Centre for Assessment of Medical Technology in Örebro

HTA-rapport 2026: 86

A Systematic Review Comparing Ambient Scribes and Conventional Documentation Methods in Outpatient Care

En systematisk översikt om AI-baserad journalföring jämfört med traditionell dokumentation av patientbesök i öppenvården

- Louise Olsson, Marek Czajkowski, Rebecka Klang, Rolf Ahlzén, Lars Breimer
  HTA-enheten Camtö

Region Örebro län

The following have contributed to this report:

Literature search: The Medical Library, Örebro university

Clinical effect: Marek Czajkowski, MD, Lars Breimer, MD, PhD, Louise Olsson, MD, PhD

Statistical review: Rebecka Klang, MSc

Ethics: Rolf Ahlzén, MD, PhD

Layout: Printing Office (Repro) Örebro University

Medical fact box: Sara Bucher, Chief Medical Officer, Region Örebro County

External review:

Stefan Jansson, GP at Brickebacken Primary Care Center, Associate Professor of Family Medicine, Örebro University

Iben Fasterholdt, Senior Researcher, Tue Kjølhede, Associate Professor, and Kristian Kidholm, Professor and research leader, HTA-Unit, Odense University Hospital and Centre for Innovative Medical Technology, University of Southern Denmark

The medical fact boxes in our reports are written by independent experts who were not involved in conducting the systematic review. External reviewers provide valuable feedback that improve the quality of our HTA reports, but responsibility for the final content rests solely with Camtö.

Marek Czajkowski is a representative of The Swedish Association of General Practice in a committee for eHealth within the Swedish Society of Medicine. All other authors and reviewers declare they have no financial or other conflict of interest related to this systematic review.

Contact: camto@regionorebrolan.se
This report is available on
https://www.regionorebrolan.se/camto

Diarienummer: 25RS13178
Centre for Assessment of Medical Technology in Örebro
University Hospital Örebro
S-701 85 Örebro
Date of publication: 2026-04-22

Abbreviations and glossary

AI	Artificial intelligence
Ambient scribe	An AI-based system that records and processes clinician–patient interactions during the consultation to generate structured medical documentation in real time or shortly thereafter.
D-i-D	Difference-in-difference analysis
EHR	Electronic health record
Epic	American electronic health record system
GDPR	General Data Protection Regulation EU regulation from 2018 on individuals' rights and requirements for how compa nies and organizations process personal data.
KLAS Research	U.S.-based healthcare research and analytics firm that evaluates health informa tion technology vendors and solutions.
MDR	Medical Device Regulation EU Regulatory framework for medical devices from 2021
Norwegian Register	A system that classifies scientific journals and publishers by academic quality and is used for research evaluation. kanalregister.hkdir.no/en
Pajama time	Informal term used in healthcare to describe the time clinicians spend working in the EHR outside of regular working hours, in the evenings or late at night, often at home.
RCTs	Randomised controlled trials
RoB2	Revised Cochrane Risk of Bias tool for randomised trials
SBU	Swedish Agency for Health Technology Assessment and Assessment of Social Services (Statens beredning för medicinsk och social utvärdering)

Innehåll

Abstract

Background

The effects of using ambient scribes for clinical documentation are not yet well understood. In this first systematic review, we aimed to evaluate the scientific evidence on time savings and the accuracy of AI-generated drafts compared with conventional documentation methods in outpatient care.

Methods

MEDLINE, Embase, and the Cochrane Library were searched by librarians for RCTs between 2022 and December 2025. The PRISMA guidelines for systematic reviews were followed. Risk of bias was assessed using the Swedish version of RoB 2. A narrative analysis followed.

Results

Four U.S. studies including between 23 to 238 physicians were identified. Participants were selected based on self-interest, recommendations, or lengthy documentation time. The control intervention (documenta-tion as usual) was not clearly defined in any study.

Two studies had severe randomization issues, implying an unacceptably high risk of bias; the remaining two had high risk of bias. One of them compared time-in-notes to baseline data among high- (n=7), moderate- (n=6) and low-frequency (n=12) users. In the largest study, two AI scribe applications were used in 33.5% and 29.5% patient visits, respectively, following a 1-month lead-in period for participants to reach proficiency.

No study reported time-in-notes that included editing time for AI-generated drafts prior to sign-off. No data on the accuracy of the AI-drafts versus other documentation methods were reported.

Conclusion

No data were identified regarding the total time used for documentation with ambient scribes, including editing, compared with conventional methods, or regarding accuracy. Factors influencing the limited uptake, essential for overall effectiveness, were not accounted for.

Plain Language Summary in Swedish/ Populärvetenskaplig sammanfattning

Bakgrund

Dokumentation av mottagningsbesök med hjälp av AI innebär att samtalet spelas in, transkriberas till text som struktureras till en journalanteckning. Anteckningen måste som vanligt granskas, vid behov korrigeras och signeras av ansvarig läkare. I dagsläget dikterar läkare ofta, varefter en medicinsk sekreterare skriver ut anteckningen. Syftet med detta projekt var att sammanställa studier som jämfört den totala tidsåtgången mellan de två metoderna samt tillförlitligheten mellan AI-genererad journaltext och andra metoder.

Metod

Projektet följde sedvanliga riktlinjer för systematiska översikter. Bibliotekarier eftersökte randomiserade studier publicerade mellan 2020 och december 2025 i tre medicinska databaser. Risk för snedvridning av resultaten i studierna (bias) granskades utifrån en särskild mall från SBU.

Resultat

Fyra studier från USA identifierades, med mellan 23 och 238 läkare i varje. Deltagarna hade valts ut baserat på eget intresse, andras rekommendation eller för att de haft problem med tidsödande journal-dokumentation. Läkarna fördelades slumpmässigt till att använda antingen AI-tekniken eller ”den vanliga metoden”. Ingen av studierna beskrev dock tydligt vad det innebar.

Två av studierna bedömdes ha oacceptabelt hög risk för snedvridna resultat och exkluderades. De övriga två hade hög risk för bias. I en av dessa använde hälften av läkarna i AI-gruppen den nya metoden för mindre än hälften av patientbesöken. I den största studien utvärderades två olika AI-system som läkarna fick träna på under en månad innan själva studien startade. Den nya AI-metoden användes sedan vid 34 % respektive 30 % av patientbesöken.

Ingen av studierna rapporterade den total tidsåtgången, det vill säga de redovisade inte tid för att korrigera de AI-genererade förslagen till journalanteckning. De rapporterade inte heller något om tillförlitligheten i den AI-genererade texten jämfört med andra metoder för journaldokumentation.

Slutsats

Det påträffades inga studier som redovisat den totala tidsåtgången med AI-baserad teknik för journal-dokumentation av patientbesök, eller jämförelse av innehållet mellan AI och annan metod för journaldokumentation. Studierna gav inga förklaringar till varför AI-tekniken användes vid en så låg andel av patientbesöken.

Medical Fact Box in Swedish/Medicinsk faktaruta

Sara Bucher, Chief Medical Officer, Region Örebro County

Ambient scribe är ett digitalt system som med hjälp av taligenkänning och artificiell intelligens (AI), baserad på stora språkmodeller, registrerar och transkriberar samtal mellan vårdpersonal och patient i realtid. Systemet strukturerar därefter informationen enligt journalmall (exempelvis strukturerade sökord för anamnes, status, bedömning och plan) och genererar ett förslag till journalanteckning som granskas av användaren, redigeras vid behov och införs i patientjournalen där den signeras. Om diagnoskoder är en del av journal kan vissa system även föreslå relevanta koder som användaren tar ställning till innan de infogas i anteckningen. Tekniken kan integreras med befintliga elektroniska journalsystem. Ljuddata och transkriptat bearbetas lokalt eller via molnbaserad tjänst beroende på systemlösning.

Tekniken omfattas av medicintekniskt regelverk (t.ex. MDR) beroende på funktion och avsedd användning och bedömning enligt GDPR för hantering av personuppgifter och ljudinspelningar. Hantering av samtycke till inspelning och användning av denna samt användning av AI och en tydlighet runt ansvar för medicinska beslut och dokumentation är viktiga områden att tydliggöra vid användningen. Potentiella fördelar är t ex minskad administrativ belastning för vårdpersonal, kortare ledtid till färdig journalanteckning och ökad tid för patientkontakt. Målet med tekniken är att få bättre kvalitet på informationen i journalerna, genom mer strukturerad och enhetlig dokumentation. Potentiella och allvarliga risker är exempelvis risk för felaktig journalföring av medicinsk information, bland annat på grund av hallucination eller semantiskt felaktigt innehåll. Risk för systematisk bias i journalföringen måste också hanteras. Det krävs också ett omsorgsfullt arbete för att säkerställa integritet och informationssäkerhet. Eftersom mycket av den hälsodata vården använder, både i den direkta vården av patienten (så kallad primäranvändning) och i verksamhetsuppföljning och forskning (så kallad sekundäranvändning), kommer från journalanteckningarna är det centralt att anteckningar håller stabil god kvalitet.

Alternativ till AI-baserad dokumentation är manuell journalföring av vårdpersonal under eller efter patientmöte, inklusive traditionell diktering med efterföljande transkribering av medicinsk sekreterare. I studier har även manuell journalföring medfört felaktigt journalförda uppgifter, men detta skall dock inte leda till att vi sänker kraven på automatiserade lösningar.

Införande av ambient scribe aktualiserar flera aspekter. Förutom potentiell tidsbesparing och effektivisering krävs utvärdering av dokumentationskvalitet, patientsäkerhet och påverkan på kliniskt beslutsfattande. AI-baserade språkmodeller kan generera text som är språkligt korrekt men fakta-mässigt felaktig, vilket ställer krav på tydlig mänsklig kontroll och ansvar. Denna kontroll åligger både den enskilda läkaren och vårdgivaren. Vidare behöver informationssäkerhet, datalagring och patientintegritet hanteras i enlighet med gällande lagstiftning. Patientens upplevelse av inspelning i vårdmötet och dess påverkan på kommunikation och förtroende är också en relevant aspekt.

Mot bakgrund av en ökande administrativ belastning och snabb teknikutveckling finns behov av systematisk utvärdering av ambient scribe avseende klinisk nytta, risker, organisatoriska konsekvenser och kostnadseffektivitet innan brett införande inom hälso- och sjukvården.

Background

Clinical documentation of outpatient visits in Sweden is primarily performed by dictation, transcribed into the electronic health record (EHR) by a medical secretary, and then reviewed and signed off by the physician. As an alternative, speech recognition technology has been introduced to automatically transcribe dictation into text, which is subsequently reviewed and edited by the physician.

A further level of automation has recently emerged, in which all spoken interaction during an outpatient visit is transcribed and an AI-based system generates a clinical note. The new technology has generated interest as it has been predicted to reduce time spent on documentation and allow for more time with patients.

At this early stage of adoption, we were primarily interested in the scientific evidence on total time savings associated with the new technology, and how the accuracy of AI-generated drafts compares with conventional documentation methods. We therefore set out to conduct a first systematic review on this topic.

Methods

This systematic review was registered in Researchweb (www.researchweb.org/is/fourol/project/286892).

Research question:

What are the effects of AI-based ambient scribe compared with conventional documentation methods on total time-in-notes and agreement of content for outpatient visits?

The following PICO was defined:

- **Population** Outpatient visits
- **Intervention** Ambient scribe technology
- **Comparison** Other documentation methods e.g., dictation+medical secretary, physician manually entering notes, voice recognition + editing.
- **Outcome** - Total documentation time until final sign-off
  - Accuracy (i.e. agreement ambient scribe technology with control)
- **Study design** Randomised controlled trials

Literature searches

Librarians at the Medical Library, Örebro University, searched for RCTs from 01-Jan-2020 until 04-Dec-2025 in MEDLINE, Embase and Cochrane Library. The search strategy, using a filter for RCT, is presented in Appendix 1.

Inclusion criteria

- Only studies including real-world patients.
- Only studies reporting total documentation time until the note is finalised and signed off in the patient record.
- Only studies published in English.
- Outpatient visits in primary and secondary care
- All healthcare professions are eligible.

Exclusion criteria

- Reviews, systematic reviews, protocols, letters, case reports, conference abstracts, pre-prints
- Studies based on simulations

Selection

An initial screening for relevance based on titles and abstracts including all publications identified from the literature search was conducted by independent reviewers (MC, LB). Any publication selected by either reviewer proceeded to the next level. At this stage, full-text versions of all selected publications were retrieved and assessed independently for relevance by the two reviewers. Any remaining discrepancies were resolved through consensus discussions.

The selection process, including the final number of relevant studies, is presented in a PRISMA flow diagram. Authors of all relevant studies were searched in Retraction Watch Database [1].

Risk of bias assessment

Three reviewers independently (LB, MC, LO) assessed the risk of bias for all relevant studies using the Swedish version of Cochrane RoB 2 [2]. Any discrepancies were resolved through discussion until consensus was reached. The findings are presented in a specific risk of bias diagram.

Statistical review

A statistician (RK) reviewed the statistical aspects of the studies, including sample size calculation, choice of statistical methods, reporting, and interpretation. Continuous baseline variables were extracted for analysis with Carlisle's method [3], where extreme values in the calculated combined p-value could indicate problem with the randomization. The cutoff values were set at a combined p-value of ≤ 0.05 or ≥ 0.95. Other inadequate or notable handling of statistics in the studies was also noted.

Conflicts of interest

The number of authors who declared financial conflicts of interest, or employment with the study sponsor was extracted and tabulated, as well as reported study funding and the professional titles of the authors.

Publishing journals

Journals that had published relevant studies were characterised by searching uptake in the Directory of Open Access Journals [4], Cabells' predatory reports [5], and the Norwegian list [6]. Journals were categorised as subscription only, Open access only and a hybrid between these two, i.e. offering both types of publication.

Data extraction

Basic characteristics of relevant studies and relevant outcome data were extracted. by one of the reviewers (MC) and double-checked by another (LB).

Analysis

A narrative analysis was planned as the primary approach.

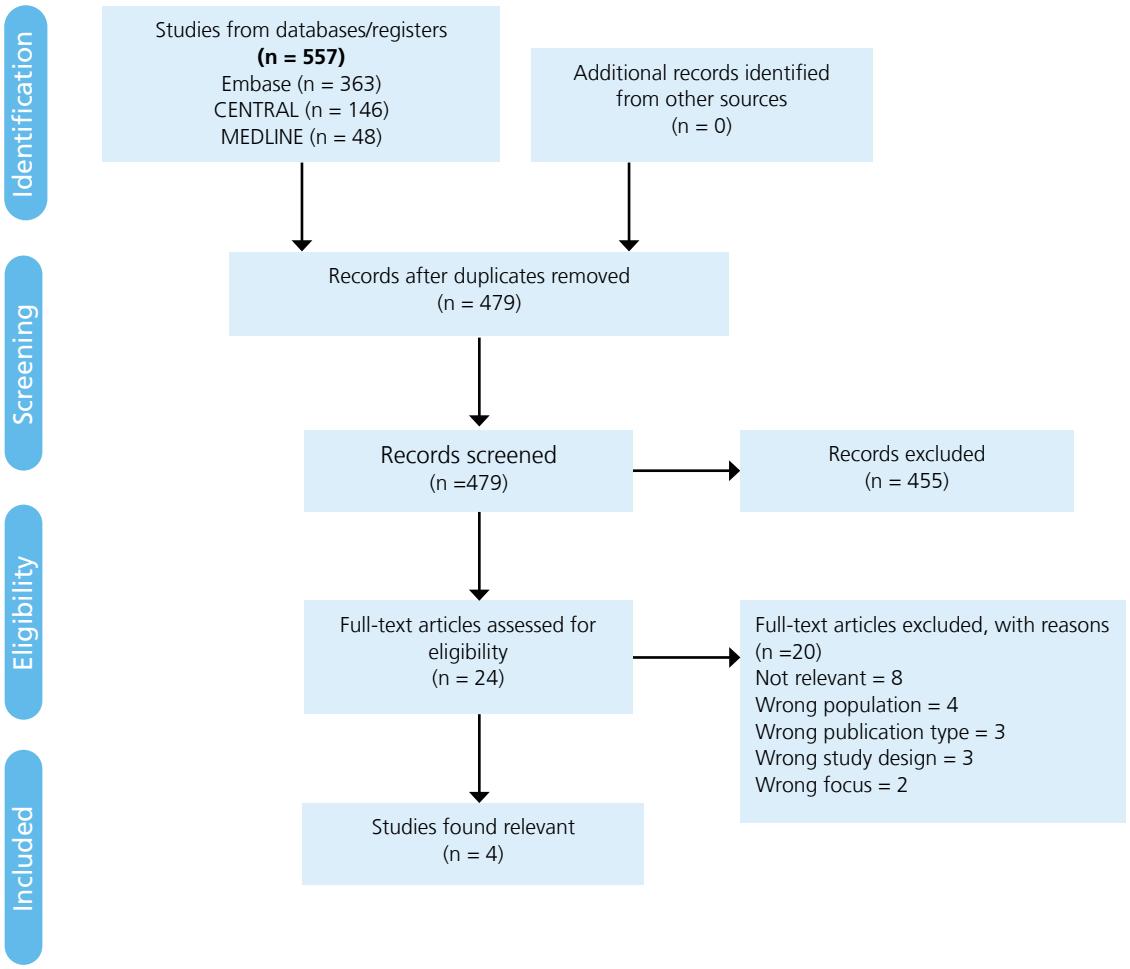
Ongoing studies

Ongoing primary studies were searched in Clinicaltrials.gov [7], the British database ISRCTN [8], and the WHO database ICTRP [9]. Ongoing systematic reviews were searched for in PROSPERO [10].

Results

Assessment of relevant studies

In all, 479 unique records remained after duplicates were removed by librarians and 24 articles remained after selection on the title/abstract. Twenty of these were not found relevant, with reasons for exclusion presented in Appendix 2. For the remaining 4 studies, all authors were searched in Retraction Watch Database, but none of them were found there.



Figur 1 Study flow chart

All relevant studies were conducted in the USA [11-14]; three included <50 participants (Table 1). Two studies focused on primary care physicians, while the other two focused on either physicians from 14 different specialties or on paediatricians. All studies involved some preselection based on participants' interest, problems with lengthy documentation, or recommendations.

Table 1 Basic characteristics of relevant studies

Author, year, country	N	Study period	Study population, setting	Intervention	Control	Outcome
Kakaday 2025 USA	45	June – August 2024	Physicians/physician assistants/nurse practitioners in community-based health-care system were invited to indicate interest in participating. Only problem-focused visits were eligible.	DAX* Copilot incl standardized note template	Traditional documentation methods.	**Primary outcome:** Documentation - time - length - composition
Lukac 2025 USA	238	Nov 4, 2024 – Jan 3, 2025	Outpatient physicians from 14 specialties recruited via department-wide emails and nominations from dept leaders.	DAX Copilot or Nabla	Usual-care control group	**Primary outcome:** - Change from baseline log writing time-in-note **Secondary outcomes:** - Mini Z 2.0 (burnout, stress) - Professional fulfilment Index - Work Exhaustion
Shin 2025 USA	23	May 1, 2024 – July 31, 2024	Pediatricians who had opportunity for improvement in pajama time and time in notes per encounter at a tertiary academic health system.	DAX Copilot	Documentation as usual	**Primary outcome** - Net EHR Experience and Experiencing Burnout - Pajama time - Time in notes per appointment - Outpatient adjusted relative work Relative Value Units per encounter
Wendt 2025 USA	24	June 1, 2022 – Dec 31, 2022	109 family medicine providers across 7 states which were most likely to benefit from automated clinical documentation. were identified; 24 signed up to participate.	DAX Copilot	Documentation as usual	**No specific primary outcome.** The following outcomes were listed: - Automated clinical documentation usage - Provider reported measures - Documentation burden (various times periods) and burnout

*Dragon Ambient eXperience Solution ®

The control group was not clearly described in any of the studies, as authors referred to "usual care" or "traditional methods". We tried to contact the authors by email but received no response regarding the papers by Kakaday or Shin. Lukac replied that the current standard is still for physicians to manually compose notes in the electronic health record. Wendt pointed out that there is no single standard and several alternatives exist.

In the study by Kakaday, the DAX representative assisted each clinician in the DAX group. In the study by Lukac, the outcome of the two AI scribes (DAX or Nabla) was only compared to the control group during the second month of intervention, i.e. the participants had one month of training to reach proficiency. For the studies by Shin and Wendt, some introductory training and support was also provided.

Regarding reported time-in-notes, Lukac noted that their study did not account for platform time, i.e. it did not include time spent editing the AI-scribe-generated draft prior to final sign-off. The time-in-notes was calculated from the total minutes a provider spent writing notes in the Epic EHR divided by the total number of notes in that week. The other studies by Kakaday, Shin and Wendt did not address this limitation, despite reporting metrics derived in the same way.

Risk of bias assessment

Risk of bias assessments are summarised in Figure 2 below.
The study by Kakaday [11] is described as an unblinded parallel randomised controlled trial, with simple random assignment without stratification. There is no information on concealment of group allocation. In table 1 there is an imbalance 1:7 for both missing age and missing gender. We find the randomisation domain is associated with high risk of bias.

Among 25 clinicians assigned to use DAX, the proportion of notes from problem-focused visits for which DAX was used ranged from 1.1% to 90.2%. Some 12/25 (48%) clinicians used DAX for <45% of all problem-focused visits and were classified as low-frequency users. 6/25 (24%) participants were classified as moderate users and 7/25 (28%) were high-frequency users. Only percentages of problem-focused visits in which DAX was used are presented, and no absolute numbers. The study participants were not blinded. We find this low uptake is associated a high risk for deviation from plan. We understand there were no dropouts among the 45 included physicians during the 3-month study period.

All primary outcomes were defined and tracked by Epic. We judge this as low risk of bias. There were no outcomes reported by the participants. The preregistration of the trial is not mentioned in the paper and the study was submitted for registration 2024-09-18; Study Details | NCT06605976 | Evaluating the Impact of Ambient AI on Documentation Efficiency and Clinician Burnout in Primary Care Settings | ClinicalTrials.gov, i.e. after the intervention period was completed (August, 2024). The authors created post hoc groups for the analysis by DAX usage. We therefore find the last domain on report is associated with high risk of bias. The overall risk of bias in this study is high.

The study by Lukac [12] used a covariate constrained random assignment based on baseline time-in-note, a single-item burnout score, and the number of self-reported clinic days per week to achieve "cohort balance" between the randomised groups. No information on concealment of allocation was found. Mainly categorical data were presented in Table 1 (see statistical review). We find the randomisation domain is associated with moderate risk of bias.

DAX was used in 8271/24,696 (33.5%) and Nabla in 6981/23,653 (29.5%) of patient visits. The study participants were not blinded. We therefore find the domain on deviation from plan is associated with high risk of bias. Approximately 15% of treatment-group physicians never used their assigned scribe. We find this is associated with moderate risk of bias based for drop-out. Measurement of the primary outcome, change from baseline log writing time-in-note, was retrieved from Epic and we find this is associated with low risk of bias. Physicians were aware of their allocation to AI or not, and we find physician-reported outcomes to be associated with high risk of bias.

It is clearly described that the study was preregistered on Study Details | NCT06792890 | A Randomized Controlled Trial of Ambient Artificial Intelligence Scribe Technologies | ClinicalTrials.gov. This work was led by a senior data scientist not included among the authors, stated under "Disclosures" at the end of the paper, The protocol was submitted on 2024-12-09, passed the quality control of the platform on 2025-01-24, and was first posted on the website of ClinicalTrials.gov on 2025-01-27, i.e. the study was in fact registered retrospectively, after completion of the inclusion period. We find this is associated with high risk of bias. The overall risk of bias of the study is considered high.

In the study by Shin et al [13], paediatricians from seven subspecialities were randomised to AI or documentation as usual. In some specialities an odd number of physicians completed the baseline survey and, in this case, a larger number of physicians were allocated to the intervention group in order to use all budgeted licences. This led to some imbalances between the intervention and control groups by subspeciality. We find this approach is associated with an unacceptably high risk of bias.

Among the group allocated to AI scribe, the usage rate of appointments during the period varied between 11 – 98%. The study participants were not blinded. We therefore find this deviation from plan is associated with high risk of bias. Of 12 participants in the intervention group, one stopped using the AI model and we find this associated with low risk of bias.

The first primary outcome listed was Net EHR Experience and Experiencing Burnout according to the KLAS EHR Experience survey. The allocation was not concealed from patients, and we find this, and other outcomes reported by the participants, to be associated with high risk of bias. Time-in-Notes per Appointment by the Epic System was judged as associated with low risk of bias. We have not been able to identify any preregistration of this specific study. In all, based on the randomisation, we find the study is associated with unacceptably high risk of bias.

The study by Wendt [14] is depicted as a "randomised, control study" with a step-wedge design. Providers were randomly assigned to an early or late implementation group. The latter was used as a control group, but some participants allocated to the late implementation group started using the AI model earlier than planned. The paper provides no information on how the random sequence was generated, or on concealment of allocation. There is no Table 1 to report the distribution between the groups. The first domain on randomisation is therefore found to be associated with unacceptably high risk of bias.

Early Implementers used ambient clinical intelligence (ACI) for 20-40% of visits. The study participants were not blinded. We find this is associated with a high risk for deviation from plan. Twenty-four participants were included and there seems to have been no dropouts, which makes for low risk of bias. There were three outcomes listed, without any hierarchy. Provider-reported measures were found associated with high risk of bias because the intervention was not blinded. Documentation time retrieved from the system were found to be at low risk of bias. We have not been able to identify any preregistration of this specific study and it is not possible to know whether the findings are selectively reported. We find this is associated with high risk of bias. Overall, based on the randomisation, we find the study is associated with unacceptably high risk of bias.

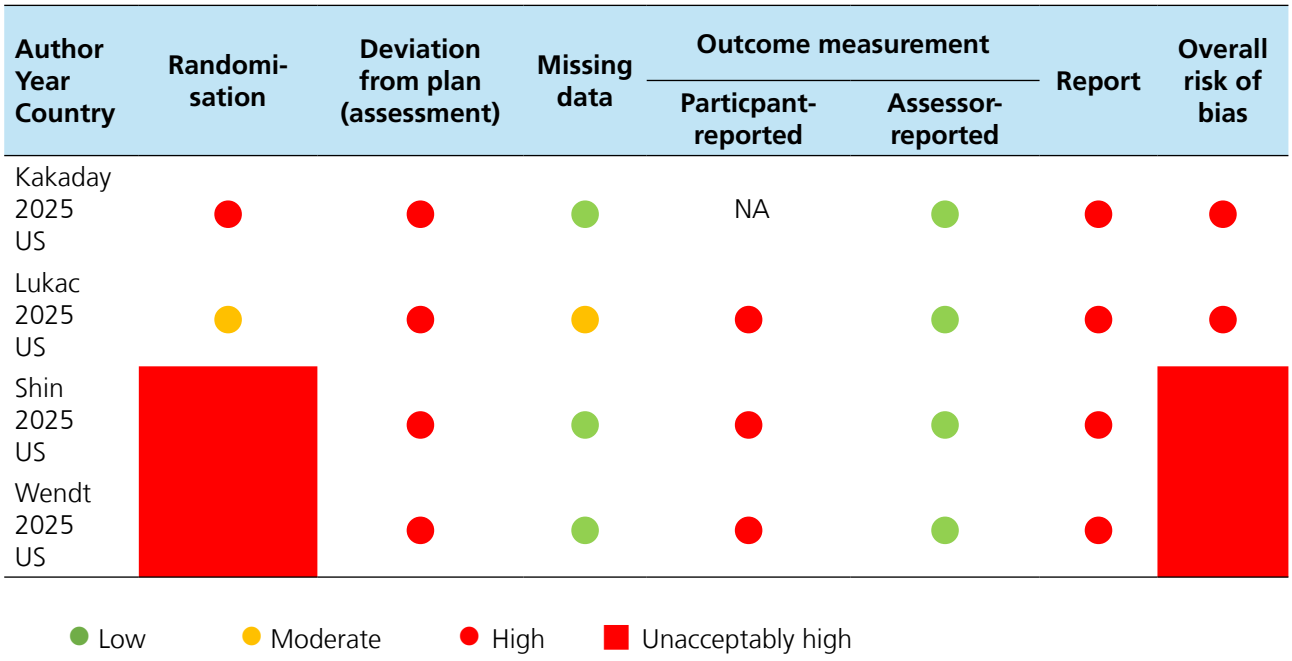


Figure 2 Risk of bias assessment of relevant studies.

Statistical review of relevant studies

The statistical review is summarised in Table 2.

The study by Kakaday [11] calculated needed sample size based on what they could afford, but presented all analyses only for subgroups and thereby reduced the statistical power. All statistical tests were limited to within-group comparisons to baseline, and no between-group comparisons were reported, despite the randomised controlled design. This approach is problematic, as testing changes from baseline separately within each group rather than comparing groups directly is known to be invalid and potentially misleading [15]. There is no mention of any correction for multiple testing, and conclusions regarding improvements are presented in the absence of statistically significant findings, and no between-group comparisons were reported, despite the randomised controlled design.

Table 2 Summary of statistical review

Author Year	Distribution of continuous baseline variables *	Planned and achieved sample size	Choice of statistical analysis	Reporting and interpretation of statistical aspects
Kakaday 2025	OK (0.10)	Effect size was based on what could be afforded 50 needed (45 randomised) Only subgroups were analyzed, reducing statistical power	No between-group testing	Only give p-values for comparisons to baseline, not between groups Draws conclusions about improvement, without any statistical significance No mention of multiplicity correction
Lukac 2025	Similar groups (0.005)	No issues noted 237 needed (238 randomised)	No issues noted	Only relative between-group effects are presented for most outcomes, not the absolute group values Likert-scale responses (ordinal data) are summarized using means instead of medians
Shin 2025	No data	Not reported - (23 randomised)	Overly complex given the sample size	No mention of multiplicity correction
Wendt 2025	OK (0.94)	Not reported - (24 randomised)	Overly complex given the sample size	No mention of multiplicity correction

* Calculation based on Carlisle's method. Interpretation: Low values (<0.05) means that groups (intervention and control) are unexpectedly similar at baseline while higher values (>0.95) means that groups are unexpectedly dissimilar.

The study by Lukac [12] had a combined p-value ≤0.05, which indicates that the groups were unexpectedly similar at baseline, but this was based on only 2 continuous variables reported as means (SD). Multiple categorical baseline variables were presented in Table 1, but, although they did not appear to be unexpectedly similar, no formal test was conducted. There are also shortcomings in the reporting of outcomes. The reduction of time-in-notes were reported as absolute values, but other outcomes were only presented with relative between-group effects, which limits transparency and interpretability. Likert-scale responses (Table 3 in the article), which are ordinal data, are summarized using means rather than medians. Because the scale consists of ordered categories, calculating averages assumes equal distances between response options, which may not be justified.

The study by Shin [13] could not be assessed using Carlisle's method due to the absence of continuous baseline variables reported with measures of dispersion (such as standard deviations). Nor was any sample size calculation reported, and no correction for multiple testing was mentioned. In addition, the study used regression methods (including difference-in-differences regression) that are likely overly com-

plex relative to the small sample size. For regression analyses it is recommended to have at least around 10 participants per parameter to reduce the risk of overfitting and unstable estimates [16].

The study by Wendt [14] did not report any sample size calculation nor any correction for multiple testing. The study used regression methods (including difference-in-differences regression) that are likely overly complex relative to the small sample size. As noted above, around 10 participants per parameter is recommended to reduce the risk of overfitting and unstable estimates.

Conflict of interest

One out of six authors declared financial conflict of interest in one study (Table 3).

Table 3 Conflict of interest as reported in relevant studies

Author, year	Authors N	Financial COI N (%) *	Employees N (%) **	Funding of the study	Professional background of authors
Kakaday 2025	5	0	0	No information on funding e.g. who funded the DAX licenses	Corresponding author: PhD
Lukac 2025	9	0	0	University of California, Los Angeles, NIH/NIA, Arnolds Venture, the Commonwealth Fund.	Reported for all authors: 6 MD, 1 PhD, 1 BS, 1 MS.
Shin 2025	6	1/6 (17%)	0	No external funding.	Corresponding author: MD
Wendt 2025	6	0	0	ACI clinical documentation software DAX® were provided by Nuance (Microsoft)	Not reported

* Proportion of authors declaring financial conflict of interest in relation to study sponsor.
** Proportion of authors employed by study sponsor.

Journals

The studies were published in three different journals (Table 4). Two of them were included in the Norwegian list.

Table 4 Summary of publishing journals

Author Year	Journal	Type of journal	DOAJ*	Predatory reports Cabells	Norwegian list**
Kakaday 2025	Applied Clinical Informatics	Hybrid	NA	No	1
Lukac 2025	NEJM AI***	Subscription	NA	No	-
Shin 2025	Applied Clinical Informatics	Hybrid	NA	No	1
Wendt 2025	Future Healthcare Journal	Open Access	Yes	No	-

* Directory of Open Access Journals
**Norwegian list is a national system that classifies academic journals and publishers into quality levels: 0 not approved; 1 the majority of legitimate, peer-reviewed scientific journals; 2 leading scientific publication channels.
***Start-up journal 2024 sponsored by Microsoft, Viz.ai, Lyric, and Elevance Health

Summary of the assessment of relevant studies

Four randomised controlled trials were identified but two of them, by Shin and by Wendt, had unacceptably high risk of bias associated with the randomisation domain and their findings are therefore not found trustworthy.

The authors of the paper by Lukac et al transparently discussed the limitation of not including platform time i.e. it is clearly stated that "the time-in-note metric does not account for time spent editing the AI scribe-generated draft within the vendors platforms." Editing time is not accounted for in any of the four studies, but only Lukac highlights it. Total documentation time including editing of the AI-derived notes was requested according to our PICO and must be acknowledged as lacking in all the available studies.

The study by Kakaday presented no data on the agreement between ambient scribes and documentation using conventional methods. The number of characters per note pre- and postintervention are reported but there is no qualitative assessment of the very content of the notes. In the study by Lukac, a post-study survey was conducted and the participants answered questions on usability, occurrence of inaccuracies and biases, and perceived risks to safety but this only involved the two AI models (DAX and Nabla). No comparison with the control group was presented. We therefore conclude we found no data for our second pre-specified outcome on agreement/accuracy of content either.

Ongoing studies

Three relevant studies were found on clinicaltrials.gov:
One ongoing study was posted 2025-09-05 and will include 105 providers and compare two AI scribes (DAX and Abridge) versus controls. Study Details | NCT07157943 | Comparison of Two Artificial Intelligence Scribe Products on Pediatric Subspecialty Provider Wellness and Experience, Patient Satisfaction, and Efficiency. | ClinicalTrials.gov. Primary outcomes are pajama time and Time in Notes per Appointment in Minutes. This study seems to be a follow-up of the Shin paper reported in this systematic review, as it will be conducted at the same clinic.

In another ongoing study from Brazil, 300 consultations across different medical specialties will be randomised to 2 groups: usual documentation (without AI) or documentation assisted by the AI scribe Study Details | NCT07302906 | Ambient AI Scribe (Voa Health) in Outpatient Clinics: Draft Notes, Documentation Burden, and Well-Being | ClinicalTrials.gov Primary outcome will be physician workload and physician well-being. The study is estimated to be completed in June, 2026.

Another study not yet recruiting is registered from Canada on family medicine or any speciality Study Details | NCT07113938 | Assessing the Efficacy and Impact of Ambient AI Scribes in Healthcare | ClinicalTrials.gov It will enrol 64 physicians and compare those using the ambient AI scribes to a group not using ambient AI scribes. Primary outcomes are workload, burnout and quality of patient-physician interaction. The trial was registered in August 2025 but is not yet recruiting according to the information on clinicaltrials.gov.

No relevant studies were found in the ISRCTN (UK) or ICTRP (WHO) databases. Eleven systematic reviews on AI scribes with different focuses are registered on PROSPERO (Appendix 4), but none of them are expected to be completed within soon.

Discussion

Four RCTs were identified as relevant for this systematic review but two of them had to be excluded due to serious problems with randomization. In fact, this was indicated already in their titles which included terms like “pilot study”, and “research article”, respectively.

The participants in all four studies were self-selected, recruited through recommendations, or had problems with lengthy documentation times from the outset. The generalisability of assessing new documentation technology in such groups is therefore questionable. Another serious shortcoming is that documentation methods in the control groups were not clearly described. Finally, we found no data on the outcomes – total time in notes and accuracy of the AI-generated drafts - predefined in our PICO.

For this our first systematic review of ambient scribes, we found these two outcomes to be most critical, but it does not imply that other outcomes are unimportant. Both time and accuracy can be assessed by blinded assessors, and this was considered a significant advantage for evaluating new technology. We noticed that several studies in the field report on outcomes like physician burnout and exhaustion but, as the participants are not blinded to the intervention, the placebo or nocebo effects should not be underestimated.

Nevertheless, when planning this systematic review, we did not anticipate substantial variation in uptake or compliance and therefore overlooked this outcome, which is highly relevant from both clinical and health-economic perspectives. In the small study by Kakaday though, 48% of participants were classified as low users. In the study by Lukac, compliance was approximately 30% for both AI models (p. 17), and the resulting reduction of time-in-notes *(editing of AI-drafts not included),* calculated according to intention-to-treat, turned out to be minimal (Appendix 3). Uptake, or compliance, must therefore be considered a key outcome in future research on ambient scribes.

Only RCTs were eligible for this systematic review, which may be considered as somewhat overly stringent. But outpatient visits for study purposes are abundant, and RCTs remain the gold standard for evaluating new methods. We found no studies assessing the accuracy of ambient scribes versus conventional documentation, and this is likely related to the request for RCTs. The extent to which patient safety aspects, including AI hallucinations, omissions or other AI-related errors have been investigated using other study designs is unknown to us and this is a limitation of this systematic review. But importantly, we did not identify any RCT evaluating any broader impact of ambient scribes, involving decision-making on management, diagnosis, treatment, use of healthcare resources, or patient outcomes, including mortality. For other potentially more far-reaching effects of using AI for clinical notes, e.g. cognitive benefits of physicians documenting clinical encounters themselves, other study designs will be needed.

The primary economic rationale for introducing ambient scribes is to reduce reliance on medical secretaries. However, if reductions in documentation time for physicians are uncertain, very small or even negative with a net increase in physicians’ time spent on documentation and more administrative tasks, the expected cost savings will quickly diminish. Moreover, any anticipated increase in spent time with patients will fail to materialize. This highlights the importance of rigorously evaluating the effects of

introducing new technologies such as ambient scribes on time use, workflow and task shifting.

The usefulness of ambient scribes may vary across clinical settings, e.g. for patients presenting to the ED who are subsequently admitted to a ward requiring immediate access to information, compared to routine follow-up of chronic disease, or between narrowly focused clinical encounters and complex psychiatric evaluations. It was therefore disappointing to notice that the number of ongoing pre-registered clinical trials is very small.

Knowledge gaps

Based on this systematic review, the following knowledge gaps were identified

- There is a lack of clinical trials comparing the total time including editing and corrections spent on documentation by physicians using ambient scribes compared to clearly described control interventions in outpatient care.
- There is a lack of clinical trials involving unselected participants (physicians, or others).
- There is a lack of clinical trials comparing the wider clinical effects of using ambient scribes for documentation, such as on patient safety and use of resources. This limitation is partly related to uncertainties regarding the accuracy of AI-generated notes.

Ethical reflections

150 years of uninterrupted medical progress have created expectations for a rapid pace of change. The economist Joseph Schumpeter described the economic system of the 20th century as characterized by "creative destruction." This observation is also relevant for modern medicine. Both the public and health professionals not only hope for continual medical breakthroughs but often take them for granted, assuming that, aside from occasional blind alleys, every new medical development is an improvement over previous methods.

This mindset creates a field of inertia for critical reflection on change. A fundamental component of evidence-based medicine is to maintain an attitude of open inquiry regarding the introduction of new methods, treatments, or organizational structures. In a manner similar to legal practice, new approaches should be assumed ineffective until proven otherwise.

The digital revolution has contributed to numerous advances in healthcare, including easier access to relevant information, more precise data handling, and the ability to conduct clinical encounters remotely. As is often the case, however, these advantages can be misapplied and produce unintended negative consequences. In retrospect, many would likely agree that digitalization often proceeded too rapidly, without sufficient critical reflection.

We are now witnessing what is frequently called an "AI revolution." AI is being rapidly introduced across multiple areas, including healthcare, where enthusiastic proponents drive change. A lack of time—or rather, a subjective sense of time pressure in daily clinical work—creates an incentive to adopt supposedly time-saving solutions, and the introduction of AI is often viewed through this lens. One such area is clinical documentation, a component of clinical work that inevitably consumes substantial time. Ambient scribes that convert spoken communication during clinical encounters directly into clinical notes for clinician review and signature are the focus of this report.

From an ethical perspective, ambient scribes in clinical settings must meet at least four criteria:

1. **Accuracy:** The transcribed text must be formally correct and faithfully represent what was said.
2. **Relevance:** The selection of information included in the medical note must be pertinent and balanced.
3. **Patient experience:** AI transcription must not undermine patient trust, compromise security, or impede dialogue.
4. **Time efficiency:** AI should deliver a genuine net gain in time without compromising the other three criteria.

Several challenges complicate these requirements:
The first point is rather technical than ethical, and depends on how techniques may be developed, but has implications for the ethical standard. Clinical dialogue in the Swedish healthcare system may be conducted in accent-free Swedish but often involves strong accents, either of the clinician or patient. Moreover, both parties—particularly the patient—may speak incoherently, interrupt themselves, or produce fragmented sentences. Ambient scribe systems face the challenge to handle such variability without demanding extensive physician editing.

Secondly, clinical dialogue cannot—and should not—be fully reproduced in medical reports. Physicians select the parts relevant to the overall clinical picture, often adding preliminary assessments not explicitly expressed during the dialogue. This necessitates additional report corrections, making it difficult to reconstruct the complete context accurately.

Furthermore, patients must be informed about AI transcription and data transfer. Responses may range from indifference to fascination, uncertainty, or discomfort. While some may refuse, confidence and assertiveness to do so are unlikely to be widespread among patients. These circumstances suggest that the likelihood of ambient scribes actually saving time is low, particularly when considering the risks described above.

The four studies analyzed in this report do not provide sufficient evidence to support the rapid introduction of these AI systems. Clinicians under time pressure, hoping to avoid dictation, may be tempted to approve inaccurate transcriptions too quickly. Our findings do not alleviate this concern. Further research needs to focus also on these considerations and not only on the possible time gain, in order to evaluate the full impact of ambient scribes.

The ethical values at stake include the patient's integrity during clinical encounters and the healthcare system's obligation to do good and avoid harm. Achieving these goals requires dialogue that is carefully compressed and selectively documented, based on trusting patient–clinician interactions.

References

1. Retraction Watch Database [cited 2025 April 30]. Available from: http://retractiondatabase.org/RetractionSearch.aspx?
2. Statens beredning för medicinsk och social utvärdering SBU. Manual för bedömning av risk för bias i randomiserade kontrollerade studier. SBU:s översättning och bearbetning av Cochranes RoB-2. 2024 [cited 2025 April 30]. Available from: https://www.sbu.se/globalassets/ebm/bedomning-av-risk-for-bias-_rct.pdf.
3. Carlisle JB, Dexter F, Pandit JJ, Shafer SL, Yentis SM. Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. Anaesthesia 2015; 70: 848-58. doi:https://10.1111/anae.13126
4. DOAJ Foundation. Directory of Open Access Journals (DOAJ) [cited 2025 May 05]. Available from: https://doaj.org/.
5. Cabells-Predatory Reports [Available from: https://cabells.com/solutions/
6. Norwegian Register for Scientific Journals SaP. Norska listan: The Norwegian Ministry of Education and Research; [cited 2025 April 30]. Available from: https://kanalregister.hkdir.no/en.
7. National Library of Medicine. ClinicalTrials [cited 2025 May 05]. Available from: https://clinicaltrials.gov
8. The UK's Clinical Study Registry. International Standard Randomised Controlled Trial Number ISRCTN [cited 2025 May 05]. Available from: https://www.isrctn.com/.
9. World Health Organization. International Clinical Trials Registry Platform, [cited 2026 Febr, 18]. Available from: https://www.who.int/tools/clinical-trials-registry-platform.
10. University of York: Centre for Reviews and Dissemination. International prospective register of systematic reviews Prospero [cited 2025 May 05]. Available from: https://www.crd.york.ac.uk/prospero/.
11. Kakaday R, Herrera EZ, Coskey O, Hertel AW, Kaiser P. The STREAMLINE Pilot Study on Time Reduction and Efficiency in AI-Mediated Logging for Improved Note-Taking Experience. Applied clinical informatics 2025; 16: 614-21. doi:https://dx.doi.org/10.1055/a-2559-5791
12. Lukac PJ, Turner W, Vangala S, Chin AT, Khalili J, Shih YT, et al. Ambient AI Scribes in Clinical Practice: A Randomized Trial. NEJM AI 2025; 2. doi:https://10.1056/aioa2501000
13. Shin HS, Williams H, Braykov N, Jahan A, Meller J, Orenstein EW. The Influence of Artificial Intelligence Scribes on Clinician Experience and Efficiency among Pediatric Subspecialists: A Rapid, Randomized Quality Improvement Trial. Applied clinical informatics 2025; 16: 1041-52. doi:https://dx.doi.org/10.1055/a-2657-8087
14. Wendt SJ, Dinh CT, Sutcliffe M, Jones K, Scanlan JM, Smitherman JS. Deploying ambient clinical intelligence to improve care: A research article assessing the impact of nuance DAX on documentation burden and burnout. Future healthcare journal 2025; 12: 100450. doi:https://dx.doi.org/10.1016/j.fhj.2025.100450
15. Bland JM, Altman DG. Comparisons against baseline within randomised groups are often used and can be highly misleading. Trials 2011; 12: 264. doi:https://10.1186/1745-6215-12-264
16. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004; 66: 411-21. doi:https://10.1097/01.psy.0000127692.23278.a9

Appendices

Appendix 1 Literature search

Database: Database(s): Ovid MEDLINE(R) ALL 1946 to December 03, 2025
Host: Ovid
Date searched: 2025-12-04

Limits applied: Publications in English language, Publication Year from 2020 to 2025

Field Codes: /:Mesh-term; exp: Exploded Mesh-term; ab: Abstract; kf: Keyword heading word; ti: Title;

Concept	#	Search	Results
Medical Record	1	exp Medical Records/ or Nursing Records/	174081
	2	("Medical Record*" or "Medical Documentation*" or "Medical Note" or "Medical Notes" or "Medical Transcription*" or "Medical Diar*" or "Clinical Record*" or "Clinical Documentation*" or "Clinical Note" or "Clinical Notes" or "Clinical Transcription*" or "Clinical Diar*" or "Patient Record*" or "Patient Documentation*" or "Patient Note" or "Patient Notes" or "Patient Transcription*" or "Patient Diar*" or "Health Record*" or "Health Documentation*" or "Health Diar*" or "Healthcare Record*" or "Healthcare Documentation*" or "Healthcare Note" or "Healthcare Notes" or "Healthcare Transcription*" or "Healthcare Diar*" or "Health Care Record*" or "Health Care Documentation*" or "Health Care Note" or "Health Care Notes" or "Health Care Transcription*" or "Health Care Diar*" or "Hospital Record*" or "Hospital Documentation*" or "Hospital Note" or "Hospital Notes" or "Hospital Transcription*" or "Hospital Diar*" or "Nurs* Record*" or "Nurs* Documentation*" or "Nurs* Note" or "Nurs* Notes" or "Nurs* Transcription*" or "Nurs* Diar*" or "Psychiatric Record*" or "Psychiatric Documentation*" or "Psychiatric Note" or "Psychiatric Notes" or "Psychiatric Transcription*" or "Psychiatric Diar*" or "Operative Record*" or "Operative Documentation*" or "Operative Note" or "Operative Notes" or "Operative Transcription*" or "Operative Diar*" or "Medical Dictation*" or "Clinical Dictation*" or "Patient Dictation*" or "Hospital Dictation*" or "Nurs* Dictation*" or "Psychiatric Dictation*" or "Medical Chart*" or "Clinical Chart*" or "Patient Chart*" or "Hospital Chart*" or "Nurs* Chart*" or "Psychiatric Chart*" or "EHR" or "Discharge summar*" or "Forms and Records Control").ab,kf,ti.	297572
	3	1 or 2	424380
Artificial Intelligence	4	exp Artificial Intelligence/ or exp Algorithms/	540710
	5	("Artificial Intel*" or "AI" or Generative or Generating or "Algorithm*" or "algorhithm*" or "algorhythm*" or "algorythm*" or "algorism*" or "Computer Reasoning" or "Machine Intel*" or "Machine Training" or "Computational Intel*" or "Computer Vision System*" or "Machine Learning" or "Deep Learning" or "Natural Language Process*" or "Neural Network" or "Large Language Model*" or "LLM" or "LLMs" or "Chatbot*" or "Chat GPT" or Gemini or Copilot or Claude or Bard or "Le Chat" or tandem or nuance or Whisper).ab,kf,ti.	1088799
	6	4 or 5	1314948

Voice or Speech	7	exp Voice/ or exp Speech/	53787
	8	(voice or vox or speech or transcription* or scribe* or Dictation* or audio or listening or recording or microphone* or speaker*).ab,kf,ti.	1218267
	9	7 or 8	1230131
Combined Sets	10	3 and 6 and 9	779
Voice or Speech recognition	11	exp Voice Recognition/ or exp Speech Recognition Software/ or Speech Perception/	32214
	12	(Speech Recognition* or voice Recognition* or Speech-to-Text or Voice-to-Text or Speech Perception* or Voice Perception*).ab,kf,ti.	15489
	13	11 or 12	38569
Combined Sets	14	3 and 13	569
Digital scribe	15	((digital or AI or Artificial or Intel* or virtual or medical) adj5 scribe*).ab,kf,ti.	291
Ambient scribe	16	((Scribe* or Digital or AI or Artificial or Intel* or Virtual or Medical or Note* or Documentation* or Record* or Transcription* or Chart* or Technolog*) adj4 Ambient).ab,kf,ti.	1288
Product names	17	(Omilon or "Dragon Medical One" or DeepScribe or "Amazon Transcribe Medical" or Leapscribe or "Accurx Scribe" or "tandem health" or "Freed AI" or Lindy or "Heidi AI" or Speechmatics or "3M*Modal" or "Express Scribe").ab,kf,ti.	25
Combined Sets	18	10 or 14 or 15 or 16 or 17	2671
RCT-filter*	19	exp randomized controlled trial/	653951
	20	controlled clinical trial.pt.	95766
	21	randomized.ab.	721722
	22	placebo.ab.	264627
	23	clinical trials as topic.sh.	206295
	24	randomly.ab.	475365
	25	trial.ti.	353257
	26	or/19-25	1735575
	27	exp animals/ not humans/	5402135
	28	26 not 27	1601583
Combined Sets	29	18 and 28	121
	30	limit 29 to (english language and yr="2020 -Current")	48

*Line 19-28 is **Cochrane Highly Sensitive Search Strategy for identifying randomized trials in MEDLINE: sensitivity- and precision-maximizing version (2023 revision); Ovid format** Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Metzendorf M-I, Noel-Storr A, Paynter R, Rader T, Thomas J, Wieland LS. Technical Supplement to Chapter 4: Searching for and selecting studies [last updated September 2024]. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editor(s). Cochrane Handbook for Systematic Reviews of Interventions version 6.5. Cochrane, 2024. Available from: cochrane.org/handbook.

Database: Embase
Host: Embase.com
Date searched: 2025-12-04

Limits applied: Publications in English language, Publication Year from 2020 to 2025, No Conference abstracts

Field Codes: /de: Emtree term; /exp: Exploded Emtree term; ab: Abstract; kw: Author Keyword; ti: Title; tt: Original non-English title

Concept	#	Search	Results
Medical Record	#1	'medical record'/exp	409198
	#2	'medical record*':ab,kw,ti OR 'medical documentation*':ab,kw,ti OR 'medical note':ab,kw,ti OR 'medical notes':ab,kw,ti OR 'medical transcription*':ab,kw,ti OR 'medical diar*':ab,kw,ti OR 'clinical record*':ab,kw,ti OR 'clinical documentation*':ab,kw,ti OR 'clinical note':ab,kw,ti OR 'clinical notes':ab,kw,ti OR 'clinical transcription*':ab,kw,ti OR 'clinical diar*':ab,kw,ti OR 'patient record*':ab,kw,ti OR 'patient documentation*':ab,kw,ti OR 'patient note':ab,kw,ti OR 'patient notes':ab,kw,ti OR 'patient transcription*':ab,kw,ti OR 'patient diar*':ab,kw,ti OR 'health record*':ab,kw,ti OR 'health documentation*':ab,kw,ti OR 'health diar*':ab,kw,ti OR 'healthcare record*':ab,kw,ti OR 'healthcare documentation*':ab,kw,ti OR 'healthcare note':ab,kw,ti OR 'healthcare notes':ab,kw,ti OR 'healthcare transcription*':ab,kw,ti OR 'healthcare diar*':ab,kw,ti OR 'health care record*':ab,kw,ti OR 'health care documentation*':ab,kw,ti OR 'health care note':ab,kw,ti OR 'health care notes':ab,kw,ti OR 'health care transcription*':ab,kw,ti OR 'health care diar*':ab,kw,ti OR 'hospital record*':ab,kw,ti OR 'hospital documentation*':ab,kw,ti OR 'hospital note':ab,kw,ti OR 'hospital notes':ab,kw,ti OR 'hospital transcription*':ab,kw,ti OR 'hospital diar*':ab,kw,ti OR 'nurs* record*':ab,kw,ti OR 'nurs* documentation*':ab,kw,ti OR 'nurs* note':ab,kw,ti OR 'nurs* notes':ab,kw,ti OR 'nurs* transcription*':ab,kw,ti OR 'nurs* diar*':ab,kw,ti OR 'psychiatric record*':ab,kw,ti OR 'psychiatric documentation*':ab,kw,ti OR 'psychiatric note':ab,kw,ti OR 'psychiatric notes':ab,kw,ti OR 'psychiatric transcription*':ab,kw,ti OR 'psychiatric diar*':ab,kw,ti OR 'operative record*':ab,kw,ti OR 'operative documentation*':ab,kw,ti OR 'operative note':ab,kw,ti OR 'operative notes':ab,kw,ti OR 'operative transcription*':ab,kw,ti OR 'operative diar*':ab,kw,ti OR 'medical dictation*':ab,kw,ti OR 'clinical dictation*':ab,kw,ti OR 'patient dictation*':ab,kw,ti OR 'hospital dictation*':ab,kw,ti OR 'nurs* dictation*':ab,kw,ti OR 'psychiatric dictation*':ab,kw,ti OR 'medical chart*':ab,kw,ti OR 'clinical chart*':ab,kw,ti OR 'patient chart*':ab,kw,ti OR 'hospital chart*':ab,kw,ti OR 'nurs* chart*':ab,kw,ti OR 'psychiatric chart*':ab,kw,ti OR 'ehr':ab,kw,ti OR 'discharge summar*':ab,kw,ti OR 'forms and records control':ab,kw,ti	526176
	#3	#1 OR #2	665172
Artificial Intelligence	#4	'artificial intelligence'/exp OR 'algorithm'/exp	897035
	#5	'artificial intel*':ab,kw,ti OR 'ai':ab,kw,ti OR generative:ab,kw,ti OR generating:ab,kw,ti OR 'algorithm*':ab,kw,ti OR 'algorhithm*':ab,kw,ti OR 'algorythm*':ab,kw,ti OR 'algorhythm*':ab,kw,ti OR 'algorism*':ab,kw,ti OR 'computer reasoning':ab,kw,ti OR 'machine intel*':ab,kw,ti OR 'machine training':ab,kw,ti OR 'computational intel*':ab,kw,ti OR 'computer vision system*':ab,kw,ti OR 'machine learning':ab,kw,ti OR 'deep learning':ab,kw,ti OR 'natural language process*':ab,kw,ti OR 'neural network':ab,kw,ti OR 'large language model*':ab,kw,ti OR 'llm':ab,kw,ti OR 'llms':ab,kw,ti OR 'chatbot*':ab,kw,ti OR 'chat gpt':ab,kw,ti OR gemini:ab,kw,ti OR copilot:ab,kw,ti OR claude:ab,kw,ti OR bard:ab,kw,ti OR 'le chat':ab,kw,ti OR tandem:ab,kw,ti OR nuance:ab,kw,ti OR whisper:ab,kw,ti	1354061
	#6	#4 OR #5	1678484

Voice or Speech	#7	'voice'/exp OR 'speech'/exp	162745
	#8	voice:ab,kw,ti OR vox:ab,kw,ti OR speech:ab,kw,ti OR transcription*:ab,kw,ti OR scribe*:ab,kw,ti OR dictation*:ab,kw,ti OR audio:ab,kw,ti OR listening:ab,kw,ti OR recording:ab,kw,ti OR microphone*:ab,kw,ti OR speaker*:ab,kw,ti	1545844
	#9	#7 OR #8	1607068
	#10	#3 AND #6 AND #9	1448
Voice or Speech recognition	#11	'voice recognition'/exp OR 'automatic speech recognition'/exp OR 'speech perception'/exp	33443
	#12	'speech recognition*':ab,kw,ti OR 'voice recognition*':ab,kw,ti OR 'speech-to-text':ab,kw,ti OR 'voice-to-text':ab,kw,ti OR 'speech perception*':ab,kw,ti OR 'voice perception*':ab,kw,ti	17529
	#13	#11 OR #12	37219
Combined Sets	#14	#3 AND #13	762
Digital scribe	#15	((digital OR ai OR artificial OR intel* OR virtual OR medical) NEAR/5 scribe*):ab,kw,ti	367
Ambient scribe	#16	(scribe* OR digital OR ai OR artificial OR intel* OR virtual OR medical OR note* OR documentation* OR record* OR transcription* OR chart* OR technolog*) NEAR/4 ambient	1537
Product names	#17	omilon:ab,kw,ti OR 'dragon medical one':ab,kw,ti OR deepscribe:ab,kw,ti OR 'amazon transcribe medical':ab,kw,ti OR leapscribe:ab,kw,ti OR 'accurx scribe':ab,kw,ti OR 'tandem health':ab,kw,ti OR 'or freed ai':ab,kw,ti OR lindy:ab,kw,ti OR 'heidi ai':ab,kw,ti OR speechmatics:ab,kw,ti OR '3m*modal':ab,kw,ti OR 'express scribe':ab,kw,ti	36
Combined Sets	#18	#10 OR #14 OR #15 OR #16 OR #17	3789
RCT-filter*	#19	'randomized controlled trial'/exp	1130209
	#20	'controlled clinical trial'/de	460979
	#21	random*:ti,ab,tt	2544281
	#22	'randomization'/de	101763
	#23	'intermethod comparison'/de	317911
	#24	placebo:ti,ab,tt	471911
	#25	compare:ti,tt OR compared:ti,tt OR comparison:ti,tt	721369
	#26	(evaluated:ab OR evaluate:ab OR evaluating:ab OR assessed:ab OR assess:ab) AND (compare:ab OR compared:ab OR comparing:ab OR comparison:ab)	3412631
	#27	(open NEXT/1 label):ti,ab,tt	192043
	#28	((double OR single OR doubly OR singly) NEXT/1 (blind OR blinded OR blindly)):ti,ab,tt	375003
	#29	'double blind procedure'/de	314352
	#30	(parallel NEXT/1 group*):ti,ab,tt	53979
	#31	crossover:ti,ab,tt OR 'cross over':ti,ab,tt	158877
	#32	((assign* OR match OR matched OR allocation) NEAR/6 (alternate OR group OR groups OR intervention OR interventions OR patient OR patients OR subject OR subjects OR participant OR participants)):ti,ab,tt	574210
	#33	assigned:ti,ab,tt OR allocated:ti,ab,tt	620491
	#34	(controlled NEAR/8 (study OR design OR trial)):ti,ab,tt	660440
	#35	volunteer:ti,ab,tt OR volunteers:ti,ab,tt	327179
	#36	'human experiment'/de	737469
	#37	trial:ti,tt	581319

	#38	#19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26 OR #27 OR #28 OR #29 OR #30 OR #31 OR #32 OR #33 OR #34 OR #35 OR #36 OR #37	7704175
	#39	((random* NEXT/1 sampl* NEAR/8 ('cross section*' OR questionnaire* OR survey OR surveys OR database OR databases)):ti,ab,tt) NOT ('comparative study'/de OR 'controlled study'/de OR 'randomised controlled':ti,ab,tt OR 'randomized controlled':ti,ab,tt OR 'randomly assigned':ti,ab,tt)	3874
	#40	'cross-sectional study' NOT ('randomized controlled trial'/exp OR 'controlled clinical trial'/de OR 'controlled study'/de OR 'randomised controlled':ti,ab,tt OR 'randomized controlled':ti,ab,tt OR 'control group':ti,ab,tt OR 'control groups':ti,ab,tt)	506115
	#41	'case control*':ti,ab,tt AND random*:ti,ab,tt NOT ('randomised controlled':ti,ab,tt OR 'randomized controlled':ti,ab,tt)	24853
	#42	'systematic review':ti,tt NOT (trial:ti,tt OR study:ti,tt)	352693
	#43	nonrandom*:ti,ab,tt NOT random*:ti,ab,tt	21571
	#44	'random field*':ti,ab,tt	3262
	#45	('random cluster' NEAR/4 sampl*):ti,ab,tt	1826
	#46	review:ab AND review:it NOT trial:ti,tt	1353700
	#47	'we searched':ab AND (review:ti,tt OR review:it)	61159
	#48	'update review':ab	156
	#49	(databases NEAR/5 searched):ab	86689
	#50	(rat:ti,tt OR rats:ti,tt OR mouse:ti,tt OR mice:ti,tt OR swine:ti,tt OR porcine:ti,tt OR murine:ti,tt OR sheep:ti,tt OR lambs:ti,tt OR pigs:ti,tt OR piglets:ti,tt OR rabbit:ti,tt OR rabbits:ti,tt OR cat:ti,tt OR cats:ti,tt OR dog:ti,tt OR dogs:ti,tt OR cattle:ti,tt OR bovine:ti,tt OR monkey:ti,tt OR monkeys:ti,tt OR trout:ti,tt OR marmoset*:ti,tt) AND 'animal experiment'/de	1329105
	#51	'animal experiment'/de NOT ('human experiment'/de OR 'human'/de)	2810687
	#52	#39 OR #40 OR #41 OR #42 OR #43 OR #44 OR #45 OR #46 OR #47 OR #48 OR #49 OR #50 OR #51	5042096
	#53	#38 NOT #52	6799757
Combined Sets	#54	#18 AND #53	977
	#55	#18 AND #53 AND [2020-2026]/py AND [english]/lim	500
	#56	#56 NOT 'conference abstract'/it	363

*Line 19-53 is **Embase RCT filter for Embase.com** 30 April 2023 revision
Glanville J. EMBASE RCT filter, The Cochrane Embase RCT filters for Embase.com and Ovid (2023 revisions) are provided on this page. ISSG filter resource; [cited 2025 Dec 4]. Available from: https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home/rcts/embase-rct-filter

Database: Cochrane Library
Host: Wiley
Date searched: 2025-12-04

Limits applied: Publications in English language, Publication Year from 2020 to 2025
Field Codes: MeSH descriptor: [] explode all trees: Exploded Mesh-term; MeSH descriptor: [] this term only: Unexploded Mesh-term; ab: Abstract; kw: Keywords; ti: Title

Concept	#	Search	Results
Medical Record	#1	MeSH descriptor: [Medical Records] explode all trees	3976
	#2	MeSH descriptor: [Nursing Records] this term only	48
	#3	(Medical NEXT Record* or Medical NEXT Documentation* or "Medical Note" or "Medical Notes" or Medical NEXT Transcription* or Medical NEXT Diar* or Clinical NEXT Record* or Clinical NEXT Documentation* or "Clinical Note" or "Clinical Notes" or Clinical NEXT Transcription* or Clinical NEXT Diar* or Patient NEXT Record* or Patient NEXT Documentation* or "Patient Note" or "Patient Notes" or Patient NEXT Transcription* or Patient NEXT Diar* or Health NEXT Record* or Health NEXT Documentation* or Health NEXT Diar* or Healthcare NEXT Record* or Healthcare NEXT Documentation* or "Healthcare Note" or "Healthcare Notes" or Healthcare NEXT Transcription* or Healthcare NEXT Diar* or Health NEXT Care NEXT Record* or Health NEXT Care NEXT Documentation* or "Health Care Note" or "Health Care Notes" or Health NEXT Care NEXT Transcription* or Health NEXT Care NEXT Diar* or Hospital NEXT Record* or Hospital NEXT Documentation* or "Hospital Note" or "Hospital Notes" or Hospital NEXT Transcription* or Hospital NEXT Diar* or Nurs* NEXT Record* or Nurs* NEXT Documentation* or Nurs* NEXT Note or Nurs* NEXT Notes or Nurs* NEXT Transcription* or Nurs* NEXT Diar* or Psychiatric NEXT Record* or Psychiatric NEXT Documentation* or "Psychiatric Note" or "Psychiatric Notes" or Psychiatric NEXT Transcription* or Psychiatric NEXT Diar* or Operative NEXT Record* or Operative NEXT Documentation* or "Operative Note" or "Operative Notes" or Operative NEXT Transcription* or Operative NEXT Diar* or Medical NEXT Dictation* or Clinical NEXT Dictation* or Patient NEXT Dictation* or Hospital NEXT Dictation* or Nurs* NEXT Dictation* or Psychiatric NEXT Dictation* or Medical NEXT Chart* or Clinical NEXT Chart* or Patient NEXT Chart* or Hospital NEXT Chart* or Nurs* NEXT Chart* or Psychiatric NEXT Chart* or "EHR" or Discharge NEXT summar* or "Forms and Records Control"):ti,ab,kw	28182
	#4	#1 or #2 or #3	29904
Artificial Intelligence	#5	MeSH descriptor: [Artificial Intelligence] explode all trees	3835
	#6	MeSH descriptor: [Algorithms] explode all trees	8702
	#7	(Artificial NEXT Intel* or AI or Generative or Generating or Algorithm* or algorhithm* or algorythm* or algorhythm* or algorism* or "Computer Reasoning" or Machine NEXT Intel* or "Machine Training" or Computational NEXT Intel* or Computer NEXT Vision NEXT System* or "Machine Learning" or "Deep Learning" or Natural NEXT Language NEXT Process* or "Neural Network" or Large NEXT Language NEXT Model* or "LLM" or "LLMs" or Chatbot* or "Chat GPT" or Gemini or Copilot or Claude or Bard or "Le Chat" or tandem or nuance or Whisper):ti,ab,kw	42942
	#8	#5 or #6 or #7	44436

Voice or Speech	#9	MeSH descriptor: [Voice] explode all trees	629
	#10	MeSH descriptor: [Speech] explode all trees	1300
	#11	(voice or vox or speech or transcription* or scribe* or Dictation* or audio or listening or recording or microphone* or speaker*):ti,ab,kw	53393
	#12	#9 or #10 or #11	53393
Combined Sets	#13	#4 and #8 and #12	131
Voice or Speech recognition	#14	MeSH descriptor: [Voice Recognition] explode all trees	1
	#15	MeSH descriptor: [Speech Recognition Software] explode all trees	33
	#16	MeSH descriptor: [Speech Perception] this term only	994
	#17	(Speech NEXT Recognition* or voice NEXT Recognition* or Speech-to-Text or Voice-to-Text or Speech NEXT Perception* or Voice NEXT Perception*):ti,ab,kw	1554
	#18	#14 or #15 or #16 or #17	1554
Combined Sets	#19	#4 and #18	38
Digital scribe	#20	((digital or AI or Artificial or Intel* or virtual or medical) NEAR/5 scribe*):ti,ab,kw	26
Ambient scribe	#21	(Scribe* or Digital or AI or Artificial or Intel* or Virtual or Medical or Note* or Documentation* or Record* or Transcription* or Chart* or Technolog*) NEAR/4 Ambient	97
Product names	#22	(Omilon or "Dragon Medical One" or DeepScribe or "Amazon Transcribe Medical" or Leapscribe or "Accurx Scribe" or "tandem health" or "Freed AI" or Lindy or "Heidi AI" or Speechmatics or "3M NEXT Modal" or "Express Scribe"):ti,ab,kw	3
Combined Sets	#23	#13 or #19 or #21 or #20 or #22	281
	#24	#23 English language, Publication Year from 2020 to 2025 in Cochrane Central Register of Controlled Trials	146

Appendix 2 Excluded publications after full-text review, with reasons.

	Year	Publication	Reason for exclusion
1	2025	Arko Iv L, Hudelson C, Kumar J, Badlani S, Stoffel M, Markowitz R, et al. Documenting Care with AI: A Comparative Analysis of Commercial Scribe Tools. Studies in health technology and informatics 2025; 329: 337-41. doi:https://10.3233/SHTI250857	**Wrong population** Not real patients
2		Hopkins BS, Dallas J, Yu J, Briggs RG, Chung LK, Cote DJ, et al. The use of generative artificial intelligence-based dictation in a neurosurgical practice: a pilot study. Neurosurgical focus 2025; 59: E8. doi:https://10.3171/2025.4.FOCUS24834	**Wrong focus** Operative reports
3		Luo M-J, Bi S, Pang J, Liu L, Tsui C-K, Lai Y, et al. A large language model digital patient system enhances ophthalmology history taking skills. NPJ digital medicine 2025; 8: 502. doi:https://dx.doi.org/10.1038/s41746-025-01841-6	**Not relevant** Educational purpose
4		Morey J, Jones D, Walker L, Lindor R, Schupbach J, Mullan A, et al. Ambient Artificial Intelligence Versus Human Scribes in the Emergency Department. Annals of emergency medicine 2025. doi:https://10.1016/j.annemergmed.2025.10.006	**Wrong study design** Not randomised
5		Morey J, Jones D, Walker L, Mullan A, Heaton H. 206 Ambient Artificial Intelligence Versus Human Scribes in the Emergency Department. 2025; 86: S88. doi:https://10.1016/j.annemergmed.2025.06.222	**Wrong publication type** Abstract
6		Olatunji T, Aka C, Okocha C, Katuka G, Tassallah A, Etori N, et al. A multi-country study comparing typed to automatic speech recognition-based medical documentation speeds among Low- and Middle-Income Country Trained Clinicians. medRxiv2025.	**Not relevant** No AI involved
7		Pelletier JH, Watson K, Michel J, McGregor R, Rush SZ. Effect of a generative artificial intelligence digital scribe on pediatric provider documentation time, cognitive burden, and burnout. JAMIA Open 2025; 8. doi:https://10.1093/jamiaopen/ooaf068	**Wrong study design** Not randomised
8		Toussi N, Zhang C, Kang J, Licitra EJ. Impact of AI medical scribes on physician productivity and satisfaction in medical oncology. 2025; 43. doi:https://www.cochranelibrary.com/central/doi/10.1002/central/CN-02920380/full	**Wrong publication type** Abstract
9	2024	Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial. The Journal of the American Academy of Orthopaedic Surgeons 2024; 32:123-9. doi:https://dx.doi.org/10.5435/JAAOSD-23-00474	**Wrong population** Not real patients: Ten standardized patient histories were generated
10		Balloch J, Sridharan S, Oldham G, Wray J, Gough P, Robinson R, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthcare Journal 2024; 11. doi:https://10.1016/j.fhj.2024.100157	**Wrong population** Not real patients
11		Sezgin E, Sirrianni JW, Kranz K. Evaluation of a Digital Scribe: Conversation Summarization for Emergency Department Consultation Calls. Applied clinical informatics 2024. doi:https://10.1055/a-2327-4121	**Wrong focus** Telephone consultations
12		Shanks D, Shah T, Hudson T, Thompson J, Filardi T, Wright K, et al. Enhancing Clinical Documentation Workflow with Ambient Artificial Intelligence: Clinician Perspectives on Work Burden, Burnout, and Job Satisfaction. medRxiv2024.	**Wrong publication type** Preprint
13		van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP. Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study. JMIR AI 2024; 3: e60020. doi:https://dx.doi.org/10.2196/60020	**Wrong population** Not real patients

14	2023	Pham CT, Visvanathan R, Strong M, Wilson ECF, Lange K, Dollard J, et al. Cost-Effectiveness and Value of Information Analysis of an Ambient Intelligent Geriatric Management (AmbIGeM) System Compared to Usual Care to Prevent Falls in Older People in Hospitals. Applied health economics and health policy 2023; 21: 315-25. doi:https://dx.doi.org/10.1007/s40258-022-00773-6	**Not relevant** Fall prevention
15		Sezgin E, Sirrianni J, Kranz K. Development and Evaluation of a Digital Scribe: Conversation Summarization Pipeline for Emergency Department Counseling Sessions towards Reducing Documentation Burden. medRxiv2023.	**Wrong study design** Preprint
16	2022	Benko S, Idarraga AJ, Bohl DD, Hamid KS. Virtual Scribe Services Decrease Documentation Burden Without Affecting Patient Satisfaction: A Randomized Controlled Trial. Foot & ankle specialist 2022; 15: 252-7. doi:https://dx.doi.org/10.1177/1938640020950544	**Not relevant** No AI involved
17	2021	Ganoe CH, Wu W, Barr PJ, Haslett W, Dannenberg MD, Bonasia KL, et al. Natural language processing for automated annotation of medication mentions in primary care visit conversations. JAMIA open 2021; 4: ooab071. doi:https://dx.doi.org/10.1093/jamiaopen/ooab071	**Not relevant** AI to search for medications
18		Mayer L, Xu D, Edwards N, Bokhart G. A Comparison of Voice Recognition Program and Traditional Keyboard Charting for Nurse Documentation. Computers, informatics, nursing : CIN 2021; 40: 90-4. doi:https://10.1097/CIN.0000000000000793	**Not relevant** No AI
19	2020	Jayasinghe L, Bittar A, Dutta R, Stewart R. Clinician-recalled quoted speech in electronic health records and risk of suicide attempt: A case-crossover study. BMJ Open 2020; 10. doi:https://10.1136/bmjopen-2019-036186	**Not relevant** AI to search for risk individuals
20		White AA, Lee T, Garrison MM, Payne TH. A Randomized Trial of Voice-Generated Inpatient Progress Notes: Effects on Professional Fee Billing. Applied clinical informatics 2020; 11: 427-32. doi:https://dx.doi.org/10.1055/s-0040-1713134	**Not relevant** No AI

Appendix 3

Time-in-notes, *editing time not included,* reported in the study by Lukac et al.

The study by Lukac reported a significantly shorter time period for time-in-notes per note for Nabla vs the control group (P=0.02), *time for editing not included* (see Table below). There was no significant difference between DAX and the control group, whereas Nabla had a significant shorter time-in notes compared to DAX -0.08 (95% CI (-0.16 – (-0.001)). There was no significant difference comparing any AI model vs the control group -0.06% (95% CI -0.12 – 0.01).

The analyses were conducted per intention-to-treat, and compliance with the AI scribes was 33.5% for DAX and 29.5% for Nabla. We understand this is the most decisive factor for the reduction of time-in-notes per note of at most 41 seconds. We emailed the authors of the paper by Lukac et al for information on a per protocol outcome (reduction of time-in-notes per note for visits when the AI scribes were actually used) and in their reply they explain this information was not available. The time in note metric was derived from the total minutes a provider spent writing notes per week and the total number of notes that week.

Table Reduction for time-in-notes per note comparing DAX and Nabla with controls, and statistical testing for relative proportional changes.
NB *Time for editing the AI-generated drafts is not included.*

Author Year	DAX N= 79	Nabla N=79	Control N=80	p-value
Lukac 2025	Pre:4 min 29 sec Post: 4 min 6 sec Reduction: 23 sec	Pre: 4 min 30 sec Post: 3 min 49 sec Reduction: 41 sec	Pre: 4 min 22 sec Post: 4 min 4 sec Reduction: 18 sec	**DAX vs control:** -1.7% (95% CI (-9.4) – 5.9); p=0.66 **Nabla vs control:** -9.5% (95% CI (-17.2) – (- 1.8)); p=0.02

These findings correspond to a physician seeing 10 outpatients in a day, of whom ambient scribes are used for documentation in 3 patients, resulting in a total reduction of documentation time (editing not included) of approximately 2 minutes (123 seconds).

Appendix 4

Systematic reviews on the use of AI scribes registered on PROSPERO, March 2, 2 2026.

1. The Use of Artificial Intelligence Scribes in Surgical Specialties: A Systematic Review 2025 CRD420251143404 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420251143404
2. Use of Ambient AI Scribe in Physicians' Clinical Documentation: A Protocol for a Systematic Review on Effectiveness, Efficiency and Satisfaction 2025 CRD420251149086 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420251149086
3. Artificial Intelligence in Nursing Practice within Palliative Care: A Systematic Review of the Literature 2025 CRD420251134129 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420251134129
4. Large language model (LLM) and AI–assisted tools for clinician performance and documentation: a systematic review and meta-analysis of randomized trials 2025 CRD420251123193 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420251123193
5. Understanding AI Scribes in Primary Care: A Systematic Review of Advantages, Disadvantages, Barriers and Facilitators 2025 CRD420251042631 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420251042631
6. Performance of Artificial Intelligence-assisted Electronic Medical/Health Records (EMR/EHR) Scribes - A Systematic Review and Meta-analysis 2025 CRD420250650612 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420250650612
7. Implementation outcomes of artificial intelligence (AI) scribes for clinical documentation in inpatient hospital settings: protocol for a systematic review. 2025 CRD420251182212 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420251182212
8. Impacts of AI Scribes on Clinical Outcomes, Efficiency, and Documentation: A Systematic Review Protocol 2024 CRD42024619680 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD42024619680
9. Efficacy, Accuracy, and Implementation of AI-Powered Clinical Documentation Tools: A Systematic Review and Meta-Analysis 2026 CRD420251274142 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420251274142
10. Effectiveness and User Experience of Artificial Intelligence-Driven Scribe Technology Among Clinicians: A Systematic Review 2024 CRD42024600076 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD42024600076
11. Artificial intelligence tools for reducing administrative burden in pre-operative surgical care: a systematic review 2025 CRD420251156704 (Ongoing) https://www.crd.york.ac.uk/PROSPERO/view/CRD420251156704


Region Örebro län


REGION
DALARNA


Region Uppsala


Region
Västmanland


Region
Värmland


Region
Gävleborg


REGION
SÖRMLAND