Spraakgestuurd rapporteren binnen de medisch-specialistische zorg

Technische rapportage

Procesleider: Roald van Leeuwen
Coördinerende onderzoeker: Sade Krijgsman
Uitvoerende onderzoeker: Katharina Abraham (KA)
In opdracht van digizo.nu

Versie 1.0
4 februari 2026

Inhoud

1 Leeswijzer

In deze technische rapportage zijn de stappen van het waardebepalend onderzoek vastgelegd. Deze rapportage moet niet worden beschouwd als een wetenschappelijk artikel maar volgt wel de standaardstructuur van een wetenschappelijke publicatie.

Allereerst zijn het onderzoeksdoel en de methoden beschreven in deze rapportage, die inzicht geven in de uitgevoerde onderzoeksstappen volgens de methodologische leidraad van Digizo.nu

In de resultatensectie zijn de geïncludeerde studies opgenomen en wordt een korte samenvatting gepresenteerd om een overzicht te geven van de resultaten per meetplanindicator. Daarnaast is een uitgebreide beschrijving opgenomen van de bevindingen met betrekking tot alle *must-have* en *should-have* indicatoren. De meetplanindicatoren zijn terug te vinden in de bijlage van deze rapportage. Voorafgaand aan iedere uitgebreide beschrijving van de bevindingen is een samenvatting opgenomen, gezien de omvang en diversiteit van de verschillende uitkomsten en indicatoren. Alle geëxtraheerde data is beschikbaar in het Excel-bestand, waarvoor een link is opgenomen aan het begin van de resultatensectie.

De beoordeling van de methodologische kwaliteit van de studies volgt aansluitend op de resultatensectie. In de discussie worden de kwaliteitsbeoordeling van de studies, de beschikbaarheid van vergelijkingen tussen interventie- en controlegroepen, en relevante studiekenmerken (zoals zorgsetting en type interventieapplicatie) behandeld.

De rapportage eindigt met de discussie en conclusie van het onderzoek. Daarnaast is in Bijlage E een overzicht opgenomen van de 'evidence gaps' per meetplanindicator: een inschatting van de mate waarin er, op basis van het meetplan, voldoende bewijs beschikbaar is. Deze inschatting is gebaseerd op de beschikbaarheid van studies, het studiedesign, de methodologische kwaliteit en de steekproefgrootte.

2 Onderzoeksdoel

Het doel van dit onderzoek was het evalueren van een nieuwe werkwijze met spraakgestuurd rapporteren (SGR) tijdens patiëntconsulten in de medisch-specialistische zorg (MSZ), met betrekking tot kwaliteit, impact op de administratieve last, werktevredenheid, gebruik, patiënt-zorgprofessionalbetrokkenheid, houding van patiënten en zorgprofessionals ten aanzien van het gebruik van SGR, en kosten.

3 Methode

De methodologische aanpak van dit onderzoek is gebaseerd op de methodiek waardebepaling van Digizo.

3.1 Zoekstrategie en screening

De zoekstrategie is op 13 juli 2025 uitgevoerd door een informatiespecialist en richtte zich op SGR binnen de domeinen sociaal domein (SD), verpleging, verzorging en thuiszorg (VVT), medisch-specialistische zorg (MSZ) en huisartsenzorg (HAZ), zonder beperkingen ten aanzien van studiedesign (zie Bijlage C voor de volledige zoekstrategie).

De titel- en abstractscreening (ti/ab) is uitgevoerd in Rayyan.ai door één onderzoeker (KA). Deze keuze is gemaakt omdat de zoekstrategie domeinoverstijgend was, waardoor het niet efficiënt was om een tweede onderzoeker als reviewer te betrekken, gezien hun beperkte inzicht in de overige domeinen.

Ter controle van de kwaliteit van de zoekstrategie en screening is een kwaliteitscheck uitgevoerd op basis van één sleutelartikel (DOI: 10.2196/60020). Indien de zoekstrategie en screening correct zijn uitgevoerd, zou dit artikel geselecteerd moeten worden in zowel de titel- en abstractscreening als in de full-tekstbeoordeling.

3.2 Inclusie- en exclusiecriteria

De onderstaande tabel geeft weer welke populaties, interventies, vergelijkingen en uitkomsten (PICO) in dit onderzoek zijn meegenomen. De PICO is gebaseerd op de meetplanindicatoren die in samenwerking met de sector MSZ zijn opgesteld (zie Bijlage A).

Studies moesten in full-tekst beschikbaar zijn en geschreven in het Engels of Nederlands. Er is in eerste instantie gekeken naar recente systematische reviews; indien deze aansloten bij de PICO, werden alleen de primaire studies die buiten de zoekperiode van deze reviews vielen, in full-tekst gescreend. Records met een van de volgende studietypen zijn niet meegenomen in het onderzoek: scoping reviews, commentaren, editorials, essays, nieuwsitems, opinieartikelen en research letters.

Voor de grijze literatuur zijn aanvullende criteria toegepast (zie Bijlage B).

PICO	Beschrijving	Operationalisatie (Engelstalig geformuleerd voor de literatuurreview).
Populatie	Zorgprofessionals & patiënten in de medische specialisten zorg	Medical specialist and/or other healthcare provider in the intramural specialist care setting; patient receiving care within the same setting.
Interventie	Spraakgestuurd rapporteren: Het consult wordt door de applicatie opgenomen, letterlijk (verbatim) uitgeschreven, vervolgens samengevat onder de juiste kopjes en opgeslagen in het EPD.	Inclusie: • Speech-to-summary: Applications that record and summarise the consultation between care professional and patient (and guardian), and (automatically) save the structured summary in the electronic health record (EHR). Exclusie: • Speech-to-text solutions • Human medical scribes Exclusie: • Interventie is getoetst voor een niet-Germaanse taal.
Vergelijking (Comparator)	Reguliere verslaglegging/rapportage Opmerking: 0ok gegevens die uitsluitend betrekking hebben op de SOLL-situatie worden meegenomen in het onderzoek.	Inclusie: • Healthcare provider typing the clinical note Exclusie: • Comparison of model performances with test data
Uitkomsten*	• Kwaliteit van klinische samenvatting • Houding patiënten • Houding zorgverleners • Gebruik • Werkplezier • Administratieve tijd • Tevredenheid patiënten • Tevredenheid zorgverleners • Kosten	

*De uitkomsten zijn in detail beschreven in het meetplan, te vinden in Bijlage A.

3.3 Data extractie

De data-extractie is uitgevoerd door één onderzoeker (KA), waarbij de studiekenmerken zijn geëxtraheerd en de uitkomsten zijn vastgelegd die relevant zijn voor het meetplan (Bijlage A). Alle geëxtraheerde data zijn beschikbaar in het Excel-bestand, desgewenst op te vragen.

3.4 Data synthese

Vanwege de heterogeniteit in studiedesigns is gekozen voor een narratieve samenvatting. De uitkomsten zijn samengevat op basis van het meetplan (zie Bijlage A).

3.5 Kwaliteitsbeoordeling

Systematische reviews worden beoordeeld met de AMSTAR-2 (Box 1). Primaire studies worden beoordeeld met de MMAT (Box 1), aangezien verschillende studiedesigns worden verwacht.

AMSTAR-2
AMSTAR-2 is een internationaal gevalideerd instrument voor de kwaliteitsbeoordeling van systematische reviews die zowel gerandomiseerde als niet-gerandomiseerde studies kunnen omvatten. Het instrument bevat 16 items die aspecten beoordelen zoals protocolregistratie, volledigheid van de literatuurzoekactie, beoordeling van risico op bias en geschiktheid van de gebruikte analysemethoden. AMSTAR-2 wordt gebruikt om de methodologische kwaliteit en daarmee de betrouwbaarheid van systematische reviews te bepalen.

MMAT
De Mixed Methods Appraisal Tool (MMAT) is een gevalideerd instrument dat wordt gebruikt voor de kwaliteitsbeoordeling van wetenschappelijke studies met uiteenlopende onderzoeksdesigns, waaronder kwalitatieve, kwantitatieve en mixed-methods studies. Het unieke van de MMAT is dat het binnen één raamwerk toelaat om verschillende typen studies systematisch en op een consistente manier te evalueren.
Het instrument bestaat uit vijf specifieke beoordelingscriteria per studiedesign, die zich richten op aspecten zoals de helderheid van de onderzoeksvraag, geschiktheid van het design, de kwaliteit van de dataverzameling, de analysemethoden en de interpretatie van de resultaten.
De MMAT wordt vooral toegepast in systematische reviews waarin verschillende soorten studies worden geïncludeerd. Het doel is om de methodologische kwaliteit van de afzonderlijke studies op een transparante en reproduceerbare manier in kaart te brengen, zodat de betrouwbaarheid van de conclusies uit de review beter kan worden ingeschat.

Kader 1

4 Resultaten

4.1 Geïncludeerde literatuur

Er zijn 1.211 records geïdentificeerd op basis van de zoekstrategie. Na het ontdubbelen zijn er 649 records overgebleven waarbij deze op basis van titel en abstract gescreend zijn op inclusie-exclusie criteria (zie tabel beneden).

83 studies zijn geïncludeerd voor full-tekst. Van 15 studies konden wij de full-tekst niet vinden. 23 studies zijn uitgesloten op basis van het studie design. Verder was de populatie niet passend (n=6), de interventie niet passend (n=21) of de taal was niet in het Engels of Nederlands (n=1).

Behalve twee preprints is er geen grijze literatuur geïncludeerd.

MSZ – Zwarte literatuur - Selectie

1. Zoekresultaten gevonden door informatiespecialist	1.211
2. Aantal resultaten na ontdubbelen Informatiespecialist Rayyan.ai	649 655 649
3. Geïncludeerd voor full-tekst na Ti/Ab screening	83
Geëxcludeerd + redenen voor exclusie	**75**
Populatie	6
Interventie	21
Studie design	23
Taal	1
No full-tekst	15
Geïncludeerd in onderzoek	**15**

Er zijn twee relevante SLR's gevonden (Ng et al., 2025; Sasseville et al., 2025). De SLR van Ng et al. hanteerde een zoekperiode tot februari 2025 en die van Sasseville et al. tot november 2024. De geïncludeerde primaire studies in de twee SLR's overlappen gedeeltelijk. Zes van de primaire studies uit beide SLR's waren volgens de PICO relevant voor dit onderzoek. Daarnaast zijn er zeven studies uit de zoekresultaten opgenomen in dit onderzoek.

In totaal zijn er dus 13 unieke primaire studies gevonden voor SGR in de MSZ waaronder:

- een RCT (nog in preprint) met een vergelijking tussen twee AI-scribes en controlegroep (Lukac et al., 2025),
- een feasibility-studie waarbij vijf real-world consulten werden opgenomen en samengevat met AI-scribe en tegelijkertijd ook handmatig door de zorgverlener (Ong et al., 2025),
- vier mixed-method(pilot)studies met pre-postvergelijking (Baldwin et al., 2025; Cao et al., 2024; Duggan et al., 2025; Nguyen et al., 2023),
- een observationele studie met pre-postvergelijking (Misurac et al., 2025),
- een implementatierapport gebaseerd op gebruiksmetrics (Basei De Paula et al., 2024)

- en vijf simulatiestudies (Anderson et al., 2025; Biro et al., 2025; Hose et al., 2025; Moryousef et al., 2025; Van Buchem et al., 2024), waarvan één in Nederland (Van Buchem et al., 2024) en één als preprint (Anderson et al., 2025).

De geïncludeerde studies zijn gepubliceerd tussen 2023 en 2025 en vonden plaats in verschillende landen: Brazilië (n=1), Verenigde Staten (n=5), Canada (n=1), en Nederland (n=1). In vijf studies werd het land van de onderzochte zorgsetting niet vermeld.

De medische specialismen die in de opgenomen studies zijn onderzocht, zijn oncologie (Baldwin, Nguyen), urologie (Ong, Moryousef), diverse medische specialismen zoals keel-, neus- en oorheelkunde, cardiologie, reumatologie, huisartsgeneeskunde, kindergeneeskunde, endocrinologie, interne geneeskunde, maag-darm-leverziekten, oncologie en spoedeisende zorg (Biro), en dermatologie (Cao). Vier studies rapporteerden geen medisch-specialistische setting (Misurac, Basei de Paula, Hose, Van Buchem). In twee studies werden respectievelijk 17 (Duggan) en 14 (Lukac) specialismen vermeld, zonder verdere informatie over de aard van de medisch-specialistische zorg. Eén studie rapporteerde poliklinische specialismen zonder nadere specificatie (Anderson).

4.2 Studieresultaten

Samenvatting

Must-have 1: Kwaliteit van de samenvatting

In kwalitatieve onderzoeken werd de kwaliteit van AI-gegenereerde notities over het algemeen als gelijkwaardig aan handmatige verslaglegging beoordeeld, met name wat betreft de accuraatheid en gedetailleerdheid.

In de kwantitatieve onderzoeken werd de algemene kwaliteit, maar ook de accuraatheid, relevantie en georganiseerdheid van onbewerkte AI-notities echter lager beoordeeld dan die van handmatige notities; na correctie door de gebruiker verdwijnen deze kwaliteitsverschillen.

Onbewerkte AI-notities bevatten regelmatig onnauwkeurigheden, vooral door ontbrekende informatie (omissies). Het foutpercentage varieert van circa 15% tot bijna 50%. Ernstige gefabriceerde informatie (hallucinaties) of toevoegingen zijn zeldzaam, en bias wordt nauwelijks gerapporteerd. Het gebruik van AI-scribe notities kan resulteren in fouten variërend van mild tot ernstig, met potentieel risico op schade voor patiënten.

Must-have 2: Adoptie door zorgverlener

Zorgprofessionals hebben een gemengde houding ten aanzien van de bruikbaarheid van AI-scribe-notities. Ze ervaren de tools als bruikbaar, haalbaar en passend, maar de tijd die nodig is om zich comfortabel te voelen met de applicatie verschilt per persoon. Zorgprofessionals staan overwegend positief tegenover toekomstig gebruik van AI-scribes, hoewel de meningen over het aanbevelen ervan aan collega's uiteenlopen.

Must-have 3: Acceptatie door patiënt

Er zijn geen inzichten in de houding van patiënten die direct bij patiënten zijn verzameld. Indirecte inzichten, via bevraging van zorgprofessionals, laten zien dat patiënten over het algemeen positief staan tegenover het gebruik van een AI-scribe.

Should-have 1: Gebruik van AI-gegenereerde samenvatting

Het gebruik van AI-scribes door zorgprofessionals in niet-Nederlandse settings laat een gemengd beeld zien. In Brazilië was er over het algemeen sprake van positieve adoptie, al fluctueerde het gebruik soms, bijvoorbeeld in het aantal zorgprofessionals dat zich registreerde en daadwerkelijk gebruikmaakte van de AI-scribe. In de Verenigde Staten werd de AI-scribe slechts bij een klein deel van de consulten gebruikt, vooral bij eerste of eenvoudige consulten. Tijdens het beperkte aantal telemedicineconsulten verliep het gebruik van AI-scribes technisch probleemloos.

Should-have 1: Werkplezier

Het gebruik van AI-notities voor verslaglegging vermindert de burn-out onder zorgprofessionals en verhoogt het werkplezier. AI-notities verlagen de mentale belasting, verbeteren de werk-privébalans en verminderen gevoelens van uitputting. Wel kan de impact op werkplezier verschillen tussen verschillende AI-scribes. Zorgprofessionals rapporteerden daarnaast een afname van mentale inspanning en tijdsdruk.

Should-have 2: Administratieve werkdruk

AI-scribes lijken de tijdsbesteding aan documentatie en de task load van zorgprofessionals te verminderen, met mogelijk ook een positief effect op documentatie buiten werktijd. Efficiëntie hangt echter af van gebruiksintensiteit en kan worden beperkt door fouten. Kwantitatieve studies tonen een duidelijke reductie in de bijdrage van zorgprofessionals aan AI-gegenereerde notities, al kan die bijdrage nog steeds aanzienlijk blijven (tot 51,7%).

Should-have 3: Aandacht voor het gesprek – zorgprofessional perspectief

Zorgprofessionals zijn tevreden over het gebruik van AI-scribes, omdat zij hierdoor meer aandacht kunnen besteden aan het gesprek met patiënten. De introductie van AI-notities leidt volgens zorgprofessionals tot minder afleiding, meer betrokkenheid bij patiënten en meer gelegenheid voor face-to-face contact, in plaats van tijd te besteden aan verslaglegging.

Should-have 4: Aandacht voor het gesprek – patiënt perspectief

Patiënten zijn ook overwegend tevreden over de patiënt-zorgprofessional interactie. Daarnaast merkten zij dat hun consult persoonlijker aanvoelde en dat de zorgprofessional minder tijd achter de computer doorbracht met typen.

Should-have 5: Kosten

Er zijn geen studies gevonden.

Must-have 1: Kwaliteit van samenvatting

Belangrijke uitkomsten die voor deze must-have zijn gerapporteerd, betreffen de algemene kwaliteit en specifieke aspecten van kwaliteit zoals accuraatheid, relevantie, correcte organisatie en juistheid (biases, fouten, hallucinaties). Deze zijn kwantitatief en kwalitatief gemeten met de PDQI-9-vragenlijst*, op een Likert-schaal, of descriptief beschreven.

* *De Physician Documentation Quality Instrument-9 (PDQI-9) is specifiek ontwikkeld voor elektronische medische documentatie, waaronder voortgangsverslagen en ontslagbrieven. De PDQI-9 beoordeelt negen kenmerken van verslaglegging: actualiteit, nauwkeurigheid, volledigheid, bruikbaarheid, organisatie, begrijpelijkheid, bondigheid, mate van synthese en interne consistentie. Elk kenmerk wordt gescoord op een vijfpuntsschaal. De som van de afzonderlijke scores resulteert in een totaalscore (maximaal 45) die de algemene kwaliteit van de documentatie weerspiegelt. Er bestaan geen officiële cut-offs binnen de PDQI-9.*

Algemene kwaliteit – kwalitatief

Samenvattend: De geïncludeerde kwalitatieve studies laten gemengde ervaringen zien: sommige zorgverleners ervaren de notities als accuraat en gedetailleerd, terwijl anderen de kwaliteit onvoldoende achten, met name vanwege fouten en omissies, ongewenste opmaak, simplistische formuleringen of juist overmatige detaillering, waardoor extra tijd nodig is voor correcties.

Zwarte literatuur

Twee studies rapporteerden over de kwaliteit van AI-gegenereerde notities (Duggan, Van Buchem).

De studie van Van Buchem liet geneeskunde studenten de verslag kwaliteit beoordelen. Vier studenten rapporteerden dat de automatisch gegenereerde samenvattingen hun verwachtingen overtroffen, terwijl vier anderen de kwaliteit onvoldoende achtten vanwege de aanwezigheid van fouten en de extra tijd die nodig was voor het aanbrengen van correcties.

De studie van Duggan liet gemengde ervaringen zien onder clinici ten aanzien van de lengte en inhoud van de automatisch gegenereerde notities. Hoewel sommige zorgverleners de notities als accuraat en gedetailleerd ervoeren, vonden anderen deze te foutgevoelig en te bewerkelijk, waardoor de beoogde tijdsbesparing deels teniet werd gedaan door de noodzaak tot correctie en redactie.

Grijze literatuur

De gerandomiseerde studie van Lukac (nog onder peer-review en daarom onderdeel van de grijze literatuur), signaleert enkele kwaliteitsproblemen. Respondenten rapporteerden het vaakst omissies in de notities (n=12), gevolgd door structurele tekortkomingen zoals ongewenste opmaak, te simplistische formuleringen of juist overmatige detaillering (n=11). Daarnaast kwamen fouten in voornaamwoordgebruik, waaronder misgendering (n=8), en problemen met het correct detecteren van bevestigingen of ontkenningen (n=5) naar voren.

Algemene kwaliteit – kwantitatief

Samenvattend: De geïncludeerde kwantitatieve wetenschappelijke studies laten zien dat de kwaliteit van onbewerkte AI-gegenereerde notities lager wordt beoordeeld dan die van handmatig opgestelde notities. Na correctie door de gebruiker is er echter geen verschil meer gevonden. De grijze literatuur suggereert daarnaast dat er kwaliteitsverschillen kunnen bestaan tussen verschillende applicaties voor spraakgestuurd rapporteren, maar ook hier bereiken de kwaliteitscores een vergelijkbaar niveau als handmatig opgestelde notities. Zorgverleners zelf zijn gemiddeld neutraal over de stelling dat AI-scribes notities van gelijke kwaliteit produceren.

Zwarte literatuur

Een simulatiestudie (Van Buchem) en een feasibility-studie met vergelijking van AI-scribe en reguliere verslaglegging binnen hetzelfde cohort (Ong) geven inzicht in de kwantitatief beoordeelde kwaliteit van medische verslaglegging. Voor het kwantitatief beoordelen van de kwaliteit van medische verslaglegging werd in beide studies gebruikgemaakt van de PDQI-9.

In de simulatiestudie (Van Buchem) werden de handmatig opgestelde notities (waarvan de best beoordeelde als referentie werd genomen) significant beter beoordeeld dan de onbewerkte AI-gegenereerde notities. Wanneer de AI-gegenereerde notities echter door de gebruiker waren aangepast, werden deze vergelijkbaar beoordeeld met de beste handmatige notities (gemiddelde totaalscore 29 versus 31). Voor de subcategorie bondigheid werden de handmatig opgestelde notities wel significant beter beoordeeld.

In de feasibility-studie (Ong), gebaseerd op vijf real-world consulten, werd de kwaliteit op de PDQI-9 als gelijkwaardig beoordeeld, zonder significante verschillen tussen handmatige en AI-scribe notities. Verder werd gevonden dat AI-gegenereerde notities in 55% van de beoordelingen de voorkeur kregen boven handmatig geschreven notities.

Grijze literatuur

In de grijze literatuur vergeleek een (preprint) studie verschillende AI-applicaties met elkaar op kwaliteit (Anderson) en een andere studie vergeleek twee AI-applicaties met elkaar en met een controlegroep (Lukac).

In de studie van Anderson werden significante verschillen tussen de toepassingen gevonden (bereik PDQI-9-score 32,3–39,4), met een gemiddelde totaalscore van 36 over alle platforms.

In de studie van Lukac waren zorgverleners gemiddeld neutraal over de stelling dat AI-scribes notities maken die even goed zijn als handmatig opgestelde notities (gemeten met een Likert-schaal).

Accuraatheid

Samenvattend: De wetenschappelijke studies die de accuraatheid van handmatige verslaglegging vergelijken met die van automatische verslaglegging laten een vergelijkbare en hoge accuraatheid zien. De simulatiestudie in de Nederlandse setting laat echter zien dat toepassingen zonder bewerking minder accuraat zijn dan handmatige verslaglegging. De grijze literatuur suggereert daarnaast dat er incidenteel fouten worden gevonden bij automatische verslaglegging.

Zwarte literatuur

In de simulatiestudie (Van Buchem) en feasibility studie (Ong) werd de accuraatheid van de (bewerkte) AI verslagen gelijk beoordeeld als de handmatige verslagen (PDQI-9 subcategorie).

In de simulatiestudie (van Buchem) werden de handmatig opgestelde notities (waarvan de best beoordeelde als referentie werd genomen) significant beter beoordeeld op accuraatheid dan de onbewerkte AI-gegenereerde notities. Wanneer de AI-gegenereerde notities echter door de gebruiker waren aangepast, werden deze vergelijkbaar beoordeeld met de beste handmatige notities (PDQI-9 subcategorie: gemiddelde score 5 versus 5).

Grijze literatuur

De RCT-studie (Lukac) vroeg de gebruikers (Likert-schaal) hoe vaak er in de afgelopen twee maanden een door een AI-gegenereerde notitie ten minste één klinisch relevante onnauwkeurigheid

('inaccuracy') bevatte, zoals een omissie, toevoeging of hallucinatie. Voor beide bestudeerde toepassingen werd dit beantwoord met 'af en toe'.

Relevantie

Samenvattend: In de simulatiestudie (Van Buchem) en feasibility studie (Ong) werd de relevantie van de (bewerkte) AI verslagen gelijk beoordeeld als de handmatige verslagen (PDQI-9 subcategorie). De simulatiestudie in de Nederlandse setting laat echter zien dat toepassingen zonder bewerking minder bruikbaar zijn dan handmatige verslaglegging.

Correct georganiseerd

Samenvattend: In de simulatiestudie (Van Buchem) en feasibility studie (Ong) werd de georganiseerdheid van de (bewerkte) AI verslagen gelijk beoordeeld als de handmatige verslagen (PDQI-9 subcategorie). De simulatiestudie in de Nederlandse setting laat echter zien dat toepassingen zonder bewerking minder correct georganiseerd zijn dan handmatige verslaglegging.

Juistheid

Samenvattend: De geïncludeerde wetenschappelijke en grijze literatuur laat zien dat AI-gegenereerde notities regelmatig onnauwkeurigheden bevatten, met name in de vorm van omissies. Hoewel de mate van fouten varieert tussen studies en toepassingen, blijkt uit meerdere bronnen dat in een aanzienlijk deel van de notities één of meerdere fouten aanwezig zijn, waarbij foutpercentages uiteenlopen van ongeveer 15% tot bijna 50%. Bias wordt daarentegen slechts zelden gerapporteerd, en ernstige hallucinaties of toevoegingen komen minder vaak voor dan omissies.

Zwarte literatuur

Vier simulatiestudies (Biro, Moryousef, Van Buchem, Hose) onderzochten de foutgevoeligheid van AI-scribes.

In de studie van Biro, waarin twee toepassingen werden vergeleken, werden in 70% van de notities fouten gevonden, met een gemiddeld aantal van 2,9 fouten per notitie (SD 2,7). De verdeling van fouttypes verschilde: omissies kwamen het vaakst voor (83% bij Product A en 54% bij Product B), gevolgd door irrelevante of verkeerd geplaatste tekst (25% bij Product B versus 6% bij Product A). Toevoegingen en 'foutieve outputs' werden minder vaak gerapporteerd (respectievelijk 4 versus 11% en 6 versus10%, Product A versus Product B). De verschillen tussen Product A en Product B waren significant.

De studie van Moryousef gebruikte een samengestelde score om het aandeel consulten met ten minste één kritieke fout vast te stellen. Zij toetsten vijf verschillende AI-applicaties en vonden dat het percentage consulten met kritieke fouten uiteenliep van 28% tot 62%.

In de studie van Van Buchem werd beschreven dat een veelvoorkomende fout het ontbreken van negatieve symptomen betrof, zoals het niet vermelden van de afwezigheid van kortademigheid. Daarnaast ondervonden gebruikers technische problemen, waarbij de AI-scribe bleef hangen of zeer traag werkte door beperkingen in de rekencapaciteit.

De studie van Hose rapporteerde in totaal 66 fouten bij 11 consulten, wat neerkomt op gemiddeld zes fouten per consult. Het merendeel betrof omissies (55; 83%), gevolgd door 'foutieve outputs' (6%), inhoud die op zichzelf correct maar klinisch ongepast was (6%), en toevoegingen (5%).

Grijze literatuur

Uit de RCT-studie van Lukac blijkt dat bias in AI-gegenereerde notities zelden voorkomt (Likert-schaal 1,6 en 1,7).

In de simulatiestudie van Anderson werden de belangrijkste elementen uit de 14 gesimuleerde gesprekken geëxtraheerd en geëvalueerd in de AI-gegenereerde notities met behulp van een referentienotitie. Gemiddeld over de vijf verschillende AI-applicaties werden in 26,3% van de gevallen (variërend van 16,7% tot 46,1%) fouten aangetroffen. Het overgrote deel hiervan betrof omissies (gemiddeld 76,3% van alle fouten), aangevuld met kleinere aantallen fouten door toevoegingen of gedeeltelijk correcte informatie.

Patiëntenveiligheid

Samenvattend: De wetenschappelijke en grijze literatuur laat zien dat AI-scribes kunnen resulteren in fouten variërend van mild tot ernstig, met potentieel risico op schade voor patiënten.

Zwarte literatuur

Een studie met als doel een error-raamwerk op te stellen en dit te testen middels een simulatie van 11 real-world consulten met een AI-scribe rapporteerde dat van de 66 gevonden error 55% van lage ernst waren en 25% van matige ernst (Hose).

Grijze literatuur

Twee studies (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerden over de ernst van fouten bij het gebruik van AI-gegenereerde notities (Lukac, Anderson).

Een RCT (Lukac) rapporteerde een event met milde ernst bij het gebruik van een van de twee AI-applicaties.

In de simulatiestudie met vijf verschillende AI-applicaties (Anderson) werd gemiddeld drie fouten per casus gevonden, met potentieel voor matige tot ernstige schade. Het laagste aantal fouten was 1,4 en het hoogste 6,2 per casus. Er werd een significant verschil gevonden tussen de commerciële en de een openbaar platform.

Must-have 2: Adoptie door zorgprofessional

Belangrijke uitkomsten die voor deze must-have zijn gerapporteerd, betreffen de bruikbaarheid en het toekomstig gebruik. Deze zijn gemeten met behulp van een Likert-schaal, NPS* en SUS score ** en beschrijvende analyses.

* *De Net Promoter Score (NPS) meet in welke mate gebruikers een product of dienst zouden aanbevelen aan anderen. De score wordt berekend als het percentage promoters (9–10) minus het percentage detractors (0–6) en loopt van –100 tot +100. Een positieve score duidt op meer promoters dan detractors en wordt geïnterpreteerd als een teken van tevredenheid en loyaliteit.*

** *De System Usability Scale (SUS) is een korte vragenlijst met tien items waarmee de gebruiksvriendelijkheid van een systeem of product wordt gemeten op een schaal van 0 tot 100.*

Bruikbaarheid

Samenvattend: De kwantitatieve uitkomsten laten een positieve houding ten aanzien van de bruikbaarheid van AI-scribes zien, terwijl de kwalitatieve uitkomsten een gemengd beeld schetsen.

Zwarte literatuur

Twee pre-postvergelijkingen van een AI-scribe (Duggan, Nguyen) en een Nederlandse simulatiestudie met mock-consultaties (Van Buchem) rapporteerden over de bruikbaarheid van AI-scribes.

De studie van Nguyen vond een gemengde houding bij zorgprofessionals. De AI-scribe werd beoordeeld als 'feasible', 'acceptable' en 'appropriate', maar zorgprofessionals hadden uiteenlopende meningen over de tijd die nodig was om zich comfortabel te voelen met de applicatie.

Duggan daarentegen rapporteerde een positieve 'ease of use-score' van AI-scribes op de gestandaardiseerde SUS. De houding van zorgprofessionals bleef positief voor en na gebruik ten aanzien van impact op 'workflow', 'appropriateness' en 'readiness', die onveranderd bleven.

In de Nederlandse studie gaven studenten een gemengd beeld over het gebruik van de AI-scribe. De helft van de 18 medische studenten gaf aan dat de AI-scribe leuk was om te gebruiken, 6 studenten vonden de tool gemakkelijk in gebruik, 4 studenten geloofden in het potentieel van de tool en 4 studenten vonden de geproduceerde samenvattingen beter dan verwacht. Daarentegen vonden 4 studenten dat de kwaliteit van de samenvattingen slecht was.

Grijze literatuur

Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde over de houding van zorgverleners ten aanzien van het gebruik van AI-gegenereerde notities (Lukac). Zorgverleners waren het gemiddeld eens dat het leren en gebruiken van de twee AI-scribes eenvoudig was (gemeten met een Likert-schaal). Er werd geen verschil gevonden tussen de twee AI-scribes.

Toekomstig gebruik

Samenvattend: De meerderheid van de studies rapporteerde een positieve houding over toekomstig gebruik.

Zwarte literatuur

Vijf studies geven inzicht in de houding van zorgverleners ten aanzien van toekomstig gebruik van AI-scribes: drie mixed-methods-studie met een pre-post-implementatievergelijking (Duggan, Cao, Baldwin), en twee simulaties (Moryousef), waarvan één in een Nederlandse setting (Van Buchem).

Vier van deze studies rapporteren een positieve houding ten aanzien van toekomstig gebruik (Cao, Baldwin, Moryousef. Van Buchem).

In de studie van Duggan waren de meningen verdeeld: ongeveer een derde was promoter van AI-scribes (35%), een derde passief (30%) en een derde detractor (35%), volgens de NPS-score die meet hoe waarschijnlijk zorgverleners een product zouden aanbevelen.

Grijze literatuur

Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde over het toekomstig gebruik van AI-scribes (Lukac). Zorgverleners konden zich voorstellen in de toekomst met de AI-scribes te werken (4,1 en 3,8 op Likert-schaal 1-5). Er werd geen verschil gevonden tussen de twee AI-scribes.

Must-have 3: Acceptatie door patiënt

Belangrijke uitkomsten die voor deze must-have zijn gerapporteerd, betreffen de bruikbaarheid en patiëntveiligheid. Deze zijn gemeten met behulp van een Likert-schaal, het aantal events en het percentage ernstige fouten.

Bruikbaarheid

Samenvattend: Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde over de houding van patiënten ten aanzien van het gebruik van AI-gegenereerde notities (Lukac). Zorgprofessionals moesten patiënten vooraf toestemming vragen voor het gebruik van SGR. De meeste patiënten stonden positief tegenover het gebruik, met een gemiddelde score van 4,4 op een 5-punt Likert-schaal. Er werd geen significant verschil gevonden tussen de twee onderzochte applicaties.

Should-have 1: Gebruik

Samenvattend: De studies laten een gemengd beeld zien van het gebruik van AI-scribes in niet-Nederlandse settings. Zo laat de wetenschappelijke literatuur in de Braziliaanse context over het algemeen een positieve adoptietrend zien, met enkele fluctuaties in bepaalde metrische data, zoals het aantal zorgverleners dat zich registreerde en binnen een week daadwerkelijk gebruikmaakte van de AI-scribe. In de oncologische context in de Verenigde Staten werd de AI-scribe daarentegen slechts in een klein deel van de consulten toegepast, vaker bij eerste consulten dan bij vervolgconsulten, en met name bij relatief eenvoudige consulten of specifieke casussen. Het gebruik van AI-scribes tijdens telemedicineconsulten werd technisch probleemloos bevonden.

Zwarte literatuur

Drie studies rapporteerden over de adoptie van SGR: een implementatierapport met metrische data van een AI-scribe (Basei de Paula) en twee mixed-method pilot pre-postvergelijkingsstudies (Baldwin, Nguyen).

De metrische data in Basei de Paula laat een duidelijke stijging zien in het aantal gebruikers over de tijd. Het aantal gebruikers begon bij 100 in begin maart en groeide tot 900 halverwege mei. Een andere gebruikte maat was de ‘weekly activation rate’: het percentage nieuwe gebruikers dat zich registreerde en binnen één week na aanmelding ten minste één anamnese genereerde. Begin mei bedroeg deze rate 45% en vertoonde vervolgens fluctuaties. Aan het eind van de studie (half mei) lag de rate boven de 50%. Volgens de auteurs kunnen verschillende factoren deze fluctuaties verklaren, zoals de introductie van nieuwe functionaliteiten, veranderingen in de trainingsprocedures of verbeteringen in de gebruikersinterface.

Daarnaast liet de data van Basei de Paula een toename zien in het aantal door zorgverleners gegenereerde documenten over een periode van 2,5 maanden. In de eerste weken na introductie werd slechts een beperkte groei waargenomen, wat door de auteurs werd geïnterpreteerd als een learning curve. Ongeveer een maand later werd een duidelijke stijging zichtbaar, hetgeen volgens de auteurs duidt op regelmatiger en intensiever gebruik van het platform. Ook het dagelijks aantal gegenereerde documenten liet een oplopende trend zien, wat wijst op toenemend gebruik van het platform. De zorgsetting waarin deze data werden verzameld is echter niet gerapporteerd.

In de studie van Baldwin, uitgevoerd met 31 deelnemers in een oncologische setting, werd AI-scribe slechts in 14% van alle consulten gebruikt. Verder pasten zorgverleners AI-scribe vaker toe bij eerste consulten dan bij vervolgconsulten (21% versus 12%).

In de studie van Nguyen, uitgevoerd met 21 zorgverleners in een oncologisch centrum, werd beschreven dat de AI-scribe vooral werd ingezet bij relatief eenvoudige consulten. Na verloop van tijd rapporteerden sommige zorgverleners de AI-scribe bij alle typen consulten te gebruiken, terwijl anderen aangaven het gebruik te beperken tot specifieke casussen, bijvoorbeeld bij het bespreken van kankerprogressie. Daarnaast werd het gebruik van de AI-scribe tijdens telemedicineconsulten onderzocht. De studie rapporteerde dat slechts een klein aantal zorgverleners hiervan gebruik maakte en werden er geen technische problemen gerapporteerd. Telemedicinepatiënten werden doorgaans gezien als relatief 'straightforward', waardoor de AI-scribe in die context toepasbaar werd geacht. Wel werden zorgen geuit over de privacy bij het gebruik van een AI-scribe tijdens telemedicineconsulten.

Grijze literatuur

Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde een gebruik van AI-scribes in 29,5% tot 33,5% van alle consulten voor twee verschillende AI-scribes. 15% procent van de zorgverleners maakte in de studie helemaal geen gebruik van een AI-scribe.

Should-have 2: Werkplezier

Belangrijke uitkomsten die voor deze must-have zijn gerapporteerd, betreffen 'burnout', 'professional fulfillment' en 'task load'. Deze zijn gemeten met behulp van een Likert-schaal, de composite Mini-Z vragenlijst* en PFI-WE**.

* *De Mini-Z is een korte, gevalideerde vragenlijst die is ontwikkeld om het welzijn van zorgverleners en het risico op burn-out in kaart te brengen. De vragenlijst bestaat uit circa 10 items (er zijn ook verkorte versies) en bestrijkt domeinen zoals arbeidstevredenheid, stress, werkomgeving, werkdruk en burn-out (één-itemmeting).*

** *De PFI-WE (Professional Fulfillment Index – Work Exhaustion subscale) is een gevalideerd instrument dat is ontwikkeld om professionele vervulling en burn-out bij zorgprofessionals te meten. De subschaal Work Exhaustion richt zich specifiek op het domein van werkgerelateerde uitputting, een kerncomponent van burn-out.*

Samenvattend: Er zijn sterke aanwijzingen dat de introductie van een AI-scribe burn-out onder zorgverleners vermindert en een positief effect heeft op het werkplezier.

Zwarte literatuur

Drie studies geven inzicht in de invloed van AI-scribes op burn-out: twee mixed-methodsstudies (Duggan, Nguyen) en één observationele studie (Misurac).

In de studie van Nguyen werd geen significant verschil gevonden in burn-out na de introductie van een AI-scribe, gemeten met de Mini-Z.

In de studie van Duggan werd daarentegen een significante verbetering gevonden na de introductie van een AI-scribe voor mentaal overbelast voelen, werk-privébalans en zich uitgeput voelen in relatie tot de documentatie van patiëntcontacten, gemeten op een Likert-schaal van 1 tot 7.

In de studie van Misurac werd een 38% lagere burn-outscore gerapporteerd na de implementatie van een AI-scribe. Verder werd een significante verbetering gevonden in werkplezier en burn-out, gemeten met de PFI-WE, na de implementatie van een AI-scribe. De score lag bovendien onder de burn-out afkapwaarde.

Grijze literatuur

Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde over de verandering in burnout en werplezier van zorgverleners met en zonder AI-scribe (Lukac). Met AI-scribe ondersteuning hadden zorgverleners significant minder burnout op de Mini-Z en significant meer

werkplezier, gemeten met de PFI-WE. Voor één van de scribes (Nbala) werd geen significant verbetering gevonden op de PFI-WE ten opzichte van geen AI-scribe.

Task load

Samenvattend: De introductie van AI-scribes lijkt de task load van zorgprofessionals te verminderen.

Zwarte literatuur

Een mixed-methodsstudie (Duggan) gaf inzicht in de ervaren task load van zorgprofessionals na de introductie van een AI-scribe. Hoewel de AI-scribe de administratieve belasting niet volledig oploste, rapporteerden zorgprofessionals wel een vermindering in mentale inspanning.

Grijze literatuur

Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde over het verschil in task load na de introductie van twee AI-scribes (Lukac). Er werd een significante verlaging gevonden voor beide AI-scribes na introductie.

Should-have 3: Administratieve werkdruk

Belangrijke uitkomsten die voor deze must-have zijn gerapporteerd, betreffen de tijd voor verslaglegging, de tijd besteed aan verslaglegging na werktijd en de aanvullingen in verslaglegging door de zorgverlener. Deze zijn gemeten met behulp van een Likert-schaal, procentuele verschillen en beschrijvende analyses.

Tijd voor verslaglegging

Samenvattend: De introductie van AI-scribes lijkt een positief effect te hebben op de tijd die aan documentatie wordt besteed. De efficiëntie van verslaglegging kan daarnaast afhankelijk zijn van de mate van gebruik van AI-scribes. Fouten werden echter genoemd als een factor die de efficiëntie kan verminderen.

Zwarte literatuur

Vier studies geven inzicht in de tijd die zorgverleners besteden aan het schrijven van notities: drie mixed-methodsstudies (Cao, Nguyen, Duggan) en een simulatiestudie in Nederland (Van Buchem).

Drie van de vier studies rapporteerden een significante vermindering van 1,7% tot 21,5% in de tijd die per casus aan verslaglegging werd besteed (Cao, Duggan, Van Buchem).

In één van de mixed-methodsstudies (Nguyen) werd beschrijvend gerapporteerd dat sommige zorgverleners wel een verbetering in documentatie-efficiëntie ervoeren, terwijl anderen geen impact merkten, onder andere omdat de AI-scribe moeite had met bijvoorbeeld het onderdeel patiënteneducatie en fouten veroorzaakte.

Grijze literatuur

Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde over het verschil in tijd besteed aan documentatie tussen twee AI-scribes en controlegroep (Lukac).

Er werd geen significant verschil gevonden in tijd besteed aan documentatie. Voor één van de AI-scribes (Nabla) werd echter wel een significante vermindering gevonden van 9,5%. De twee AI-scribes

verschilden significant in de mate van tijdsreductie. De auteurs merkten op dat een hoger gebruik correleerde met een sterkere vermindering van de tijd besteed aan documentatie, wat ook gold voor Nabla. Zorgverleners gaven daarnaast aan neutraal te zijn over een vermindering in tijd, gemeten op een 1–5 Likert-schaal (3,7 en 3,5). Er werd geen verschil gevonden tussen de twee AI-scribes.

Verslaglegging na werktijd

Samenvattend: De introductie van AI-scribes lijkt geen effect of mogelijk een positief effect (lees: vermindering) te hebben op de tijd die aan documentatie buiten werktijd wordt besteed.

Zwarte literatuur

Drie mixed-methods studies (Cao, Baldwin, Duggan) geven inzicht in de tijd die zorgverleners besteden aan het schrijven van notities ná werktijd, vóór en na de introductie van AI-scribes.

Twee studies rapporteerden een significante vermindering van de tijd in het EHR ná werktijd van 21% en 30% (Cao, Duggan).

In de andere studie rapporteerde 44% van de zorgverleners een vermindering van de tijd besteed aan documentatie buiten werktijd, terwijl 17% juist een toename rapporteerde (Baldwin).

Grijze literatuur

Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde over het verschil in tijd besteed aan documentatie ná werktijd (Lukac). Er werd geen significant verschil gevonden in de tijd besteed aan documentatie ná werktijd tussen de AI-scribes en reguliere verslaglegging. Zorgverleners waren neutraal over de ervaren vermindering in tijd voor documentatie ná werktijd, gemeten op een 1–5 Likert-schaal.

Bijdrage van zorgprofessionals aan samenvatting

Samenvattend: De kwantitatieve studies laten een significante vermindering zien in de bijdrage van zorgprofessionals aan documentatie opgesteld door een AI-scribe, terwijl de kwalitatieve studie een gemengd beeld toont. De absolute bijdrage kan echter nog steeds hoog zijn, tot wel 51,7%.

Zwarte literatuur

Drie studies (Cao, Duggan, Nguyen) geven inzicht in de bijdrage van zorgprofessionals aan documentatie opgesteld door een AI-scribe.

In twee studies werd een significante verlaging in de bijdrage van zorgprofessionals aan documentatie gevonden: de ene rapporteerde een vermindering van 46,5% (Duggan) en de andere een vermindering van 29,6% (Cao) ten opzichte van de periode vóór de introductie van AI-scribes. De procentuele bijdrage van zorgprofessionals voor de introductie van AI-scribes bedroeg 51,7% in de studie van Cao en was aanzienlijk lager in de studie van Duggan, namelijk 7,9%.

In de derde studie (Nguyen) rapporteerden alle zorgprofessionals dat het noodzakelijk was de notities aan te passen voordat deze ondertekend konden worden, variërend van kleine tot grote aanpassingen.

Should-have 4: Tijdigheid zorgverleners

Samenvattend: De inzet van AI-scribes lijkt een positieve impact te hebben op de tevredenheid van zorgverleners doordat zij meer aandacht kunnen besteden aan het gesprek met hun patiënten.

Zwarte literatuur

Een mixed-methodsstudie met pre-postvergelijking (Duggan) rapporteerde een significant verschil in de tevredenheid van zorgverleners na de introductie van AI-scribes, met betrekking tot afleiding, mate van betrokkenheid bij patiënten en de mogelijkheid om face-to-face met patiënten te praten in plaats van tijd te besteden aan documentatie.

Grijze literatuur

Een RCT (nog onder peer review en daarom onderdeel van de grijze literatuur) rapporteerde over de tevredenheid van zorgprofessionals met consulten die werden gefaciliteerd door een AI-scribe (Lukac). Zorgverleners gaven aan het ermee eens te zijn dat de AI-scribe hen in staat stelde meer met hun patiënten in gesprek te gaan, gemeten op een 1–5 Likert-schaal. De mate van instemming verschilde significant tussen de twee scribes.

Should-have 5: Tevredenheid patiënten

Samenvattend: Meer dan 65% van de patiënten gaf aan tevreden te zijn met de patiënt-zorgverlener interactie.

Zwarte literatuur

De pre-postvergelijkingsstudie van Cao geeft inzicht in de tevredenheid van patiënten met consulten die gefaciliteerd werden door een AI-scribe. Van de patiënten was 65% het ermee eens dat de zorgprofessional tijdens het consult meer aandacht voor hen had. Daarnaast vond 73,9% dat hun consult persoonlijker was en was 78,3% het ermee eens dat de zorgprofessional minder tijd achter de computer doorbracht met typen.

Should-have 6: Kosten

Er zijn geen studies gevonden.

4.3 Kwaliteit van studies

Amstar-2

De systematische literatuurreview (SLR) van Ng et al. is volgens de AMSTAR-2 van lage kwaliteit, waardoor de review mogelijk geen accuraat en volledig overzicht biedt van de beschikbare studies die de onderzoeksvraag adresseren. Aangezien slechts 6 van de 29 geïncludeerde studies in Ng et al. relevant zijn voor onze PICO, is de algemene kwaliteitsbeoordeling van de review zelf van minder belang. De items 1 t/m 9 van de AMSTAR-2 geven echter wel inzicht in de nauwkeurigheid van de verzameling van de studies voor de review, zoals de volledigheid van de inclusie van relevante studies en de beoordeling van hun kwaliteit. Hier zien we een kritieke tekortkoming, namelijk het ontbreken van een lijst met geëxcludeerde studies. Daarnaast wordt niet toegelicht waarom bepaalde studiedesigns voor inclusie zijn gekozen.

Voor ons onderzoek zijn de uiteindelijk geïncludeerde studies (prospectieve en retrospectieve studies, pilot- en feasibility-studies) passend. Het ontbreken van een lijst met uitgesloten fulltekststudies maakt het echter onmogelijk te verifiëren of de selectie zorgvuldig is uitgevoerd.

Verder hebben Ng et al. de voor ons relevante zes studies beoordeeld met behulp van de QUADAS-2. Deze tool is primair ontwikkeld voor diagnostic accuracy studies en is daarom minder geschikt voor ons type interventieonderzoek.

Daarom is aanvullend de MMAT toegepast op deze zes studies om de methdologische kwaliteit te beoordelen.

MMAT

13 primaire studies zijn geëvalueerd met de MMAT. Acht van deze studies werden door de beoordelaar als voldoende van kwaliteit beschouwd, aangezien zij op meer dan 50% van de MMAT-vragen over kwaliteit positief scoorden (zie de tabel hieronder voor een overzicht). In sommige studies was het niet duidelijk of een vraag over kwaliteit positief of negatief beantwoord moest worden; deze gevallen zijn als negatief beschouwd.

Onvoldoende kwaliteit werd vooral gevonden in de simulatiestudies. Hoewel sampling belangrijk zou zijn om de onderzoeksvraag (kwaliteit van AI-gegenereerde notities) te beantwoorden, is gekozen voor een simulatiedesign met óf mockconsulten óf real-world consulten waarbij niet te beoordelen is hoe de selectie is uitgevoerd. Ook de selectie van beoordelaars van de AI-notities was onduidelijk of niet representatief (bijvoorbeeld studenten). Daardoor is de steekproef als geheel niet representatief. In één van de studies (Misurac) was geen full-tekst beschikbaar om de kwaliteit van de studie te beoordelen, maar het was wel mogelijk om de relevante data te extraheren. Deze studie werd dus ook als van onvoldoende kwaliteit beschouwd.

(Eerste) auteur	Studiedesign	Kwaliteitsbeoordeling*
Lukac	RCT	Voldoende
Cao	Mixed-methods pilot study	Voldoende
Ong	Feasibility study (quantitative)	Voldoende
Misurac	Pre-post observational study	Onvoldoende
Moryousef	Simulation study	Voldoende
Van Buchem	Simulation study	Onvoldoende
Hose	Simulation study	Onvoldoende
Anderson	Simulation study	Onvoldoende
Basei de Paula	Experience/implementation based on usage metrics	Onvoldoende
Biro	Simulation study	Voldoende
Baldwin	Mixed-methods pilot study	Voldoende
Nguyen	Mixed-methods pilot study	Voldoende
Duggan	Mixed-methods study	Voldoende

* *Studies met meer dan 50% positieve antwoorden op de kwaliteitsvragen werden als voldoende van kwaliteit beschouwd. Dit is geen score op basis van de MMAT, maar dient alleen om een algemene indicatie te geven van de kwaliteit van de studie.*

5 Discussie

Voor alle relevante meetplanindicatoren behalve de should-have-indicator 'Kosten', werden uitkomsten gevonden. De meest gerapporteerde thema's in de geïncludeerde literatuur waren de kwaliteit van de samenvattingen (n=8), de houding van zorgverleners (n=7) en de impact op administratieve tijd (n=6). Minder vaak werd gerapporteerd over het werkplezier van zorgverleners (n=4), de houding van patiënten (n=3), gebruik (n=4) en patiënt-zorgverlenercommunicatie vanuit het perspectief van zorgverleners (n=2) en patiënten (n=1).

De geïncludeerde studies verschilden in bewijskracht op basis van hun studiedesign. Slechts één studie had een design met hoge bewijskracht: een gerandomiseerde klinische trial waarin AI-scribes werden vergeleken met elkaar en met een controlegroep, maar deze studie was nog niet peer-reviewed. De resterende studies hadden een design met lage bewijskracht, zoals simulatiestudies, pre-post observationele studies, implementatiestudie en mixed-methodsstudies. De beoordeling van de methodologische kwaliteit laat zien dat vooral de simulatiestudies van onvoldoende kwaliteit werden bevonden, aangezien deze niet representatief zijn voor de doelgroep. De overige acht studies werden wel als van voldoende kwaliteit beschouwd. Hoewel het studiedesign van de meeste studies zwak was, lijkt de uitvoering in veel gevallen degelijk te zijn en kunnen de bevindingen uit de studies als betrouwbaar worden beschouwd.

In het merendeel van de studies is het nieuwe proces van rapporteren onderzocht ten opzichte van de reguliere verslaglegging. Dit maakt het mogelijk de toegevoegde waarde van SGR voor de relevante uitkomsten te beoordelen. Omdat een derde van de studies echter is uitgevoerd in de Amerikaanse zorgsetting, waar reguliere verslaglegging vaak het gebruik van een menselijke scribe omvat, kan het zijn dat de bevindingen van de vergelijking tussen reguliere verslaglegging en SGR niet volledig overeenkomen met de vooraf bepaalde PICO voor dit onderzoek. Of de reguliere verslaglegging daadwerkelijk een menselijke scribe omvatte, werd echter niet expliciet vermeld in de opgenomen studies en is daardoor lastig vast te stellen.

De studies geven inzicht in een breed spectrum van specialismen. In vier studies werd gerapporteerd dat een groot aantal specialismen werd geïncludeerd. Deze vier studies geven gezamenlijk inzicht in alle relevante uitkomsten, behalve patiënt-zorgverlenerbetrokkenheid vanuit het perspectief van de patiënt en kosten. In vijf studies werd daarnaast het medisch specialisme specifiek genoemd: oncologie, dermatologie en urologie. De studies dekken dus een breed spectrum aan medische specialismen voor het merendeel van de uitkomsten. De bevindingen lijken daarmee representatief te zijn voor de medisch-specialistische zorgsetting.

Maar een deel van de studies gaf inzicht in welk product werd gebruikt. Het meest genoemde product was de Dragon Ambient eXperience (DAX) (Microsoft), dat in Nederland (nog) niet wordt toegepast. In de Braziliaanse setting werd de AI-scribe Voa onderzocht en in de Nederlandse studie werd Autoscriber onderzocht. Verder onderzocht één studie vijf vrij toegankelijke AI-scribes, namelijk Scribe MD, Heidi, Scribeberry, Tali en Nabla. AI-scribes gericht op de Nederlandse of Europese zorgsetting zijn dus nog nauwelijks onderzocht in de wetenschappelijke literatuur. Uit de opgenomen literatuur is onduidelijk of de verschillende AI-scribes significant van elkaar verschillen in hun prestaties. In de RCT van Lukac werd namelijk geen significant verschil gevonden in de kwaliteit van AI-gegenereerde notities tussen DAX en Nabla, terwijl in de simulatiestudie van Anderson wel een significant verschil werd aangetoond tussen vijf verschillende AI-scribes. Ook in de simulatiestudie van Biro werd een significant verschil gevonden tussen twee AI-scribeproducten met betrekking tot foutgevoeligheid. Dit verschil werd echter niet gevonden voor DAX en Nabla in de studie van Lukac. Het is onduidelijk of er tussen de AI-applicaties die in dit onderzoek zijn opgenomen en de applicaties die in de Nederlandse setting worden gebruikt, een significant verschil te verwachten is.

6 Conclusie

Naar de meeste indicatoren lijkt voldoende onderzoek te zijn gedaan, behalve voor de kosten en de objectieve kwaliteit van AI-gegeneerde notities. Maar door verschillen in AI-scribe producten onderzocht in de literatuur en de Nederlandse zorgsetting is het lastig de bevindingen één-op-één te vertalen naar de Nederlandse context. AI-scribes kunnen mogelijk verschillen in kwaliteit en daarmee in de tijd die besteed moet worden aan het bewerken van notities, het ervaren werkplezier en de houding van patiënten en zorgverleners ten opzichte van het gebruik ervan. Ook worden de AI-modellen in AI-scribes steeds doorontwikkeld en komen er nieuwe versies beschikbaar, wat maakt dat de bevindingen over de tijd kunnen verschillen.

In Bijlage F staat een overzicht van toepassingen die momenteel in de Nederlandse MSZ getest worden, maar waarvoor op dit moment nog geen gepubliceerde gegevens over de uitkomsten beschikbaar zijn.

7 Referenties

Anderson, T. N., Mohan, V., Dorr, D. A., Ratwani, R. M., Biro, J. M., & Gold, J. A. (2025). Evaluating the Quality and Safety of Ambient Digital Scribe Platforms Using Simulated Ambulatory Encounters. https://doi.org/10.2139/ssrn.5255300

Baldwin, A., Lipitz-Snyderman, A., Goyal, S., Kuperman, G., Stetson, P. D., Chimonas, S., Watson, A., & Damstrom, M. (2025). Early insights into the impact of an ambient AI scribe solution on clinical documentation to reduce clinician burnout in oncology. Journal of Clinical Oncology, 43(16_suppl). https://doi.org/10.1200/JCO.2025.43.16_suppl.e13722

Basei De Paula, P. A., Berger, M. N., Bruneti Severino, J. V., Ribeiro, K. D. P., Loures, F. S., Todeschini, S. A., Roeder, E. A., & Lenci Marques, G. (2024). Improving Documentation Quality and Patient Interaction with AI: A Tool for Transforming Medical Records — An Experience Report. https://doi.org/10.32388/FFZGE5

Biro, J., Handley, J. L., Cobb, N. K., Kottamasu, V., Collins, J., Krevat, S., & Ratwani, R. M. (2025). Accuracy and Safety of AI-Enabled Scribe Technology: Instrument Validation Study. Journal of Medical Internet Research, 27, e64993. https://doi.org/10.2196/64993

Cao, D. Y., Silkey, J. R., Decker, M. C., & Wanat, K. A. (2024). Artificial intelligence-driven digital scribes in clinical documentation: Pilot study assessing the impact on dermatologist workflow and patient encounters. JAAD International, 15, 149–151. https://doi.org/10.1016/j.jdin.2024.02.009

Duggan, M. J., Gervase, J., Schoenbaum, A., Hanson, W., Howell, J. T., Sheinberg, M., & Johnson, K. B. (2025). Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency. JAMA Network Open, 8(2), e2460637. https://doi.org/10.1001/jamanetworkopen.2024.60637

Hose, B.-Z., Handley, J. L., Biro, J., Reddy, S., Krevat, S., Hettinger, A. Z., & Ratwani, R. M. (2025). Development of a Preliminary Patient Safety Classification System for Generative AI. BMJ Quality & Safety, 34(2), 130–132. https://doi.org/10.1136/bmjqs-2024-017918

Lukac, P. J., Turner, W., Vangala, S., Chin, A. T., Khalili, J., Shih, Y.-C. T., Sarkisian, C., Cheng, E. M., & Mafi, J. N. (2025). A Randomized-Clinical Trial of Two Ambient Artificial Intelligence Scribes: Measuring Documentation Efficiency and Physician Burnout. https://doi.org/10.1101/2025.07.10.25331333

Misurac, J., Knake, L. A., & Blum, J. M. (2025). The Effect of Ambient Artificial Intelligence Notes on Provider Burnout. Applied Clinical Informatics, 16(02), 252–258. https://doi.org/10.1055/a-2461-4576

Moryousef, J., Nadesan, P., Uy, M., Matti, D., & Guo, Y. (2025). Assessing the Efficacy and Clinical Utility of Artificial Intelligence Scribes in Urology. Urology, 196, 12–17. https://doi.org/10.1016/j.urology.2024.11.061

Ng, J. J. W., Wang, E., Zhou, X., Zhou, K. X., Goh, C. X. L., Sim, G. Z. N., Tan, H. K., Goh, S. S. N., & Ng, Q. X. (2025). Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: A systematic review. BMC Medical Informatics and Decision Making, 25(1), 236. https://doi.org/10.1186/s12911-025-03061-0

Nguyen, O. T., Turner, K., Charles, D., Sprow, O., Perkins, R., Hong, Y.-R., Islam, J. Y., Khanna, N., Alishahi Tabriz, A., Hallanger-Johnson, J., Bickel Young, J., & Moore, C. E. (2023). Implementing Digital Scribes to Reduce Electronic Health Record Documentation Burden Among Cancer Care Clinicians: A Mixed-Methods Pilot Study. JCO Clinical Cancer Informatics, 7, e2200166. https://doi.org/10.1200/CCI.22.00166

Ong, J. H. W., Tung, J. Y. M., Sng, G. G. R., Lim, D. Y. Z., Yang, X., Mohamad Rafi, I. B., Loke, M. Z. Y., Kumar, A., Tan, J. Y. E., Lew, E. Y. L., Heng, M. K. H., & Ho, H. S. S. (2025). A pilot study using ambient artificial intelligence scribes in clinical documentation in a urology outpatient clinic. BJU International, bju.16784. https://doi.org/10.1111/bju.16784

Sasseville, M., Yousefi, F., Ouellet, S., Naye, F., Stefan, T., Carnovale, V., Bergeron, F., Ling, L., Gheorghiu, B., Hagens, S., Gareau-Lajoie, S., & LeBlanc, A. (2025). The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. Healthcare, 13(12), 1447. https://doi.org/10.3390/healthcare13121447

Van Buchem, M. M., Kant, I. M. J., King, L., Kazmaier, J., Steyerberg, E. W., & Bauer, M. P. (2024). Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study. JMIR AI, 3, e60020. https://doi.org/10.2196/60020

Bijlagen

Bijlage A – Meetplan

Hoofddomein	Waardebepalings-criterium (indicator)	Algemene omschrijving van de indicator	MOSCOW	Specificatie indicator op zorgproces
1. Kwaliteit	Kwaliteit	De mate waarin SOLL veilig is, zowel op gebied van zorg en gezondheid.	1. M	De gestructureerde samenvatting is medisch inhoudelijk volledig, accuraat, relevant (geen onnodige informatie, gebruik van gestructureerde en gecodeerde data, mate van granulariteit), en correct georganiseerd (informatie komt in de juiste velden/kopjes terecht).
1. Kwaliteit	Adoptie (zorgprofessionals /ondersteuners)	Hoe zorgprofessionals reageren op SOLL (in vergelijking met IST) in gedrag en SGR adopteren.	2. S	Het aantal zorgprofessionals dat SGR gebruikt neemt toe.
1. Kwaliteit	Adoptie (zorgprofessionals /ondersteuners)	Hoe zorgprofessionals reageren op SOLL (in vergelijking met IST) in gedrag en SGR ervaren.	1. M	SGR gebruikende zorgprofessionals staan neutraal of positief tegenover het gebruik SGR.
1. Kwaliteit	Acceptatie	Hoe de beoogde individuele eindgebruikers reageren op SOLL (in vergelijking met IST): ervaringen/percepties (bijv. nut, gemak, vertrouwen, houding, intentie, aanraden van het proces aan anderen).	1. M	Patiënten waarbij SGR is toegepast in het consult, staan neutraal of positief niet negatief tegenover het gebruik SGR.
4. Professioneel welbevinden	Werkplezier, werktevredenheid, of werkdruk (focus werkplezier)	Hoe de zorgverlener de arbeidsomstandigheden ervaart.	2. S	"Door gebruik van SGR ervaart de zorgprofessional tijdens het consult meer werkplezier. "
4. Professioneel welbevinden	Werkplezier, werktevredenheid, of werkdruk (focus werkdruk)	Hoe de zorgverlener de arbeidsomstandigheden ervaart.	2. S	Inzet van SGR vermindert administratieve druk voor de zorgprofessionals tijdens en na het consult.
1. Kwaliteit	Tijdigheid	De mate waarin eindgebruikers (van SOLL) de juiste zorg of ondersteuning op het juiste moment door de juiste professional en op de juiste plek ontvangen.	2. S	Door inzet van SGR hebben zorgprofessional en patiënt meer aandacht voor het gesprek. De zorgprofessionals ervaren door SGR een verbetering van de kwaliteit van het gesprek.**
1. Kwaliteit	Tijdigheid	De mate waarin eindgebruikers (van SOLL) de juiste zorg of ondersteuning op het juiste moment door de juiste professional en op de juiste plek ontvangen.	2. S	Door inzet van SGR hebben zorgprofessional en patiënten meer aandacht voor het gesprek. De patiënten zijn tevreden over de kwaliteit van het gesprek. **
2. Betaalbaarheid	Kosten binnen de gezondheidszorg	De mate waarin de structurele kosten van zorg en ondersteuning die verband houden met de uitvoering van het proces na afronding van SOLL (in vergelijking met IST).	2. S	De structurele kosten wegen op tegen verbetering werkomgeving van zorgprofessional en verbetering in kwaliteit van zorg.

Bijlage B – Criteria grijze literatuur

Grijze literatuur – Criteria voor inclusie/exclusie

Selectiecriteria (inclusie)	
Het document bevat resultaten/ uitkomsten van een (pilot/implementatie/onderzoeks) traject*	Ja/Nee
Het document bevat een methode**	Ja/Nee
Het document beschrijf de juiste setting (sector en SOLL)	Ja/Nee
Het document heeft een auteur / afzender***	Ja/Nee
Het document is recent (< 5 jaar)	Ja/Nee
Inclusie bij 5x JA	
* mag ook kwalitatieve data beschrijven vanuit interviews of focusgroepen	
** het is helder waar de data vandaan komt en wat ermee gedaan is	
*** kan ook een instelling, bedrijf of leverancier zijn	
Exclusie	
Documenten voor promotiedoeleinden zonder data/onderbouwing	
Documenten zonder uitkomsten/resultaten	
Documenten/presentaties met uitkomsten/resultaten die niet herleidbaar zijn tot het brondocument	
Handleidingen bij toepassingen	
Documenten die de productontwikkeling/verantwoording beschrijven	
Studieprotocollen	
Achtergrond informatie (bijvoorbeeld bij werkingsmaganisme techniek SGR)	
Hypothetische business case beschrijvingen	

Bijlage C – Zoekstrategieën

OVID/Medline 13 jul 2025

Ovid MEDLINE(R) ALL <1946 to July 11, 2025>

1	exp "Speech Recognition Software"/ or (google-speech or ambient-scribe* or digital-scribe* or ai-scribe* or virtual-scribe* or deepscribe* or deep-scribe* or ai-generated-scribe* or machine-learning-scribe* or automatic-scribe* or artificial-intelligence-scribe* or automatic-transcri* or ai-transcri* or virtual-transcri* or deeptranscri* or deep-transcri* or ambient-transcri* or machine-learning-transcri* or artificial-intelligence-transcri* or ai-generated-transcri* or sassi or vocera or ((speech or command* or voice-recogn*) adj3 (interface* or software* or system* or automat* or smartwatch or smart-watch or technolog*)) or ((dictat* or speech) adj3 (text or command*) adj3 (executi* or functional* or interface*)) or voice-activated-system* or "attendi" or "tell james" or "medendo assistent" or "juvoly quickconsult" or "digitale assistent" or "autoscriber" or "health talk" or "ontzorgd" or "verticai" or "(nuance adj3 microsoft)" or "tonos" or "cato" or "elanza" or "smart planning" or "youplan" or "voca zorg" or "epic assistent" or "robin scribe" or "consult assistent" or "g2speech" or "ourmind" or "wellcom health").ti,ab,kf.	**6496**
2	"Electronic Health Records"/ or "Medical Records"/ or "Medical Records Systems, Computerized"/ or "Documentation"/ or (medical-record* or electronic-medical-record* or electronic-health-record* or computerized-medical-record* or computerised-medical-record* or computerized-patient-medical-records* or computerised-patient-medical-records* or electronic-health-record* or automated-medical-record* or electronic-patient-record* or electronic-client-record* or electronic-client-documentation* or electronic-patient-documentation* or clinical-documentation* or physician-documentation* or pdqi* or medical-note* or nurse-documentation* or medical-information-management* or automated-report*).ti,ab,kf.	**306443**
3	1 and 2	**399**

Embase.com 13 jul 2025

No.	Query	Results
#4	#3 NOT ('Conference Abstract'/it OR 'Conference Paper'/it)	**415**
#3	#1 AND #2	**478**
#2	'electronic health record'/de OR 'electronic medical record'/de OR 'electronic patient record'/de OR 'medical record'/de OR 'electronic medical record system'/de OR 'medical documentation'/de OR 'medical record*':ti,ab,kw OR 'electronic medical record*':ti,ab,kw OR 'computerized medical record*':ti,ab,kw OR 'computerised medical record*':ti,ab,kw OR 'computerized patient medical records*':ti,ab,kw OR 'computerised patient medical records*':ti,ab,kw OR 'electronic health record*':ti,ab,kw OR 'automated medical record*':ti,ab,kw OR 'electronic patient record*':ti,ab,kw OR 'electronic client record*':ti,ab,kw OR 'electronic client documentation*':ti,ab,kw OR 'electronic patient documentation*':ti,ab,kw OR 'clinical documentation*':ti,ab,kw OR 'physician documentation*':ti,ab,kw OR pdqi*:ti,ab,kw OR 'medical note*':ti,ab,kw OR 'nurse documentation*':ti,ab,kw OR 'medical information management*':ti,ab,kw OR 'automated report*':ti,ab,kw	**561666**
#1	'automatic speech recognition'/exp OR 'google speech':ti,ab,kw OR 'ambient scribe*':ti,ab,kw OR 'digital scribe*':ti,ab,kw OR 'ai scribe*':ti,ab,kw OR 'virtual scribe*':ti,ab,kw OR deepscribe*:ti,ab,kw OR 'deep scribe*':ti,ab,kw OR 'ai generated scribe*':ti,ab,kw OR 'machine learning scribe*':ti,ab,kw OR 'automatic scribe*':ti,ab,kw OR 'artificial intelligence scribe*':ti,ab,kw OR 'automatic transcri*':ti,ab,kw OR 'ai transcri*':ti,ab,kw OR 'virtual transcri*':ti,ab,kw OR deeptranscri*:ti,ab,kw OR 'deep transcri*':ti,ab,kw OR 'ambient transcri*':ti,ab,kw OR 'machine learning transcri*':ti,ab,kw OR 'artificial intelligence transcri*':ti,ab,kw OR 'ai generated transcri*':ti,ab,kw OR sassi:ti,ab,kw OR vocera:ti,ab,kw OR (((speech OR command* OR 'voice recogn*') NEAR/3 (interface* OR software* OR system* OR automat* OR smartwatch OR 'smart watch' OR technolog*)):ti,ab,kw) OR (((dictat* OR speech) NEAR/3 (text OR command*) NEAR/3 (executi* OR functional* OR interface*)):ti,ab,kw) OR 'voice activated system*':ti,ab,kw OR 'attendi':ti,ab,kw OR 'tell james':ti,ab,kw OR 'medendo assistent':ti,ab,kw OR 'juvoly quickconsult':ti,ab,kw OR 'digitale assistent':ti,ab,kw OR 'autoscriber':ti,ab,kw OR 'health talk':ti,ab,kw OR 'ontzorgd':ti,ab,kw OR 'verticai':ti,ab,kw OR ('nuance' NEAR/3 'microsoft'):ti,ab,kw OR 'tonos':ti,ab,kw OR 'cato':ti,ab,kw OR 'elanza':ti,ab,kw OR 'smart planning':ti,ab,kw OR 'youplan':ti,ab,kw OR 'voca zorg':ti,ab,kw OR 'epic assistent':ti,ab,kw OR 'robin scribe':ti,ab,kw OR 'consult assistent':ti,ab,kw OR 'g2speech':ti,ab,kw OR 'ourmind':ti,ab,kw OR 'wellcom health':ti,ab,kw	**8111**

Elsevier/Scopus 13 jul 2025

3	**#1 AND #2**	**258**
2	TITLE-ABS (medical-record* OR electronic-medical-record* OR electronic-health-record* OR computerized-medical-record* OR computerised-medical-record* OR computerized-patient-medical-records* OR computerised-patient-medical-records* OR electronic-health-record* OR automated-medical-record* OR electronic-patient-record* OR electronic-client-record* OR electronic-client-documentation* OR electronic-patient-documentation* OR clinical-documentation* OR physician-documentation* OR pdqi* OR medical-note* OR nurse-documentation* OR medical-information-management* OR automated-report*) OR AUTHKEY (medical-record* OR electronic-medical-record* OR electronic-health-record* OR computerized-medical-record* OR computerised-medical-record* OR computerized-patient-medical-records* OR computerised-patient-medical-records* OR electronic-health-record* OR automated-medical-record* OR electronic-patient-record* OR electronic-client-record* OR electronic-client-documentation* OR electronic-patient-documentation* OR clinical-documentation* OR physician-documentation* OR pdqi* OR medical-note* OR nurse-documentation* OR medical-information-management* OR automated-report*)	**249600**
1	TITLE-ABS(google-speech OR ambient-scribe* OR digital-scribe* OR ai-scribe* OR virtual-scribe* OR deepscribe* OR deep-scribe* OR ai-generated-scribe* OR machine-learning-scribe* OR automatic-scribe* OR artificial-intelligence-scribe* OR automatic-transcri* OR ai-transcri* OR virtual-transcri* OR deeptranscri* OR deep-transcri* OR ambient-transcri* OR machine-learning-transcri* OR artificial-intelligence-transcri* OR ai-generated-transcri* OR sassi OR vocera OR ((speech OR command* OR voice-recogn*) W/3 (interface* OR software* OR system* OR automat* OR smartwatch OR smart-watch OR technolog*)) OR ((dictat* OR speech) W/3 (text OR command*) W/3 (executi* OR functional* OR interface*)) OR voice-activated-system* OR "attendi" OR "tell james" OR "medendo assistent" OR "juvoly quickconsult" OR "digitale assistent" OR "autoscriber" OR "health talk" OR "ontzorgd" OR "verticai" OR ("nuance" W/3 "microsoft") OR "tonos" OR "cato" OR "elanza" OR "smart planning" OR "youplan" OR "voca zorg" OR "epic assistent" OR "robin scribe" OR "consult assistent" OR "g2speech" OR "ourmind" OR "wellcom health") OR AUTHKEY(google-speech OR ambient-scribe* OR digital-scribe* OR ai-scribe* OR virtual-scribe* OR deepscribe* OR deep-scribe* OR ai-generated-scribe* OR machine-learning-scribe* OR automatic-scribe* OR artificial-intelligence-scribe* OR automatic-transcri* OR ai-transcri* OR virtual-transcri* OR deeptranscri* OR deep-transcri* OR ambient-transcri* OR machine-learning-transcri* OR artificial-intelligence-transcri* OR ai-generated-transcri* OR sassi OR vocera OR ((speech OR command* OR voice-recogn*) W/3 (interface* OR software* OR system* OR automat* OR smartwatch OR smart-watch OR technolog*)) OR ((dictat* OR speech) W/3 (text OR command*) W/3 (executi* OR functional* OR interface*)) OR voice-activated-system* OR "attendi" OR "tell james" OR "medendo assistent" OR "juvoly quickconsult" OR "digitale assistent" OR "autoscriber" OR "health talk" OR "ontzorgd" OR "verticai" OR ("nuance" W/3 "microsoft") OR "tonos" OR "cato" OR "elanza" OR "smart planning" OR "youplan" OR "voca zorg" OR "epic assistent" OR "robin scribe" OR "consult assistent" OR "g2speech" OR "ourmind" OR "wellcom health")	**77119**

Clarivate Analytics/Web of Science Core Collection 13 jul 2025

3	**#2 AND #1**	**139**
2	TS=("medical record*" OR "electronic medical record*" OR "electronic health record*" OR "computerized medical record*" OR "computerised medical record*" OR "computerized patient medical records*" OR "computerised patient medical records*" OR "electronic health record*" OR "automated medical record*" OR "electronic patient record*" OR "electronic client record*" OR "electronic client documentation*" OR "electronic patient documentation*" OR "clinical documentation*" OR "physician documentation*" OR "pdqi*" OR "medical note*" OR "nurse documentation*" OR "medical information management*" OR "automated report*")	**203319**
1	TS=("google speech" OR "ambient scribe*" OR "digital scribe*" OR "ai scribe*" OR "virtual scribe*" OR "deepscribe*" OR "deep scribe*" OR "ai generated scribe*" OR "machine learning scribe*" OR "automatic scribe*" OR "artificial intelligence scribe*" OR "automatic transcri*" OR "ai transcri*" OR "virtual transcri*" OR "deeptranscri*" OR "deep transcri*" OR "ambient transcri*" OR "machine learning transcri*" OR "artificial intelligence transcri*" OR "ai generated transcri*" OR "sassi" OR "vocera" OR (("speech" OR "command*" OR "voice recogn*") NEAR/3 ("interface*" OR "software*" OR "system*" OR "automat*" OR "smartwatch" OR "smart watch" OR "technolog*")) OR (("dictat*" OR "speech") NEAR/3 ("text" OR "command*") NEAR/3 ("executi*" OR "functional*" OR "interface*")) OR "voice activated system*" OR "attendi" OR "tell james" OR "medendo assistent" OR "juvoly quickconsult" OR "digitale assistent" OR "autoscriber" OR "health talk" OR "ontzorgd" OR "verticai" OR ("nuance" NEAR/3 "microsoft") OR "tonos" OR "cato" OR "elanza" OR "smart planning" OR "youplan" OR "voca zorg" OR "epic assistent" OR "robin scribe" OR "consult assistent" OR "g2speech" OR "ourmind" OR "wellcom health")	**21445**

Zoekresultaten overzicht

Database	Aantal treffers vóór ontdubbelen (2024)	Aantal treffers ná ontdubbelen (2024)	Aantal treffers vóór ontdubbelen (2025)	Aantal treffers ná ontdubbelen (2025)
Medline	16 SR + 343 niet-SR	16 SR + 341 niet-SR	399	35
Embase	362	107	415	32
Scopus	219	97	258	17
WoS	115	8	139	2
Totaal	1055	569	1211	86

Bijlage D – Studies per uitkomst

ID	Eerste auteur	Kwaliteit van samenvatting	Houding patienten	Houding zorgverlener	Gebruik	Werkplezier	Administratieve tijd	Patient-zorg-verlener-relatie (patient)	Patient-zorg-verlener-relatie (zorgverlener)	Kosten
1	Lukac	1	1	1	1	1	1	0	1	0
2	Baldwin	0	0	1	1	0	1	0	0	0
3	Basei de Paula	0	0	0	1	0	0	0	0	0
4	Ong	1	0	0	0	0	0	0	0	0
5a	Biro	1	0	0	0	0	0	0	0	0
6a	Cao	0	0	1	0	0	1	1	0	0
7a	Duggan	1	0	1	0	1	1	0	1	0
8a	Misurac	0	0	0	0	1	0	0	0	0
9a	Moryousef	1	0	1	0	0	0	0	0	0
10a	Van Buchem	1	0	1	0	0	1	0	0	0
11b	Nguyen	0	0	1	1	1	1	0	0	0
12	Hose	1	1	0	0	0	0	0	0	0
13	Anderson	1	1	0	0	0	0	0	0	0

Bijlage E – Evidence gap

Wat is een 'evidence gap' en hoe is deze in dit onderzoek bepaald?

Er is óf helemaal geen bewijs, óf het bestaande bewijs is onvoldoende, van lage kwaliteit, of niet relevant voor de onderzoeksvraag.

De 'evidence gap' in de huidige literatuur werd bepaald op basis van de bewijskracht van de studiedesigns, de methodologische kwaliteit van de studies, het al dan niet aanwezig zijn van vergelijkingen en de steekproefgrootte. Als er voldoende bewijs beschikbaar is, is aanvullend onderzoek niet nodig om de indicatoren te beoordelen op basis van een vooraf vastgestelde acceptatiegrens.

Uiteindelijk werd de 'evidence gap' gezamenlijk met het veld vastgesteld. Dit overzicht dient ter informatie.

Voor alle relevante indicatoren voor SGR in de MSZ werd wetenschappelijke literatuur gevonden, behalve voor de should-have-indicator 'Kosten'. Het onderzoek laat verder zien dat de indicatoren die met de sector MSZ zijn opgesteld om de waarde van SGR in de MSZ te evalueren, overeenkomen met de uitkomsten in de wetenschappelijke literatuur. Dit suggereert dat deze indicatoren goed meetbaar zijn en dat de evidence gap daardoor niet is ontstaan.

De kwaliteit van AI-gegenereerde notities werd beoordeeld op algemene kwaliteit en op specifieke aspecten: accuraatheid, relevantie, organisatie en correctheid van de samenvatting (inclusief bias, fouten en hallucinaties), in acht studies; vijf daarvan hadden een voldoende methodologische kwaliteit (zie tabel). Van de studies met voldoende methodologische kwaliteit werd in twee studies de kwaliteit van AI-scribes vergeleken met reguliere verslaglegging. Maar één studie van deze twee had een design met hoge bewijskracht (RCT) en een grote steekproef (n=238). Daarnaast was er een mixed-methodsstudie van voldoende methodologische kwaliteit met een redelijk aantal deelnemers (n=46), maar zonder vergelijking.

De studies onderzochten echter vooral de subjectieve kwaliteit van de door AI gegenereerde samenvattingen, bepaald door zorgverleners via handmatige beoordeling, bijvoorbeeld met instrumenten zoals de PDQI-9 of een Likert-schaal. Maar de onderliggende LLM's zijn niet-deterministisch (de output kan bij dezelfde input telkens verschillen) en zijn onderhevig aan frequente versiewijzigingen, die de resultaten in de tijd sterk kunnen beïnvloeden. Hierdoor is het onrealistisch om uitsluitend te vertrouwen op continue feedback van eindgebruikers als waarborg voor kwaliteit. Sterker nog, dit kan de werkdruk van zorgprofessionals juist verhogen. De validatie van AI-scribes verschuift hiermee dus grotendeels naar de implementatiefase. Er zijn nog onvoldoende studies bekend dat LLM's zelf (gedeeltelijk) de kwaliteitscontrole doet, door automatisch de kwaliteit van gegenereerde samenvattingen te toetsen. In Nederland wordt er daarom ook door het RIGH:T consortium onderzoek gedaan naar een geschikt validatieframework om de kwaliteit en veiligheid van verschillende AI-scribes te kunnen beoordelen. Bron: https://bio.link/rightconsortium

Dus, ook al is er mogelijk straks (na peerreview van de RCT) voldoende onderzoek gedaan naar de subjectieve kwaliteit, is er onvoldoende (geen) bewijs over de objectieve kwaliteit van door AI-scribes gegenereerde samenvattingen. Voor welke kwaliteitsonderdelen dit van belang is, moet verder door het veld worden bepaald.

Kwaliteit van AI-gegenereerde notities

(Eerste) auteur	Studiedesign	Vergelijking	Steekproefgrootte	Kwaliteits-beoordeling*
Lukac	RCT	Ja*	238	Voldoende
Ong	Feasibility study (quantitative)	Ja	5	Voldoende
Biro	Simulation study	Nee	11	Voldoende
Duggan	Mixed-methods study	Nee	46	Voldoende
Moryousef	Simulation study	Nee	4	Voldoende
Van Buchem	Simulation study	Ja	27	Onvoldoende
Hose	Simulation study	Nee	11	Onvoldoende
Anderson	Simulation study	Nee	14	Onvoldoende

* *Likert-schaal met een stelling over de kwaliteit van AI-notities ten opzichte van reguliere verslaglegging.*

De houding van patiënten werd geëvalueerd aan de hand van algemene houding en patiëntveiligheid in drie studies; een daarvan had een voldoende methodologische kwaliteit, een design van hoge bewijskracht en een groot steekproef (n=238). De studie is momenteel alleen als preprint beschikbaar. Verder waren de bevindingen over de algemene houding gebaseerd op het perspectief van zorgverleners die toestemming aan patiënten moesten vragen en kwamen dus niet rechtstreeks van de patiënten zelf.

Praktijkonderzoek naar de houding van patiënten, rechtstreeks vanuit de patiënten zelf, ten aanzien van AI-scribes kan meer inzicht geven in de informatiebehoeften rondom SGR in de MSZ vanuit patiëntperspectief.

Houding van patiënten

(Eerste) auteur	Studiedesign	Vergelijking	Steekproefgrootte	Kwaliteits-beoordeling*
Lukac	RCT	Nee	238	Voldoende
Hose	Simulation study	Nee	11	Onvoldoende
Anderson	Simulation study	Nee	14	Onvoldoende

De houding van zorgverleners werd geëvalueerd op basis van bruikbaarheid en toekomstig gebruik in zeven studies. Zes daarvan hadden een voldoende methodologische kwaliteit; één studie met een design van hoge bewijskracht (preprint-RCT) en vijf studies met een design van lagere bewijskracht. Alleen in één studie werd een voor-en-na-vergelijking gemaakt om te bepalen of de houding van zorgverleners veranderde na de introductie van een AI-scribe. De omvang van de geïncludeerde zorgverleners is redelijk representatief, met 238 deelnemers in de RCT en in totaal 108 deelnemers in alle mixed-methodsstudies.

Er lijkt voldoende onderzoek te zijn gedaan naar om de houding van zorgverleners ten opzichte van AI-scribes.

Houding van zorgverleners

(Eerste) auteur	Studiedesign	Vergelijking	Steekproef-grootte	Kwaliteits-beoordeling*
Lukac	RCT	Nee	238	Voldoende
Duggan	Mixed-methods study	Ja	46	Voldoende
Moryousef	Simulation study	Nee	4	Voldoende
Van Buchem	Simulation study	Nee	27	Onvoldoende
Baldwin	Mixed-methods pilot study	Nee	31	Voldoende
Cao	Mixed-methods pilot study	Nee	10	Voldoende
Nguyen	Mixed-methods pilot study	Nee	21	Voldoende

Het gebruik van SGR in de MSZ werd vooral geëvalueerd op basis van gebruiksstatistieken (m.b.t. het percentage gebruik van de AI-scribe over alle consulten, niet-gebruik, cumulatieve gebruikers, wekelijkse activatiegraad, aantal gegenereerde documenten in de tijd en dagelijks gegenereerde documenten) in vier studies. Drie daarvan hadden een voldoende methodologische kwaliteit; één studie met een design van hoge bewijskracht (preprint-RCT) en in twee studies met een design van lagere bewijskracht. In deze studies werd gekeken naar het gebruik van AI-scribes (niet van toepassing binnen de reguliere verslaglegging); daarom werd geen vergelijking gemaakt. De omvang van de geïncludeerde groep zorgverleners is redelijk representatief, met 238 deelnemers in de RCT en in totaal 52 deelnemers in de twee mixed-methodsstudies.

Er is mogelijk voldoende onderzoek gedaan naar het gebruik van AI-scribes maar aangezien de RCT nog een preprint is moet verder door het veld worden bepaald of gebruik voldoende onderzocht is.

Gebruik door zorgverleners

(Eerste) auteur	Studiedesign	Vergelijking	Steekproef-grootte	Kwaliteits-beoordeling*
Lukac	RCT	Nee	238	Voldoende
Baldwin	Mixed-methods pilot study	Nee	31	Voldoende
Basei de Paula	Experience/implementation based on usage metrics	Nee	n.g.	Onvoldoende
Nguyen	Mixed-methods pilot study	Nee	21	Voldoende

Het werkplezier van zorgverleners werd geëvalueerd aan de hand van uitkomsten voor burn-out, werkplezier en 'task load' in vier studies. Drie daarvan hadden een voldoende methodologische kwaliteit; één studie met een design van hoge bewijskracht (preprint-RCT) en twee studies met een met een design van lagere bewijskracht (mixed-methodsstudies). In deze studies werd een vergelijking

gemaakt met reguliere verslaglegging. De omvang van de geïncludeerde zorgverleners is redelijk representatief, met 238 deelnemers in de RCT en in totaal 67 deelnemers in de mixed-methodsstudies.

Er lijkt voldoende onderzoek te zijn gedaan naar het werkplezier bij het gebruik van AI-scribes.

Werkplezier

(Eerste) auteur	Studiedesign	Vergelijking	Steekproef-grootte	Kwaliteits-beoordeling*
Lukac	RCT	Ja	238	Voldoende
Duggan	Mixed-methods study	Ja	46	Voldoende
Misurac	Pre-post observational study	Ja	35	Onvoldoende
Nguyen	Mixed-methods pilot study	Ja	21	Voldoende

De impact van AI-scribes op **administratieve werkdruk** werd geëvalueerd op basis van de tijd voor verslaglegging, de tijd voor verslaglegging na werktijd en de bijdrage van zorgverleners voor de bewerking van AI-gegenereerde notities in zes studies. Vijf daarvan hadden een voldoende methodologische kwaliteit; één studie met een design van hoge bewijskracht (preprint-RCT) en vier studies met een design van lagere bewijskracht (mixed-methodsstudie). In deze studies werd een vergelijking gemaakt met reguliere verslaglegging. De omvang van de geïncludeerde zorgverleners is redelijk representatief, met 238 deelnemers in de RCT en in totaal 108 deelnemers in de mixed-methodsstudies.

Er lijkt voldoende onderzoek te zijn gedaan naar de impact van AI-scribes op administratieve werkdruk.

Administratieve werkdruk

(Eerste) auteur	Studiedesign	Vergelijking	Steekproef-grootte	Kwaliteits-beoordeling*
Lukac	RCT	Ja	238	Voldoende
Baldwin	Mixed-methods pilot study	Ja (descriptive)	31	Voldoende
Cao	Mixed-methods pilot study	Ja	10	Voldoende
Duggan	Mixed-methods study	Ja	46	Voldoende
Van Buchem	Simulation study	Ja	27	Onvoldoende
Nguyen	Mixed-methods pilot study	Ja (descriptive)	21	Voldoende

Het perspectief van zorgverleners op de patiënt-zorgverlenerrelatie werd geëvalueerd op basis van betrokkenheid op een Likert-schaal in één studie met een design van hoge bewijskracht (preprint-RCT) en één studie met lagere bewijskracht (mixed-methodsstudie), allebei van voldoende

methodologische kwaliteit. In een studie werd een vergelijking gemaakt met reguliere verslaglegging. In totaal werden 284 zorgverleners bevraagd.

Er lijkt voldoende onderzoek te zijn gedaan naar het perspectief van zorgverleners op hun betrokkenheid bij patiënten.

Tevredenheid zorgverleners

(Eerste) auteur	Studiedesign	Vergelijking	Steekproef-grootte	Kwaliteits-beoordeling*
Lukac	RCT	Nee	238	Voldoende
Duggan	Mixed-methods study	Ja	46	Voldoende

Het perspectief van patiënten op de patiënt-zorgverlenerrelatie werd geëvalueerd in een studie met een lager bewijskrachtniveau (mixed-methodsstudie) maar voldoende methodologische kwaliteit. Het doel van de studie was het verkennen van het gebruik van een AI-scribe in een dermatologische setting, waarbij 23 patiënten werden bevraagd over hun tevredenheid met de patiënt-zorgverlenerrelatie. Er is geen vergelijking gedaan met reguliere verslaglegging voor deze uitkomst.

Het patiëntperspectief in andere specialistische settings lijkt relevant om een breder beeld te krijgen van de houding van patiënten binnen de medisch-specialistische zorg.

Tevredenheid patiënten

(Eerste) auteur	Studiedesign	Vergelijking	Steekproefgrootte	Kwaliteits-beoordeling*
Cao	Mixed-methods pilot study	Nee	23	Voldoende

Voor **kosten** zijn geen relevante uitkomsten gevonden.

Praktijkonderzoek naar de kosten van SGR in de MSZ kan meer inzicht bieden in deze indicator.

Bijlage F – Overzicht Spraakgestuurd rapporteren in de MSZ

Ziekenhuis	Leverancier	Referentie
Reinier de graaf ziekenhuis	Autoscriber	https://www.chipsoft.com/nl-nl/nieuws/tijdwinst-in-het-consult-door-ai-in-het-epd/ https://www.zorgvisie.nl/reinier-de-graaf-kan-als-eerste-ziekenhuis-autoscriber-in-chipsoft-epd-gebruiken/ https://www.rtl.nl/nieuws/tech/artikel/5438671/minder-werkdruk-artsen-door-kunstmatige-intelligentie
Amsterdam UMC	Autoscriber	Nieuwsbrief Autoscriber oktober 2024
LUMC	Autoscriber	Nieuwsbrief Autoscriber oktober 2024 https://www.lumc.nl/over-het-lumc/maatschappelijke-rol/waarde--en-datagedreven-zorg/cairelab-ai/de-toekomst-van-consultdocumentatie/
Deventer Ziekenhuis	Juvoly	https://www.ai-inventarisatie-zorg.nl/
Maasstad-Ziekenhuis	Juvoly	https://www.ad.nl/rotterdam/ziekenhuis-zet-ai-assistent-in-tijdens-spreekuur-ik-heb-nu-veel-meer-tijd-voor-mijn-patient~affaed6b/
Wilhelmina Ziekenhuis Assen	Autoscriber	Nieuwsbrief Autoscriber oktober 2024
Treant	Autoscriber	Product reviews op website https://nl.autoscriber.com/about Nieuwsbrief Autoscriber oktober 2024
St Antonius Ziekenhuis	Autoscriber	Nieuwsbrief Autoscriber oktober 2024
Jeroen Bosch Ziekenhuis	Autoscriber	Nieuwsbrief Autoscriber oktober 2024 www.jeroenboschziekenhuis.nl/nieuws/meer-aandacht-voor-patient-door-inzet-ai-in-de-spreekkamer
Rijnstate ziekenhuis	Autoscriber	Nieuwsbrief Autoscriber oktober 2024 https://www.rijnstate.nl/over-rijnstate/nieuws/2025/meer-aandacht-voor-de-patient-door-inzet-ai-in-de-spreekkamer/ *Het gebruik van spraakherkenning met AI in de spreekkamer is één van de transformatieprojecten binnen hetmProve zorgtransformatieprogramma (Rijnstate, Albert Schweitzer ziekenhuis, Isala, Jeroen Bosch Ziekenhuis, Maxima MC, Noordwest Ziekenhuisgroep en Zuyderland Medisch Centrum).
Isala	Autoscriber	Nieuwsbrief Autoscriber oktober 2024
Elkerliek ziekenhuis	Autoscriber	Nieuwsbrief Autoscriber oktober 2024
Diakonessenhuis	Autoscriber	Nieuwsbrief Autoscriber oktober 2024
CWZ	Autoscriber	https://www.cwz.nl/naar-het-ziekenhuis/digitale-zorg/gespreksverslag/
UMCU	Autoscriber	https://www.kinase.nl/actueel/ai-in-de-praktijk
Nij Smellinghe Drachten	Ourmind	https://www.nijsmellinghe.nl/verwijzers/nieuws-regionijs/minder-typwerk-en-een-open-en-rustig-gesprek-met-de-patient

Erasmus MC	Digizorg	https://icthealth.nl/nieuws/ai-gegenereerde-patientensamenvatting-met-digizorg-assistent (en mogelijk andere ziekenhuizen in de regio Rotterdam)
Radboudumc	Autoscriber	https://autoscriber.com/resources/radboudumc-launches-innovative-autoscriber-pilot-integrated-into-epic-ehr
St Jansdal	Ourmind	https://www.ourmind.ai/nl/
Streekziekenhuis Koningin Beatrix	Ourmind	https://www.ourmind.ai/nl/

* *Dit is geen complete inventarisatie en is gebaseerd op beschikbare (openbare) informatie*