools: *Dragon Medical 10.1, 360, One,* ■ Note-taking documentation tools: *PhenoPad, IBM Watson* ■ General-purpose LLMs: *ChatGPT, FLAN, Llama*
Indicative MDR classification	■ Class I to Class IIa (depending on whether clinician oversight is mandatory or optional)[9]
Anticipated benefits [23]	■ Scalable, potentially cost saving ■ Faster, more accurate documentation ■ Improved workflow ■ May reduce clinician burnout ■ Better patient care.
Implementation & costs [21]	■ Subscription/licensing fees ■ Integration with EHR systems ■ Infrastructure (servers, cloud, security) ■ Staff training, change management ■ Governance/monitoring
Evaluation metrics [21]	**Process:** ■ documentation time (in notes/EHR/visits/after-hours) ■ note completion (time to closure, % closed same day) ■ adoption & usage (number of encounters with AI scribe, provider counts) ■ integration with EHR/workflows **Quality:** ■ documentation accuracy (quality scores, error rate, medical term recall, precision of transcription) ■ documentation completeness (% retained) **Financial:** ■ productivity (RVUs, encounters per period) ■ coding quality for billing ■ cost-effectiveness **Experience:** ■ clinician well-being (burnout, turnover, satisfaction), usability ■ patient experience ■ technological retention

[8] Identified through the included reviews.

[9] If the AI output is reviewed and validated by a clinician before being saved or used in care, it is administrative or assistive, not directly influencing care decisions → Class I (low risk, informational support only). If the AI output is used directly in the EHR or clinical workflow without mandatory human review, it can influence diagnosis, treatment, or care decisions, even indirectly → Class IIa.

Evaluation metrics [21] *(continuation)*	**Fairness & Subgroup-Level Performance** ■ stratified accuracy/error rates by clinician accent, gender, native language ■ subgroup-level precision/recall/F1 for key medical entities ■ completeness differences across specialties or encounter types ■ documentation performance across patient demographics (where legitimate and privacy-compliant) ■ defined disparity thresholds (<5% acceptable; 5-10% needs monitoring; >10% significant gap) **Transparency & Reproducibility** ■ availability of anonymised validation datasets/data descriptors ■ availability of preprocessing scripts ■ clarity of evaluation pipelines ■ peer-verifiable validation reproducibility **Bias mitigation logs & continuous monitoring** ■ routine assessment of subgroup performance (e.g., accuracy, completeness, hallucination rates across clinician groups, specialties, encounter types) ■ predefined disparity thresholds to flag potential bias ■ documentation of detected disparities in a bias-mitigation log ■ recorded corrective actions (model updates, prompt adjustments, workflow changes) ■ scheduled re-evaluation cycles and tracking of outcomes ■ traceability of fairness issues across validation and deployment phases.
Deployment strategies [24].	■ Human oversight required for all outputs; formal evaluation and potential regulatory approval for generative summaries

Abbreviations: EHR … electronic health record, GPT … Generative Pre-trained Transformer; FLAN … Finetuned Language Net; LLM … large language model, MDR … Medical Device Regulation, RVU … relative value unit

Evidence snapshot

Seven reviews (four SR, three ScR) covering 36 primary studies were included. Most studies come from the U.S., with some studies from Europe (UK, the Netherlands) and Asia (China, Bangladesh, South-Korea). Clinical settings included hospitals, specialized healthcare services (surgical departments, tertiary cancer centres, children's clinics), emergency departments, and primary care. A broad range of specialities are represented (internal medicine, surgery, psychiatry, dermatology, radiology, pathology). Detailed information extracted from the reviews can be found in Appendix B Table A-2.

7 Reviews, 36 Primärstudien, internationale Studien; Fachrichtungen und Settings vielfältig

A synthesis of outcome-related findings is provided in Vignette 1.

Technical performance and documentation quality

Findings on performance metrics were heterogeneous and context dependent. Reported accuracy ranged from moderate to high and often varied by document type. Recall values indicated frequent omissions.

Leistung und Dokumentationsqualität heterogen

Assessments of documentation quality were likewise inconsistent: some studies using formal instruments (e.g., PDQI-9 or comparable scales) described higher completeness/readability or perceived improvements, whereas other evaluations – frequently based on non-standardised assessments – reported declines or loss of narrative detail. Deficiency outcomes were highly inconsistent (from reductions through no change to increased billing submission deficiencies).

Genauigkeit mittel bis hoch, häufige Auslassungen; Qualitätsergebnisse inkonsistent

Clinician- and patient-reported outcomes

Findings on clinician-reported outcomes were variable. All reviews reported high clinician satisfaction, moderate to high perceived usability of the AI scribe and reductions in perceived documentation burden. However, quantitative effects were inconsistent across – and sometimes within – reviews. Documentation time varied widely: several studies reported substantial savings, whereas others found no gains or increases, including more after-hours EHR work. Burnout findings were similarly mixed, ranging from no change to reductions. Concerns raised by clinicians included accuracy, and potential loss of narrative nuance. From the patient perspective, encounters were often described as more personal and less screen-focused, most patients agreed to participate, although a minority expressed discomfort with recordings. Patient safety incidents were not documented in the included studies.

hohe Zufriedenheit, variable Zeitersparnis, gemischte Burnout-Effekte; Bedenken zu Genauigkeit und Narrativverlust; Sicherheitsvorfälle nicht berichtet

Organisational outcome

Findings at the organisational level suggest potential for productivity improvements (e.g., shorter consultation or capacity to see more patients); however, concerns were also raised that it may translate into expectations to increase patient throughput. Potential cost savings relative to human scribes were also noted, with estimates of USD 13,000-14,000 per provider per year.

mögliche Produktivitätspotenzial bei gleichzeitigen Bedenken zu erhöhtem Patientendurchsatz

Table 3-3: Vignette 1 – AI scribe

Evidence base	**7 reviews** (SR: [1], [23], [22], [25]; ScR: [5], [26], [27]) covering **36 primary studies.** **Study designs:** 1 RCT (scribe vs typing vs dictation), 1 controlled (scribe vs typing), observational (retrospective, prospective cohort, cross-sectional) and mixed-methods studies
Reported findings	**Technical performance and documentation quality (5 reviews):** 🔶 *Accuracy*: moderate to high, depending on document type (68-97%). ⚠ *Recall*: frequent omissions. 🔶 *Documentation quality*: mixed (higher in some studies, declines in others). 🔶 *Deficiency rates*: inconsistent (significant reductions ↔ no effects). ⚠ *Errors*: moderate to high error rates (*RCT*: 36% notes contained erroneous/fictitious info). ⚠ *Reproduction*: 50% failed, 35% of notes fully reproduced. **Clinician-reported outcomes (7 reviews):** ✅ *Satisfaction*: generally high. 🔶 *Burnout*: mixed (no change ↔ some reduction). ✅ *Documentation burden*: reduced, moderate to high usability scores. 🔶 *Documentation time*: highly variable (19-92% decrease to 13-50% increase; sometimes more after-hours work) (*RCT*: AI scribes 2.7x faster than typing, 2x faster than dictation for patient history sections, and 3x faster for physical exam). ⚠ *User experience*: concerns about accuracy, reliability, loss of narrative. **Patient-reported outcomes (4 reviews):** ✅ *Patient experience*: less screen-focused; low opt-out, some discomfort with recordings. ✅ *Safety*: no incidents reported. **Organisational outcomes (4 reviews):** ✅ *Consultation length*: up to 26% shorter with AI. 🔵 *Productivity* (RVUs, patient load): modest (~4%) but significant increase in RVUs, no change in patient volume. Concerns about patient load expectations. ✅ *Cost efficiency*: ~$13-$14,000 annual saving/provider vs. in-person scribes.

Abbreviations: AI … artificial intelligence; RCT … randomised controlled trial; RVU … relative value unit, ScR … scoping review, SR … systematic review.

Legend: ✅ *positive findings,* 🔶 *mixed findings,* 🔵 *no difference,* ⚠ *caution.*

Structuring unstructured text

Technology

KI-basierte Textstrukturierung: keine Spracherkennung, automatische Annotation und Integration ins EHR

In contrast to an AI scribe, these systems do not rely on speech recognition for automatic recording and transcription. Instead, physician notes – written by hand or entered electronically – are processed by the tool, which structures the content and integrates it into the EHR. Annotation is an inherent step in this process rather than a stand-alone functionality: metadata or semantic tags (e.g., SNOMED codes, concept labels) are assigned to the text, enabling downstream use such as retrieval, coding, or secondary analysis [25, 27].

Übersicht zu Produkten, Nutzen, Kosten, Evaluierung und Implementierung

Key characteristics, available products, anticipated benefits, implementation costs, recommended evaluation metrics and deployment category and strategies are summarized in Table 3-4.

Table 3-4: Key characteristics, implementation costs, evaluation metrics of structuring unstructured text use case

Core function [25, 27]	■ Note processing (incl. annotation) and integration into the EHR ■ *No* speech recognition and recording.
Target group	■ Physicians and other healthcare professionals.
AI products[10]	■ Rule-based NLP, machine learning and deep learning models, ■ General purpose LLMs (e.g., GPT-3.5, GPT-4), and ■ Other open-source or fine-tuned LLMs (e.g., FLAN-T5, FLAN-UL2, Llama-2, Vicuna, Alpaca, and Med-Alpaca) ■ Cloud-based AI service platforms (e.g., Microsoft Azure AI) and commercial applications (e.g. informed.360) [28, 29].
Indicative MDR classification	■ NA to Class IIa (depending on whether used admin only vs clinical decision support)[11]
Anticipated benefits	■ Largely the same as AI scribes (reduced documentation burden; better structure/quality; workflow support). ■ *Differences*: no reduction in on-screen time during the encounter; benefits are typically post-encounter (structuring/coding) rather than real-time.
Implementation & costs	■ Largely the same as for AI scribes. ■ *Differences*: less infrastructure requirements (no microphones/recording hardware or high-bandwidth audio pipelines); lower audio-specific privacy risks (no continuous voice capture); still require secure text pipelines/storage/compute.
Evaluation metrics	■ Same *core metrics* as to AI scribes: extraction accuracy/correctness (precision, recall, F1), field-level completeness (% correctly populated fields), and error types (omissions, mismatches, false positives), time required for reviewing/correcting extracted fields, editing workload, task load, and clinician satisfaction, adoption/usage patterns, clinician-reported usability and workflow impact. ■ *Not applicable*: speech-to-text metrics (WER, transcription recall/precision), patient-reported outcomes generally not applicable, PDQI-9, readability.
Deployment strategies	■ Integration with EHR modules; benchmarking and validation per dataset

Abbreviations: EHR … electronic health record, GPT … Generative Pre-trained Transformer; FLAN … Finetuned Language Net; MDR … Medical Device Regulation, NA … not applicable, NLP … natural language processing, LLM … large language model, PDQI … Physician Documentation Quality Instrument, T5 … Text-to-Text Transfer Transformer; WER … word error rate

[10] Identified through the included reviews and internet search.

[11] If the structured output is used only for administrative, research, or analytics purposes (e.g. for reporting, workload tracking, or coding support), then it is not a medical device and no risk class applies. If its structured output feeds into the EHR or supports clinical analytics or decision-making, (e.g. structured fields used in clinical dashboards or decision support) then under MDR it is likely Software as a Medical Device (SaMD) and the same rules apply as for AI scribe and AI-generated clinical documentation.

Evidence snapshot

Two reviews (one SR and one ScR) were included. The primary studies span a wide range of clinical settings, including hospital wards, emergency departments, operating theatres, specialized services (e.g., HIV clinics), tertiary centres, outpatient clinics, and primary care. Medical specialties represented include general medicine, internal medicine, surgery, psychiatry, dermatology, oncology, paediatrics, and orthopaedics. Geographically, studies were conducted predominantly in the U.S., with additional evidence from Europe (UK, Germany, the Netherlands), Asia (South Korea, China, Japan), and Israel. Detailed information extracted from the reviews can be found in Appendix B Table A-3.

2 Reviews, vielfältige klinische Settings und Fachrichtungen;

überwiegend USA, zusätzlich Europa, Asien, Israel

A synthesis of outcome-related findings is provided in Vignette 2.

Technical performance and documentation quality

Across studies reporting technical performance, AI models generally showed high scores on standard performance metrics (accuracy, F score, PPV, AUC), though performance varied by task, dataset, and modelling approach. For symptom-labelling and similar classification tasks, LLMs were reported to perform better on common symptoms than on rarer ones. Evidence on completeness was limited: in one study of rule-based models, precision exceeded recall, indicating fewer incorrect extractions but missed relevant items (precision/recall). In another study, the neural network model converting free text into structured records achieved moderate coherence. Documentation quality using validated tools or bespoke quality scores was not reported in the included reviews.

technische Leistung meist hoch, variiert nach Aufgabe und Datensatz;

Vollständigkeit begrenzt, Qualitätsbewertungen selten berichtet

Clinician- and patient-reported outcomes

Clinician-reported outcomes were limited but suggestive of faster documentation, alongside reductions in documentation time. One study noted that time gains coincided with a slight decrease in quality. Patient-reported outcomes were not reported in the included evidence.

schnellere Dokumentation, Zeitersparnis; mögliche Qualitätsminderung;

Documentation speed, reported on in one study, increased by 15%, while two studies reported documentation time, which decreased. Patient-reported outcomes were not assessed in the included reviews.

keine Patient:innen Ergebnisse

Organisational outcomes

None of the reviews reported on outcomes in this category.

keine organisatorischen Ergebnisse

Table 3-5: Vignette 2 – Structuring unstructured text

Evidence base	**2 reviews** (1 SR: [25], 1 ScR: [27]) covering **~90 primary studies** **Study designs:** retrospective and prospective observational, cross-sectional, technical benchmarking pilots, qualitative, and mixed-methods studies
Reported findings	**Technical performance and documentation quality (2 reviews):** ☑ *Accuracy*: generally high (>90%), with some tasks reaching 100%. ☑ *F-score:* up to 0.984 (e.g., race classification) ☑ *PPV*: 0.95-0.97 (e.g., patient safety events, social factors) ☑ *AUC*: up to 0.876 (e.g., actionable findings in radiology) ◆ *Sensitivity*: high for common symptoms (0.85-1.00), moderate/low for less common ones (0.20-1.00). ☑ *Specificity*: high for all symptoms for labelling tasks by GPT-4 (0.947-1.000). ⚠ *Precision/recall* (showing *completeness*): limited (e.g., phenotype recognition 83%/51%)

Reported findings *(continuation)*	⚠ *Coherence*: limited (e.g., 69% with neural networks). 🔶 *Performance* varied by model (ChatGPT-3.5 vs. GPT-4), task and symptom type. **Clinician-reported outcomes (1 review):** ✅ *Documentation speed/time*: reported gains (rule-based models increased speed by 15%, documentation time decreased up to 56%, but quality also decreased slightly).

Abbreviations: AUC … area under the curve, GPT … Generative Pre-trained Transformer; PPV … positive predictive value, ScR … scoping review, SR … systematic review

Legend: ✅ *positive findings,* 🔶 *mixed findings,* ⚠ *caution.*

AI-generated documentation (without speech recognition)

Technology

From structured or semi-structured data from the EHR (e.g., lab results, diagnoses, medications, procedures) or transcripts of encounters, the AI tool generates a coherent medical document (e.g., discharge summary or handover note between hospital departments) without automatic speech recognition element. Template-/rule-based natural language generation, traditional machine learning (ML), or prompt-engineered LLMs are AI technologies used for this purpose. Annotation is an inherent step in this process: metadata or semantic tags (e.g., SNOMED codes, concept labels) are assigned to the text, enabling downstream use. Medical text summarisation also belongs to this category [1, 5, 26, 27].

KI-generierte Dokumentation: aus EHR-Daten oder Transkripten, keine Spracherkennung; Annotation und Metadaten für Weiterverwendung, inkl. Textzusammenfassung

Key characteristics, available products, anticipated benefits, implementation costs, recommended evaluation metrics and deployment category and strategies are summarized in Table 3-6.

Übersicht zu Produkten, Nutzen, Kosten, Evaluierung und Implementierung

Table 3-6: Key characteristics, implementation costs, evaluation metrics of AI-generated documentation

Core function [1, 5, 26, 27]	■ Generation of clinical notes, summaries, discharge letters from structured/semi-structured input. ■ *No* speech recognition and recording.
Target group	■ Physicians and other healthcare professionals.
AI products[12]	■ Template/rule-based NLP, ML generators (i.e. non-LLMs, typically in earlier studies), and ■ General purpose LLMs (e.g., GPT-3.5, GPT-4) configured for clinical summarization ■ Commercial applications (e.g. informed.360, Notable, Abridge) [24, 28].
Indicative MDR classification	■ Class I to Class IIa (depending on whether clinician oversight is mandatory or optional)[13]
Anticipated benefits	■ Largely the same as AI scribes (reduced documentation burden; better structure/quality; workflow support). ■ *Differences*: no reduction in on-screen time during the encounter; benefits are typically post-encounter (coherent document generation) rather than real-time.
Implementation & costs	■ Largely the same as for AI scribes. ■ *Differences*: less infrastructure requirements (no microphones/recording hardware or high-bandwidth audio pipelines); lower audio-specific privacy risks (no continuous voice capture), ■ Stronger emphasis on model governance (bias, hallucination risk, legal/ethical safeguards) and possible higher compute/storage costs if large LLMs are used.

[12] Identified through the included reviews and internet search.

[13] If the AI output is reviewed and validated by a clinician before being saved or used in care, it is administrative or assistive, not directly influencing care decisions → Class I (low risk, informational support only). If the AI output is used directly in the EHR or clinical workflow without mandatory human review, it can influence diagnosis, treatment, or care decisions, even indirectly → Class IIa.

Evaluation metrics	■ Same core metrics as AI scribes: documentation quality, efficiency, usability, user experience. ■ Emphasis on accuracy of generated content, hallucination/error rates. ■ *Not applicable*: speech-to-text metrics (WER, transcription recall/precision).
Deployment strategies	■ EHR integration under clinician supervision; technical customisation.

Abbreviations: AI … Artificial intelligence; EHR … electronic health record, GPT … Generative Pre-trained Transformer; NLP … natural language processing, LLM … large language model, ML … machine learning, WER … word error rate

Evidence snapshot

Three reviews were included (one SR, two ScR). The included studies were conducted in hospitals and specialized oncology clinics, covering medical specialties such as general medicine, internal medicine, surgery, paediatrics, and oncology. Geographically, most evidence came from the U.S., with additional studies from Europe (UK, the Netherlands) and South Korea. Detailed information extracted from the reviews can be found in Appendix B Table A-4.

3 Reviews, Krankenhäuser und Onkologie; Fachrichtungen vielfältig; v. a. USA, teils Europa und Südkorea

A synthesis of outcome-related findings is provided in Vignette 3.

Technical performance and documentation quality

Documentation quality, when assessed with validated instruments (e.g., PDQI-9), was rated moderate to high. In head-to-head comparisons, best adapted LLMs produced higher quality medical summaries than dictation or typing in quality and were often rated equivalent or superior to medical experts, with significantly fewer errors, greater conciseness and higher completeness. In some cases, they even captured information missed by the experts, although clinician oversight remained necessary. Other studies measured factual correctness as moderate to high, depending on the document type. At the same time, limitations were noted: in some assessments a substantial proportion of AI-generated notes required clinician edits or showed high error counts, and the prevalence of omissions and fabricated content (“hallucinations”) ranged from none detected to notable levels of missing or fictitious elements. Technical performance metrics were not reported.

Dokumentationsqualität moderat bis hoch; LLMs meist besser als manuell; Bearbeitung nötig, Halluzinationen möglich

Clinician- and patient-reported outcomes

Clinician-reported time and effort for documentation varied widely from no reduction to considerably less effort and time. The acceptance of the AI-generated documents was high in some settings (e.g., general practitioners (GPs) fully accepted AI-generated discharge summaries), yet concerns persisted about accuracy, medico-legal liability, privacy and data security, timing mismatches with clinical workflows, missing clinical details (e.g., differential diagnoses, pertinent negatives), and over-general action plans.

Zeit- und Arbeitsaufwand variabel, hohe Akzeptanz, Bedenken zu Genauigkeit, Haftung, Datenschutz;

Patient-reported outcomes were rarely assessed; one review of potential harm from summarisation errors found that AI-generated summaries showed a lower estimated likelihood and extent of harm than expert written summaries. Nevertheless, review authors emphasised that clinician oversight remined essential.

patient:innenrelevante Endpunkte selten erhoben; geringeres Schadenspotenzial berichtet, dennoch Kontrolle nötig

Organisational outcomes

Reported organisational evidence focused on implementation challenges rather than quantified impacts. Studies highlighted the need for technical improvements and customisation for EHR integration and extensive training to achieve effective use.

organisatorische Evidenz v. a. zu Implementierungshürden

Table 3-7: Vignette 3 – AI-generated medical documentation without speech recognition

Evidence base	**3 reviews** (1 SR: [1]; 3 ScR: [26], [27],) covering **24 primary studies** **Study designs:** 1 controlled study (LLM vs medical experts), prospective and retrospective cohort studies, mixed-methods and quasi-experimental studies.
Reported findings	**Technical performance and documentation quality (4 reviews):** ✅ *Overall performance*: LLM vs medical experts: LLM is equivalent (45%) or superior (36%). ✅ *Correctness, conciseness, completeness:* ■ LLM vs medical experts: LLM significantly better; in radiology tasks similar to experts. ■ Non-contr ■ olled studies: varying median factual correctness (81 to 85% in discharge summaries and 71 to 79% in surgical notes). ✅ *Overall quality:* ■ LLM vs medical experts: LLM-generated notes scored higher than typing/dictation. ■ Non-controlled studies: moderate-to-high (PDQI-9 scores up to 48/50, in most studies 30-36/50). 🔶 *Hallucinations*: ■ LLM vs medical experts: 5% vs. 12%. ■ Non-controlled studies: mixed (from no hallucinations up to 10%, depending on the type of document). 🔶 *Errors*: ■ LLM vs medical experts: 2% vs 4%. ■ Non-controlled studies: mixed (low to high error rate). ⚠ *Omissions*: high rate (up to 86%). **Clinician-reported outcomes (2 reviews):** 🔶 *Time/effort reduction*: up to 43% less time/33% less effort, 2-5 min faster compared to dictation; some studies found no significant time or effort difference. ✅ *Acceptance*: outputs were fully accepted (GPs rated ChatGPT summaries as good as or better vs junior doctors', adherence comparable to junior doctors). 🔶 Clinician *experience*: lower perceived effort with ChatGPT, but persistent *concerns* about accuracy, errors, missing information, patient safety, patient privacy, liability risks, vague action plans (e.g., "follow up on pending results") and perception that existing templates may be easier to use. **Patient-reported outcomes (1 review):** ✅ *Safety*: GPT-4 mistakenly generated several absent conditions, still slightly lower likelihood and extent of its potential harm than that of expert summaries (12-16% vs 14-22%). **Organisational outcomes (1 review):** ⚠ *Implementation challenges*: technical improvements and customisation for effective integration into existing EHR systems, extensive training of personnel needed.

Abbreviations: EHR … electronic health record, GPs … general practitioners; HPI … history of present illness, LLM … large language model, ScR … scoping review, SR … systematic review, Error detection/note quality assessment

Legend: ✅ *positive findings,* 🔶 *mixed findings,* ⚠ *caution.*

Patient-friendly summaries

Technology

LLMs and other transformer-based models, along with rule-based and hybrid approaches, have been applied to link medical terms in clinical notes to lay definitions, thereby improving comprehension among non-specialists. Ontology-based algorithms have also been used to convert medical language into simplified sentences in plain terms [5, 25-27].

KI vereinfacht medizinische Sprache für Patient:innen

Key characteristics, available products, anticipated benefits, implementation costs, recommended evaluation metrics and deployment category and strategies are summarized in Table 3-8.

Übersicht zu Produkten, Nutzen, Kosten, Evaluierung und Implementierung

Table 3-8: Key characteristics, implementation costs, evaluation metrics of patient-friendly summaries

Core function	▪ Convert clinical notes into plain-language summaries by linking medical terms to lay definitions and simplifying sentences.
Target group	▪ Patients or caregivers primarily, ▪ Clinicians indirectly via clearer post-visit communication.
AI products[14]	▪ Ontology/rule-based NLP, ▪ General purpose LLMs (e.g., GPT-3.5, GPT-4), and ▪ Hybrid approaches combining both.
Indicative MDR classification	▪ NA (communication/education function)
Anticipated benefits	▪ Improved patient comprehension, ▪ Clearer post-visit instructions; potential support for shared decision-making.
Implementation & costs	▪ Largely the same as for AI scribes. ▪ *Differences*: less infrastructure requirements (no microphones/recording hardware or high-bandwidth audio pipelines), ▪ Stronger emphasis on model governance (bias, hallucination risk, legal/ethical safeguards), term-mapping/ontology maintenance and review workflows for patient-facing content.
Evaluation metrics	▪ Same core metrics as AI scribes: expert-assessed documentation quality, efficiency, usability, user experience, clinician review time (if applicable). ▪ Emphasis on readability/comprehension, correctness/absence of fabricated content, alignment with intended meaning, patient satisfaction/acceptability. ▪ *Not applicable*: speech-to-text metrics (WER, transcription recall/precision).
Deployment strategies	▪ Embedded in patient portals or discharge letters

Abbreviations: AI … Artificial intelligence; NA … not applicable, NLP … natural language processing, LLM … large language model, WER … word error rate

Evidence snapshot

Three reviews were included (one RS and two ScR). The included studies were conducted in hospitals and outpatient clinics, covering specialties such as internal medicine, pediatrics, and oncology. Most of the evidence originated from the U.S., with additional studies from China and the Netherlands. Detailed information extracted from the reviews can be found in Appendix B Table A-5.

3 Reviews, Krankenhäuser und Ambulanzen;

Fachrichtungen vielfältig; USA, China und NL

A synthesis of outcome-related findings is provided in Vignette 4.

Technological performance and documentation quality

Outcomes in this category were not measured in the studies.

keine Ergebnisse

Clinician-reported and patient-reported outcomes

Studies described improved patient understanding and health literacy, high acceptance of AI-assisted after-visit/discharge summaries, and perceived improvements in patient–clinician interactions.

besseres Verständnis, hohe Akzeptanz, positive Interaktionen

Organisational outcomes

These outcomes are not applicable for this case vignette.

keine Ergebnisse

[14] Identified through the included reviews.

Table 3-9: Vignette 4 – Patient engagement and communication

Evidence base	**3 reviews** (1 SR: [25], 3 ScR: [5], [27]) covering **8 primary studies** **Study designs:** RCTs, prospective observational, and cross-sectional studies
Reported findings	**Clinician-reported and patient-reported outcomes (4 reviews)** ✅ Significantly improved *readability* and *understandability* of LLM-transformed discharge summaries. ✅ Enhanced *patient understanding* and health literacy. ✅ Patient *acceptance*: 96% of patients recommended AI-assisted after-visit summaries. ✅ Improved *patient–clinician relationship* and interactions.

Abbreviations: *LLM … large language model, RCTs … randomised controlled trials; ScR … scoping review, SR … systematic review*

Legend: ✅ *positive findings.*

3.2.4 Error detection, clinical note quality assessment

Technology

A frequent component of clinical documentation improvement initiatives is manual chart review to assess clinical notes for timeliness, completeness, precision, and clarity. AI tools can assist in that end by recognising the presence or absence of knowledge domains, social determinants of health, performance status, and topic discussion, prompting clinicians to make additional notes relating to a domain when needed. In addition to these domains, note unclarity and redundant information comprise major problems in clinical documentation.

KI unterstützt Qualitätsprüfung und Vollständigkeit klinischer Dokumentation

Key characteristics, available products, anticipated benefits, implementation costs, recommended evaluation metrics and deployment category and strategies are summarized in Table 3-10.

Übersicht zu Produkten, Nutzen, Kosten, Evaluierung und Implementierung

Table 3-10: Key characteristics, implementation costs, evaluation metrics of error detection and note quality assessment

Core function	■ Automatical detection of errors, omissions, redundancies, and unclear content in clinical notes. ■ Flagging missing knowledge domains (e.g., social determinants of health, performance status) and prompt clinicians to add relevant information.
Target group	■ Clinicians or administrative staff.
AI products[15]	■ Rule-based systems, NLP and ML (incl. deep learning and neural networks) ■ Research prototypes in the literature [25], some commercial products already available (e.g., informed.360, 3M, Optum) [24, 28, 29].
Indicative MDR classification	■ NA (purely administrative)
Anticipated benefits	■ Improved documentation quality (timeliness, completeness, precision, clarity) ■ Reduction of redundant or unclear content, ■ Potential to enhance patient safety by reducing incomplete notes.
Implementation & costs	■ Similar to other documentation tools (integration with EHR, training, governance). ■ *Differences:* greater reliance on quality assurance and feedback workflows, with need for clinician acceptance of automated prompts; resources required for ongoing fine-tuning and validation of models against clinical standards.

[15] Identified through the included reviews and internet search.

Evaluation metrics	■ *Core performance metrics*: correctness of identified errors and quality issues compared with a clinician-annotated gold standard (e.g., agreement rates, false-positive/false-negative balance), and accuracy in detecting omissions, inconsistencies, contradictions, redundant content, or clinically relevant missing elements. ■ *Quality assessment indicators*: domain completeness (presence/absence of required sections), structural quality (organisation, sectioning, coherence), clarity/readability indices (e.g., readability grade level), and measures of redundancy or copy-paste patterns. ■ *Usability and workflow impact*: clinician workload for reviewing flagged issues, perceived usefulness of suggestions, and influence on documentation efficiency.
Deployment strategies	■ Integrated audit feedback tools; internal quality assurance processes.

Abbreviations: EHR … electronic health records NA … not applicable, NLP … natural language processing, ML … machine learning

Evidence snapshot

One review (SR) was included. The primary studies with the SR were carried out in hospital wards, specialised healthcare services, and outpatient clinics. They covered medical specialties such as general medicine, oncology, and paediatrics. Geographically, most evidence came from the U.S. and Europe, with some additional studies from Asia. Detailed information extracted from the reviews can be found in Appendix B Table A-6.

1 Review; Krankenhäuser, Ambulanzen; Fachrichtungen vielfältig; v. a. USA/Europa, teils Asien

A synthesis of outcome-related findings is provided in Vignette 5.

Technical performance and documentation quality

Technical performance was assessed using standard performance metrics (accuracy, PPV, F1-scores). AI models generally achieved high scores with some variability across tasks and datasets. F1 score, a metric that combines precision and recall was found to be moderate in some datasets, and high in others, indicating varying but generally solid balance between correctly identifying and missing documentation issues.

technische Leistung: meist hoch, variiert nach Aufgabe/Dataset

Clinician- and patient-reported outcomes

Outcomes in this category were not reported.

keine Kliniker:innen/ Patient:innen und organisatorischen Ergebnisse

Organisational outcomes

Outcomes in this category were not reported.

Table 3-11: Vignette 5 – Error detection in clinical notes, assessing note quality

Evidence base	**1 review** (SR: [25]) covering **20 primary studies**
Reported findings	**Technical performance and documentation quality (1 review):** ☑ *Accuracy:* high (91-93%) ☑ *F1-score:* moderate to high performance (0.68-0.94) ☑ *PPV* up to 0.93.

Abbreviations: PPV … positive predictive value, SR … systematic review
Legend: ☑ *positive findings*

3.2.5 Billing codes

Technology

AI supported tools used to automatically extract and assign billing codes (e.g., ICD codes) from clinical notes using NLP and machine learning models.

KI-gestützte Zuweisung von Abrechnungscodes aus klinischen Notizen

Key characteristics, available products, anticipated benefits, implementation costs, recommended evaluation metrics and deployment category and strategies are summarized in Table 3-12.

Table 3-12: Key characteristics, implementation costs, evaluation metrics of assigning billing codes

Core function	■ Automatically extract and assign billing codes from clinical notes.
Target group	■ Primarily administrative staff, hospital billing departments.
AI products[16]	■ NLP and ML models (random forest, deep learning, recurrent neural networks). ■ Prototype models in the literature (no commercial products identified).
Indicative MDR classification	■ NA (purely administrative)
Anticipated benefits	■ Improved coding accuracy and completeness, ■ Potential time savings compared to manual coding.
Implementation & costs	■ Similar requirements to other documentation tools (integration with EHR and billing systems, staff training, governance). ■ Emphasis on alignment with billing regulations, local coding standards, and continuous updates (e.g., ICD revisions).
Evaluation metrics	■ Standard performance metrics for classification: accuracy, precision, recall, F-score, and error rates compared with human coders, agreement on code sets or DRG assignment; analysis of undercoding and overcoding patterns and their potential financial impact
Deployment strategies	■ Validation phase ■ Use in supervised billing environments before full automation

Abbreviations: EHR … electronic health record, ICD … International Classification of Diseases, NA … not applicable, NLP … natural language processing, ML … machine learning

Evidence snapshot

One review was included (ScR). The included studies were conducted in hospitals and outpatient clinics and covered mainly specialties such as HIV care, orthopaedics, and dermatology. Geographically, the evidence came from China and South Korea. Detailed information extracted from the reviews can be found in Appendix B Table A-7.

1 Review; Krankenhäuser/ Ambulanzen; China/Südkorea

A synthesis of outcome-related findings is provided in Vignette 6

Technical performance and documentation quality

The included studies focused exclusively on technical performance. Automated coding systems improved coding accuracy and completeness and reduced ICD-related errors. However, performance varied by modelling approach and dataset. Risks or implementation challenges were not reported.

Genauigkeit und Vollständigkeit verbessert; Leistung variabel; Risiken/Implementierung nicht berichtet

Clinician- and patient-reported outcomes

Outcomes in this category were not reported.

keine Kliniker:innen/ Patient:innen und …

[16] Identified through the included reviews.

Organisational outcomes

Outcomes in this category were not reported.

… keine organisatorischen Ergebnisse

Table 3-13: Vignette 6 – AI-generated billing codes

Evidence base	**1 review** (1 ScR: [26]) covering **2 primary studies**
Reported findings	**Technical performance and documentation quality (1 review):** ◆ *Accuracy* ranged 59-87% depending on model (deep learning: 59%; random forest: 87%) vs human coders. ☑ *Errors:* fewer ICD-related errors vs human coders. ☑ *Completeness*: improved with AI.

Abbreviations: AI … artificial intelligence; ICD … International Classification of Diseases, ScR … scoping review
Legend: ☑ *positive findings,* ◆ *mixed findings.*

3.3 Piloting the AIHTA and ASSESS DHT guidance documents

The ASSESS DHT guidance document on the taxonomy [15] was piloted first to test its applicability to the documentation support use cases. According to the ASSESS DHT classification, documentation support systems were categorised as non-medical, operational tools, as their primary function concerns information management rather than direct diagnosis, monitoring, or treatment. However, as highlighted in other frameworks, such tools may still indirectly influence patient care by affecting documentation quality, completeness, and clinical decision-making. The ASSESS DHT manual for assessment methods [14] itself could not be applied to those use cases that fall outside the taxonomy's medical scope, since evidence requirements and evaluation dimensions are defined only for DHTs with a "medical purpose." The ASSESS DHT manual on real-world data validation methods for AI-based decision support systems [16] provides structured guidance on validation and monitoring throughout the AI lifecycle. Although primarily intended for clinical decision support systems, its principles are relevant for documentation support tools as well. The manual proposes a risk-based approach that distinguishes pre-deployment validation, localised real-world testing, and ongoing post-deployment monitoring. Key elements include bias and fairness assessment, explainability, and transparent performance reporting. These concepts can inform proportionate validation strategies for documentation support systems, ensuring that implementation in hospital settings is accompanied by continuous quality assurance, user feedback mechanisms, and mechanisms to detect performance drift over time.

ASSESS DHT-Taxonomie: Einordnung als nicht-medizinische, operationale Tools mit indirektem Einfluss auf Versorgungsqualität

ASSESS DHT-Validierungsleitfaden: risikobasierter Ansatz für kontinuierliche Qualitätssicherung und Leistungsüberwachung

The AIHTA procurement checklist [13] (Appendix C) was subsequently piloted to examine its relevance for documentation support applications. The checklist is fully applicable to functionalities that qualify as medical devices under the MDR definition. For documentation support not qualifying as medical device, only a subset of items is pertinent – primarily those addressing purpose, data protection and privacy, and AI-specific technical and organisational considerations (e.g. dataset quality, bias mitigation, human oversight), monitoring and performance. By contrast, items directly linked to medical device classification, CE marking, HTA evaluation are not applicable. See

AIHTA-Beschaffungs-checkliste: teilweise anwendbar für nicht-medizinische KI-Dokumentationstools, Fokus auf Datenschutz, Bias und Monitoring

Section 3.4 for a decision flow diagram (Figure 3-3) and modified checklist for hospital procurement decisions about AI-enabled documentation support systems.

In summary, both the ASSESS DHT guidance documents and the AIHTA procurement checklist provided useful reference points but require adjustment to better capture the borderline nature of documentation support tools, which often fall between administrative and clinically relevant functions.

bestehende Leitfäden teils unzureichend für hybride KI-Dokumentationstools zwischen Administration und klinischer Relevanz

3.4 Implementation and procurement considerations for hospital managers

From a hospital management perspective, the pathway toward implementing AI-enabled documentation tools follows two interlinked stages: 1. **ensuring organisational readiness** and 2. **applying proportionate validation and governance**. A sequential and adaptable structure begins with functionality-specific implementation readiness (Figure 3-2), followed by risk-adjusted validation and governance to ensure safe, effective, and compliant deployment (Figure 3-3).

Krankenhausperspektive: zweistufiger Implementierungspfad: Bereitschaft & Validierung

Figure 3-2 outlines cross-cutting implementation enablers – such as workflow integration, data security, bias mitigation, cost management, and trust building – that provide the structural foundation for any AI deployment [24].

zentrale Implementierungsfaktoren für KI-Einführung

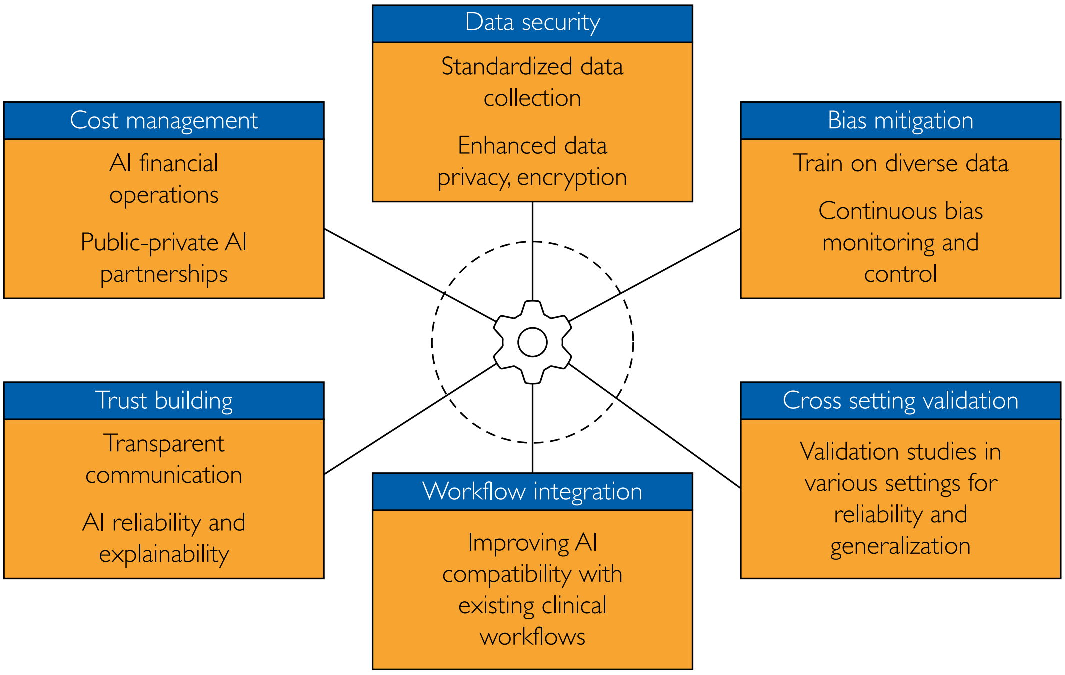
Data security
Standardized data collection
Enhanced data privacy, encryption
Cost management
AI financial operations
Public-private AI partnerships
Bias mitigation
Train on diverse data
Continuous bias monitoring and control
Trust building
Transparent communication
AI reliability and explainability
Cross setting validation
Validation studies in various settings for reliability and generalization
Workflow integration
Improving AI compatibility with existing clinical workflows

Figure 3-2: Implementation enablers (Source: [24])

A first set of factors relates to *technical and organisational integration*. Seamless interoperability with existing EHR systems and workflow compatibility are prerequisites for uptake. In addition to interoperability, robust version control and traceability mechanisms are important. Validation datasets and model artefacts should carry persistent version identifiers, accompanied by preprocessing metadata and stored outputs from each validation cycle. Such records make it possible to determine whether observed performance changes arise from model updates, data modifications, or altered workflows, and thereby support reproducibility, auditability, and regulatory review in hospital environments. Systems that require additional manual steps or complex interfaces tend to generate resistance among clinicians and reduce potential efficiency gains. Successful adoption also depends on adequate *infrastructure and resourcing*, including secure data storage (on-premise or cloud), sufficient computational capacity, and well-defined maintenance responsibilities [16, 30, 31].

technische & organisatorische Integration: Interoperabilität, Infrastruktur, Ressourcen

Data governance and security are consistent concerns, particularly for cloud-based or third-party solutions that process identifiable health data. Compliance with data protection standards, encryption, and clear accountability for data access and storage are key. Closely linked is *bias mitigation*, which requires both diverse training data and continuous post-deployment monitoring to prevent systematic performance disparities [16, 30, 32].

Datenmanagement & Sicherheit: Datenschutz, Verschlüsselung, Verantwortlichkeit; Bias-Minderung

Implementation also involves *change management and capacity building*. User training, feedback loops, and transparent communication about AI functionality help build clinician trust. Beyond informal feedback, structured channels should be established for clinicians and administrative staff to flag unexpected model behaviours, such as omitted content, incoherent phrasing, or workflow disruptions. These observations should feed into scheduled validation and governance reviews, closing the loop between day-to-day use and oversight. In this way, stakeholder communication becomes a core governance mechanism that converts individual experiences into systematic quality control. Several sources emphasise that AI documentation tools should augment rather than replace clinician judgment, maintaining human oversight throughout the documentation process [16, 24, 31].

Change-Management & Schulung: Vertrauen aufbauen, KI unterstützt, ersetzt nicht

Finally, *cost and sustainability* considerations extend beyond licensing fees to include integration costs, staff training, governance, and evaluation. Public–private partnerships and procurement models that account for ongoing monitoring and updating of AI systems were identified as critical for long-term viability [30, 32].

Kosten & Nachhaltigkeit: Integration, Integration, Schulung, Governance, Monitoring

It must be noted that the specific organisational and technical requirements may vary by functionality. For instance, a coding-assistance tool may primarily require integration with billing and coding workflows and safeguards for data accuracy, whereas an AI scribe may demand closer alignment with clinical documentation practices, real-time processing capacity, and stronger oversight of output accuracy [24].

Anforderungen variieren je Funktion: Abrechnungstools vs. KI-Scribes

Once the organisational and infrastructural conditions are addressed, decision-makers must determine the tool's regulatory status and potential clinical impact. Figure 3-3 illustrates a risk-based validation and governance framework, distinguishing between lower-risk, operational applications and those that may influence clinical decision-making. This enables a proportionate approach to evidence generation, oversight, and post-deployment monitoring [16, 31].

risikobasierte Validierung: regulatorischer Status, klinischer Einfluss, proportionierte Überwachung

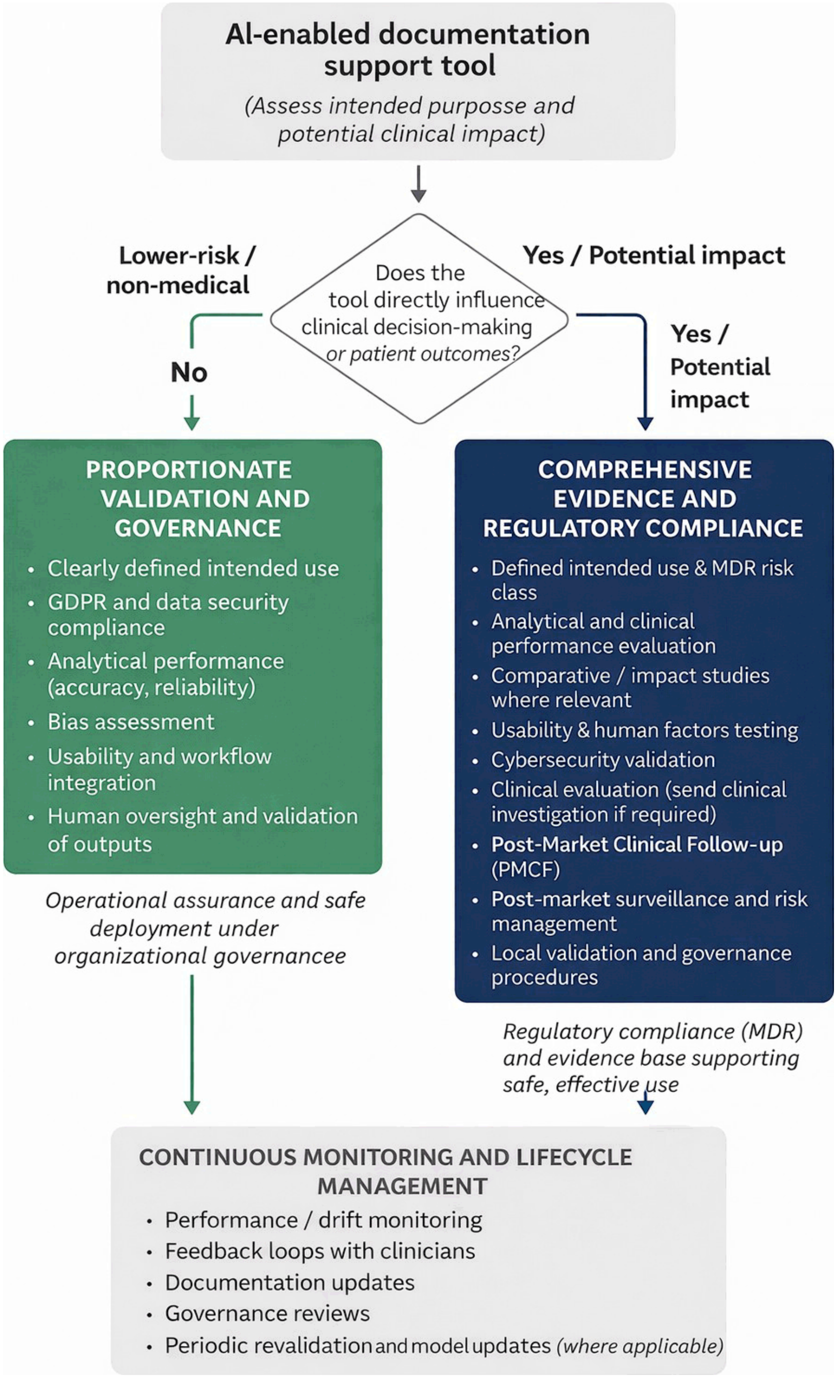
AI-enabled documentation support tool
(Assess intended purposse and potential clinical impact)
Lower-risk / non-medical
Does the tool directly influence clinical decision-making or patient outcomes?
Yes / Potential impact
No
Yes / Potential impact
PROPORTIONATE VALIDATION AND GOVERNANCE
• Clearly defined intended use
• GDPR and data security compliance
• Analytical performance (accuracy, reliability)
• Bias assessment
• Usability and workflow integration
• Human oversight and validation of outputs
Operational assurance and safe deployment under organizational governancee
COMPREHENSIVE EVIDENCE AND REGULATORY COMPLIANCE
• Defined intended use & MDR risk class
• Analytical and clinical performance evaluation
• Comparative / impact studies where relevant
• Usability & human factors testing
• Cybersecurity validation
• Clinical evaluation (send clinical investigation if required)
• Post-Market Clinical Follow-up (PMCF)
• Post-market surveillance and risk management
• Local validation and governance procedures
Regulatory compliance (MDR) and evidence base supporting safe, effective use
CONTINUOUS MONITORING AND LIFECYCLE MANAGEMENT
• Performance / drift monitoring
• Feedback loops with clinicians
• Documentation updates
• Governance reviews
• Periodic revalidation and model updates (where applicable)

Figure 3-3: Decision flow diagram for procurement decisions of AI-supported documentation tools (Source: Review authors' concept; visual generated via ChatGPT from author-provided prompts)

Externe Validierung,

Bias-Checks

As part of external validation, explainability and interpretability checks should be incorporated where technically feasible. Model-agnostic techniques (e.g. feature-attribution or local explanation methods) can help clarify which input sections or variables most strongly influence the generated text and whether the system systematically gives disproportionate weight to certain keywords or document segments. Such analyses can reveal subtle biases and support verification that model behaviour remains stable across validation contexts and aligns with plausible clinical reasoning rather than coincidental language patterns [16].

Revalidierung und definierte Trigger

In operational terms, a proportionate approach benefits from clearly defined validation intervals and triggers for re-evaluation. Even for lower-risk documentation support systems, hospitals should establish routine revalidation cycles – such as quarterly reviews – and specify conditions that require earlier reassessment. Typical triggers include identifiable data drift, changes in clinical workflows, deployment of new model versions, or a predefined drop in key performance metrics (e.g., >5%). Setting such thresholds ensures that validation becomes a continuous and predictable process rather than a one-off approval, fully consistent with real-world, risk-proportionate governance principles [16].

4 Discussion

4.1 Main findings

This scoping review identified six use cases for AI-enabled documentation support in hospitals: (1) AI scribes that generate draft notes from spoken encounters; (2) tools that structure unstructured text by extracting key data into fields or codes; (3) systems that auto-draft clinical documents (e.g., discharge letters, operation notes) from existing digital inputs without speech recognition; (4) patient-friendly summaries; (5) error detection and note-quality assessment that flag inaccuracies, inconsistencies, or omissions; and (6) automated assignment of billing or classification codes. According to the included sources, for clinicians the expected benefits centre on reduced documentation burden and more consistent clinical documentation, leading to increased time available for patient care, quicker access to key information and potentially fewer errors. For organisations, anticipated benefits include productivity improvements and downstream efficiency in quality assurance and revenue cycle processes, contingent on successful integration with EHRs and appropriate training and oversight. The evidence reported in reviews broadly reflects these expectations but is heterogeneous and unevenly documented across use cases.

erwartete Vorteile für 6 KI-Use-Cases: weniger Dokumentationsaufwand, konsistentere Notizen, Zeitgewinn für Patient:innen;

Organisation: Produktivität, Effizienz; Evidenz heterogen

For AI scribes, clinician perceptions of reduced burden and good usability were common, yet quantitative effects on documentation time, quality and productivity varied by dataset and documentation type, with reports of additional after-hours work and substantial editing in some settings. Concerns recurred about accuracy, omissions and fabricated content, medico-legal liability, discomfort with audio recording, data protection, and the risk that perceived efficiency gains could translate into expectations to see more patients. For AI-generated clinical documents without speech recognition, several comparisons favoured LLM-drafted text over typing or dictation – and sometimes even over expert summaries – on conciseness and completeness, but omissions and fabricated elements were also reported, reinforcing the need for clinician review. Structuring unstructured text often showed strong performance on standard metrics, while completeness and clinical context could be limited for certain extraction tasks. Patient-friendly summaries were typically judged clearer and well accepted. Error-detection and note-quality tools generally demonstrated good discriminative performance. Automated coding showed early signals of improved accuracy and completeness, yet the evidence base is small and non-technical impacts, especially on the organisational level (e.g., coding turnaround time, integration and maintenance, cost metrics) are largely unexamined.

KI-Scribes und LLM-Drafting: positive Ergebnisse, aber klinische Prüfung nötig

Strukturierung/ Qualitäts-Tools: technische Leistung meist gut, Kontext/ Vollständigkeit limitiert

Patient:innen-Summaries: Verständlichkeit/ Akzeptanz hoch

Organisational outcomes were mainly reported for AI scribes and included shorter consultations, modest-to-significant productivity improvements, and potential cost savings relative to human scribes. All other use cases present plausible pathways to improve documentation processes, but real-world benefits will depend on implementation quality (integration, training, monitoring, governance) and should be examined in targeted evaluations rather than inferred from early studies.

organisatorische Ergebnisse: v. a. bei Scribes; sonst weitgehend unklar

4.2 Regulatory context and evidence expectations at EU level

Three legislative instruments are principally relevant to the classification and obligations of AI-enabled documentation support tools: the *MDR (Reg. 2017/745) [8]*, the *European Health Data Space (EHDS) (Reg. 2025/327)* [33] and the *EU AI Act (Reg. 2024/1689) [7]*. However, the regulatory field remains an evolving and sometimes ambiguous. While under the MDR software for administrative tasks (without a medical purpose) is typically not a medical device, the boundaries are not always clear-cut. The distinction depends on whether the tool has a potential impact on medical decision-making or patient-relevant outcomes. This creates a borderline for some documentation support functions: simple, verifiable transcription is generally not a medical device, whereas summarisation/structuring that shapes the clinical record used for care may fall within MDR scope. This is reflected in the UK regulatory considerations as well, where National Health Service (NHS) England, however not a regulator, issued guidance [6] on AI-enabled ambient scribing saying that simple, verifiable transcription is generally not a medical device, whereas generative summarisation is likely to be.

regulatorischer Rahmen: MDR, EHDS, AI Act;

Abgrenzung schwierig, einfache Transkription meist nicht, generative Zusammenfassungen potenziell Medizinprodukt

Regardless of the regulatory status of use cases, hospitals currently need to strike a sensitive balance between rapid implementation of AI tools in (routine) administrative and overly cautious restrictive use. In practice this means distinguishing use cases with no direct patient impact (purely administrative) and low-risk cases – for which proportionate evidence (defined intended use, privacy/security compliance, basic performance and bias checks, usability, and post-deployment monitoring) is sufficient – from higher-risk, decision-influencing functions that may constitute medical devices. For the latter, evidence requirements may extend beyond analytical validation to include clinical evaluation under the MDR. Local validation, human-in-the-loop controls, and change-management are needed in both categories, with the depth of evidence scaled to risk and novelty.

Implementierung vs. Risiko: proportionierte Validierung je nach Funktion und Risiko

The EHDS [33] sets common services/specifications for primary and secondary use of health data. For hospitals, this translates into documentation that is interoperable by design (structured, coded, provenance-tracked) and ready for reuse under EHDS governance. Additionally, the EU AI Act [7] introduces governance expectations (e.g., transparency, incident reporting, post-market monitoring and risk management), even where documentation support is treated as purely operational, which must be implemented at national level.

EHDS & AI Act: Interoperable, Transparenz, Monitoring, Risikomanagement erforderlich

In parallel to commercial AI documentation tools, several hospitals and research networks are exploring *open-source or locally trained models* as privacy-preserving alternatives. Such approaches can mitigate concerns about data transfer to third-party cloud systems and allow adaptation to local language, documentation standards, and clinical workflows. Newer architectures such as *retrieval-augmented generation (RAG)*, which combine local data retrieval with language-model reasoning, further strengthen transparency and accuracy by grounding generated text in verified clinical sources. Locally governed models also facilitate auditability, aligning with the EU AI Act's principles of explainability and human oversight. However, maintaining these systems requires dedicated technical capacity, continuous retraining with institution-specific data, and robust quality assurance processes to ensure consistent performance and compliance with the MDR and data-protection regulations.

lokale/open-source KI: datenschutzfreundlich, anpassbar, auditierbar; erfordert technische Kapazität, kontinuierliches Training, QA; Evidenzanforderungen abhängig von Risiko und Funktion

The current regulatory landscape suggests that evidence expectations will diverge depending on whether a tool is judged to shape the clinical record or to affect patient outcomes, with corresponding implications for the type and depth of evidence expected. Assessment frameworks address evidence needs for effectiveness, safety, and other HTA domains in addition to regulatory conformity: under ASSESS-DHT [14, 15], purely operational documentation support typically sits outside "medical purpose" and therefore has no listed evidence requirements, whereas under NICE's Evidence Standards Framework, evidence expectations scale with risk and function – lower for lower-risk/system-impact tools and progressively stronger for higher-risk, decision-shaping functions such as summarisation, coding, or records that inform decisions [6].

Beyond Europe, regulatory authorities are also exploring how to evaluate AI-enabled technologies in real-world settings. The U.S. Food and Drug Administration (FDA) recently issued a *Request for Public Comment on Measuring and Evaluating Artificial Intelligence-Enabled Medical Device Performance in the Real World*, emphasising post-market monitoring, performance drift, and transparency in continuous learning systems [34]. This initiative reflects growing international attention to the methodological and governance challenges of evaluating continuously learning AI systems.

internationale Perspektive: FDA betont Post-Market-Monitoring, Performance Drift, Transparenz bei lernenden KI-Systemen

4.3 Limitations

Limitations of our scoping review

The original research question aimed to assess the clinical and organisational impacts and resource needs for the implementation of AI-enabled DHTs in documentation support. During the early phase of the project, it became evident that the available evidence was insufficient to quantitatively or qualitatively assess impacts across technologies. The approach was therefore adapted to a scoping review with an evidence snapshot, aiming to identify and describe relevant use cases and summarise the existing evidence base rather than to evaluate outcomes or effects.

ursprüngliche RQ: Klinische/organisatorische Auswirkungen unzureichend belegt; Scoping Review durchgeführt

As a mapping exercise, we did not undertake formal risk-of-bias appraisal or meta-analysis and relied solely on reviews, with a small portion of overlapping primary studies, and potential selective reporting. Search limits (databases, dates, languages) and our inclusion criteria may have missed relevant sources. The survey intended to address stakeholder priorities in Austria had a low response rate, consequently no conclusions can be drawn about which functions are currently viewed as most relevant, and broader, more representative engagement is needed to address this gap. Findings should be interpreted as a description of the landscape rather than an assessment of effect.

Mapping ohne Bias-Bewertung; Umfrage schwach; Ergebnisse beschreiben nur Landschaft

Terminology and scope also constrain interpretation. "*Documentation support*" lacks standardised definition: in the literature it is often used synonymously with AI scribes (or ambient scribe), yet underlying functions vary widely – from basic transcription of clinician–patient encounters to systems that also summarize or structure data. Several reviews grouped heterogeneous primary studies and fundamentally different functions, complicating interpretation and synthesis.

uneinheitliche Definition von „Dokumentationssupport" erschwert Interpretation und Synthese

In defining the scope of this review, we focused on AI tools for *documentation support* as an administrative function, i.e., systems that transcribe, summarise, or structure clinical notes. Tools providing diagnostic, prognostic, or treatment recommendations – functions qualifying as medical devices under the MDR with correspondingly stricter approval and risk assessment – were excluded. AI-enabled DHTs assessed here are generally classified as low-risk DHTs, used as adjuncts to clinician oversight and do not directly drive medical decisions.

administrative KI-Dokumentation, Low-Risk, ohne direkte medizinische Entscheidungen

Although our review covered a broad range of documentation support functions, it may not fully reflect emerging or niche developments in this rapidly evolving field.

spiegelt neue oder Nischenentwicklungen möglicherweise nicht vollständig wider

Limitations of the included evidence

The available evidence is dominated by pilots, single-centre evaluations, and small observational studies, limiting generalisability. Outcome measures are heterogeneous: some studies use validated instruments, while others rely on ad hoc or subjective ratings. Technical performance outcomes are frequently underreported, hindering comparisons across use cases. Many studies also apply imprecise or overlapping productivity definitions. These factors warrant caution when interpreting performance outcomes and support ongoing site-level performance monitoring (e.g., routine error/omission audits, checks for data and concept drift, periodic revalidation against local reference standards, and documented feedback loops with clinicians).

Evidenz überwiegend Pilot-/Einzelstudien; heterogene Ergebnisse; begrenzte Vergleichbarkeit; kontinuierliches Monitoring empfohlen

Observed effects on technical efficiency appear highly dependent on outcome choice (e.g., error rate vs hallucinations) and measurement approach. Cost and productivity data are incomplete, often omitting hiring, training, maintenance, and supervision costs. Transferability of findings to different settings is uncertain, and there was little systematic information on variation by medical specialty or care contexts.

Effekte variabel, Kosten- und Produktivitätsdaten lückenhaft, Übertragbarkeit unsicher

Even low-risk AI applications can pose safety risks (e.g., hallucinations, factual inaccuracies), underscoring the need for clearer and more proportionate regulatory guidance and reporting requirements. Such guidance should clarify not only when AI-enabled documentation tools qualify as medical devices under the MDR or AI Act, but also when they do not – particularly for administrative or workflow-support functions. For low-risk or non-medical applications, formal certification or randomized evaluations may not be appropriate or feasible; however, transparent reporting, internal validation, and post-deployment monitoring remain essential to ensure safe use in clinical environments. Recent work proposes structured frameworks for assessing and documenting risks [35]. In addition, several studies [36, 37] stress that hallucinations and factual errors remain a persistent problem, compounded by outdated knowledge in static training datasets. Furthermore, published studies may be skewed toward successful implementations, with less reporting on failed or abandoned deployments.

auch Low-Risk-KI birgt Sicherheitsrisiken;

klare, proportionierte Regulierung, transparente Berichte und Monitoring erforderlich

Deployment considerations were largely absent in the included reviews, whereas they would be important to understand implementation constraints. However, a review identified through hand search [30] provides valuable insights into the practical challenges of implementing AI systems in hospital environments. The authors describe AI deployment as a complex socio-technical process that requires alignment between technology, workflows, and human factors. They highlight barriers such as inadequate interoperability with existing EHR, poor data quality, lack of performance monitoring mechanisms, and

Bereitstellungskriterien selten untersucht; erfolgreiche Implementierung hängt von Workflow-Integration, Infrastruktur, Governance und kontinuierlicher Überwachung ab

limited technical infrastructure in hospitals. Organisational constraints, including insufficient leadership engagement, clinician distrust, and the absence of established governance frameworks, were also identified as critical impediments. Overall, the study underscores that successful deployment depends less on algorithmic performance than on the hospital's capacity to integrate AI tools into routine clinical workflows and ensure continuous evaluation and oversight.

Studien überwiegend US-basiert; Übertragbarkeit nach Ö begrenzt

Finally, most included studies were conducted in the U.S. While clinician- and patient-reported outcomes and technical performance may be less geography-sensitive, organisational outcomes are context-dependent; differences in health-system organisation, financing, and regulation may limit transferability to European – and specifically Austrian – settings.

Evidence gaps and future research

Evidenz heterogen, methodisch limitiert; RCTs nötig für klinikrelevante Fälle; patient:innenzentrierte Ergebnisse, Kosten-Nutzen, Equity und Bias unterforscht

The current evidence is heterogeneous and methodologically limited. For use cases that are non-medical in purpose or purely administrative, technological benchmarking or technical performance evaluations are generally sufficient. By contrast, for use cases that inform clinical decision-making or may affect patient care, rigorous designs (like RCTs, prospective multi-site studies, and long-term evaluations) would be needed. Although a recent RCT [38] on AI scribes has begun to address this gap, robust comparative evidence remains scarce overall, leaving the sustainability of reported benefits uncertain. Patient-centred outcomes such as health status, safety incidents, or satisfaction remain underexplored, with most studies focusing on clinician or documentation endpoints. No cost-effectiveness analyses were identified, creating uncertainty about the economic value compared to alternatives such as human scribes or workflow redesign. Additional gaps include equity, algorithm and data collection biases considerations in training data, impacts on interprofessional workflows, and generalisability across languages and clinical contexts.

MDR-KI: fehlende Integration, geringe Diversität, Bedarf an realen Studien, Stakeholder-Perspektiven wichtig

Other recent reviews (e.g., [37]) on the use cases which are potentially fall under the MDR, also highlight the lack of multimodal integration (e.g., combining text with imaging or laboratory data), insufficient testing in diverse patient populations, and the need for rigorous real-world prospective trials. Alongside these methodological issues, future work should extend beyond literature synthesis to systematically capture perspectives from healthcare professionals and hospital managers regarding the practical utility of AI-enabled documentation tools. Such input would be particularly valuable for identifying priorities for implementation, unmet needs in everyday clinical practice, and potential barriers in hospital workflows.

4.4 Implementation context in Austria

Österreich: ELGA-Integration für KI-Dokumentation erforderlich, CDA/FHIR-konform

In Austria, hospital documentation is primarily managed within local hospital information systems, while selected clinical document types (such as discharge letters, imaging reports, etc.) are transferred to ELGA *(Elektronische Gesundheitsakte)*, the national shared EHR that enables exchange between healthcare providers and patients [39]. Historically, ELGA has used HL7 CDA (Clinical Document Architecture) for document exchange; FHIR (Fast Healthcare Interoperability Resources) is being introduced progressively as a more granular, API-based standard that supports resource-level data sharing. Austria's eHealth Strategy 2024-2030 [10] positions ELGA as the core platform and sets milestones for the expansion of medical documents to be uploaded to the ELGA. From 1 July 2025 ambulatory radiology practices and private laboratories must upload reports (including images) to ELGA; from 1 January 2026 additional datasets (including hospital reports) are phased in [40, 41]. For hospital-facing AI documentation tools, this means AI-generated or assisted notes must be ELGA-compatible (compliance with CDA/FHIR, standardised formats/codes with clear provenance and auditable export) to ensure future readiness as ELGA upload obligations expand [41], otherwise providers will struggle to meet upload obligations.

EHDS & EU AI Act: interoperable, nachverfolgbare Dokumentation erforderlich

As EHDS and EU AI Act obligations phase in [7, 33], hospitals will increasingly need documentation that is interoperable by design and ready for primary and secondary use under EHDS rules [27] – building on ELGA and upcoming upload duties. Austria's roll-out of the EU AI Act is supported by the national AI Service Desk at Rundfunk und Telekom Regulierungs-GmbH (RTR), which provides guidance during implementation [42].

AI Act & ELGA: Integration, Governance, Human-in-Command Pflicht

While many documentation support functions are currently treated as operational, the AI Act's governance (e.g., transparency, incident reporting, post-market monitoring) and Austria's emerging implementation arrangements signal growing expectations around risk management, logging, and oversight. In practical terms for Austrian hospitals, this entails upfront investment in infrastructure, hospital information system (Krankenhaus Information System, KIS) and EHR (ELGA) integration, staff training, and governance [1, 21-23, 26, 27]. AI outputs must be ELGA-compatible (standardised data formats with clear provenance) and support secure exchange. Regardless of use case, hospitals should appoint clinical and patient-safety leads and complete data protection impact assessments (DPIA) under the General Data Protection Regulation (GDPR) for audio capture and automated processing (including clear consent/notice, retention, and roles for controller/processor in vendor contracts). A human-in-command process is needed so clinicians validate outputs before they enter the record, backed by logging, audit trails, incident reporting, and performance/drift monitoring [6, 43].

Schulung, Interoperabilität, Datenschutz, Risikomanagement beachten

Typical barriers include time-intensive training, short-term adoption burden, handling German language/dialect and medical terminology, interoperability with local KIS, privacy and medico-legal concerns [4]. In practice, organisations can address these through phased roll-outs with a small group of expert users who coach colleagues, allocation of temporary cover or protected time for training, early collaboration with ELGA/KIS vendors to ensure interoperability, adoption of standardised clinical data coding (e.g., support in using ICD-10 and mapping local terms), and clear governance with defined responsibilities and audit trails [6, 43]. As EHDS requirements and EU AI Act national arrangements mature, hospitals should expect increasing expectations around risk management, technical documentation, and post-market monitoring, alongside Austria's eHealth Strategy priorities through 2030 [10].

5 Conclusion

Potenzial über Effizienz hinaus: Effekte auf Wohlbefinden und Nutzer:innen-Ergebnisse

Our review indicates that the AI-supported documentation may not only impact administration efficiency but could also impact patient-reported and clinician-reported outcomes, with numerous reviews investigating a potential for a decrease in clinician burnout through reducing administrative burden and potential benefits on clinician well-being and workflow efficiency.

risikobasierter Ansatz, Kontrolle und Überwachung erforderlich

Regulatory and implementation challenges remain substantial. A proportionate, risk-based approach is needed – avoiding both excessive caution that stalls useful innovation and uncritical optimism that risks premature or unsafe deployment. Practical safeguards should include local validation, human-in-the-loop oversight, transparent performance reporting, bias and privacy checks, interoperability with EHRs, and post-deployment monitoring.

klare Ziele, Ressourcen und Evaluationsplanung notwendig

If evidence supports wider adoption, programme-level rollouts will require clear objectives (e.g., reducing documentation burden, improving satisfaction, enhancing quality, or increasing throughput), adequate resourcing, and explicit evaluation plans using predefined metrics across process, experience, financial, and quality domains. Decisions should also account for organisational readiness, barriers and facilitators to integration, and comparison with alternatives (e.g., human scribes or workflow redesign).

Weiterentwicklung methodischer Rahmenwerke für KI-Bewertung

Continued methodological development – such as work within the ASSESS-DHT consortium – will be important to establish robust, transparent, and context-sensitive frameworks for evaluating AI-enabled documentation technologies.

6 References

[1] Bracken A., Reilly C., Feeley A., Sheehan E., Merghani K. and Feeley I. Artificial Intelligence (AI) – Powered Documentation Systems in Healthcare: A Systematic Review. Journal of Medical Systems. 2025;49(1):28. DOI: https://dx.doi.org/10.1007/s10916-025-02157-4.

[2] Canada's Drug Agency (CDA). 2025 Watch List: Artificial Intelligence in Health Care. Canadian Journal of Health Technologies. 2025;5(3).

[3] DeChant P. F., Acs A., Rhee K. B., Boulanger T. S., Snowdon J. L., Tutty M. A., et al. Effect of Organization-Directed Workplace Interventions on Physician Burnout: A Systematic Review. Mayo Clinic Proceedings Innovations, Quality & Outcomes. 2019;3(4):384-408. DOI: https://dx.doi.org/10.1016/j.mayocpiqo.2019.07.006.

[4] Ghatnekar S., Faletsky A. and Nambudiri V. E. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. Health and Technology. 2021;11(4):803-809. DOI: 10.1007/s12553-021-00568-0.

[5] Lee C., Britto S. and Diwan K. Evaluating the Impact of Artificial Intelligence (AI) on Clinical Documentation Efficiency and Accuracy Across Clinical Settings: A Scoping Review. Cureus. 2024;16(11):e73994. DOI: https://dx.doi.org/10.7759/cureus.73994.

[6] NHS England. Guidance on the use of AI-enabled ambient scribing products in health and care settings. 2025 [cited 03.11.2025]. Available from: https://www.england.nhs.uk/long-read/guidance-on-the-use-of-ai-enabled-ambient-scribing-products-in-health-and-care-settings/.

[7] Regulation (EU) 2024/1689 of the European Parliament and of the Council. In: European Parliament and the Council, editor. 2024.

[8] Regulation (EU) 2017/745 of the European Parliament and of the Council. In: European Parliament and the Council, editor. 2017.

[9] Wiesmüller M., Hegny I., Triska M., Banfield-Mumb-Mühlhaim A., Prem E. and Dachs B. AIM AT 2030 Artificial Intelligence Mission Austria 2030. Die Zukunft der Künstlichen Intelligenz in Österreich gestalten In: Bundesministerium für Verkehr I. u. T. B., Bundesministerium für Digitalisierung und Wirtschaftsstandort (BMDW),, editor. Wien2018.

[10] Federal Ministry of Social Affairs H., Care and Consumer Protection (BMSGPK),. eHealth Strategy Austria. v1.0 In: Federal Ministry of Social Affairs H., Care and Consumer Protection, , editor. Vienna2024.

[11] Erdos J. Artificial Intelligence in Health Care: Evaluation of the Clinical and Organizational Impacts of selected AI Applications in Hospitals. Project Plan. 2025 [cited 10.07.2025]. Available from: https://aihta.at/page/kuenstliche-intelligenz-im-gesundheitswesen-bewertung-der-klinischen-und-organisatorischen-effekte-ausgewaehlter-ki-anwendungen-in-krankenhaeusern/en.

[12] Erdos J. Project Protocol Registration: Artificial Intelligence in Health Care: Evaluation of the Clinical and Organizational Impacts of selected AI Applications in Hospitals. 2025 [cited 20.09.2025]. Available from: https://osf.io/eb7hj/.

[13] Riegelnegg M G. D., Goetz G. Artificial Intelligence in Health Care with a Focus on Hospitals: Methodological Considerations for Health Technology Assessment. A scoping review. AIHTA Project Report No.: 142. 2024 [cited 15.07.2025]. Available from: https://eprints.aihta.at/1546/1/HTA-Projektbericht_Nr.164.pdf.

[14] ASSESS-DHT. D4.5 Manual for piloting assessment methods for digital health technologies (interim/piloting version of the full manual D4.6). 2025.

[15] ASSESS-DHT. Taxonomy of DHT, mapping to the assessment framework. Deliverable 2.1: Taxonomy of DHT for assessment purposes. V 1.0. 2025.

[16] ASSESS-DHT. Real-World Data Validation methods for AI-based decision support systems. 2025.

[17] Würflinger D., Tanriverdi E. F. and Degelsegger-Marquez A. Künstliche Intelligenz im intramuralen Bereich Österreichs. In: Gesundheit Österreich GmbH, editor. Wien2025.

[18] EUnetHTA JA2. HTA Core Model Version 3.0 for the full assessment of Diagnostic Technologies, Medical and Surgical Interventions, Pharmaceuticals and Screening Technologies. 2016.

[19] ASSESS-DHT. Glossary of Terms for AI Validation in Healthcare. 2025.

[20] Tricco A. C., Lillie E., Zarin W., O'Brien K. K., Colquhoun H., Levac D., et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Annals of Internal Medicine. 2018;169(7):467-473. DOI: 10.7326/M18-0850.

[21] Peterson Health Technology Institute (PHTI). Adoption of Artificial Intelligence in Healthcare Delivery Systems: Early Applications and Impacts. 2025 [cited 20.08.2025]. Available from: https://phti.org/wp-content/uploads/sites/3/2025/03/PHTI-Adoption-of-AI-in-Healthcare-Delivery-Systems-Early-Applications-Impacts.pdf.

[22] Hassan H., Zipursky A. R., Rabbani N., You J. G., Tse G., Orenstein E., et al. Special Topic on Burnout: Clinical Implementation of Artificial Intelligence Scribes in Healthcare: A Systematic Review. Applied Clinical Informatics. 2025;30:30. DOI: https://dx.doi.org/10.1055/a-2597-2017.

[23] Sasseville M., Yousefi F., Ouellet S., Naye F., Stefan T., Carnovale V., et al. The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. Healthcare. 2025;13(12):16. DOI: https://dx.doi.org/10.3390/healthcare13121447.

[24] Bongurala A. R., Save D., Virmani A. and Kashyap R. Transforming Health Care With Artificial Intelligence: Redefining Medical Documentation. Mayo Clin Proc Digit Health. 2024;2(3):342-347. Epub 20240522. DOI: 10.1016/j.mcpdig.2024.05.006.

[25] Perkins S. W., Muste J. C., Alam T. and Singh R. P. Improving Clinical Documentation with Artificial Intelligence: A Systematic Review. Perspectives in Health Information Management. 2024;21(2):1d.

[26] Lumbiganon S., Abou Chawareb E., Moukhtar Hammad M. A., Azad B., Shah D. and Yafi F. A. Artificial Intelligence as a Tool for Creating Patient Visit Summary: A Scoping Review and Guide to Implementation in an Erectile Dysfunction Clinic. Current Urology Reports. 2025;26(1). DOI: 10.1007/s11934-024-01237-1.

[27] Vrdoljak J., Boban Z., Vilovic M., Kumric M. and Bozic J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare. 2025;13(6):10. DOI: https://dx.doi.org/10.3390/healthcare13060603.

[28] KI-Funktionen von informed.360 [cited 30.09.2025]. Available from: https://www.infomed.at/ki.

[29] Verband Österreichischer Software Innovationen (VÖSI). KI-Landkarte. 2025 [cited 30.09.2025]. Available from: https://voesi.or.at/voesi-aktiv/ki-landkarte/.

[30] Kamel Rahimi A., Pienaar O., Ghadimi M., Canfell O. J., Pole J. D., Shrapnel S., et al. Implementing AI in Hospitals to Achieve a Learning Health System: Systematic Review of Current Enablers and Barriers. Journal of Medical Internet Research. 2024;26:e49655. Epub 20240802. DOI: 10.2196/49655.

[31] Kelly C. J., Karthikesalingam A., Suleyman M., Corrado G. and King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. Epub 20191029. DOI: 10.1186/s12916-019-1426-2.

[32] Nasef D., Nasef D., Sawiris V., Weinstein B., Garcia J. and Toma M. Integrating artificial intelligence in clinical practice, hospital management, and health policy: literature review. Journal of Hospital Management and Health Policy. 2025;9.

[33] Regulation (EU) 2025/327 of the European Parliament and of the Council. In: European Parliament and the Council, editor. 2025.

[34] Food and Drug Administration (FDA). Request For Public Comment: Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real-World. 2025 [cited 30.09.2025]. Available from: https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device.

[35] Bednarczyk L., Reichenpfader D., Gaudet-Blavignac C., Ette A. K., Zaghir J., Zheng Y., et al. Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review. Journal of Medical Internet Research. 2025;27:e68998. DOI: https://dx.doi.org/10.2196/68998.

[36] Preiksaitis C., Ashenburg N., Bunney G., Chu A., Kabeer R., Riley F., et al. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. JMIR Medical Informatics. 2024;12:e53787. DOI: https://dx.doi.org/10.2196/53787.

[37] Yang Z., Wang D., Zhou F., Song D., Zhang Y., Jiang J., et al. Understanding natural language: Potential application of large language models to ophthalmology. Asia-Pacific Journal of Ophthalmology. 2024;13(4):100085. DOI: https://dx.doi.org/10.1016/j.apjo.2024.100085.

[38] Lukac P. J., Turner W., Vangala S., Chin A. T., Khalili J., Shih Y. T., et al. A Randomized-Clinical Trial of Two Ambient Artificial Intelligence Scribes: Measuring Documentation Efficiency and Physician Burnout. medRxiv. 2025. Epub 20250711. DOI: 10.1101/2025.07.10.25331333.

[39] Federal Ministry of Social Affairs H., Care and Consumer Protection (BMSGPK),. ELGA. 2025 [cited 02.09.2025]. Available from: https://www.sozialministerium.gv.at/Themen/Gesundheit/eHealth/ELGA.html.

[40] Federal Ministry of Social Affairs H., Care and Consumer Protection (BMSGPK),. Radiologie- und Laborbefunde ab Juli in der ELGA verfügbar. 2025 [cited 02.09.2025]. Available from: https://www.sozialministerium.gv.at/Services/Aktuelles/Archiv-2025/ELGA-befunde.html.

[41] Federal Ministry of Social Affairs H., Care and Consumer Protection (BMSGPK). Speicherverpflichtungen in ELGA. 2025 [cited 02.09.2025]. Available from: https://www.arztnoe.at/fileadmin/Data/Documents/pdfs/e-Services/ELGA/Brosch%C3%BCre_Speicherverpflichtung__Stand_07.07.25.pdf.

[42] Service Desk for Artificial Intelligence. 2025 [cited 03.09.2025]. Available from: https://www.rtr.at/rtr/service/ki-servicestelle/KI-Servicestelle.en.html.

[43] NHS Wales. The safe and responsible adoption of ambient voice technologies (AI scribes) in clinical settings (WHC/2025/026). 2025 [cited 30.08.2025]. Available from: https://www.gov.wales/sites/default/files/pdf-versions/2025/8/2/1754404899/safe-and-responsible-adoption-ambient-voice-technologies-ai-scribes-clinical-settings-whc2025026-0.pdf.

[44] European Commission. Digital health and care. 2025 [cited 30.11.2025]. Available from: https://health.ec.europa.eu/ehealth-digital-health-and-care/digital-health-and-care_en.

[45] Ting K. M. Precision and Recall. In: Sammut C. and Webb G. I., editors. Encyclopedia of Machine Learning. Boston, MA.: Springer; 2011.

Appendix

Appendix A: Glossary of terms

Digital Health Technology (DHT): There is no single, uniform definition of "digital health technology." According to the European Commission, DHTs are tools and services that use information and communication technologies to improve prevention, diagnosis, treatment, monitoring and management of health [44].

AI-enabled DHT: a DHT that incorporates artificial intelligence. When such a technology has an intended medical purpose, it may fall under the EU Medical Device Regulation (*Regulation 2017/745*); AI used as (SaMD) or in medical devices (SiMD) is generally treated as high-risk under the EU AI Act (*Regulation 2024/1689*), triggering requirements such as risk management, data governance, transparency, and human oversight.

Accuracy: Proportion of correct predictions. Accuracy measures the percentage of all predictions (both positive and negative) that are correct. In imbalanced data sets, it can be misleading if it is a ruling class. Example: In a model that predicts whether a patient has a chronic disease, if the model correctly predicts 950 out of 1,000 patients, the accuracy would be 95%. However, if only 50 patients actually have the disease, this metric would not correctly reflect the model's performance in detecting the disease [19].

AUC (Area Under the Curve): A measure of the performance of a classification model. It represents the area under a curve that plots the trade-off between true positive and false positive rates. A larger area indicates better performance in distinguishing between positive and negative classes. A value of 0.5 indicates random performance, while a value of 1 indicates perfect discrimination. Example: For a model that predicts the probability of relapse in cancer patients, a high value indicates that the model is very effective at distinguishing between patients who will relapse and those who will not [19].

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): A performance metric for classification models that measures their ability to distinguish between classes. The ROC curve plots the true positive rate (recall) against the false positive rate (1 – specificity) at various classification thresholds. The AUC-ROC summarizes the model's overall discriminative ability. It is widely used for evaluating classification models, although it may be less informative than the AUC-PR in highly imbalanced datasets. Example: In a model predicting heart attacks, the AUC-ROC shows how well the model distinguishes between patients who will and will not experience a heart attack, balancing sensitivity (recall) and the rate of false alarms [19].

F1 Score: A performance metric for classification models, especially useful when dealing with imbalanced classes. It is the harmonic mean of precision and recall, providing a single score that balances both false positives and false negatives. Precision measures the proportion of true positives among predicted positives, while recall measures the proportion of true positives among actual positives. Example: In a model designed to detect a rare disease, overall accuracy may be misleading due to the low prevalence of positive cases. A high F1 Score would indicate that the model is effectively identifying true cases while limiting false alarms [19].

Recall (True Positive Rate, TPR): Also known as sensitivity, this is a performance metric used in classification models that measures the proportion of true positive cases correctly identified out of all actual positive cases. Recall reflects the model's ability to detect positive instances and is especially important in contexts where missing positive cases has serious consequences. Example: In a breast cancer detection model, if there are 100 actual cancer cases and the model correctly identifies 90 of them, the recall is 90%, indicating the model effectively captures most cancer cases [19].

Specificity (True Negative Rate): A performance metric used in classification models that measures the proportion of true negative cases correctly identified out of all actual negative cases. Specificity reflects the model's ability to correctly identify individuals who do not have a particular condition. High specificity means the model accurately excludes those without the condition, minimizing false positives. Example: In a disease screening model, high specificity indicates that healthy individuals are rarely misclassified as having the disease [19].

Precision (Positive Predictive Value, PPV): is the proportion of retrieved instances (or predicted positives) that are actually relevant (or correctly labelled). In formula form: Precision = True Positives/ (True Positives + False Positives). It shows how many of the items identified by the system as positive are indeed correct [45].

Word Error Rate (WER) is a standard metric for evaluating the accuracy of speech recognition systems. It measures how many errors a system makes when converting speech into text, compared with a reference (the "correct" transcript).The formula is: WER=$(S+D+I\ )/N$, where S = substitutions (wrong word instead of correct one), D = deletions (missed word), I = insertions (extra word added), N = total number of words in the reference. The result is expressed as a proportion of errors per word, 0.0 = perfect recognition (no errors) [19].

Appendix B: Extraction tables

Table A-1: Characteristics of included reviews

Author, year	Perkins, 2024 [25]	Bracken, 2025 [1]	Sasseville, 2025 [23]	Hassan, 2025 [22]	Lee, 2024 [5]	Lumbiganon, 2025 [26]	Vrdoljak, 2024 [27]
Publication type/ study design	**Systematic (narrative) review**	**Systematic review**	**Systematic review**	**Systematic review**	**Scoping review**	**Scoping review**	**Scoping review**
Study aim(s)	To summarize existing research and explain how AI tools could be used broadly to improve documentation efficiency.	To evaluate the efficiency, quality, and stakeholder opinion regarding the use of AI-driven documentation systems (generative and ambient AI) to inform policymakers on the viability of adopting AI-driven documentation solutions in clinical practice.	To evaluate AI tools designed to streamline clinical documentation for healthcare providers across all clinical settings.	To propose an evaluation framework for future AI scribe implementations. To evaluate the impact of AI scribes on clinicians, patients, and organizations. To identify knowledge gaps in the AI scribe implementation literature.	To explore the impact of natural language processing (NLP), machine learning (ML), and speech recognition (SR) on the accuracy and efficiency of clinical documentation across various clinical settings, including hospital wards, emergency departments, and outpatient clinics.	To examine the evidence on AI-generated patient summary and evaluated their implementation in ED clinics.	To examine how large language models (LLMs) are currently applied in medical education, clinical decision support and knowledge retrieval, and healthcare administration: (1) summarize the breadth of LLM-based tools and their efficacy, (2) highlight challenges related to reliability, bias, and safety, and (3) discuss emerging techniques that might mitigate these limitations.
Eligibility criteria	New AI tool or a new way of using an existing AI tool specifically for improving clinical documentation.	AI technology for clinical documentation generation by healthcare professionals in healthcare settings, assessing outcomes such as documentation quality, efficiency, and stakeholder opinion.	AI-based interventions such as real-time transcription, automated EHR data entry, NLP-based clinical summarization, and tools that transform spoken interactions into organized clinical notes. Study design: all interventional study designs (RCTs, quasi-experimental designs, prospective cohorts, pre-post studies, observational studies, and mixed-methods studies).	Ambient AI scribes using NLP and automatic speech recognition in healthcare settings, English-language studies. Excluded: review articles, simulation studies, pre-implementation opinions, non-AI scribes, direct transcription tools, and LLMs without ambient listening.	Application and impact of AI technologies in clinical documentation: NLP, ML, SR, and other AI technologies used in various clinical settings, including inpatient units, emergency departments, and outpatient clinics. Study type: empirical research articles (quantitative, qualitative, or mixed methods), case studies, evaluations, experience reports, observational studies, systematic reviews, scoping reviews, meta-analyses, and conference papers. English language studies or with an English translation published within the last five years.	AI in creating visit summary (studies involving real clinical usage or at least evaluation with mock patient data). Excluded: studies lacking clinical application or not written in English.	LLMs in medical or healthcare settings, LLM-based interventions or workflows in education, clinical decision-making, or administration, English-language studies.

Author, year	Perkins, 2024 [25]	Bracken, 2025 [1]	Sasseville, 2025 [23]	Hassan, 2025 [22]	Lee, 2024 [5]	Lumbiganon, 2025 [26]	Vrdoljak, 2024 [27]
Number of included studies	129	11	8	11	36[1]	11	5 in the healthcare administration section of the review
Included studies (author, year)	Please see the list in the original publication. Duplicates were checked and 4 identified: Clough 2023, Lindvall 2022, Marshall 2023, Zhang 2021.	Barak-Corren 2024, Baker 2024, Balloch 2024, Clough 2023, Dos Santos 2024, Dubinski 2024, Galloway 2024, Kernberg 2024, Owens 2024, Robinson 2023, Tung 2024.	Haberle 2024, Hudelson 2024, Islam 2024, Kernberg 2024, Nguyen 2023, Sezgin 2024, Van Buchem 2024, Wang 2021.	Albrecht 2024, Bundy 2024, Cao 2023, Galloway 2024, Haberle 2024, Liu 2024, Misurac 2024, Nguyen 2023, Owens 2024, Shah 2025, Tierney 2024.	Ahuja 2019, Ando 2022, Baughman 2024, Chen 2020, Clough 2023, Duffourc 2023, Florig 2021, Gaffney 2022, Giorgi 2023, Goss 2019, Kim 2024, Kernberg 2024, Krishna 2021, Lin 2020, Lindvall 2022, Liu 2023, Liu 2024, Marshall 2023, Meng 2024, Nayak 2023, Patel 2023, Preiksaitis 2023, Roberts 2024, Sushil 2024, Tang 2023, Tierney 2024, Tran 2020, Van Veen 2024, Waisberg 2023, Warner 2024, Williams 2024, Zaretsky 2024, Zhang 2021, Zernikow 2023.	Bala 2020, Barack-Corren 2024, Cho 2022, Clough 2023, Ganoe 2021, Hyun 2003, Kim 2022, Krishna 2005, Wang 2021, Wang 2022, Young 2023.	Huang 2024, Liu 2024, Van Veen 2024, Wei 2024, Zaretsky 2024.
Year of publication of included studies	2005-2024 (33% of the studies published after 2020)	2023-2024	2021-2024	2021-2024	2019-2024	2003-2024 (80% of the studies published after 2020)	2024
Medical specialty in the included studies	General	General	General	Primary care, Internal medicine, Surgical medicine, Psychiatry, Dermatology, Emergency medicine.	Multispecialty/not specified (radiology, internal medicine, hospital medicine, etc.) in majority of the studies. General medicine, Pediatrics.	Emergency medicine, general, spine surgery, Pediatrics, Dermatology.	Pathology, general and multispeciality (radiology, internal medicine, hospital medicine, etc.)
Setting of the included studies	Not reported.	Hospitals (wards, clinic, emergency department (ED), operating theatre), Primary care.	University center and medical college hospitals (multiple departments), Specialized healthcare services (tertiary care and above: National Cancer Institute-designated Comprehensive Cancer Center, Nationwide Children's Hospital Physician Consult and Transfer Center).	Primarily outpatient/ ambulatory clinics.	Hospital wards, Emergency department, Outpatient clinics.	Hospital (tertiary care), Emergency department, Specialty clinics (dermatology, orthopedics, HIV clinic), Primary care.	Hospitals (n=1), Not reported in the other studies.

Author, year	Perkins, 2024 [25]	Bracken, 2025 [1]	Sasseville, 2025 [23]	Hassan, 2025 [22]	Lee, 2024 [5]	Lumbiganon, 2025 [26]	Vrdoljak, 2024 [27]
Country of the included studies	Not reported.	Not reported.	U.S. (n=6) NL (n=1), Bangladesh (n=1).	U.S. (n=11)	U.S. (n=~20) Asia (China, Japan, Korea) (n=~5) Europe (Germany, UK, NL) (n=~5) International dataset (n=~5)	U.S (n=6), UK (n=1), South-Korea (n=2), Israel (n=1), China (n=1)	China (n=2), U.S. (n=3)
AI function/use[2]	Structuring free-text data (88 studies) Increasing patient understanding (2 studies) Speech recognition and error detection (4 studies) Integrative documentation assistant[3] (6 studies) Assessing clinical note quality (23 studies) Identifying documentation trends[4] (6 studies)	Generation of all types of clinical documentation such as progress notes, discharge summaries, handover documents, clinic letters, operation notes[5]	Ambient AI scribe[6]	Ambient AI scribe	Review authors included both performance-focused empirical studies and context-only studies (commentaries, surveys, dataset papers, or scoping reviews). The context-related studies are not listed here. Ambient AI scribe (4 studies) Text summarization (discharge summaries) (17 studies) Transforming medical text to patient-friendly language summaries (9 studies)[7]	Ambient AI scribe (AI assisted notes vs physician-generated notes) (5 studies)[8] Clinical note creation with NLP (3 studies)[9] Assigning billing codes (2 studies) AI-assisted clinical documentation platform for note-taking (1 study)	Extracting structured data from clinical notes (from free-text notes) Text summarization Medical dialogue (between patient and doctor) summarization Transforming medical text to patient-friendly language summaries Medical education, clinical decision support and knowledge retrieval[10]

AI methodology and model type	Natural Language Processing (NLP) incl. Large Language Models (LLM);
	Machine Learning (ML);
	Deep Learning
		NLP incl. LLM;
		ML;
		SR
			NLP incl. LLM;
			ML;
			SR
				NLP incl. LLM;
				ML;
				SR
					NLP incl. LLM;
					ML;
					SR;
					Deep Learning
						NLP incl. LLM;
						ML;
						SR;
						Deep Learning
							NLP incl. LLM;
							ML,
							Deep Learning

Specific tools named in the included studies	No specific tool named.	**Voice-based ambient AI medical scribe** (incl. SR, NLP, ML) (n=2): Dragon Ambient eXperience (Nuance), Tortus **LLM:** Chat GPT (n=9)	**Voice-based ambient AI medical scribe:** Dragon Ambient eXperience (Nuance) **Clinical NLP:** Autoscriber **Pre-trained LLM:** T5-small, T5-base, PEGASUS-PubMed, and BART-Large-CNN	**Voice-based ambient AI medical scribe:** Dragon Ambient eXperience (Nuance) (n=7), Abridge (n=2), Nabla Copilot (n=1)	**Voice-based ambient AI medical scribe:** Dragon Medical 10.1 and Dragon Medical 360 (Nuance) integrated with Epic EHR (n=2), IBM Watson (n=1) **LLM:** GPT-4, GPT-3.5, FLAN-T5, FLAN-UL2, Llama-2, Vicuna, Alpaca, Med-Alpaca (n=9), BERT-based/transformer variants (n=4)	**Voice-based ambient AI medical scribe:** Nabla Copilot Sunoh.ai Amazon Web Services HealthScribe Dragon Medical One **NLP:** ChatGPT (n=3) **Note-taking:** PhenoPad (n=1)	**LLM:** ChatGPT (n=4), other LLMs, e.g., FLAN-T5, FLAN-UL2, Llama-2, Vicuna, Alpaca, Med-Alpaca (n=1)
Analysed outcomes	Documentation efficiency/reduction of documentation burden/ time savings Documentation quality	Documentation efficiency/time-savings Documentation quality Impact on HCP workflow	Clinician outcomes Healthcare system efficiency metrics Documentation outcomes Patient outcomes	Clinician outcomes (efficiency, wellness/ burnout, experience) AI scribe performance Patient experience	Efficiency and user experience, Accuracy and error management, Clinical utility and safety, Patient-centered care,	Accuracy, Documentation time, Burden, User experience,	Accuracy: automatic metrics and human expert evaluation (readability, clinical relevance, completeness, correctness, conciseness),

Author, year	**Perkins, 2024 [25]**	**Bracken, 2025 [1]**	**Sasseville, 2025 [23]**	**Hassan, 2025 [22]**	**Lee, 2024 [5]**	**Lumbiganon, 2025 [26]**	**Vrdoljak, 2024 [27]**
Analysed outcomes *(continuation)*	Opportunity to reduce costs Reimbursement improvement Error rates in AI-generated notes Quality of care improvement Clinician satisfaction and burnout Provider disengagement	Presence of "hallucinations" or fictitious information Stakeholder opinion/usability		Business/healthcare system efficiency Documentation outcomes Equity considerations	Liability and ethical considerations.	GP acceptance of hospital discharge summaries, Completion rate, Documentation quality, Challenges and opportunities.	Patient safety.
Study design of the included studies	Primary studies	All types of peer-reviewed primary studies (quantitative, qualitative, and mixed methods)	Peer-matched controlled cohort study (n=1) Mixed methods pilot study (n=2) AI system development process with post-test questionnaire (n=1) Comparative study (n=1) Usability study (n=2) Quantitative descriptive (n=1)	Observational (n=8), Non-randomized controlled trial (n=3)	Not reported for each included study. The following categories were identified by the present review authors: Cross-sectional survey (n=3) Quantitative model development/evaluation (n=8) Experimental/comparative (n=8) Systematic/scoping reviews (n=2) Experience reports (n=2) Editorials (n=5)	Not reported for each included study. The following categories were identified by the present review authors: Retrospective study (n=5) Prospective study (n=5) Quasi-experimental study (n=1)	Not reported for each included study. The following categories were identified by the present review authors: Retrospective study (n=3) Prospective study (n=1) Benchmarking technical evaluation (n=1)
Assessment of the quality of the included studies by review authors	Not assessed formally. Limitations listed: Relevance of studies determined by the authors. Efficacy was not objectively compared. AI tools/algorithms not published in peer-reviewed journals could not be included.	Mixed Methods Appraisal Tool (MMAT) was used to assess the quality of the included studies: High quality in 7 studies (80-100% of criteria met). 4 studies met 60% of the criteria due to concerns regarding the appropriateness of the chosen sample population or sampling strategy.	MMAT was used to assess the quality of included studies: varying methodological quality, 1 study 5/5, 2 studies 4/5, 4 studies 3/5 criteria met. Common limitations: Incomplete or biased data, Small or un-representative samples, Poor integration of mixed methods, Limited generalizability.	Newcastle-Ottawa Scale was used to assess the quality of included studies. Scores ranged from 4/8 to 9/9. High quality: 4 studies Moderate quality: 5 studies Low quality: 3 studies.	Not assessed formally. Limitations listed: Heterogeneous study types (experience reports, cross-sectional studies), lack of longitudinal studies. Lack of long-term data. Not all AI technology currently in use or emerging are covered. Cultural and geographical bias (only English-language articles and developed countries covered). Articles included only from the last 5 years.	Not assessed formally. Limitations listed: Small, heterogeneous studies, often limited to single specialties, prototypes or simulations rather than real-world implementations, no RCTs. Outcomes varied widely across studies. Already outdated AI uses included given the rapid evolution of AI. Overall, evidence on effectiveness, safety, and integration into clinical workflows remains limited.	Not assessed formally. Limitations listed: Small, heterogeneous studies (settings, tasks, outcomes, and evaluation metrics varied widely). Inconsistent reporting in the studies (insufficient detail on methods, datasets, or evaluation criteria). Possible study omissions due to fast-changing evidence.

Author, year	Perkins, 2024 [25]	Bracken, 2025 [1]	Sasseville, 2025 [23]	Hassan, 2025 [22]	Lee, 2024 [5]	Lumbiganon, 2025 [26]	Vrdoljak, 2024 [27]
Conclusion/ recommendation of the review authors	While current AI tools offer targeted improvements to clinical documentation processes, moderately high error rates preclude the broad use of a comprehensive AI documentation assistant. While LLMs have the potential to greatly reduce error rates, many of these models are proprietary and not well-studied in peer-reviewed literature. In the future, this hurdle may be overcome with further rigorous tool evaluation and development in direct consultation with physicians, as well as robust discussion of the legal and ethical ramifications of AI clinical decision support tools.	AI technologies like Chat GPT and ambient AI show promise in enhancing the efficiency and quality of clinical documentation, significant challenges remain. The variability in documentation quality undermines efficiency gains. Continued research and development are needed to refine AI tools, improve their reliability, and ensure that they can consistently meet the high standards required in clinical documentation. Careful consideration of the benefits and limitations will be crucial for a successful integration into clinical practice.	AI scribes can reduce documentation time and clinician burden, especially with tailored workflows and training. Documentation quality and efficiency improved most consistently; effects on patient outcomes and system efficiency were mixed. Evidence gaps remain regarding patient perspectives, data privacy, bias, workforce impacts, and long-term outcomes, underscoring the need for robust, real-world evaluation and careful implementation planning.	AI scribes represent a promising tool for improving clinical efficiency and alleviating documentation burden. This systematic review highlights the potential benefits of AI scribes, including reduced documentation time and enhanced clinician satisfaction, while also identifying critical challenges such as variable adoption, and evaluation gaps.	AI has potential to ease documentation burden through summarization, discharge notes, and ambient scribing, with early evidence of improved efficiency and patient communication. Outputs are often accurate and readable, and plain-language summaries may support health literacy, but risks such as hallucinations and missing details remain. Most studies are small or pilot-level, with limited specialty coverage and little large-scale, real-world evidence, leaving important gaps on safety, accuracy, and long-term impact.	AI has shown promising results in creating accurate and efficient patient visit summary. Supervision by clinicians remains crucial to address medico-legal concerns and ensure patient safety.	LLMs show potential for generating accurate, coherent, and readable administrative documents and may streamline documentation workflows and reduce time burden for clinicians. However, evidence is still early, small-scale, and heterogeneous, with reliance on simulated or retrospective data. Risks include incomplete capture of clinical details, hallucinations, and lack of standard evaluation metrics.

Abbreviations: AI … artificial intelligence; LLM … large language model; MMAT … Mixed Methods Appraisal Tool.

Notes:

[1] *The review authors reported and included 36 studies in the PRISMA tree, however, in the absence of a comprehensive data extraction table, we could identify only 34 studies from the results section of the review.*

[2] *As reported/categorized in the review.*

[3] *According to our definition, this covers ambient AI scribe*

[4] *This review had a broader scope, including topics which are outside the scope of our review, i.e. identifying documentation trends. Hence, results related to this AI function are not extracted in the results table.*

[5] *According to our definition and upon examining the individual studies, the categories of ambient AI scribe and text summarization and medical text generation using ChatGPT.*

[6] *The included studies span across AI scribes, which transcribe and summarize speech real-time and those which convert audio recordings into text. Some of them uses NLP to correct the summaries and also can be used to give command to create e-prescriptions.*

[7] *The review did not categorise the included studies according to AI function but provided an overview of the study findings. We extracted the AI function from Table 3 summary of key findings.*

[8] *Two of these studies fall under the category of "Transforming medical text to patient-friendly language summaries (translating/explaining text in plain language)" and three used AI for text summarization after patient visits.*

[9] *Medication extraction from EHR or visit transcript (i.e. annotation task) and creating structured notes from unstructured data.*

[10] *This review had a broader scope, including topics which are outside the scope of our review: medical education, clinical decision support and knowledge retrieval.*

Table A-2: Use case AI scribes (part 1)

Perkins, 2024 [25]	Bracken, 2025 [1]	Sasseville, 2025 [23]	Hassan,2025 [22]
Clinician-reported outcomes			
Outcomes: Clinician satisfaction, provider (dis)engagement, burnout, documentation time **Evidence**: 4 studies **Findings:** *Satisfaction*: average clinician satisfaction increased (1 study), *Provider disengagement* decreased, but *burnout* score did not decrease (1 study).	**Outcomes:** clinician experience, burnout, concerns, documentation time **Evidence:** 3 studies **Findings:** *Clinician experience:* opinion was generally positive, with users highlighting ease of use and reduced task load as primary benefits (2 studies). *Burnout:* increased use of ambient AI was associated with improved burnout scores (1 study). *Concerns:* reliability and validity of AI-generated documentation, inaccuracies, and potential loss of narrative (23 studies). *Documentation time:* average documentation time per encounter was reduced by 28.8% (1 study).	**Outcome:** clinician engagement, documentation burden, burnout, clinician experience, documentation time **Evidence**: 6 studies **Findings:** *Clinician engagement*: AI scribe users vs. non-users indicated a score of 3.62 vs. 3.37 on a scale of 5 (1 study). *Documentation burden*: decreased for some clinicians (1 study). *Clinician experience:* ■ Mixed feedback: ease-of-use but concerns over training and quality (2 studies). ■ Feasibility 16.0, acceptability 16.3, usability 68.6 scores on a REDCap survey (1 study). *Burnout*: no significant change (p = 0.081), but perceived documentation time improved (p = 0.005) (1 study). *Documentation time*: mostly time reductions ■ Per patient decreases from 5.3 minutes to 4.54 minutes for AI scribe users (1 study). ■ AI scribes 2.7 times faster than typing and 2x faster than dictation for history sections, and 3x faster for physical exams (1 study). ■ Efficiency improved, with a median time for manual summarization at 202 s vs. editing automatic summaries at 186 s (1 study). ■ After-hours EHR work increased for AI scribe users by 4.69%, vs. a decrease of 0.945% for the control group (1 study).	**Outcomes**: documentation time, clinician wellness and burnout, clinician experience. **Evidence**: 10 studies **Findings:** *Documentation time:* ■ Improvement in at least 1 efficiency metric (9 studies). ■ Total time spent in EHR: no change (2 studies) to significant decreases (from 90.1 to 70.3 minutes/day, p < 0.001) (2 studies). ■ Time outside working hours: reductions in EHR time outside typical hours (from 35.5 to 28.1 min/day, p = 0.005) (3 studies). ■ Time to write notes: decreased time per note or per appointment, with reductions ranging from 7% to 29% or 1.4 to 2.2 minutes per note (5 studies) ■ Provider contribution to note content: decreased from 97% to 52% (p < 0.001) (2 studies). *Clinician wellness and burnout:* positive effect of AI scribes (7 studies) and a mixed positive and neutral effect (3 studies). *Clinician experience:* mixed results: ■ Favourable improvements (9 studies) and favourable clinician perceptions of patients' experience with AI scribes (3 studies). ■ Both positive and negative elements of scribe use (3 studies).
Organisational outcomes			
Not reported.	**Outcome:** *Consultation time* **Evidence**: 1 study **Findings:** *Consultation time:* consultations using AI were significantly shorter equalling to 26.3% time saving.	**Outcomes:** Productivity (work relative value unit/wRVU), panel size in value-based care (VBC) **Evidence**: 1 study **Findings:** *Productivity*: Statistically significant but modest increase in wRVU productivity among AI scribe users (94.2% vs. 90.6%). *Panel size (number of patients) assigned VBC:* the number for VBC providers did not significantly change.	**Outcomes:** Business efficiency (wRVU), costs, patient flow **Evidence**: 6 studies **Findings:** *wRVUs/Revenue per visit:* Mixed results: from no change in wRVUs or gross revenue per visit (1 study) to a significant increase in annualized wRVUs (from 90.6% to 94.2%, p < 0.001) (1 study). *Cost efficiency:* Estimated cost savings of $13,400-$14,400 per user/year compared to in-person scribes (1 study). *Patient volume/productivity:* Mixed findings: ■ 48-58% of providers reported perceived ability to see more patients (5 studies). ■ Objective data showed no increase in monthly patient visits or panel size in most studies (3 studies). ■ Some providers expressed concern that AI scribes might increase patient load expectations (1 study).

Perkins, 2024 [25]	Bracken, 2025 [1]	Sasseville, 2025 [23]	Hassan,2025 [22]
Technical performance and documentation quality			
Outcomes: accuracy, WER, recall **Evidence**: 5 studies **Models**: automatic speech recognition, deep learning **Findings**: ■ *Word Error Rate (WER):* 0.5318.4-22.5% in 1 study ■ *Note accuracy*: 96% (manual validation) – 97% (2 studies) ■ Quality: decreased slightly in 1 study (measurement instrument NR), 62% of notes met standard in another study	**Outcomes:** accuracy/hallucinations, quality **Evidence**: 1 study **Findings**: ■ Ambient AI improved quality greater than 2x compared to traditional EHR use in outpatient letters. ■ No hallucinations.	**Outcomes:** accuracy, quality, deficiency **Evidence**: 5 studies **Findings:** *Accuracy:* ■ ChatGPT-4 showed substantial variability in errors, accuracy, and note quality (1 study). ■ System-generated outputs showed similarity rates to manually created notes of 87.5% for scribes and 96.2% for prescriptions (1 study). ■ Pre-trained model's performance was ROUGE-1 F1 = 0.49; recall = 71.4%, accuracy = 67.7%; performance dropped significantly in zero-shot settings (1 study). *Documentation quality*: Automatically generated summaries sometimes had lower PDQI-9 scores, and higher word counts compared to manual summaries (1 study). *Deficiency rate:* statistically significant decrease in the 24 h documentation deficiency rate (from 8.6% to 6.3%, meaning that less clinicians failed to complete documentation within 24-h). However, there was a statistically significant increase in the 24-hour billing submission deficiency rate (from 27.9% to 30.0%, meaning that more clinicians failed to submit the required codes for billing within 24-h) (1 study).	**Outcomes:** accuracy (quantitative assessment using a modified PDQI-9 and qualitative assessment), quality, deficiency rate of billing **Evidence:** 6 studies **Findings:** *Accuracy:* positive findings: ■ *Quality*: high scores (modified PDQI-9: average score 48/50 in 1 study). ■ AI scribe-generated notes rated 4.3/5.0 stars in (1 study). ■ Perceived quality improvement (2 studies) (no exact values provided in 1 study, 52% of users reported improvement in another study). ■ *Deficiency rate of billing outcomes:* No significant impact on billing submission or timeliness of documentation (1 study).
Patient-reported outcomes			
Not reported.	**Outcome:** impact of the clinicians' current documentation process on the patient experience **Evidence**: 1 study **Findings:** *Patient experience:* Improved, 35.5% of clinicians responded negatively pre-AI compared to 6.5% with AI use about the impact of their documentation practice.	**Outcomes**: patient safety, patient experience with AI and care **Evidence**: 3 studies **Findings:** *Safety*: no documented patient safety events (1 study). *Experience*: enhancing patient-provider communication while maintaining effective documentation (1 study), some patients expressing discomfort with smartphone recordings in another study.	**Outcome:** *Patient experience*. **Evidence**: 3 studies **Findings:** No studies used standardized or validated patient experience questionnaires. ■ 81-91% of patients perceived that providers spent less time looking at the screen or typing (2 studies) ■ 65-83% felt the visit was more personable, more focused on the patient (2 studies) ■ 100% reported that the AI scribe had no negative effect on the visit (1 study) ■ No significant change in likelihood to recommend scores (1 study) ■ Opt-out rate 0.014% (1 study).

Abbreviations: AI … artificial intelligence; EHR … electronic health record; GPT … Generative Pre-trained Transformer; NR … not reported; PDQI-9 … Physician Documentation Quality Instrument-9; REDCap … Research Electronic Data Capture; ROUGE … Recall-Oriented Understudy for Gisting Evaluation; VBC … value-based care; WER … word error rate; wRVU … work relative value unit

Table A-2: Use case AI scribes (part 2)

Vrdoljak, 2024 [27]	Lee, 2024 [5]	Lumbiganon, 2025 [26]
Clinician-reported outcomes		
Models evaluated: BART, ChatGPT, BERTSUM **Outcomes:** Automatic metrics and human expert evaluation (readability, accuracy, fluency, clinical relevance) **Evidence**: 1 study (Liu) **Findings:** - *BART*: low human evaluation scores (except for readability <30%) - *BERTSUM*: failed human evaluation - *ChatGPT*: preferred overall by medical experts over BERTSUM and BART, summaries judged more comprehensible than some human-written results, but generated some clinically incorrect content (e.g., test results that did not occur). Performance is sensitive to prompt design and fine-tuning.	**Outcomes:** Documentation burden, documentation time, user experience, satisfaction **Evidence**: 2 studies (Goss, Tran) **Findings:** - *Documentation time*: 77% of clinicians reported SR saved time and improved efficiency. 21% spent $\geq$25% of documentation time on editing. Improved efficiency linked to fewer errors ($p < 0.001$) and less editing ($p = 0.02$). More clinically relevant errors correlated with increased editing time ($p < 0.001$) (1 study). Reduced time on clerical tasks and improved workflow efficiency (2 studies). - *Burden*: 62% felt SR reduced administrative burden; 35.9% disagreed or were neutral (1 study). - *User experience*: 86% rated SR system as easy to use; 79% were satisfied, 6% very unsatisfied (1 study). - *Satisfaction* positively associated with efficiency ($p < 0.001$), fewer errors ($p < 0.001$), and less editing time ($p = 0.006$). Satisfaction highest in providers seeing 55-70 patients/week, lowest in >100/week (1 study).	**Outcomes**: time needed for record completion, completion rate, accuracy **Evidence**: 1 study (Cho) **Findings:** - *Time for record completion:* 204 (IQR 155, 277) seconds with AI and 231 (IQR 180, 313) seconds using manual input by EMR. The difference between the 2 methods was statistically significant ($P<.001$) - *Completion rate:* AI achieved 81.8% completion for the first chief concern, vital signs mostly >50% completion (except respiratory rate). AI had lower completion than manual notes for most fields. Higher completion with AI for additional chief concerns and past medical history ($p<0.001$)
Organisational outcomes		
Not reported.	**Outcomes**: challenges of implementation **Evidence**: 5 studies **Findings:** *Challenges:* technical improvements and customization are needed for effective integration into existing EHR systems (2 studies). Extensive training of personnel is required (3 studies).	**Model evaluated:** ChatGPT **Outcomes**: challenges and opportunities of implementing ChatGPT in paediatric emergency medicine **Evidence**: 1 study (Barak-Corren) **Findings:** *Challenges***:** - Concerns about patient privacy and HIPAA compliance - Timing mismatch between ChatGPT summaries and resident notes - Accuracy, patient safety, and liability risks - Perception that existing templates may be easier to use *Opportunities***:** - Potential to develop improved documentation templates using ChatGPT - Widespread belief that efficiency gains outweigh concerns, especially in high-burden settings like emergency medicine

Vrdoljak, 2024 [27]	Lee, 2024 [5]	Lumbiganon, 2025 [26]
Technical performance and documentation quality		
Not reported.	**Outcome:** accuracy and error management **Evidence**: 7 studies **Findings:** *Accuracy*: LLM-generated summaries contain sometimes misinterpretations, fabricated information and errors; therefore, they need editing by physicians for corrections.	**Model evaluated:** ChatGPT **Outcomes**: documentation quality, accuracy **Evidence**: 4 studies (Barak-Corren, Young, Clough, Cho) **Findings***: Quality* mean ratings (0-10 scale): completeness 7.6, accuracy 8.6, efficiency 8.2, readability 8.7 (1 study). Higher quality produced by ChatGPT than junior doctors (1 study). *Accuracy*: 19% of ChatGPT-generated summaries required physician-edit due to incorrect and incomplete information (1 study). 4 of 9 variables had >50% accuracy, chief concern (most important variable) failed reproduction in 50% and 35% complete reproduction (1 study).
Patient-reported outcomes		
Not reported.	Not reported.	Not reported.

Abbreviations: AI … artificial intelligence; BART … Bidirectional and Auto-Regressive Transformers; BERTSUM … Bidirectional Encoder Representations from Transformers Summarization; EHR … electronic health record; GPT … Generative Pre-trained Transformer; HIPAA … Health Insurance Portability and Accountability Act; IQR … interquartile range; LLM … large language model; SR … speech recognition

Table A-3: Use case structuring free-text data

Perkins, 2024 [25]	Vrdoljak, 2024 [27]
Clinician-reported outcomes	
Outcomes: efficiency (speed, time) **Evidence**: 15 studies **Models**: NLP, ML, rule-based (3 studies), AI-SR (12 studies) **Findings:** *Efficiency*: 1. rule-based model: documentation speed increased by 15% (1 study), documentation time decreased (2 studies; in 1 study by 56%, in 1 study no values reported but the study reported that quality also decreased slightly). 2. AI-SR's model: mixed results, 19-92% decrease in mean documentation time (5 studies), increases of 13-50% (4 studies), and no significant difference (3 studies).	Not reported.
Organisational outcomes	
Not reported.	Not reported.
Technical performance and documentation quality	
Outcomes: automatic performance metrics, accuracy, precision/recall **Evidence**: 88 studies **Models**: Rule-based, NLP, machine learning, deep learning, neural networks **Findings**: automatic performance metrics, accuracy, precision/recall ■ *Accuracy*: mainly >0.90 ■ *F-score*: up to 0.984 (e.g., race classification) ■ *PPV*: 0.95-0.97 (e.g., patient safety events, social factors) ■ *AUC*: up to 0.876 (e.g., actionable findings in radiology) ■ *Coherence* (text structuring): 69% (neural network) ■ *Precision/recall*: e.g., phenotype recognition: 83% precision, 51% recall **Task: Annotating clinical notes** **Outcomes**: accuracy, automatic performance metrics **Evidence**: 1 study **Models:** ML, NLP, neural networks **Findings**: ■ *Accuracy*: up to 0.95 ■ *AUC*: up to 0.90 ■ *F1-score*: up to 0.85	**Model evaluated:** ChatGPT 3.5 **Outcomes:** accuracy **Evidence:** 1 study (Huang) **Findings:** *Accuracy*: ChatGPT 3.5 in extracting pathological classifications from lung cancer and paediatric osteosarcoma pathology reports: 89% to 100% accuracy across different datasets. **Model evaluated**: GPT-4 **Outcomes**: specificity, sensitivity **Evidence**: 1 study (Wei) **Findings:** *Specificity, sensitivity*: GPT-4 achieved high specificity (0.947 [95% binCI: 0.894-0.978]-1.000 [95% binCI: 0.965-0.988, 1.000]) for all symptoms, high sensitivity for common symptoms (0.853 [95% binCI: 0.689-0.950]-1.000 [95% binCI: 0.951-1.000]), and moderate sensitivity for less common symptoms (0.200 [95% binCI: 0.043-0.481]-1.000 [95% binCI: 0.590-0.815, 1.000]) (using zero-shot prompting, i.e. no examples). Few-shot prompting (i.e. few examples) increased sensitivity and specificity. GPT-4 outperformed GPT-3.5 in response accuracy and consistent labelling.
Patient-reported outcomes	
Not reported.	Not reported.

Abbreviations: AI-SR … artificial intelligence supported speech recognition; AUC … area under the curve; binCI … binominal confidence interval (estimates the uncertainty around a proportion that is derived from binary outcomes); GPT … Generative Pre-trained Transformer; ML … machine learning; NLP … natural language processing; PPV … positive predictive value

Table A-4: Use case AI-generated medical documentation

Bracken, 2025 [1]	**Lumbiganon, 2025 [26]**	**Vrdoljak, 2024 [27]**
Clinician-reported outcomes		
Outcomes: documentation time, clinician experience, concerns **Evidence**: 3 studies **Models**: ChatGPT **Findings**: *Documentation time:* Mixed results with ChatGPT (3 studies): ■ Time savings in 2 studies (discharge summaries generated by Chat GPT were 2.3-4.6 min faster compared to dictation, and operation notes were 2.7-5.1 min faster compared to dictation (1 study) ■ Inpatient progress notes mean 2 min time saving with ChatGPT in another (1 study). ■ No statistically significant difference was found in efficiency score between ChatGPT and dictation (1 study). *Clinician experience:* Less effort needed with ChatGPT (2 studies), but *concerns* related to inaccuracies (1 study)	**Model evaluated:** ChatGPT **Outcomes**: documentation time, burden, user experience, GP acceptance of hospital discharge summaries **Evidence**: 5 studies (Barak-Corren, Clough, Cho, Bala, Young) **Findings:** ■ *Time* reduction: 0-43% ■ *Effort* reduction: minimal to 33% ■ User-reported *concerns* 1. Missing differential diagnosis (intentionally excluded in prompt design) 2. Lack of pertinent negatives in HPI and physical exam, especially those critical for ruling out serious conditions 3. Overly vague action plans (e.g., "follow up on pending results") 4. Unmarked AI-generated interpretations, raising accountability concerns ("if ChatGPT is wrong, it's on me") *GP acceptance:* 100% of ChatGPT summaries accepted vs. 92% of junior doctor summaries (mean scores: 1.00 vs. 0.92; P = 0.15). Adherence to minimum dataset: Both groups had a median score of 19/20; 97% mean adherence; no significant difference (P = 0.78)	Not reported.
Organisational outcomes		
Not reported.	Not reported.	Not reported.
Technical performance and documentation quality		
Outcomes: overall documentation quality, accuracy, hallucinations **Evidence**: 8 studies **Findings:** *Overall documentation quality:* ■ moderate to high quality: PDQI-9 scores of 30-36 (2 studies) and Likert ratings 7-9/10 (2 studies) ■ ChatGPT-generated notes had higher PDQI-9 scores than typing and dictation (1 study) ■ No difference between ChatGPT vs junior doctors (97% adherence to minimal dataset) (1 study)	**Outcomes:** accuracy **Evidence:** 3 studies (Ganoe, Hyun, Krishna) **Findings:** High *accuracy* (>95% in 1 study and no exact details but statement about higher accuracy than existing models in 1 study), but grammar and lexical issues identified (1 study)	**Task:** ***clinical text summarization*** **Models evaluated**: LLM **Outcomes:** completeness, correctness, conciseness, hallucinations **Evidence**: 1 study (Van Veen) **Findings:** Summaries from the best-adapted LLMs (GPT-4, ICL) were deemed either equivalent (45%) or superior (36%) to those produced by medical experts. ■ *Completeness*: best model summaries vs. medical expert summaries were more complete across all 3 summarization tasks (radiology report, patient question summary, progress notes) ($p < 0.001$). Lengths of summaries were similar between the model and medical experts for all 3 tasks. The model correctly identified conditions that were missed by the medical expert, but it also missed historical context.

Bracken, 2025 [1]	**Lumbiganon, 2025 [26]**	**Vrdoljak, 2024 [27]**
■ Factually correct (similar to gold standard) documentation produced by ChatGPT (2 studies) ■ Operation notes meeting gold standard to lesser extent (2 studies) ■ Ambient AI improved quality greater than 2x compared to traditional EHR use in outpatient letters (1 study). *Hallucinations/Accuracy:* ■ Mixed results with ChatGPT (3 studies): mean 23.6 errors per clinical case, omission (86%), addition errors (10.5%), and incorrect facts (3.2%) (1 study), median factual correctness 81 to 85% in discharge summaries and 71 to 79% in surgical notes (1 study) and 36% of ChatGPT generated notes contained fictitious elements (1 study). ■ No hallucinations in 2 studies. ■ No hallucinations with Ambient AI (1 study).		■ *Correctness*: the best model generated significantly fewer errors (p < 0.001) compared to medical expert summaries overall and on 2 of 3 summarization tasks. E.g., on the radiology report summarization task, it avoided common medical expert errors related to lateral distinctions (right versus left). For the problem list summarization task, the physician reader erroneously assumed that a hallucination was made by the model. In this case, the medical expert was responsible for the hallucination. This underscores the point that even medical experts, not just LLMs, can hallucinate. The model was not perfect across all tasks, e.g., the model mistakenly generated several absent conditions. ■ *Conciseness*: the best model performed significantly better than medical experts (p < 0.001) overall and on 2 tasks, whereas, for radiology reports, it performed similarly to medical experts. The model's summaries are more concise while concurrently being more complete. ■ *Hallucinations*, *inaccuracies*: LLM model committed misinterpretations, inaccuracies and hallucinations on 6%, 2% and 5% of samples, compared to 9%, 4% and 12%, by medical experts.
Patient-reported outcomes		
Not reported.	Not reported.	**Task**: ***clinical text summarization*** **Outcomes:** patient safety **Evidence**: 1 study (Van Veen) **Findings:** summarization errors in relation to *medical harm* (harm study): the medical expert summaries would have both a higher likelihood (14%) and a higher extent (22%) of possible harm compared to the summaries from the best model (12% and 16%, respectively).

Abbreviations: AI … artificial intelligence; EHR … electronic health record; GP … general practitioner; GPT … Generative Pre-trained Transformer; HP …- history of present illness; LLM … large language model; PDQI-9 … Physician Documentation Quality Instrument-9

Table A-5: Use case AI-generated billing codes

Lumbiganon, 2025 [26]
Clinician-reported outcomes
Not reported.
Organisational outcomes
Not reported.
Technical performance and documentation quality
Model evaluated: NLP bidirectional recurrent neural network, Phyton-based NLP tool **Outcomes**: accuracy, automatic metrics (AUROC, AUPRC area under the precision-recall curve) **Evidence**: 2 studies (Kim, Wang) **Findings***: Accuracy*: various models 59-87% accuracy compared to human coders (random forest model 87% accuracy, deep learning model 59% accuracy) (1 study), reduced coding errors in ICD code extraction (1 study).
Patient-reported outcomes
Not reported.

Abbreviations: AUPRC … area under the precision-recall curve; AUROC … area under the receiver operating characteristic curve; ICD … International Classification of Diseases; NLP … natural language processing

Table A-6: Use case AI-generated patient-friendly summaries

Lee, 2024 [5]	**Perkins, 2024 [25]**	**Vrdoljak, 2024 [27]**
Clinician-reported outcomes		
Not reported.	Not reported.	Not reported.
Organisational outcomes		
Not reported.	Not reported.	Not reported.
Technical performance and documentation quality		
Not reported.	Not reported.	Not reported.
Patient-reported outcomes		
Outcomes: patient understanding, quality of care, safety **Evidence**: 9 studies **Findings**: more time is claimed to be available for patient care if AI helps alleviate documentation burden, however fabricated information poses *safety* risks (6 studies). Improved *health literacy* and treatment *adherence*, improved *understanding*, improved *readability* (5 studies), improved *patient-physician interactions* (3 studies).	**Outcome**: patient understanding **Evidence**: 2 studies **Findings**: improvement of lay understanding (2 studies)	**Outcomes**: patient understanding, readability **Evidence**: 1 study (Zaretsky) **Findings**: LLM-transformed discharge summaries were significantly more *readable* and *understandable* when compared to original summaries.

Abbreviations: AI … artificial intelligence; LLM … large language model.

Table A-7: Use case Error detection & note quality assessment

Perkins, 2024 [25]	
Clinician-reported outcomes	
Not reported.	
Organisational outcomes	
Not reported.	
Technical performance and documentation quality	
Task: ***Error detection*** **Outcomes**: accuracy, automatic performance metrics **Evidence**: 4 studies **Models**: Rule-based, NLP, neural networks **Findings**: ■ *Accuracy*: 0.91-0.93 ■ *F1-score:* 0.68-0.94 ■ *PPV*: Up to 0.93	**Task:** ***Assessing clinical note quality*** **Outcomes**: accuracy, automatic performance metrics **Evidence**: 8 studies **Models**: NLP, rule-based, hybrid approaches **Findings**: ■ *Accuracy*: 0.91 ■ *F1-score*: Up to 0.92
Patient-reported outcomes	
Not reported.	

Abbreviations: NLP … natural language processing; PPV … positive predictive value

Appendix C: Procurement checklist for decision-makers [13]

Checklist	
Purpose	
	What is the main purpose of the AI and what is the main utility?
	Which specific healthcare processes will be affected?
	Who are the intended users (healthcare professionals, patients, administrators)?
Regulatory Requirements	
Medical Device Classification	
	Is it considered a medical device under MDR?
	What is its risk classification under MDR (Class I, IIa, IIb, or III)
	What is its risk classification under EU AI Act (high-risk, low-risk)?
	Does the AI-system adhere to high-risk AI systems transparency and safety requirements? (see MDR, EU AI Act)
	Is a valid CE marking present?
Data Protection and Privacy	
	Does the AI-enabled DHT comply with GDPR requirements?
	Are there procedures for patient consent and data rights?
	Consider the EHDS once fully implemented.
HTA Evaluation	
	Reflect on who will conduct the assessment, if HTA-reports are not yet available
AI relevant considerations (covered in standard methodology[17])	
CUR	What are the main characteristics of the health problem, including the proposed AI solution, and the specific patient populations and clinical settings where it can be implemented?
TEC	What are the main characteristics of the AI-enabled DHT?
EFF	What are the clinical benefits and quality of life impact of the AI-enabled DHT, and are the benefits superior to those of existing alternatives?
SAF	Are there risks or possible undesirable effects caused by the AI-enabled DHT that could lead to physical or psychological harm to patients or professionals?
ETH	Does the AI-enabled DHT have an impact on inequalities?
SOC	What is the user experience of the AI-enabled DHT?
ORG	Does the implementation of the AI-enabled DHT involve the training of the professional team?
ECO	What are the costs of acquiring, maintaining and using the AI-enabled technology at the patient and health system level?
AI-specific considerations (not covered in standard methodology)	
TEC	Which data sets were used for training and validating the DHT? Is there a strategy how to handle incomplete data? What is the type of machine learning? How will the performance be measured?
SAF	Are there strategies on data risk management foreseen? How can anomalies of the AI-enabled DHT in operational use be detected?
ETH	Are there strategies to mitigate algorithmic bias in the AI-enabled DHT?
ORG	What is the level of professional oversight? Is staff's approval needed for action, proposed by the AI-enabled DHT? Has the output been cross-checked by a qualified human?
ECO	Is it clear what ongoing support is available for adopters and what it would cost?
Monitoring of performance	
	Define strategies on post-deployment for the AI-enabled DHT.
	How often will the AI-enabled DHT be monitored and by whom?
	How will changes in performance be detected and measured?
	When should a re-assessment of the AI-enabled DHT be conducted?
Check again in case of changes in performance and purpose	

Abbreviations: AI … Artificial Intelligence, CUR … Current Use, DHT … Digital Health Technology, ECO … Economic, EFF … Effectiveness, EHDS … Electronic Health Data Space, ETH … Ethical, EU … European Union, GDPR … General Data Protection Regulation, HTA … Health Technology Assessment, MDR … Medical Device Regulation, ORG … Organisational, SAF … Safety, SOC … Social; TEC … Technical.

[17] E.g. the EUnetHTA Core Model

Appendix D: Search strategies

Search strategy for Medline via Ovid

Database: Ovid MEDLINE(R) ALL <1946 to June 30, 2025>	
Search date: June 30, 2025	
ID	Search
1	exp Artificial Intelligence/ (240659)
2	artificial intelligence.mp. (102423)
3	AI.mp. (80096)
4	Large language model*.mp. (6593)
5	LLM.mp. (2733)
6	LLMs.mp. (3264)
7	exp Natural Language Processing/ (8131)
8	natural language processing.mp. (15820)
9	generative AI.mp. (1518)
10	Gen?AI.mp. (227)
11	Gen-AI.mp. (27)
12	exp Generative Artificial Intelligence/ (657)
13	Chat?GPT.mp. (6580)
14	Chat-GPT.mp. (215)
15	GPT.mp. (7912)
16	generative multimodal model*.mp. (0)
17	(automat* adj3 ((report* or note* or record* or discharg* or document*) adj generat*)).mp. (137)
18	scribe*.mp. (826)
19	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 (337018)
20	exp Medical Records/ (166620)
21	medical record*.mp. (245455)
22	(discharg* adj3 (letter* or summar*)).mp. (3292)
23	((medical or clinical or health or discharg*) adj3 (record* or report* or note* or document*)).mp. (522450)
24	(automat* adj3 (record* or report* or note* or document* or discharg*)).mp. (6937)
25	EHR.mp. (15005)
26	EHRs.mp. (6247)
27	20 or 21 or 22 or 23 or 24 or 25 or 26 (573591)
28	19 and 27 (14489)
29	*Documentation/ (8982)
30	document*.mp. (553086)
31	generat*.mp. (1745852)
32	automat*.mp. (364831)
33	prepar*.mp. (1319687)
34	29 or 30 or 31 or 33 (3484655)
35	28 and 34 (4612)
36	limit 35 to (meta analysis or "systematic review") (69)
37	(((comprehensive* or integrative or systematic*) adj3 (bibliographic* or review* or literature)) or (meta-analy* or metaanaly* or "research synthesis" or ((information or data) adj3 synthesis) or (data adj2 extract*))).ti,ab. or (cinahl or (cochrane adj3 trial*) or embase or medline or psyclit or (psycinfo not "psycinfo database") or pubmed or scopus or "sociological abstracts" or "web of science").ab. or ("cochrane database of systematic reviews" or evidence report technology assessment or evidence report technology assessment summary).jn. or Evidence Report: Technology Assessment*.jn. or ((review adj5 (rationale or evidence or safety or effectiveness)).mp. and review.pt.) or meta-analysis as topic/ or Meta-Analysis.pt. (893517)

38	35 and 37 (474)
39	36 or 38 (478)
40	remove duplicates from 39 (476)
Total hits: 476	

Search strategy for Cochrane

Search Name: AI to support clinical documentation	
Last Saved: 01/07/2025 17:24:54	
Comment: JE	
ID	Search
#1	MeSH descriptor: [Artificial Intelligence] this term only
#2	("artificial intelligence")
#3	(AI):ti,ab,kw
#4	(large NEXT language NEXT model*) (Word variations have been searched)
#5	(LLM):ti,ab,kw
#6	(LLMs):ti,ab,kw
#7	MeSH descriptor: [Natural Language Processing] explode all trees
#8	("natural language processing") (Word variations have been searched)
#9	("generative AI") (Word variations have been searched)
#10	(Gen?AI) (Word variations have been searched)
#11	(Gen-AI) (Word variations have been searched)
#12	MeSH descriptor: [Generative Artificial Intelligence] explode all trees
#13	(Chat?GPT) (Word variations have been searched)
#14	(Chat-GPT) (Word variations have been searched)
#15	(GPT):ti,ab,kw
#16	(generative NEXT multimodal NEXT model*):ti,ab,kw (Word variations have been searched)
#17	(automat* NEAR ((report* OR note* OR record* OR discharg* OR document*) NEAR generat*)) (Word variations have been searched)
#18	(scribe*) (Word variations have been searched)
#19	#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16 OR #17 OR #18
#20	MeSH descriptor: [Medical Records] explode all trees
#21	(medical NEXT record*):ti,ab,kw (Word variations have been searched)
#22	(discharg* NEAR (letter* OR summar*)) (Word variations have been searched)
#23	((medical OR clinical OR health OR discharg*) NEAR (record* OR report* OR note* OR document*)):ti,ab,kw
#24	(automat* NEAR (record* OR report* OR note* OR document* OR discharg*)):ti,ab,kw (Word variations have been searched)
#25	(EHR):ti,ab,kw (Word variations have been searched)
#26	(EHRs):ti,ab,kw (Word variations have been searched)
#27	#20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26
#28	#19 AND #27 in Cochrane Reviews, Cochrane Protocols
Total hits: 11	

Search strategy for Embase

Search Name: AI to support clinical documentation	
Search date: 2025-07-01	
ID	Search query,"Hits","Searched At"
41	(((prepar*) OR (automat*) OR (generat*) OR (document*) OR ("Documentation"[mhe])) AND (((EHRs) OR (EHR) OR ((automat*) AND (record* OR report* OR note* OR document* OR discharg*)) OR ((medical OR clinical OR health OR discharg*) AND (record* OR report* OR note* OR document*)) OR ((discharg*) AND (letter* OR summar*)) OR ("medical records") OR ("medical record") OR ("Medical Records"[mhe])) AND ((scribe*) OR ((automat*) AND ((report* OR note* OR record* OR discharg* OR document*) AND generat*)) OR (GPT) OR (Chat-GPT*) OR ("Chat GPT") OR ("Generative artificial intelligence") OR ("Generative Artificial Intelligence"[mhe]) OR (Gen-AI*) OR (GenAI*) OR ("generative AI") OR ("Natural language processing") OR ("Natural Language Processing"[mhe]) OR (LLMs) OR (LLM) OR ("Large language models") OR ("Large language model") OR (AI) OR ("artificial intelligence") OR ("Artificial Intelligence"[mhe])))) AND (English OR German)[Language],"21","2025-07-01T16:12:16.000000Z"
40	((prepar*) OR (automat*) OR (generat*) OR (document*) OR ("Documentation"[mhe])) AND (((EHRs) OR (EHR) OR ((automat*) AND (record* OR report* OR note* OR document* OR discharg*)) OR ((medical OR clinical OR health OR discharg*) AND (record* OR report* OR note* OR document*)) OR ((discharg*) AND (letter* OR summar*)) OR ("medical records") OR ("medical record") OR ("Medical Records"[mhe])) AND ((scribe*) OR ((automat*) AND ((report* OR note* OR record* OR discharg* OR document*) AND generat*)) OR (GPT) OR (Chat-GPT*) OR ("Chat GPT") OR ("Generative artificial intelligence") OR ("Generative Artificial Intelligence"[mhe]) OR (Gen-AI*) OR (GenAI*) OR ("generative AI") OR ("Natural language processing") OR ("Natural Language Processing"[mhe]) OR (LLMs) OR (LLM) OR ("Large language models") OR ("Large language model") OR (AI) OR ("artificial intelligence") OR ("Artificial Intelligence"[mhe]))),"35","2025-07-01T16:11:58.000000Z"
39	((prepar*) OR (automat*) OR (generat*) OR (document*) OR ("Documentation"[mhe])) AND (((EHRs) OR (EHR) OR ((automat*) AND (record* OR report* OR note* OR document* OR discharg*)) OR ((medical OR clinical OR health OR discharg*) AND (record* OR report* OR note* OR document*)) OR ((discharg*) AND (letter* OR summar*)) OR ("medical records") OR ("medical record") OR ("Medical Records"[mhe])) AND ((scribe*) OR ((automat*) AND ((report* OR note* OR record* OR discharg* OR document*) AND generat*)) OR (GPT) OR (Chat-GPT*) OR ("Chat GPT") OR ("Generative artificial intelligence") OR ("Generative Artificial Intelligence"[mhe]) OR (Gen-AI*) OR (GenAI*) OR ("generative AI") OR ("Natural language processing") OR ("Natural Language Processing"[mhe]) OR (LLMs) OR (LLM) OR ("Large language models") OR ("Large language model") OR (AI) OR ("artificial intelligence") OR ("Artificial Intelligence"[mhe]))),"35","2025-07-01T16:11:00.000000Z"
38	((prepar*) OR (automat*) OR (generat*) OR (document*) OR ("Documentation"[mhe])) AND (((EHRs) OR (EHR) OR ((automat*) AND (record* OR report* OR note* OR document* OR discharg*)) OR ((medical OR clinical OR health OR discharg*) AND (record* OR report* OR note* OR document*)) OR ((discharg*) AND (letter* OR summar*)) OR ("medical records") OR ("medical record") OR ("Medical Records"[mhe])) AND ((scribe*) OR ((automat*) AND ((report* OR note* OR record* OR discharg* OR document*) AND generat*)) OR (GPT) OR (Chat-GPT*) OR ("Chat GPT") OR ("Generative artificial intelligence") OR ("Generative Artificial Intelligence"[mhe]) OR (Gen-AI*) OR (GenAI*) OR ("generative AI") OR ("Natural language processing") OR ("Natural Language Processing"[mhe]) OR (LLMs) OR (LLM) OR ("Large language models") OR ("Large language model") OR (AI) OR ("artificial intelligence") OR ("Artificial Intelligence"[mhe]))),"35","2025-07-01T16:10:54.000000Z"
37	(prepar*) OR (automat*) OR (generat*) OR (document*) OR ("Documentation"[mhe]),"1999","2025-07-01T16:10:43.000000Z"
36	prepar*,"417","2025-07-01T16:10:27.000000Z"
35	automat*,"189","2025-07-01T16:10:14.000000Z"
34	generat*,"570","2025-07-01T16:09:58.000000Z"
33	document*,"1072","2025-07-01T16:09:45.000000Z"
32	"Documentation"[mhe],"17","2025-07-01T16:09:30.000000Z"
31	((EHRs) OR (EHR) OR ((automat*) AND (record* OR report* OR note* OR document* OR discharg*)) OR ((medical OR clinical OR health OR discharg*) AND (record* OR report* OR note* OR document*)) OR ((discharg*) AND (letter* OR summar*)) OR ("medical records") OR ("medical record") OR ("Medical Records"[mhe])) AND ((scribe*) OR ((automat*) AND ((report* OR note* OR record* OR discharg* OR document*) AND generat*)) OR (GPT) OR (Chat-GPT*) OR ("Chat GPT") OR ("Generative artificial intelligence") OR ("Generative Artificial Intelligence"[mhe]) OR (Gen-AI*) OR (GenAI*) OR ("generative AI") OR ("Natural language processing") OR ("Natural Language Processing"[mhe]) OR (LLMs) OR (LLM) OR ("Large language models") OR ("Large language model") OR (AI) OR ("artificial intelligence") OR ("Artificial Intelligence"[mhe])),"60","2025-07-01T16:08:43.000000Z"
30	((EHRs) OR (EHR) OR ((automat*) AND (record* OR report* OR note* OR document* OR discharg*)) OR ((medical OR clinical OR health OR discharg*) AND (record* OR report* OR note* OR document*)) OR ((discharg*) AND (letter* OR summar*)) OR ("medical records") OR ("medical record") OR ("Medical Records"[mhe])) AND ((scribe*) OR ((automat*) AND ((report* OR note* OR record* OR discharg* OR document*) AND generat*)) OR (GPT) OR (Chat-GPT*) OR ("Chat GPT") OR ("Generative artificial intelligence") OR ("Generative Artificial Intelligence"[mhe]) OR (Gen-AI*) OR (GenAI*) OR ("generative AI") OR ("Natural language processing") OR ("Natural Language Processing"[mhe]) OR (LLMs) OR (LLM) OR ("Large language models") OR ("Large language model") OR (AI) OR ("artificial intelligence") OR ("Artificial Intelligence"[mhe])),"60","2025-07-01T16:08:35.000000Z"
29	(EHRs) OR (EHR) OR ((automat*) AND (record* OR report* OR note* OR document* OR discharg*)) OR ((medical OR clinical OR health OR discharg*) AND (record* OR report* OR note* OR document*)) OR ((discharg*) AND (letter* OR summar*)) OR ("medical records") OR ("medical record") OR ("Medical Records"[mhe]),"5140","2025-07-01T16:08:25.000000Z"
28	(scribe*) OR ((automat*) AND ((report* OR note* OR record* OR discharg* OR document*) AND generat*)) OR (GPT) OR (Chat-GPT*) OR ("Chat GPT") OR ("Generative artificial intelligence") OR ("Generative Artificial Intelligence"[mhe]) OR (Gen-AI*) OR (GenAI*) OR ("generative AI") OR ("Natural language processing") OR ("Natural Language Processing"[mhe]) OR (LLMs) OR (LLM)

	OR ("Large language models") OR ("Large language model") OR (AI) OR ("artificial intelligence") OR ("Artificial Intelligence"[mhe]),"158","2025-07-01T16:07:51.000000Z"
27	EHRs,"6","2025-07-01T16:06:56.000000Z"
26	EHR,"7","2025-07-01T16:06:52.000000Z"
25	(automat*) AND (record* OR report* OR note* OR document* OR discharg*),"102","2025-07-01T16:05:38.000000Z"
24	(medical OR clinical OR health OR discharg*) AND (record* OR report* OR note* OR document*),"5103","2025-07-01T16:04:38.000000Z"
23	(discharg*) AND (letter* OR summar*),"38","2025-07-01T16:03:28.000000Z"
22	"medical records","27","2025-07-01T16:02:55.000000Z"
21	"medical record","12","2025-07-01T16:02:49.000000Z"
20	"Medical Records"[mhe],"56","2025-07-01T16:02:28.000000Z"
19	scribe*,"0","2025-07-01T16:01:55.000000Z"
18	(automat*) AND ((report* OR note* OR record* OR discharg* OR document*) AND generat*),"17","2025-07-01T16:00:55.000000Z"
17	GPT,"3","2025-07-01T15:59:27.000000Z"
16	Chat-GPT*,"0","2025-07-01T15:59:15.000000Z"
15	"Chat GPT","0","2025-07-01T15:59:02.000000Z"
14	"Generative artificial intelligence","0","2025-07-01T15:58:38.000000Z"
13	"Generative Artificial Intelligence"[mhe],"0","2025-07-01T15:58:18.000000Z"
12	Gen-AI*,"0","2025-07-01T15:57:53.000000Z"
11	GenAI*,"0","2025-07-01T15:57:49.000000Z"
10	"generative AI","0","2025-07-01T15:57:29.000000Z"
9	"Natural language processing","2","2025-07-01T15:57:08.000000Z"
8	"Natural Language Processing"[mhe],"0","2025-07-01T15:56:47.000000Z"
7	LLMs,"1","2025-07-01T15:56:18.000000Z"
6	LLM,"1","2025-07-01T15:56:09.000000Z"
5	"Large language models","0","2025-07-01T15:55:56.000000Z"
4	"Large language model","1","2025-07-01T15:55:43.000000Z"
3	AI,"0","2025-07-01T15:55:04.000000Z"
2	"artificial intelligence","32","2025-07-01T15:54:44.000000Z"
1	"Artificial Intelligence"[mhe],"133","2025-07-01T15:54:19.000000Z"
Total hits: 21	

aihta

HTA Austria
Austrian Institute for
Health Technology Assessment
GmbH