AALBORG UNIVERSITY
STUDENT REPORT

Patient Comprehension and Perception of AI-Generated Clinical Notes in Musculoskeletal Healthcare

An Experimental Proof-of-Concept Study

Master's Thesis

Written by:

Daniel Keskin, Katrine Vilja Sand & Mathias Louie Hansen

February 2025 - June 2025

Titel: Patienters forståelse og opfattelse af AI-genererede journalnotater inden for det muskuloskeletale sundhedsområde: Et eksperimentelt proof-of-concept-studie

Semester: 4. semester, Klinisk Videnskab & Teknologi

Semestertema: Kandidatspeciale

Projektperiode: februar 2025 - juni 2025

ECTS: 30

Vejledere: Stine Hangaard & Jannie Damsgaard Nørlev

Projektgruppe: Kvt-f25-4-gr10514

Daniel Keskin

Katrine Vilja Sand

Katrine Vilja Sand

Mathias Louie Hansen

Antal sider: 71

Bilag: 5

AALBORG UNIVERSITET
STUDENTERRAPPORT

Baggrund: Muskuloskeletale lidelser påvirker omkring 1.71 milliarder mennesker globalt og medfører betydelige udfordringer i form af funktionsnedsættelse, reduceret livskvalitet og store sundhedsøkonomiske omkostninger. Effektiv kommunikation mellem patienter og sundhedsprofessionelle er afgørende for håndteringen af disse lidelser. Selvom patientadgang til elektroniske patientjournaler (EPJ) kan øge engagementet i egen behandling, indeholder kliniske journalnotater typisk komplekst medicinsk sprog, der vanskeliggør forståelsen, især hos personer med lav sundhedskompetence. Nyere fremskridt inden for generativ kunstig intelligens (AI) har potentiale til at skabe forenklede, patientvenlige kliniske notater.

Formål: Dette proof-of-concept studie havde til formål at evaluere effekten af AI-genererede, patientvenlige kliniske notater på deltagernes objektive forståelse og deres opfattelse sammenlignet med originale notater, med særligt fokus på forskelle relateret til niveau af sundhedskompetence.

Metode: I alt 19 deltagere (gennemsnitsalder 55.7 ± 19.9 år) evaluerede originale og AI-genererede versioner af seks muskuloskeletale kliniske notater. Notaterne blev genereret med GPT-4o ved brug af zero-shot prompting-teknikker målrettet personer med lav sundhedskompetence. Objektiv forståelse blev målt ved hjælp af selvudviklede checklister, mens deltagernes opfattelse blev vurderet med et seks-delt spørgeskema målt på en fem-punkts Likert-skala. Sundhedskompetencen blev målt med den danske HLS-EU-Q16 og DS-TOFHLA. De statistiske analyser omfattede Wilcoxon signed-rank test, Mann-Whitney U test, parrede t-tests samt lineære mixed-models.

Resultater: Deltagerne opnåede signifikant højere objektiv forståelse ved læsning af AI-genererede notater (median 80 %, IQR 44 %) sammenlignet med originale notater (median 38 %, IQR 44.5 %; z=-3.823, P<.001). De subjektive evalueringer favoriserede ligeledes de AI-genererede notater signifikant på samtlige dimensioner (alle P<.001). Der blev fundet en signifikant interaktion mellem niveau af sundhedskompetence (HLS) og notattype, hvilket indikerede størst forbedring af forståelsen hos deltagere med lav sundhedskompetence (F(1, 85.8)=8.9, P=.004).

Konklusion: Dette studie fremhæver AI's potentiale som et effektivt værktøj til at imødekomme forståelsesforskelle og styrke patienternes muligheder for aktiv deltagelse i egen behandling. På baggrund af disse lovende resultater vil integration af AI-baserede, forenklede notater i klinisk praksis kunne øge patientinddragelsen betydeligt. Fremtidige studier bør prioritere udviklingen og valideringen af værktøjer, der er specifikt designet til at måle patienters forståelse af kliniske journalnotater.

Title: Patient Comprehension and Perception of AI-Generated Clinical Notes in Musculoskeletal Healthcare: An Experimental Proof-of-Concept Study

Semester: 4th semester, Clinical Science & Technology

Semester theme: Master's thesis

Project period: February 2025 - June 2025

ECTS: 30

Supervisors: Stine Hangaard & Jannie Damsgaard Nørlev

Project group: Kvt-f25-4-gr10514

Daniel Keskin

Katrine Vilja Sand

Katrine Vilja Sand

Mathias Louie Hansen

Number of pages: 71

Appendix: 5

AALBORG UNIVERSITY
STUDENT REPORT

Background: Musculoskeletal disorders affect approximately 1.71 billion people globally, posing significant burdens in terms of disability, reduced quality of life, and healthcare costs. Effective patient-provider communication is critical for managing these conditions. Although patient access to electronic health records (EHRs) enhances engagement, clinical notes typically contain complex medical jargon that impedes understanding, especially among individuals with limited health literacy. Recent advances in generative artificial intelligence (AI) have potential for creating simplified, patient-friendly clinical notes.

Objective: This proof-of-concept study aimed to evaluate the impact of AI-generated, patient-friendly clinical notes on participants' objective comprehension and their perceptions compared to original notes, with particular attention to variations based on health literacy levels.

Methods: A total of 19 participants (mean age 55.7 ± 19.9) evaluated original and AI-generated versions of six musculoskeletal clinical notes. Notes were generated using GPT-4o with zero-shot prompting techniques tailored to low health literacy individuals. Objective comprehension was assessed via self-developed checklists, and perception via a six-item questionnaire measured on a five-point Likert-scale. Health literacy was measured using the Danish HLS-EU-Q16 and DS-TOFHLA questionnaires. Statistical analyses involved Wilcoxon signed-rank tests, Mann-Whitney U tests, paired t-tests, and linear mixed models.

Results: Participants demonstrated significantly higher objective comprehension scores for AI-generated notes (median 80%, IQR 44%) compared to original notes (median 38%, IQR 44.5%; z=-3.823, P<.001). Subjective evaluations favored AI-generated notes significantly across all dimensions (all P<.001). A significant interaction was found between health literacy levels (HLS) and note version, indicating greater comprehension benefits for individuals with lower health literacy (F(1, 85.8)=8.9, P=.004).

Conclusions: This study supports AI's potential as a powerful tool to bridge comprehension gaps and foster patient empowerment. Given these promising results, integrating AI-driven note simplification into routine clinical practice could significantly enhance patient engagement. Future studies should prioritize the development and validation of tools specifically designed for measuring patient comprehension of clinical notes.

Table of Contents

Preface

This master's thesis was conducted by three students in their 4th semester of the Master's programme in Clinical Science and Technology at Aalborg University, Denmark, from February to June 2025.

The aim of this thesis was to investigate comprehension and perception of original clinical notes versus AI-generated simplifications in musculoskeletal healthcare among individuals with varying health literacy levels.

Scientific paper

The thesis is presented in the form of a scientific paper, prepared according to international standards and based on the author guidelines of JMIR AI, a leading journal in digital health and patient communication.

Systematic literature search

As a foundation for the project, a systematic literature search was conducted to identify relevant research and inform both the study design and interpretation of results. Further details on the literature search are included in the supplementary worksheets.

Worksheets and appendices

Supplementary worksheets are provided after the scientific paper to meet the semester's learning objectives. Additional appendices offer further methodological details and supporting materials.

The authors wish to express their sincerest gratitude to their supervisor Stine Hangaard and co-supervisor Jannie Damsgaard Nørlev for their invaluable support, guidance, and encouragement throughout the project. Special thanks are extended to the staff at Aalborg Rehabilitation for assistance during data collection, and to all study participants for their valuable time and contributions.

The Vancouver citation style has been used throughout this thesis.

Original Paper

Patient Comprehension and Perceptions of AI-Generated Clinical Notes in Musculoskeletal Healthcare: An Experimental Proof-of-Concept Study

Abstract

Background: Musculoskeletal disorders affect approximately 1.71 billion people globally, posing significant burdens in terms of disability, reduced quality of life, and healthcare costs. Effective patient-provider communication is critical for managing these conditions. Although patient access to electronic health records (EHRs) enhances engagement, clinical notes typically contain complex medical jargon that impedes understanding, especially among individuals with limited health literacy. Recent advances in generative artificial intelligence (AI) have potential for creating simplified, patient-friendly clinical notes.

Objective: This proof-of-concept study aimed to evaluate the impact of AI-generated, patient-friendly clinical notes on participants' objective comprehension and their perceptions compared to original notes, with particular attention to variations based on health literacy levels.

Methods: A total of 19 participants (mean age 55.7 ± 19.9) evaluated original and AI-generated versions of six musculoskeletal clinical notes. Notes were generated using GPT-4o with zero-shot prompting techniques tailored to low health literacy individuals. Objective comprehension was assessed via self-developed checklists, and perception via a six-item questionnaire measured on a five-point Likert-scale. Health literacy was measured using the Danish HLS-EU-Q16 and DS-TOFHLA questionnaires. Statistical analyses involved Wilcoxon signed-rank tests, Mann-Whitney U tests, paired t-tests, and linear mixed models.

Results: Participants demonstrated significantly higher objective comprehension scores for AI-generated notes (median 80%, IQR 44%) compared to original notes (median 38%, IQR 44.5%; *z*=-3.823, *P*<.001). Subjective evaluations favored AI-generated notes significantly across all dimensions (all *P*<.001). A significant interaction was found between health literacy levels (HLS) and note version, indicating greater comprehension benefits for individuals with lower health literacy (F(1, 85.8)=8.9, *P*=.004).

Conclusions: This study supports AI's potential as a powerful tool to bridge comprehension gaps and foster patient empowerment. Given these promising results, integrating AI-driven note simplification into routine clinical practice could significantly enhance patient engagement. Future studies should prioritize the development and validation of tools specifically designed for measuring patient comprehension of clinical notes.

Keywords: Patient communication; Medical records; Large Language Model; LLM; Musculoskeletal disorders; Health literacy; Generative AI

Introduction

Musculoskeletal disorders are a major global health issue, affecting around 1.71 billion people and ranking as one of the leading causes of disability worldwide (1). The term covers a wide range of conditions, including low back and neck pain, fractures, osteoarthritis, rheumatoid arthritis, amputations, and other injuries involving joints, bones, muscles, and connective tissues. These disorders affect people of all ages and are linked to long-term disability, reduced quality of life, and high healthcare costs. Collectively, these factors place a considerable burden on both individuals and healthcare systems (2,3).

Given the high global prevalence and burden of musculoskeletal disorders, ongoing management requires active patient engagement and collaboration with healthcare providers. Effective communication is essential not only for ensuring understanding of diagnoses and treatments but also for empowering patients to participate in their own care. However, research has shown that much health-related communication remains overly complex, with medical jargon and low readability presenting barriers to patient comprehension and engagement (4–8).

In recent years, many countries have implemented systems that grant patients direct access to their electronic health records (EHRs). While this transparency has been shown to enhance patient empowerment and foster a greater sense of ownership over one's health, it also presents challenges. Because clinical documentation is still primarily written for professionals, many patients encounter difficulties understanding their records, further underlining the need for patient-friendly communication tools in the EHR context (8–12). These difficulties are particularly pronounced among individuals with limited health literacy, defined as the ability to access, understand, appraise, and apply health information to make informed decisions about one's health (13). For these patients, medical jargon and unclear language often result in misunderstanding of essential health information. This often leads to increased anxiety, a higher likelihood of medical errors, and greater utilization of healthcare services and online sources driven by follow-up questions and concerns arising from misinterpretation (10–12,14,15). Therefore, while patient access to EHRs holds promise for enhancing engagement, it also necessitates the implementation of strategies to ensure that all patients, regardless of their health literacy levels, can effectively understand and utilize their health information.

Recent advances in generative artificial intelligence (AI), particularly Large Language Models (LLMs) such as OpenAI's GPT-4, present a promising solution for enhancing patient comprehension of complex medical information (16). The majority of published research on LLMs in healthcare focuses on their use as medical chatbots (94.4%), while a smaller share investigates their application in clinical documentation and the simplification of complex medical texts (17).

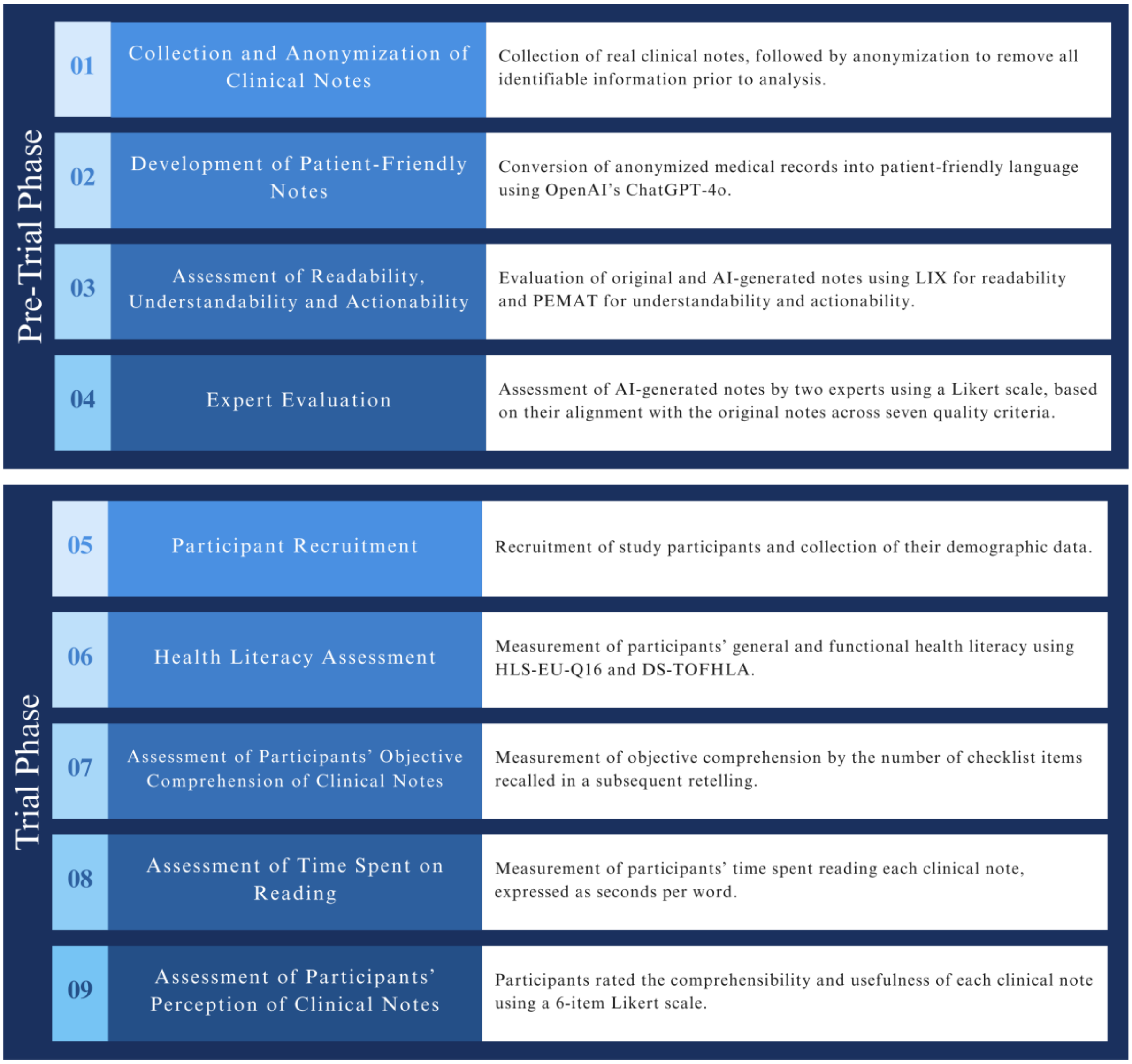
Existing tools integrated into EHRs typically provide fragmented, word-by-word definitions (18–20), whereas generative AI has the potential to provide meaningful summaries that translate access into comprehension. Research, especially in radiology reports and discharge summaries, has demonstrated the potential of LLMs to improve readability and patient understanding significantly through simplification and reduction of medical jargon (16,21,22). These AI-generated summaries are intended as transformations of original clinical notes and discharge summaries into clearer, more accessible language for patients (23–25). For example, Zaretsky et al. (2024) showed that discharge summaries rephrased by AI achieve significantly improved readability (lower Flesch-Kincaid scores) and higher patient understanding (higher PEMAT scores) (16), while Tang et al. (2024) reported success in simplifying radiological findings using ChatGPT-based methods (26).

However, despite these benefits, AI-generated summaries may also introduce inaccuracies or hallucinations, which occur when the AI produces information that appears plausible but is factually incorrect or fabricated (27), posing potential risks for patient misunderstanding or harm (16,28,29). Existing assessments of AI-generated medical texts have predominantly been conducted without direct patient involvement, limiting insights into patients' experiences and perceptions of these summaries, as well as a lack of focus on patients' actual comprehension of AI-generated content. Previous studies have largely relied on quantitative readability metrics and expert-based evaluations (16,22–26,28,30).

To the best of the authors' knowledge, no prior studies have directly compared laypeople's comprehension of original clinical notes with that of AI-generated versions. This experimental proof-of-concept study aims to evaluate both participants' objective comprehension and their perceptions of both original and AI-generated notes within musculoskeletal healthcare. Special attention is given to differences according to health literacy levels.

Methods

The following methods section outlines the procedures carried out prior to and during the trial, as well as the measures employed. An overview of all phases of the workflow in chronological order is presented in Figure 1.



Pre-Trial Phase	01	Collection and Anonymization of Clinical Notes	Collection of real clinical notes, followed by anonymization to remove all identifiable information prior to analysis.
	02	Development of Patient-Friendly Notes	Conversion of anonymized medical records into patient-friendly language using OpenAI's ChatGPT-4o.
	03	Assessment of Readability, Understandability and Actionability	Evaluation of original and AI-generated notes using LIX for readability and PEMAT for understandability and actionability.
	04	Expert Evaluation	Assessment of AI-generated notes by two experts using a Likert scale, based on their alignment with the original notes across seven quality criteria.
Trial Phase	05	Participant Recruitment	Recruitment of study participants and collection of their demographic data.
	06	Health Literacy Assessment	Measurement of participants' general and functional health literacy using HLS-EU-Q16 and DS-TOFHLA.
	07	Assessment of Participants' Objective Comprehension of Clinical Notes	Measurement of objective comprehension by the number of checklist items recalled in a subsequent retelling.
	08	Assessment of Time Spent on Reading	Measurement of participants' time spent reading each clinical note, expressed as seconds per word.
	09	Assessment of Participants' Perception of Clinical Notes	Participants rated the comprehensibility and usefulness of each clinical note using a 6-item Likert scale.

Figure 1: Overview of the study process, including pre-trial and trial phases.

Pre-Trial Procedures

Collection and Anonymization of Clinical Notes

Initially, 14 clinical notes were collected through the networks of three authors (D.K., K.V.S., and M.L.H.). Contributors received written information about the purpose of the study and were explicitly instructed to anonymize all personal and sensitive information from their clinical notes. Contributors consented through email correspondence and the clinical notes were received via the authors' encrypted Microsoft Outlook email addresses, hosted on Aalborg University's secure network. The clinical notes were then moved to a secure OneDrive folder administered by Aalborg University and later deleted from the mail correspondence. The three authors reviewed submissions to confirm anonymity, and that the information was not traceable to an individual. As a result, the data were not considered personal data under the General Data Protection Regulation (GDPR) or the Danish Data Protection Act (31,32).

All clinical notes were systematically screened according to predefined exclusion criteria to ensure patient relevance and clinical appropriateness of the study materials. Specifically, notes were excluded if the patient was under 18 years old at the time the note was written (n=3), if there were repetitive entries from the same patient and for the same diagnosis (n=4), or if the note was not related to musculoskeletal care in the orthopaedic or rheumatology departments within the secondary care sector (n=1). The application of these criteria ensured that only relevant and unique clinical documentation was included for further analysis in the study.

Six notes were selected to represent various musculoskeletal conditions, complexities, and note types (two outpatient, two surgical, and two imaging), reflecting the diversity encountered by patients accessing their clinical documentation via EHRs. The selection of clinical notes aimed to encompass a realistic range of clinical documentation complexity typically encountered by patients. By incorporating different clinical note types, lengths, and readability levels of musculoskeletal notes, the study sought to assess whether AI-generated notes could consistently improve patient comprehension across different clinical documentation scenarios.

AI Model and Development of Patient-Friendly Notes

AI-generated patient-friendly notes were produced using OpenAI's flagship model GPT-4o, accessed via the Plus-version of the platform at chatgpt.com (33). GPT-4o was selected as the AI-model of this study based on prior research showing that its predecessor, GPT-4, performed better than both earlier OpenAI models and other widely used AI models (21,23,24). Likewise, evidence shows that GPT-4 scores high on the United States Medical Licensing Examination and outperforms models specifically fine-tuned on medical knowledge (34). GPT-4o builds on this foundation with improved efficiency and a greater ability to communicate in a more natural and emotionally nuanced manner (33).

A zero-shot prompting technique was used, involving iterative refinement across nine cycles to optimize the prompt for clarity, accuracy and readability. Prompt quality was reviewed using two test notes with different readability levels through discussions among the authors (D.K., K.V.S., and M.L.H.). The prompt criteria included clear and simple layman's terms, accuracy, completeness, avoidance of fabricated information, a structured format and tailored to individuals with low health literacy. The development and refinement of these criteria drew upon methodologies described in previous literature (16,23,25,30). The detailed final prompt and examples of an original clinical note and its corresponding AI version are provided in Appendix 1 and Appendix 2.

Participant Recruitment

To be eligible for participation, individuals had to be at least 18 years old, able to read, speak, and understand Danish, and capable of providing informed consent prior to enrolment. To highlight the patient perspective and a broad demographic, both patients and laypeople were recruited to represent the broader population. Participants were recruited from two sources: a Danish municipal rehabilitation unit serving individuals post-hospitalization (i.e. patients), and through the authors' networks (D.K., K.V.S., and M.L.H.). This dual strategy ensured diversity in health competencies and socio-demographic backgrounds. Potential participants received written information about the study purpose and their role of participating prior to enrolment. Data collection was conducted in private consultation rooms at the municipal rehabilitation unit as well as in a designated group room at Aalborg University. One author (D.K.) primarily facilitated the sessions, with two others (K.V.S. and M.L.H.) acting as supplementary evaluators.

Trial Procedures

Upon arrival, participants received thorough verbal and written explanations regarding study purpose, procedures, confidentiality, and voluntary participation. Written consent was then obtained using the official consent form for competent individuals (S1: Consent Form for Competent Persons) from the Danish National Committee on Health Research Ethics in compliance with Danish national guidelines, ensuring ethical integrity and participant rights protection (35).

Participants then completed a brief questionnaire on demographic information (age, gender, education, employment status, marital status, ethnicity, and dyslexia diagnosis) via REDCap electronic data capture tool hosted at Aalborg University (36,37).

Subsequently, participants filled out two health literacy instruments: the validated Danish version of the 16-item European Health Literacy Survey Questionnaire (HLS-EU-Q16; hereafter referred to as HLS) (38) and the (non-validated) Danish adaptation of the Short Test of Functional Health Literacy in Adults (DS-TOFHLA; hereafter referred to as TOFHLA) (39).

Next, block randomization was used to ensure balanced exposure to note versions and note types. Each participant was randomly assigned to one of twelve predefined note bundles, each comprising a total of six clinical notes. Each bundle included one version (either original or AI) of every clinical note corresponding to two of every note type (two outpatient, two surgical, and two imaging), ensuring participants did not receive both the original and AI versions of the same note. Each participant evaluated three AI-generated notes and three original notes (total N=12), maintaining a balanced exposure across both note version and note type within each session. Additionally, the sequencing of notes within each bundle was randomized to mitigate potential order bias.

After reading each note, participants were asked to retell what they had read, using their own words, imagining they had to explain the clinical note to a friend or a family member, a process inspired by the clinically recognized teach-back technique (40). During retelling, participants were permitted to have the document in front of them, rather than relying solely on memory. This modification was chosen to better simulate the real-world scenario of patients accessing their own EHRs via an eHealth portal. The facilitator intervened as little as possible to avoid influencing responses. However, if participants used medical terminology, they were asked to elaborate or explain these terms in their own words. At the end of each session, the facilitator asked whether the participant had covered everything they found relevant, to ensure a complete account of their interpretation and evaluation. This was done systematically with each clinical note to ensure a standardization of interaction. Immediately following each retelling, participants completed a six-item questionnaire measured on a five-point Likert-scale, assessing their subjective experience of reading the clinical note.

Measures

Health Literacy Assessment

The HLS-EU-Q16 is a short version of the original HLS-EU-Q47 questionnaire and was developed by the HLS-EU Consortium. Scores range from 0 to 16 and are categorized as likely inadequate (0-8), likely problematic (9-12), and likely sufficient (13-16) health literacy (38,41). The DS-TOFHLA is a Danish adaptation of the S-TOFHLA, originally developed by Baker et al. as part of the Literacy in Health Care Project (42). Scores range from 0 to 100, with cut-offs defining inadequate (0-59), marginal (60-74), and adequate (75-100) functional health literacy (39). Participants' health literacy was assessed after the session, using the official scoring manuals for both instruments.

Assessment of Readability, Understandability and Actionability

To assess the readability, understandability, and actionability of the two note versions (both AI and original) two validated tools were applied.

Readability was assessed using the LIX (abbreviation of Swedish *läsbarhetsindex*, "readibility index") formula, a Swedish- developed tool designed to measure objective readability based on sentence length and word complexity in the Scandinavian languages (43,44). The LIX score was calculated using an online tool (45) based on the following formula:

$$LIX = \left(\frac{number\ of\ words}{number\ of\ periods}\right) + \left(\frac{number\ of\ long\ words * \times\ 100}{number\ of\ words}\right)$$

*$Long\ words \geq 7\ letters$ (45)

The LIX scale consists of five readability categories:

- ≥ 55: Very difficult, e.g. academic-level text and legal documents
- 45 - 54: Difficult, e.g. factual book, popular science works and academic publications
- 35 - 44: Medium, e.g. newspapers and journal
- 25 - 34: Easy for experienced readers, e.g. weekly magazines and light fiction
- ≤ 24: Easy for all readers, e.g. children's literature (45)

Understandability and actionability were evaluated using the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P), a tool designed to assess how easily the content can be understood and acted upon by readers. PEMAT scores are expressed as percentages, indicating the proportion of applicable criteria met within each domain. Higher PEMAT scores reflect greater understandability or actionability (46).

Assessment of Participants' Objective Comprehension of Clinical Notes

Comprehension, as defined by the PubMed Medical Subject Headings (MESH) as "the act or fact of grasping the meaning, nature, or importance of information; includes understanding by a patient or research participant of information disclosed orally or in writing," (47) was measured using self-developed checklists containing five to nine essential items tailored to each clinical note (for example of a checklist, see Appendix 3). These items were determined by the authors to represent key messages that a patient should ideally comprehend after reading. If an item was solely expressed with the original medical terminology without showing an understanding of the terms the item was not checked as correct.

Three authors (D.K., K.V.S., and M.L.H.) independently scored participants' responses in real-time during the retelling. To minimize the risk of omitting relevant content in participants' retellings, the authors subsequently reviewed their independent checklists of participants' answers after each session and reached consensus for every note to ensure completeness and accuracy. Objective comprehension scores were calculated as the percentage of checklist items marked during participants' retellings.

Assessment of Participants' Perception of Clinical Notes

To complement the objective measures, participants' perceptions of the clinical notes was measured using a six-item questionnaire measured on a five-point Likert-scale (Appendix 4). This aimed to assess participants' evaluation of the notes in terms of comprehensibility, confidence in understanding, clarity, informativeness, actionability, and personal relevance. The questionnaire was designed specifically for this study, as previous research has not examined the patient perspective; however, the categories were inspired by those used in expert evaluations of AI-generated notes in earlier studies and were adapted to be more relevant and understandable to patients (16,22,23,28,30).

Assessment of Time Spent on Reading

Participants' reading time for each note was recorded and subsequently normalized as seconds per word. Notes were coded as not fully read if the participant either (a) explicitly gave up due to perceived difficulty, (b) unintentionally skipped part of the text (e.g., forgetting to read page two), or (c) chose not to read the entire text because it was perceived as very easy to understand

and could be retold based on key bullet points alone. Notes that were not fully read were excluded from the subsequent analysis of reading time. For all remaining readings, the total number of seconds spent reading each note was divided by the number of words in that specific note, resulting in a normalized measure of reading time expressed as seconds per word.

Expert Evaluation

Independent of participant evaluations, two musculoskeletal experts assessed the quality and potential safety risks of the AI-generated notes relative to the original clinical notes using an online questionnaire administered via REDCap (36,37). The experts assessed each note on seven predefined dimensions: completeness, conciseness, factual accuracy, clarity, presence of hallucinations, potential risks, and overall usability, adapted from prior research (16,22,23,28,30). For each dimension, experts provided evaluations using a structured seven-item questionnaire measured on a five-point Likert scale, where a score of one indicated optimal fulfilment of the respective criterion and a score of five indicated the lowest level of fulfilment (Appendix 5). Voluntary open-ended text fields were included for each note, allowing the experts to elaborate on their evaluations and offer more nuanced qualitative feedback.

Statistical Analysis

Statistical analyses were conducted using IBM SPSS Statistics (Version 29.0.0.0) (48). Descriptive statistics were used to summarize participant demographics and study outcomes. Normality of continuous data was assessed using the Shapiro-Wilk test. Based on these results, appropriate parametric or non-parametric statistical tests were selected for each outcome variable. For normally distributed data, t-tests were used as parametric tests; for non-normally distributed or ordinal data, non-parametric alternatives such as Mann-Whitney U or Wilcoxon signed-rank tests were applied. For the statistical analysis of the assessment of health literacy interaction, a linear mixed model with random intercepts for both participants and each unique note was conducted. Inspection of residual plots indicated no substantial deviations from normality; both histogram and Q-Q plot confirmed that model residuals were approximately normally distributed. The model included fixed effects for note version (AI-generated vs. original), HLS score, and their interaction. Afterwards, an additional analysis was conducted using the TOFHLA score as a standardized covariate in place of the HLS, applying the same mixed model structure.

Results

Participants

The study sample comprised 19 participants with a mean age of 55.7 years (SD 19.9). Of the participants, 8 (42%) were male and 11 (58%) were female. 5 (26%) were recruited through the authors network and 14 (74%) through the municipal rehabilitation unit.

Regarding educational attainment, 1 participant (5%) had completed primary school or 10th grade, 4 (21%) had completed upper secondary education, and 3 (16%) had completed vocational education. Additionally, 3 participants (16%) had completed vocational academy or other higher adult education, while 7 (37%) had a bachelor's degree or diploma education. 1 participant (5%) held a master's degree.

Regarding dyslexia, 3 participants (16%) had been formally diagnosed, 1 (5%) suspected having dyslexia, and the remaining 15 participants (79%) reported no dyslexia. Full demographic characteristics can be found in Table 1.

Table 1: Demographic Characteristics of the Study Sample

Variable	Value* (N=19)
Age, mean (SD)	55.7 (19.9)
Sex	
Male	8 (42)
Female	11 (58)
Highest Completed Education	
Primary school or 10th grade	1 (5)
Upper secondary education	4 (21)
Vocational education	3 (16)
Vocational academy education or other higher adult education	3 (16)
Bachelor's degree or diploma education	7 (37)
Master's degree or graduate education	1 (5)
Current Employment Status	
Employee	10 (53)
Student	2 (11)
Unemployed	1 (5)
Retired	5 (26)
Disability pensioner	1 (5)
Marital Status	
Married or living with a partner	11 (58)
Single (unmarried, divorced or widow/widower)	8 (42)
Ethnicity	
Ethnically Danish	19 (100)
Dyslexia	
Diagnosed dyslexia	3 (16)
Suspected dyslexia	1 (5)
No dyslexia	15 (79)

** Values are presented as n (%), except age, which is shown as mean (SD).*

Health Literacy Assessment

Health literacy levels among participants were broadly distributed, with roughly one-third classified as having likely inadequate (n=7), likely problematic (n=5), or likely sufficient (n=7) health literacy based on the HLS (mean 11.1, SD 3.3). For functional health literacy measured by the TOFHLA, most participants were classified as having adequate literacy (n=12), while a minority fell into the marginal (n=5) or inadequate (n=2) categories (median 80, IQR 30). Further details on score distributions are presented in Table 2.

Table 2: Overview of Health Literacy Levels According to HLS-EU-Q16 and DS-TOFHLA

Measure	Category	Value* (N=19)
HLS-EU-Q16		
	Likely inadequate health literacy	7 (37)
	Likely problematic health literacy	5 (26)
	Likely sufficient health literacy	7 (37)
	Mean (SD)	11.1 (3.3)
	Range	6-16
DS-TOFHLA		
	Inadequate	2 (11)
	Marginal	5 (26)
	Adequate	12 (63)
	Median (IQR)	80 (30)
	Range	55-95

** Values are presented as n (%), except mean values shown as mean (SD), median values shown as median (IQR), and score ranges shown as minimum-maximum.*

Assessment of Readability, Understandability, and Actionability

When looking at readability expressed through LIX-scores, the original clinical notes had a median of 47 (IQR 7), placing them within the threshold for texts classified as difficult. The AI-generated notes had a median of 33 (IQR 8), corresponding to a readability level considered easy for experienced readers (Table 3). The analysis revealed a statistically significant difference in readability between the two groups (*U*=4, *P*=.026), with the AI-generated notes demonstrating significantly lower LIX scores (Figure 2).

For the PEMAT assessments, the original notes had a median of 27% (IQR 17.8%) in the understandability domain, while the AI-generated notes had a median of 83% (IQR 8%). In the actionability domain, the original notes had a median of 10% (IQR 20%), while the AI-generated notes had a median of 50% (IQR 25%) (Table 3). A Mann–Whitney U test indicated a significant difference between the note versions in both understandability (*U*=36, *P*=.002) and actionability (*U*=34.5, *P*=.004) (Figure 3).

Table 3: Medians and Interquartile Ranges of Readability, Understandability, and Actionability Measures Across Note Types for Original and AI-Generated Texts

Note type	Measure	Original, median (IQR)	AI, median (IQR)
Outpatient			
	Word count	155.5	251.5
	LIX	39.5	29
	PEMAT Understandability %	34.5	84
	PEMAT Actionability %	20	60
Surgical			
	Word count	390	314
	LIX	47	36.5
	PEMAT Understandability %	19	84
	PEMAT Actionability %	10	50

Imaging			
	Word count	107	209.5
	LIX	47.5	35
	PEMAT Understandability %	23	77
	PEMAT Actionability %	0	30
Total			
	Word count	166.5 (154)	258 (115)
	LIX	47 (7)	33 (8)
	PEMAT Understandability %	27 (17.8)	83 (8)
	PEMAT Actionability %	10 (20)	50 (25)

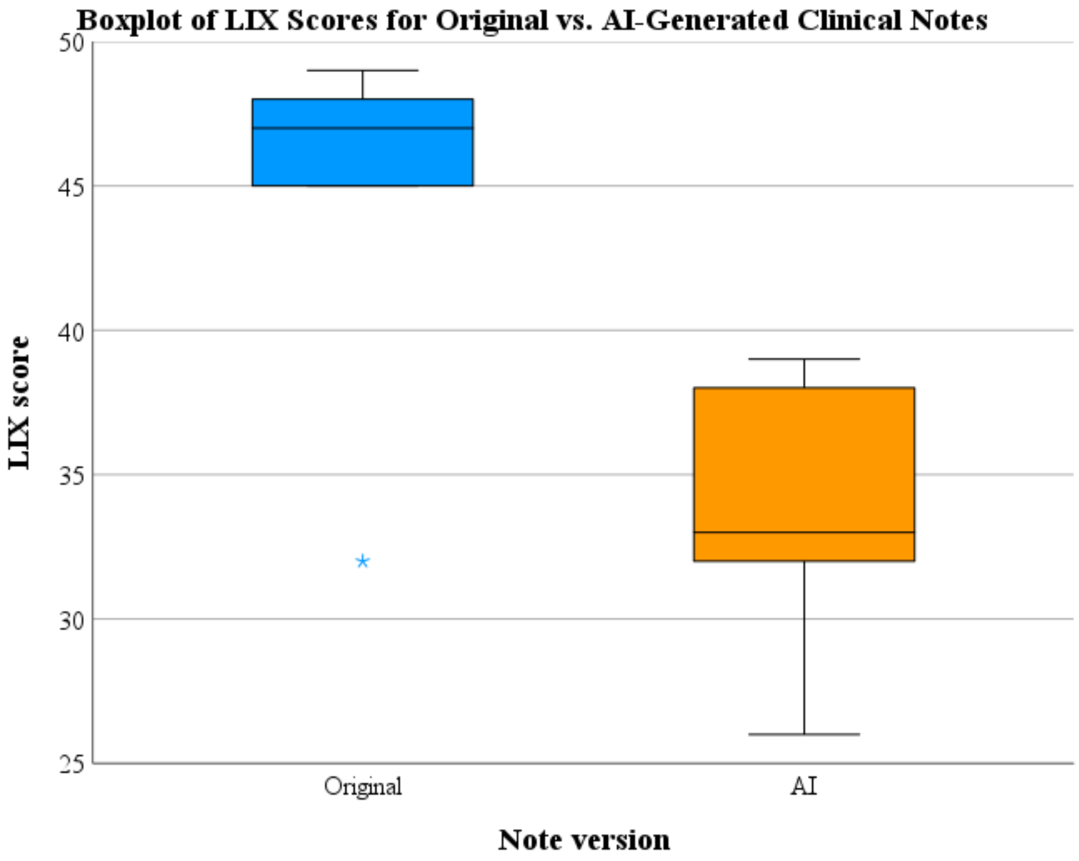


Figure 2: Boxplots of LIX scores for original versus AI-generated clinical notes. Lower LIX scores reflect higher objective readability.

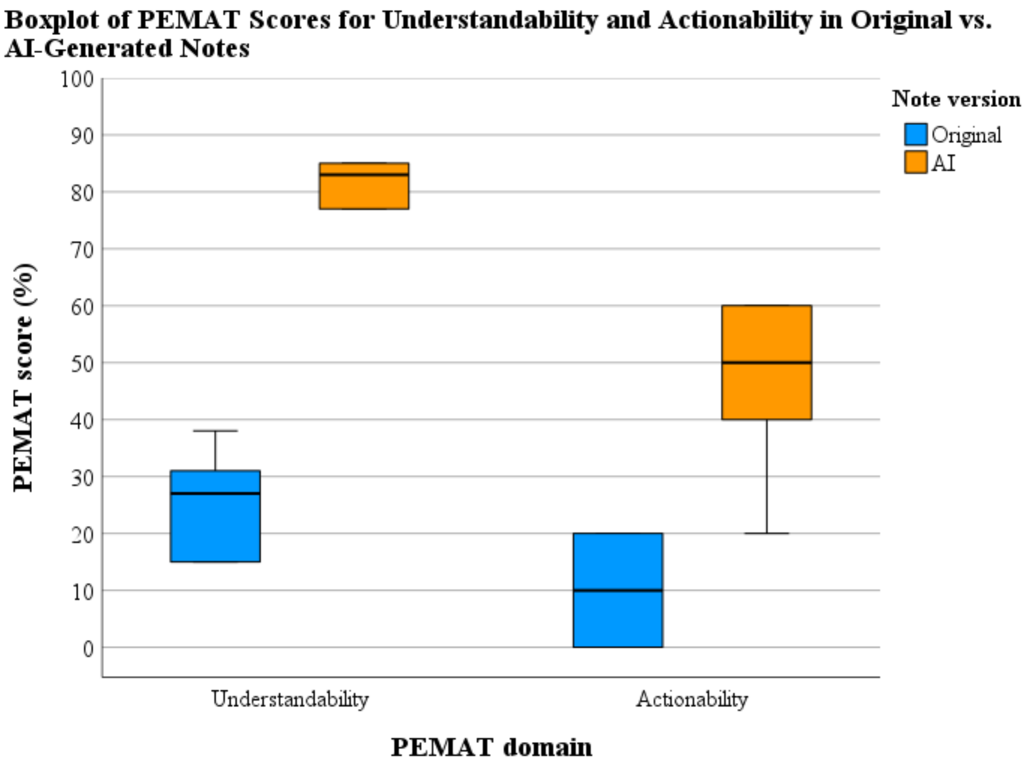


Figure 3: Boxplots of PEMAT ratings for understandability and actionability in original versus AI-generated clinical notes. Ratings reflect the percentage of applicable criteria met within each PEMAT domain.

Assessment of Participants' Objective Comprehension of Clinical Notes

The self-developed checklists used to assess objective comprehension revealed a median score of 38% (IQR 44.5%) for original notes, compared to 80% (IQR 44%) for AI-generated notes. A Wilcoxon signed-rank test showed a statistically significant difference between comprehension scores of original and AI-generated notes (*z*=-3.823, *P*<.001), with AI-generated notes exhibiting a higher mean comprehension score compared to original notes (Figure 4).

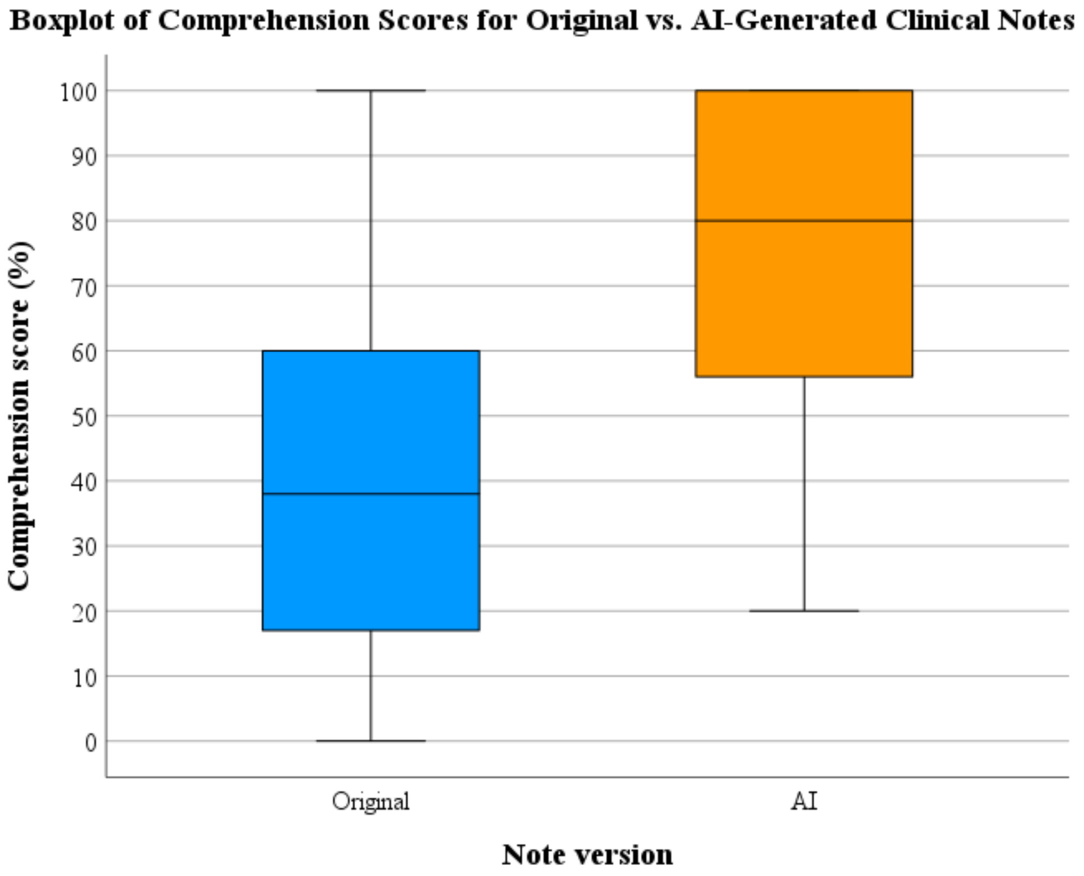


Figure 4: Boxplots of comprehension scores for original versus AI-generated clinical notes. Scores reflect the percentage of checklist items marked for each note version.

The median comprehension score for outpatient notes was 50% (IQR 37%) for the original notes and 63% (IQR 50%) for the AI-generated notes. The median comprehension score for surgical notes was 33% (IQR 33%) for the original notes and 89% (IQR 44%) for the AI-generated notes. For imaging notes, the median comprehension score was 20% (IQR 33%) for the original notes and 80% (IQR 33%) for the AI-generated notes.

A paired samples t-test revealed no significant difference in comprehension scores between original and AI-generated outpatient notes (*t*(18)=-1.63, *P*=.121). A Wilcoxon signed-rank test revealed a statistically significant increase in comprehension scores for AI-generated compared to original surgical notes (*z*=-3.83, *P*<.001). Similarly, a paired-samples t-test indicated a significant improvement in comprehension scores for AI-generated imaging notes relative to the originals (*t*(18)=-6.37, *P*<.001) (Figure 5).

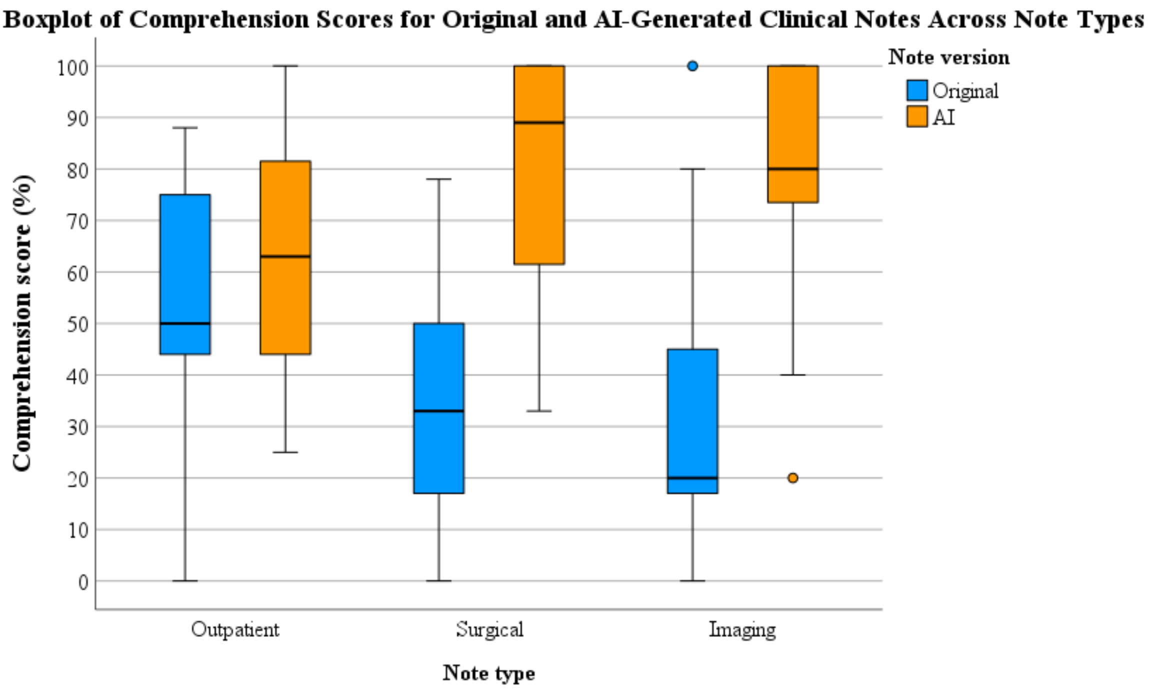


Figure 5: Boxplots of comprehension scores for original and AI-generated clinical notes across three note types (outpatient, surgical, and imaging). Scores reflect the percentage of checklist items marked for each note version and type.

Assessment of Health Literacy Interaction

When accounting for repeated measures and random effects, participants scored significantly higher on objective comprehension of AI-generated notes with an estimated marginal mean of 74.7% (SE 5.9%) compared to original notes with an estimated marginal mean of 40.1% (SE 5.9%, *F*(1, 9.8)=26.7, *P*<.001), resulting in an average improvement of 34.6% (95% CI 19.7-49.6).

When using HLS as a measure for health literacy, no significant main effect was shown (*F*(1, 16.8)=.2, *P*=.637), indicating that higher health literacy was not associated with better comprehension across both note versions. However, a significant interaction between note version and health literacy was observed (*F*(1, 85.8)=8.9, *P*=.004), suggesting that the effect of AI-generated notes on comprehension was moderated by health literacy. This interaction and corresponding values can be seen in Figure 6 and Table 4.

Model fit statistics showed that fixed effects explained 33% of the variance in comprehension (marginal R^2), while the full model including participant and unique note variation explained 66% (conditional R^2).

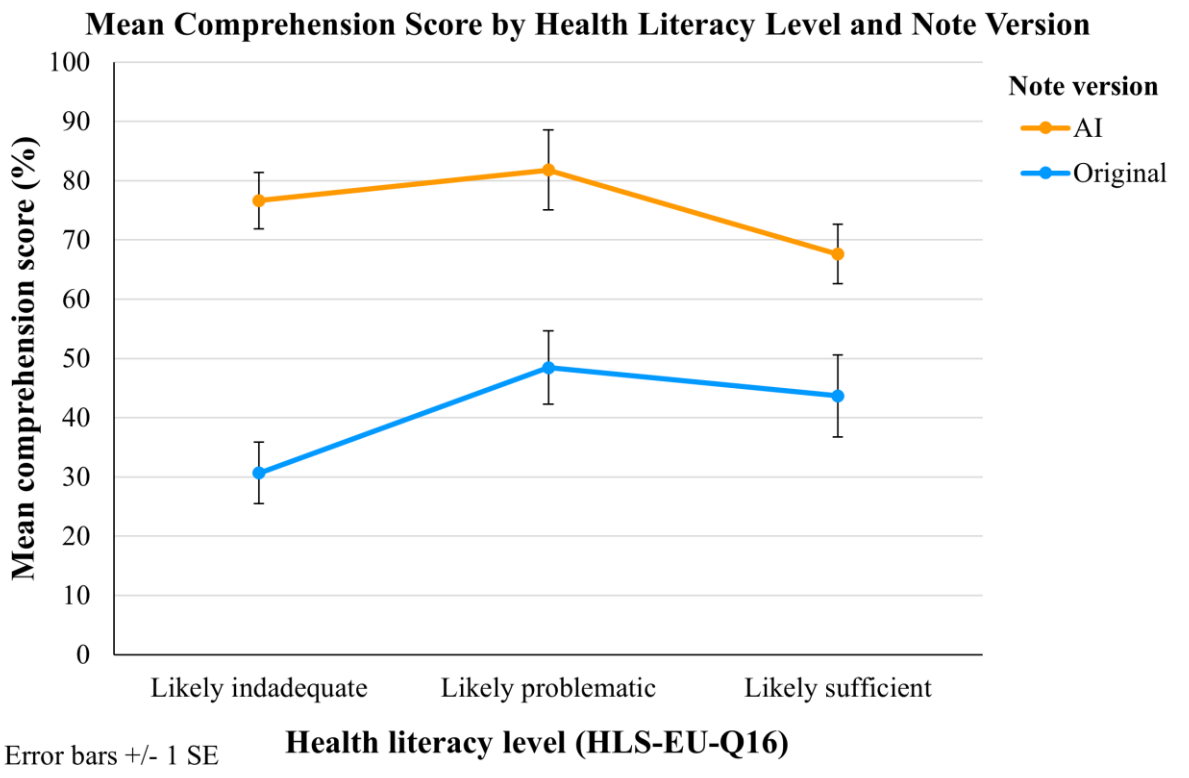


Figure 6: Mean comprehension scores with standard error bars for each health literacy level (HLS-EU-Q16: inadequate, problematic, sufficient), stratified by note version (AI-generated vs. original). Error bars indicate standard error of the mean.

Table 4: Mean Comprehension Scores by Health Literacy Group (HLS-EU-Q16) and Note Version

Health Literacy Level	Original Mean Comprehension Score (SE)	AI Mean Comprehension Score (SE)
Likely inadequate (n=7)	30.6% (5.1%)	76.6% (4.7%)
Likely problematic (n=5)	48.4% (6.1%)	81.8% (6.7%)
Likely sufficient (n=7)	43.6% (6.8%)	67.6% (5%)

Substituting HLS with TOFHLA as the covariate in the model yielded nearly identical results for the mean comprehension scores. Participants showed significantly higher comprehension scores for AI-generated notes with an estimated marginal mean of 74.7% (SE 5.7%) compared to original notes having an estimated marginal mean of 40.1% (SE 5.7%, *F*(1, 9.7)=26.3, *P*<.001), with a mean difference of 34.6% (95% CI 19.5-49.7).

The main effect of TOFHLA was not statistically significant, *F*(1, 17)=3.6, *P*=.076, and no significant interaction between note version and TOFHLA score was found (*F*(1, 88.2)=.31, *P*=.578).

Model fit statistics showed that fixed effects explained 35% of the variance in comprehension (marginal R^2), while the full model including participant and unique note variation explained 63% (conditional R^2).

Assessment of Participants' Perception of Clinical Notes

The participant ratings used to assess their perceptions of the original clinical notes revealed consistently low scores across all dimensions, with median scores of 1 (IQR 1) for ease of understanding, confidence in comprehension, clarity and precision of language, and personal relevance. The median scores for perceived informativeness and the ability to act on the information provided were scored at 1 (IQR 2) and 2 (IQR 2), respectively.

In contrast, AI-generated notes had a median score of 5 (IQR 1) for ease of understanding, clarity and precision of language, perceived informativeness, ability to act on information, and personal relevance. Confidence in comprehension had a median score of 4 (IQR 1).

Statistical comparisons between original and AI-generated notes revealed significant differences across all six questionnaire items (*P*<.001). AI-generated notes were rated significantly higher in ease of understanding (*U*=167, *z*=-8.519), confidence in comprehension (*U*=220.5, *z*=-8.178), and clarity and precision of language (*U*=157.5, *z*=-8.556). Similar significant differences were observed for perceived informativeness (*U*=153, *z*=-8.595), the ability to act on provided information (*U*=377.5, *z*=-7.274), and personal relevance (*U*=172, *z*=-8.533). These finding are summarized in Table 5.

Table 5: Subjective Evaluation of AI-Generated Notes Compared to Original Notes

Statement	AI-generated notes Median (IQR)	Original notes Median (IQR)	Mann-Whitney U	Z-score	P-value
S1: Ease of understanding	5 (1)	1 (1)	167	-8.519	<.001
S2: Confidence in comprehension	4 (1)	1 (1)	220.5	-8.178	<.001
S3: Clarity and precision of language	5 (1)	1 (1)	157.5	-8.556	<.001
S4: Perceived level of information	5 (1)	1 (2)	153	-8.595	<.001
S5: Ability to act on information	5 (1)	2 (2)	377.5	-7.274	<.001
S6: Personal relevance of note	5 (1)	1 (1)	172.0	-8.533	<.001

Assessment of Time Spent on Reading

In total, 6 out of 114 note readings (5%) were classified as not fully read. Of these, 4 (67%) were original notes where participants gave up due to difficulty, while the remaining 2 (33%) were AI-generated notes - one due to an accidental omission, and one because the participant deemed full reading unnecessary because they felt it could be retold based on the bullet points alone.

The median reading time per word was 0.52 seconds (IQR 0.26) for the original notes and 0.36 seconds (IQR 0.14) for the AI-generated notes. Statistical analysis of the time spent reading revealed a significant difference between note versions (*U*=525, *P*<.001), with participants spending more seconds per word on the original notes (mean rank=72.1) than on the AI-generated notes (mean rank=37.5) (Figure 7).

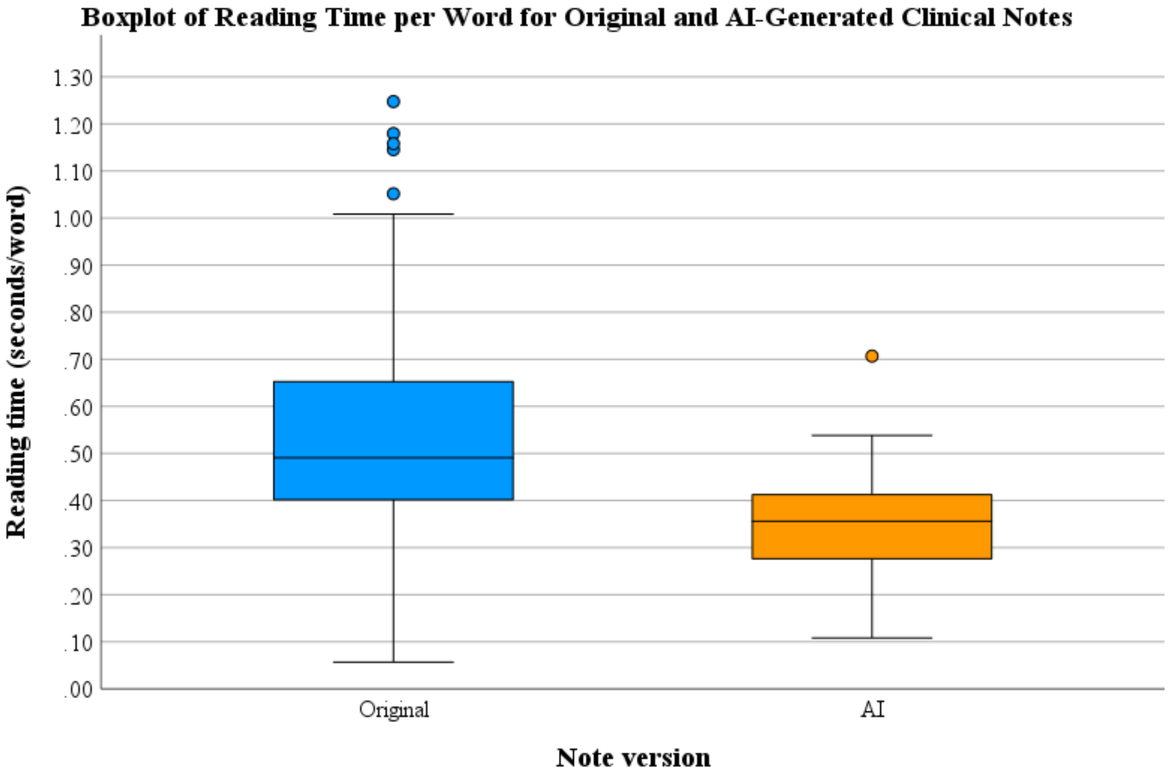


Figure 7: Boxplots of reading time per word for original and AI-generated clinical notes. The boxplots display the distribution of reading times (in seconds per word) for each note version.

Expert Evaluation

AI-generated notes were evaluated by experts across seven quality dimensions. Most notes were rated as highly complete, with a median completeness score of 1 (IQR 1), indicating that nearly all relevant information was included. The AI-generated notes were also generally clear and logically organized (median 2, IQR 1) and typically usable for patient communication with minimal changes required (median 2, IQR 2). Minor issues were observed for conciseness (median 2, IQR 2) and factual accuracy (median 2, IQR 2), suggesting the presence of some irrelevant information or minor ambiguities. The greatest concern was found in risk identification, where ratings showed the most variation (median 3, IQR 3), reflecting differences in expert opinion regarding the potential for patient misunderstanding or safety risks. A detailed overview of ratings for each dimension is shown in Figure 8.

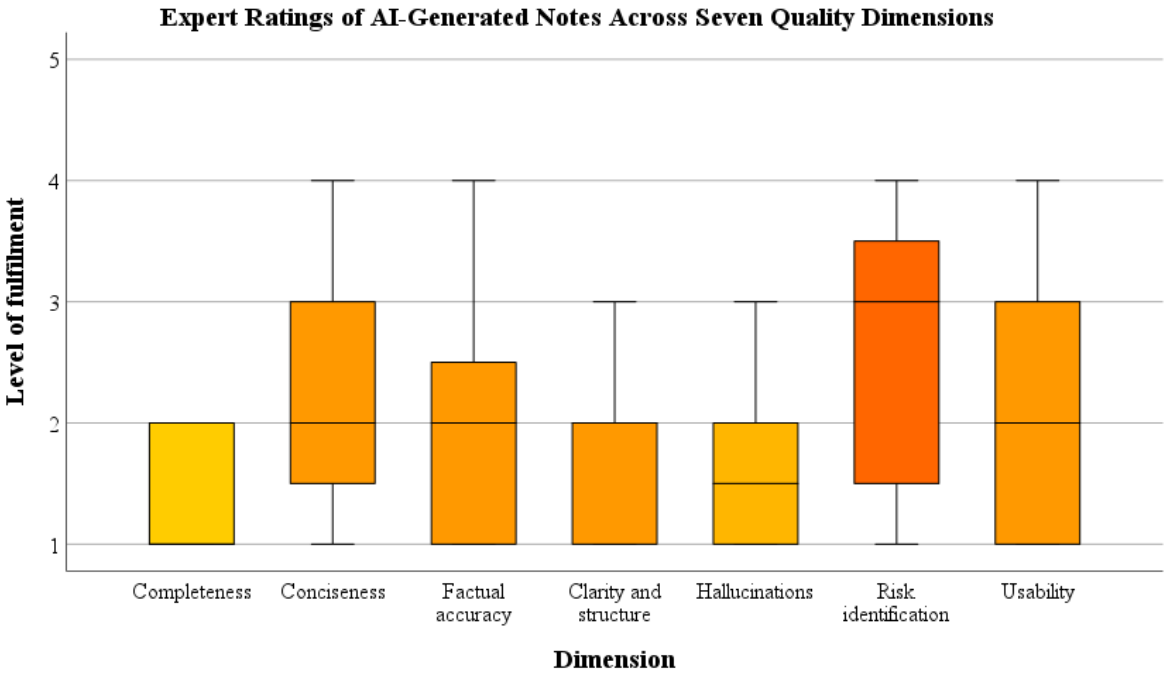


Figure 8: Expert ratings of AI-generated notes across seven quality dimensions. Boxplots display the distribution of scores for completeness, conciseness, factual accuracy, clarity and structure, hallucinations, risk identification, and usability (1=optimal fulfilment; 5=lowest fulfilment). Warmer colours indicate higher median scores.

Analysis of Free-Text Responses

The thematic analysis of the experts' free-text responses revealed five recurrent themes:

Language Clarity and Terminology

Several comments indicated that the notes contained unclear or overly literal translations of medical expressions, which could lead to misunderstanding. Terms such as "your feet are soft [pes planus]" and "upper part of the elbow [proximal ulna]" were flagged as confusing, while certain colloquial phrases e.g., "small shoulder pain package [a standard prescription medicine bundle for post-operative shoulder pain]" were perceived as odd or misleading.

Lack of Detail or Explanation

The experts noted missing information or insufficient elaboration on clinical findings and advice. Examples included unclear references to spinal misalignment and the absence of details about exercise or rehabilitation programs.

Need for Practical Guidance

The notes were described as lacking actionable instructions, especially regarding the use of medication or self-care tools. Experts emphasized the importance of clear dosage information and cautions about over-the-counter drugs such as Paracetamol and Ibuprofen.

Ambiguity and Speculative Statements

Comments also pointed to vague or uncertain formulations that could cause confusion or be perceived as hallucinations. Phrases like "you will probably receive information …" when no future plans are mentioned in the original notes were highlighted as particularly problematic due to their lack of clarity.

Scope of Patient Information

One expert reflected on the broader question of how much information patients need, particularly in relation to surgical notes. The comment expressed uncertainty about whether all procedural details are relevant or helpful for patients and suggested that the appropriate level of detail might be better determined by patients themselves.

Discussion

Principal Results

This proof-of-concept study demonstrated that AI-generated clinical notes significantly improved participants' objective comprehension, as reflected by higher scores on self-developed checklists. Subjective evaluations also consistently favoured AI notes. These findings are consistent with prior research demonstrating that AI-generated simplifications enhance the accessibility and clarity of complex medical documentation (16,21,22,24). This study further contributes to the existing literature by incorporating patients and measuring their comprehension rather than relying solely on readability formulas or expert evaluations. Additionally, this trial showed the greatest improvements in comprehension scores in surgical and imaging notes, which typically contain complex terminology, supporting prior studies emphasizing simplification of radiology reports (21–26,29).

An important contribution of this study was the integration of the HLS assessment, which allowed for a stratified analysis across literacy levels. Although higher health literacy scores were not consistently associated with better overall comprehension, a significant interaction between HLS levels and note versions was identified. These findings suggest that the benefit of AI-generated notes depended on participants' literacy level, with particularly strong effects observed among those with likely inadequate or problematic health literacy. These findings indicate that AI-generated notes notably enhance comprehension for individuals with lower health literacy, emphasizing the importance of tailoring health communication to patients' literacy levels. This observed effect aligns with previous studies demonstrating that individuals with low health literacy experience greater gains in self-efficacy and improved clinical outcomes when actively engaged in their care compared to individuals with higher health literacy (49). Furthermore, while patients with high health literacy generally report greater satisfaction with their physician interactions, research has shown that patients who received explanations in accessible language and had basic knowledge of their own medical history also reported higher satisfaction with their care (50,51). These findings highlight the potential for simplified, patient-centred communication to benefit all patients, particularly those with limited health literacy, by improving both understanding and the overall care experience. Consequently, these results reinforce the necessity for health systems to prioritize health literacy-sensitive communication strategies, especially for populations at risk of misunderstanding or low engagement (48).

Despite tailoring AI-generated notes specifically toward readers with low health literacy, substantial variability in comprehension persisted across different literacy groups. However, the relationship between self-reported health literacy and comprehension may not be entirely straightforward. The HLS captures perceived ability (41) rather than tested functional capacity, and it is therefore plausible that some participants with lower HLS scores demonstrated high comprehension, while others with high HLS scores exhibited a lower comprehension score, as seen in Figure 6 and Table 4. This possible mismatch between perceived literacy and actual performance is supported by previous research. Subjective instruments such as the HLS, while valuable for capturing participants' experiences and perspectives, can be influenced by social desirability and feelings of shame or embarrassment, potentially leading individuals with low health literacy to overestimate their abilities or underreport difficulties (39). Furthermore, recent evidence indicates that individuals with low objective health literacy may display high confidence in their own health knowledge, which can further widen the gap between self-perception and actual capacity (52). These findings underscore the importance of using both subjective and objective measures when evaluating health literacy, as relying solely on self-assessment may obscure key differences in how individuals process and retain clinical information.

This study attempted to combine HLS' subjective and TOFHLA's objective measures of health literacy to compensate for the limitations of HLS, as the instrument relies on self-reported data, which may introduce response biases or inaccuracies (38). However, when using DS-TOFHLA to analyse the interaction between health literacy and comprehension, the results were not significant. This could be attributed to the instruments' lack of formal validation, raising uncertainty about its precision in capturing functional literacy accurately (39), or as previously mentioned discrepancies between self-perception and actual capacity.

Objective readability assessment revealed that LIX scores for AI-generated notes were markedly lower compared to original notes, indicating significantly improved readability. These results are consistent with previous studies that have reported similar improvements in readability using automated methods (16,21,26). However, it should be noted that those studies primarily applied the Flesch-Kincaid and Flesch Reading Ease formulas, which are designed for English-language texts, and are therefore not directly comparable to the Scandinavian readability formula used in this study (53). In addition, PEMAT assessments demonstrated significantly higher scores for understandability and actionability in the AI-generated notes compared to the originals. These findings also align with prior research, further supporting the observed improvements in readability, understandability, and actionability (16).

Expert evaluations indicated that the AI-generated clinical notes either fully met or required only minor revisions to satisfy most of the assessed criteria. These findings are consistent with previous studies reporting that AI-simplified clinical texts can generally meet professional standards with minimal adjustments (24,28). However, a few discrepancies were noted, including instances that could potentially compromise patient safety, reflecting similar concerns about occasional inaccuracies and hallucinations raised in earlier research on AI-generated medical documentation (22,29). These concerns were echoed in the experts' free-text comments, which reinforced their quantitative assessments and collectively point to the necessity of human oversight in validating AI-generated outputs. This aligns with recent findings in the literature, which emphasize that, despite ongoing improvements in large language models, human validation remains essential to ensure clinical accuracy, contextual appropriateness, and patient safety before AI-generated content can be distributed to patients (17).

Limitations

Given the nature of a proof-of-concept study, certain limitations must be taken into consideration. The comprehension checklist and subjective evaluation questionnaire were developed specifically for this trial due to the novelty of the research field. Although the subjective evaluation questionnaire was developed with inspiration from established literature and expert evaluations (16,28), these tools have not undergone formal psychometric validation, which could affect their reliability and generalizability.

The use of self-developed checklists to assess objective comprehension may have favoured AI-generated notes. Since the AI-generated notes were deliberately formulated using patient-friendly language and minimal medical jargon, the correct answers on the checklists often closely matched the simplified phrasing and structure of the AI versions. Consequently, when participants were encouraged to use their own words during retelling, these "own words" frequently mirrored the wording and terminology already present in the AI-generated notes. This alignment could have increased the likelihood of higher comprehensions scores for AI-generated notes, particularly when participants referenced the text directly during retelling. Thus, the potential for bias in favour of AI-generated content should be considered when interpreting the magnitude of the observed effect.

The key messages included in the checklists were determined by the study authors, all of whom had formal training in the musculoskeletal field, and with one author possessing three years of practical experience in this area. This expertise was considered sufficient for the purpose of this study; however, the selection process could potentially have been strengthened by involving independent external experts from the musculoskeletal specialty. In future studies, such an approach may enhance the objectivity and generalizability of the comprehension assessment.

Another limitation of this study concerns the use of a single prompt designed specifically for individuals with low health literacy, despite inclusion of participants with a broad range of health literacy levels. While this approach aligns with policy recommendations advocating for plain language as a strategy to address low health literacy and improve access to health information (51,54), it raises the possibility that some individuals with higher health literacy may have perceived the notes as overly simplistic or even condescending (51). The purpose of plain language is intended to promote clarity and meaning rather than "dumbing down" information or compromising accuracy (56). Moreover, research suggests that simple language and materials intended for a lower literacy audience is generally well accepted by people with adequate literacy levels (51).

A key strength of the prompt used in this study is the patient-centred focus: it ensures simple language, avoidance of medical jargon, and structured organization. The clear rules about fidelity to the original content and in-text explanations for technical terms help reduce the risk of inaccuracies and support consistent, empathetic communication. However, the strict adherence to the original note can also limit the amount of helpful background or context provided, particularly when the source notes are sparse or technical. Moreover, the standard template may not always fit the varied structure and complexity of clinical notes, and any relevant information missing in the original text will also be absent in the AI-generated note. In summary, while the prompt provides a strong and reproducible foundation for generating patient-friendly notes, future efforts should explore adapting prompts to clinical scenarios.

Conclusions

This experimental proof-of-concept study indicated that AI-generated versions of original clinical notes within musculoskeletal healthcare can improve participants' objective comprehension, while also being consistently preferred in subjective evaluations. Furthermore, the study showed that participants with low health literacy benefited the most from AI-generated notes in terms of comprehension.

AI-generated, patient-friendly notes may help shift clinical documentation toward a more inclusive and accessible format that supports patient understanding and engagement. By reducing linguistic barriers and presenting information in a coherent, patient-centred manner, AI can decrease reliance on informal interpretation and reduce information-related anxiety. As patient access to EHRs continues to expand, AI-based simplification may offer a scalable solution, particularly for those with limited health literacy. Future studies should prioritize the development and formal validation of assessment tools specifically designed for measuring patient comprehension of clinical notes.

Acknowledgements

The authors would like to thank Stine Hangaard and Jannie Damsgaard Nørlev for providing supervision and support. The authors would also like to express our gratitude to Aalborg Rehabilitation for facilitating participant recruitment and providing support and utilities during data collection. Additionally, the authors would like to thank Lisa Korsbakke Emtekær Hæsum for providing both the HLS-EU-Q16 and DS-TOFHLA instruments and for offering guidance on the application of these tools.

Conflicts of Interest

The authors do not have any conflicts of interests to declare.

References

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet [Internet]. 2018 Nov 10 [cited 2025 May 29];392(10159):1789–858. Available from: https://www.sciencedirect.com/science/article/pii/S0140673618322797

2. Gill TK, Mittinty MM, March LM, Steinmetz JD, Culbreth GT, Cross M, et al. Global, regional, and national burden of other musculoskeletal disorders, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol [Internet]. 2023 Nov 1 [cited 2025 May 26];5(11):e670–82. Available from: https://www.sciencedirect.com/science/article/pii/S2665991323002321

3. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet [Internet]. 2020 Dec 19 [cited 2025 May 26];396(10267):2006–17. Available from: https://www-sciencedirect-com.zorac.aub.aau.dk/science/article/pii/S0140673620323400

4. Okuhara T, Furukawa E, Okada H, Yokota R, Kiuchi T. Readability of written information for patients across 30 years: A systematic review of systematic reviews. Patient Educ Couns [Internet]. 2025 Jun 1 [cited 2025 May 26];135:108656. Available from: https://www-sciencedirect-com.zorac.aub.aau.dk/science/article/pii/S0738399125000230

5. Gotlieb R, Praska C, Hendrickson MA, Marmet J, Charpentier V, Hause E, et al. Accuracy in Patient Understanding of Common Medical Phrases. JAMA Netw Open [Internet]. 2022 Nov 30;5(11):e2242972–e2242972. Available from: https://doi.org/10.1001/jamanetworkopen.2022.42972

6. Imoisili OE, Levinsohn E, Pan C, Howell BA, Streiter S, Rosenbaum JR. Discrepancy Between Patient Health Literacy Levels and Readability of Patient Education Materials from an Electronic Health Record. HLRP: Health Literacy Research and Practice [Internet]. 2017 Oct [cited 2025 May 29];1(4). Available from: https://doi.org/10.3928/24748307-20170918-01

7. Thai P, Flores-Cruz G, Roque NA. Leveraging Healthcare Technology to Improve Patient-Doctor Communication. Proceedings of the Human Factors and Ergonomics Society Annual Meeting [Internet]. 2023 Sep 1;67(1):2317–22. Available from: https://doi.org/10.1177/21695067231192641

8. Sætre LMS, Jarbøl DE, Raasthøj IP, Seldorf SA, Rasmussen S, Balasubramaniam K. Examining health literacy in the Danish general population: a cross-sectional study on the associations between individual factors and healthcare-seeking behaviour. Eur J Public Health [Internet]. 2024 Dec 1;34(6):1125–33. Available from: https://doi.org/10.1093/eurpub/ckae150

9. Sham S, Shiwlani S, Kirshan Kumar S, Bai P, Bendari A. Empowering Patients Through Digital Health Literacy and Access to Electronic Medical Records (EMRs) in the Developing World. Cureus [Internet]. 2024 Apr 3 [cited 2025 May 29]; Available from: https://doi.org/10.7759/cureus.57527

10. Nøhr C, Parv L, Kink P, Cummings E, Almond H, Nørgaard JR, et al. Nationwide citizen access to their health data: analysing and comparing experiences in Denmark, Estonia and Australia. BMC Health Serv Res [Internet]. 2017;17(1):534. Available from: https://doi.org/10.1186/s12913-017-2482-y

11. Zarcadoolas C, Vaughon WL, Czaja SJ, Levy J, Rockoff ML. Consumers' Perceptions of Patient-Accessible Electronic Medical Records. J Med Internet Res [Internet]. 2013 Aug 26;15(8):e168. Available from: http://www.jmir.org/2013/8/e168/

12. Wass S, Vimarlund V, Ros A. Exploring patients' perceptions of accessing electronic health records: Innovation in healthcare. Health Informatics J [Internet]. 2017 Apr 30;25(1):203–15. Available from: https://doi.org/10.1177/1460458217704258

13. Sørensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, et al. Health literacy and public health: A systematic review and integration of definitions and models. BMC Public Health [Internet]. 2012;12(1):80. Available from: https://doi.org/10.1186/1471-2458-12-80

14. Wolf MS, Davis TC, Shrank W, Rapp DN, Bass PF, Connor UM, et al. To err is human: Patient misinterpretations of prescription drug label instructions. Patient Educ Couns [Internet]. 2007 Aug 1 [cited 2025 May 29];67(3):293–300. Available from: https://doi.org/10.1016/j.pec.2007.03.024

15. Reynolds TL, Ali N, McGregor E, Longhurst C, Rosenberg AL, Rudkin SE, et al. Understanding Patient Questions about their Medical Records in an Online Health Forum: Opportunity for Patient Portal Design.

16. Zaretsky J, Min Kim J, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. JAMA Netw Open [Internet]. 2024 Mar 11 [cited 2025 May 29];E240357. Available from: https://doi.org/10.1001/jamanetworkopen.2024.0357

17. Busch F, Hoffmann L, Rueger C, van Dijk EHC, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Communications Medicine [Internet]. 2025 Jan 21;5(1):26. Available from: https://doi.org/10.1038/s43856-024-00717-2

18. van Mens HJT, van Eysden MM, Nienhuis R, van Delden JJM, de Keizer NF, Cornet R. Evaluation of lexical clarification by patients reading their clinical notes: a quasi-experimental interview study. BMC Med Inform Decis Mak [Internet]. 2020;20(10):278. Available from: https://doi.org/10.1186/s12911-020-01286-9

19. Lalor JP, Levy DA, Jordan HS, Hu W, Smirnova JK, Yu H. Evaluating Expert-Layperson Agreement in Identifying Jargon Terms in Electronic Health Record Notes: Observational Study. J Med Internet Res [Internet]. 2024 Oct 15 [cited 2025 May 29];26:e49704. Available from: https://doi.org/10.2196/49704

20. Chen J, Druhl E, Polepalli Ramesh B, Houston TK, Brandt CA, Zulman DM, et al. A Natural Language Processing System That Links Medical Terms in Electronic Health Record Notes to Lay Definitions: System Development Using Physician Reviews. J Med Internet Res [Internet]. 2018 Jan 22 [cited 2025 May 29];20(1):e26. Available from: https://doi.org/10.2196/jmir.8669

21. Doshi R, Amin KS, Khosla P, Bajaj S, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. Radiology [Internet]. 2024 Mar 1 [cited 2025 May 29];310(3). Available from: https://doi.org/10.1148/radiol.231593

22. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol [Internet]. 2024 May 1 [cited 2025 May 29];34(5):2817–25. Available from: https://doi.org/10.1007/s00330-023-10213-1

23. Can E, Uller W, Vogt K, Doppler MC, Busch F, Bayerl N, et al. Large Language Models for Simplified Interventional Radiology Reports: A Comparative Analysis. Acad Radiol [Internet]. 2024 Feb 1 [cited 2025 May 29]; Available from: https://doi.org/10.1016/j.acra.2024.09.041

24. Kuckelman IJ, Wetley K, Yi PH, Ross AB. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. Skeletal Radiol [Internet]. 2024 Aug 1 [cited 2025 May 29];53(8):1621–4. Available from: https://doi.org/10.1007/s00256-024-04599-2

25. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art [Internet]. 2023 Dec 1 [cited 2025 May 29];6(1). Available from: https://doi.org/10.1186/s42492-023-00136-5

26. Tang CC, Nagesh S, Fussell DA, Glavis-Bloom J, Mishra N, Li C, et al. Generating colloquial radiology reports with large language models. Journal of the American Medical Informatics Association [Internet]. 2024 Nov 1 [cited 2025 May 29]; Available from: https://doi.org/10.1093/jamia/ocae223

27. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation [Internet]. Vol. 55, ACM Computing Surveys. Association for Computing Machinery; 2023 [cited 2025 May 29]. Available from: https://doi.org/10.1145/3571730

28. Kim H, Jin HM, Jung Y Bin, You SC. Patient-Friendly Discharge Summaries in Korea Based on ChatGPT: Software Development and Validation. J Korean Med Sci [Internet]. 2024 [cited 2025 May 29];39(16). Available from: https://doi.org/10.3346/jkms.2024.39.e148

29. Shen Y, Xu Y, Ma J, Rui W, Zhao C, Heacock L, et al. Multi-modal large language models in radiology: principles, applications, and potential [Internet]. Abdominal Radiology. Springer; 2024 [cited 2025 May 29]. Available from: https://doi.org/10.1007/s00261-024-04708-8

30. Aali A, Van Veen D, Arefeen YI, Hom J, Bluethgen C, Reis EP, et al. A dataset and benchmark for hospital course summarization with adapted large language models. Journal of the American Medical Informatics Association [Internet]. 2025 Mar 1 [cited 2025 May 29];32(3):470–9. Available from: https://doi.org/10.1093/jamia/ocae312

31. European Union. Official Journal of the European Union. 2016 [cited 2025 May 26]. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj

32. Danish Ministry of Justice. Retsinformation. 2018 [cited 2025 May 26]. Lov om supplerende bestemmelser til forordning om beskyttelse af fysiske personer i forbindelse med behandling af personoplysninger (Databeskyttelsesloven). Available from: https://www.retsinformation.dk/eli/lta/2018/502

33. OpenAI. OpenAI. 2024 [cited 2025 May 29]. Hello GPT-4o. Available from: https://openai.com/index/hello-gpt-4o/

34. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. 2023 Mar 20 [cited 2025 May 29]; Available from: https://doi.org/10.48550/arxiv.2303.13375

35. National Committee on Health Research Ethics. researchethics.dk. 2024 [cited 2025 May 26]. Consent Declarations and Forms. Available from: https://researchethics.dk/information-for-researchers/consent-declarations-and-forms

36. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform [Internet]. 2009 Apr 1 [cited 2025 May 29];42(2):377–81. Available from: https://doi.org/10.1016/j.jbi.2008.08.010

37. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform [Internet]. 2019 Jul 1 [cited 2025 May 29];95:103208. Available from: https://doi.org/10.1016/j.jbi.2019.103208

38. Nielsen MG, Svendsen MT, Sørensen K, Grønborg TK, Torp-Pedersen C, Bøggild H. Psychometric properties of the Danish version of the European Health Literacy Survey Questionnaire. Eur J Public Health [Internet]. 2020 Sep 1;30(Supplement_5):ckaa165.581. Available from: https://doi.org/10.1093/eurpub/ckaa165.581

39. Emtekær Hæsum LK, Ehlers L, Hejlesen OK. Validation of the Test of Functional Health Literacy in Adults in a Danish population. Scand J Caring Sci [Internet]. 2015 Sep 1;29(3):573–81. Available from: https://doi.org/10.1111/scs.12186

40. Yen PH, Leasure AR. Use and Effectiveness of the Teach-Back Method in Patient Education and Health Outcomes. Fed Pract [Internet]. 2019 Jun [cited 2025 May 29];36(6):284–9. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6590951/

41. Pelikan JM, Ganahl K, Van den Broucke S, Sørensen K. Measuring health literacy in Europe: Introducing the European Health Literacy Survey Questionnaire (HLS-EU-Q). In: Okan O, Pinheiro P, Levin-Zamir D, Bauer U, Sørensen K, editors. International Handbook of Health Literacy. Bristol, UK: Policy Press; 2019. p. 115–38.

42. Baker DW, Williams M V, Parker RM, Gazmararian JA, Nurss J. Development of a brief test to measure functional health literacy. Patient Educ Couns [Internet]. 1999 Sep [cited 2025 May 29];38(1):33–42. Available from: https://doi.org/10.1016/S0738-3991(98)00116-5

43. Meier F, Fugl Eskjær M. Analysis of Textual Complexity in Danish News Articles on Climate Change. Digital Humanities in the Nordic and Baltic Countries Publications [Internet]. 2024 Sep 18 [cited 2025 May 29];6(1). Available from: https://doi.org/10.5617/dhnbpub.11490

44. Skrzypczak T, Mamak M. Assessing the Readability of Online Health Information for Colonoscopy — Analysis of Articles in 22 European Languages. Journal of Cancer Education [Internet]. 2023 Jul 26;38(6):1865–70. Available from: https://doi.org/10.1007/s13187-023-02344-2

45. Niels Gamborg. Lixtals beregner [Internet]. Nielsgamborg.dk. Niels Gamborg; 2025 [cited 2025 May 29]. Available from: https://www.nielsgamborg.dk/indhold/lixberegner.htm

46. Agency for Healthcare Research and Quality. Agency for Healthcare Research and Quality (AHRQ). 2020 [cited 2025 May 29]. Patient Education Materials Assessment Tool (PEMAT). Available from: https://www.ahrq.gov/health-literacy/patient-education/pemat.html

47. National Library of Medicine. National Library of Medicine. 2003 [cited 2025 May 29]. MeSH Browser: Comprehension. Available from: https://www.ncbi.nlm.nih.gov/mesh/68032882

48. IBM Corp. IBM SPSS Statistics for Windows [Internet]. Armonk, NY: IBM Corp.; 2023 [cited 2025 May 29]. Available from: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-29

49. Ishikawa H, Yano E. The relationship of patient participation and diabetes outcomes for patients with high vs. low health literacy. Patient Educ Couns [Internet]. 2011 Sep [cited 2025 May 29];84(3):393–7. Available from: https://doi.org/10.1016/j.pec.2011.01.029

50. Altin SV, Stock S. The impact of health literacy, patient-centered communication and shared decision-making on patients' satisfaction with care received in German primary care practices. BMC Health Serv Res [Internet]. 2016 Dec 30 [cited 2025 May 29];16(1):450. Available from: https://doi.org/10.1186/s12913-016-1693-y

51. Meppelink CS, Edith G. S, Bianca M. B, and van Weert JCM. Should We Be Afraid of Simple Messages? The Effects of Text Difficulty and Illustrations in People With Low or High Health Literacy. Health Commun [Internet]. 2015 Dec 2;30(12):1181–9. Available from: https://doi.org/10.1080/10410236.2015.1037425

52. Canady BE, Larzo M. Overconfidence in Managing Health Concerns: The Dunning–Kruger Effect and Health Literacy. J Clin Psychol Med Settings [Internet]. 2023 Jun 29 [cited 2025 May 29];30(2):460–8. Available from: https://doi.org/10.1007/s10880-022-09895-4

53. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: current role in orthopaedics. Clin Orthop Relat Res. 2010 May 22;468(10):2572–80.

54. Stableford S, Mettger W. Plain Language: A Strategic Response to the Health Literacy Challenge. J Public Health Policy [Internet]. 2007 [cited 2025 May 29];28(1):71–93. Available from: https://doi.org/10.1057/palgrave.jphp.3200102

Appendices

Appendix 1 - Final Prompt

“Transform the detailed technical results in the following original medical note into a simplified, patient-friendly version based on the following guidelines:

- Use clear and simple language without medical jargon. If a technical term cannot be simplified, add a brief explanation in brackets.
- Assume the reader has low health literacy and no background knowledge of medical terms or treatments.
- Maintain accuracy - don't remove critical information or add speculative or elaborate explanations that are not explicitly stated in the original medical record.
- Only use information that is directly stated in the original medical record. Do not add detailed explanations of treatment principles, symptoms or prognosis if they are not mentioned.
- If a guide is referenced in the medical record, it can be mentioned - but if no guide is provided or mentioned, do not include additional explanations.
- Use a reassuring and empathetic tone. Avoid wording that may cause undue concern.
- Clearly state any recommendations for further evaluation, follow-up, guidance or treatment if it is included in the original medical record.

Structure the patient-friendly version according to the following template. If there is no information in the record note that fits a specific heading, omit both the heading and the corresponding section. Keep the order and headings, but only for those sections where there is relevant content in the original record note. Do not write that something is not mentioned. If information is missing, simply leave it out without comment.

[Insert an appropriate heading]

What has happened?

(Description of what happened (examination, interview, treatment, surgery, etc.). Must be presented in a coherent prose text).

What was done or found?

(Description of observations, findings or actions etc. For examination: use bullet points for findings/results. For surgery/procedure: describe in an easy-to-understand form what was done and possibly found).

What does it mean to you?

(Explanation of the significance of what was found or done - if the information is provided. Must be presented in prose text).

What should happen next?

(Information about next steps, plan, follow-up, guidance or treatment, etc. Use bullet points if the information is structured. Use prose if the information is concise).

This is the original journal entry:”

Appendix 2 - Example of Original and Corresponding AI-Generated Clinical Note

Case 5 - Back X-ray

Original Clinical Note	**AI-Generated Note**
X-ray Outpatient **1/1: RU columna totalis - E10P** X-ray examination of the thoracic spine: There is a slight, elongated sinistroconvex scoliosis with a mid-thoracic apex. The scoliosis angle is measured at 6°. Normal vertebrae and i.v. spaces. On the lateral projection, several open intervertebral joints are seen. Diagnosis: Very mild scoliosis. X-ray examination of the lumbar spine: There is a slight dextroconvex scoliosis with the apex at L2. The scoliosis angle is measured at 8°. The 5th disc is moderately degenerated. At several levels low lumbar bilaterally, slight spondylarthrosis is seen. The sacroiliac joints have normal joint space width. They are sharp and well-defined. Absence of bone reaction in the area, and overall, no sacroiliitis. Thus, no signs of Morbus Bechterew. Diagnosis: Mild scoliosis. Moderate degenerative changes in the lower lumbar region.	**X-ray Examination of the Back** **What has happened?** You have had an X-ray examination of your back. The images show both the upper (thoracic) and lower (lumbar) parts of the spine. **What was done or found?** The examination showed: • In the upper part of the back: A very mild sideways curvature to the left (scoliosis) with a small angle of 6 degrees. The spinal joints and discs otherwise appear normal. • In the lower part of the back: A mild sideways curvature to the right with a small angle of 8 degrees. • The lowest disc (between the vertebrae) is somewhat worn. • Mild wear is seen in some of the small joints in the lower part of the back (spondylarthrosis). • The joints between the spine and the pelvis appear completely normal, and there are no signs of inflammatory conditions such as Morbus Bechterew (a rheumatic disease of the spine). **What does this mean for you?** You have a mild spinal curvature (scoliosis), but to a very small degree. There are also early signs of wear in the lower part of the back. This is not unusual and can occur with age. No signs of serious disease or inflammation were found in the back. **What happens next?** No follow-up plans are mentioned in the medical note, but you are welcome to get in touch if you experience symptoms or have questions.

Appendix 3 - Example of Checklist (Objective Comprehension)

Words marked with red and italic indicate medical terms. If mentioned by participants, a follow-up question ("What does that mean?") was used to assess understanding of the term.

Case 5 – Back X-ray

Time: [xx.xx]

Scan type:

□ X-ray

Location:

□ Upper (*thoracic*) part of the spine
□ Lower (*lumbar*) part of the spine

Findings:

□ Mild curvature of the spine (*scoliosis*)
□ Mild to moderate degeneration in the lower back (*spondyloarthrosis*)
□ No signs of serious disease or inflammation/rheumatic disease in the back have been found

Notes:

Appendix 4 - Six-Item Questionnaire with Likert Scale (Participants' Perception of Clinical Notes)

To what extent do you agree with the following statements?

Statement	1 Strongly disagree	2 Disagree	3 Neither agree nor disagree	4 Agree	5 Strongly agree
1. The note was easy to understand					
2. I feel confident that I have understood it correctly					
3. The language was clear and precise					
4. I feel well informed after reading the note					
5. I would be able to act based on the information received in the note					
6. I felt that the note was addressed to me as a patient					

Appendix 5 - Seven-Item Questionnaire with Likert Scale (Expert Evaluation)

Completeness

To what extent does the note contain all relevant information that the patient should have?

1. The note contains all relevant and necessary information that the patient should have.
2. The note lacks some non-critical information, but the overall content is comprehensive.
3. The note is missing several essential pieces of information that the patient should know, which may cause confusion or concern.
4. The note lacks many important details, making it difficult for the patient to gain a comprehensive understanding.
5. The note is irrelevant or confusing and lacks essential content for the patient.

Conciseness

To what extent is irrelevant or unnecessary information omitted from the note?

1. The note is very focused and contains only relevant information.
2. The note contains only a small amount of irrelevant information.
3. The note contains some irrelevant information, but it does not significantly interfere.
4. The note contains a fair amount of irrelevant information, making it less focused.
5. The note contains a lot of information that confuses the reader and should be omitted.

Factual Accuracy

To what extent is the information in the note correctly represented compared to the original text?

1. The note is correct and clearly presented in full accordance with the original text.
2. The note is generally correct, but some formulations may cause doubt about the original content.
3. The note contains some errors or inaccuracies that alter the meaning of the information.
4. The note contains several significant errors in translation, formulation, or interpretation of the content.
5. The note contains many incorrect representations or distortions of meaning compared to the original.

Clarity and Structure

To what extent is the note logically constructed and easy to understand for patients?

1. The note is clearly and logically structured, and easy for patients to understand.
2. The note is generally well-written and structured, with some areas that could be improved.
3. The note is partly understandable but has some unclear or illogical sections.
4. The note has an unclear structure and several confusing formulations.
5. The note is poorly organized and difficult to understand.

Hallucinations

To what extent does the note contain information that has no basis in the original text, i.e., content added by the AI without reference?

1. The note contains only information that can be verified in the original text.
2. The note is almost entirely free of hallucinations, but one or two unclear additions occur.
3. The note contains some new statements that cannot be found or verified in the original.
4. The note contains substantial new content that is not documented in the original text.
5. The note contains several serious claims that are completely fabricated and not found in the original.

Risks

To what extent can the note lead to misunderstandings that negatively affect the patient's physical or mental health?

1. The note is clear and correct, with no risk of health-related misunderstandings.
2. The note is mostly clear, but some parts may cause uncertainty.
3. The note may lead to some minor misunderstandings with limited health impact.
4. The note may lead to significant misunderstandings that could negatively affect the patient's health.
5. The note is highly misleading and may lead to serious misunderstandings with health consequences.

Usability

To what extent can the note be immediately given to the patient in its current form?

1. The note can be given directly to the patient without any changes.
2. The note can be given with minor changes that do not significantly affect understanding.
3. The note can be given to the patient but requires adjustments in several places.
4. The note requires significant changes before it can be given to the patient.
5. The note is unsuitable for patient distribution even with extensive changes.

Worksheets

These supplementary worksheets are provided to enhance understanding of central themes and concepts that underpin the research project but extend beyond the scope of the main scientific paper. The worksheets are furthermore intended to support the semester's learning objectives.

Worksheet 1: Health Literacy

Definition

Health literacy is a multifaceted concept that significantly impacts both individuals and society. It involves the ability to understand and use health-related information, with two key components: personal health literacy, which refers to an individual's capacity to find, understand, and apply health information, and organizational health literacy, which focuses on how well organizations enable this process (1).

While health literacy is often framed as an individual responsibility, it is heavily influenced by social and organizational factors. These factors shape access to resources and the ability to understand health information. The Health Promotion Glossary 2021 emphasizes that health literacy is "the personal knowledge and competencies accumulated through daily activities, social interactions, and across generations," which are mediated by organizational structures and available resources (2). Beyond individual competencies, health literacy is essential for empowering both individuals and communities to make informed health decisions. It relies on equitable access to education and lifelong learning, serving as an observable outcome of health education within broader health promotion efforts. However, cultural and situational factors, alongside the organizations and societies that shape health communication, further impact health literacy. Governments, civil society, and healthcare services share the responsibility of facilitating access to trustworthy, understandable health information, while social resources, such as media regulations, are vital in ensuring individuals can effectively obtain and use this information (2). Improving health literacy is crucial for promoting informed health decisions and reducing health disparities. It requires a comprehensive approach, addressing both individual competencies and the systemic factors that support or hinder access to reliable health information (3).

Global Health Literacy

In 18 OECD countries, over one-third of the population struggles with inadequate health literacy, with this proportion exceeding 50% in 12 of those countries (4). In Europe, over 10% of the population in several countries exhibit inadequate health literacy, with this proportion varying between 1.8% and 26.9%. Furthermore, almost half of the population in some countries is affected by limited health literacy, with rates ranging from 29% to 62% (5). In Europe, individuals face particular challenges when it comes to understanding and evaluating health information: 47% have difficulty assessing the reliability of health information from media sources, while 41% struggle to judge the benefits and risks of different medical treatments (4). Certain groups, such as those facing financial deprivation, lower social status, limited education, and older age, are particularly vulnerable to limited health literacy (5).

Health Literacy in the Danish Population

Even though equal and free access to public healthcare has been a cornerstone of Danish healthcare policy for many years, ensuring that everyone has access to healthcare services based on their needs, not all individuals are equally capable of utilizing these resources effectively. A significant factor in this disparity is health literacy, which remains a critical issue in the Danish population. A population-based study found that 39% of Danish individuals had inadequate health literacy, with the highest prevalence observed among those receiving unemployment benefits (around 54%) and the lowest among those on voluntary early retirement pensions (around 32%) (6). Additionally, individuals who smoke, live alone, belong to a non-Danish ethnic group, or report poor self-rated health continue to face challenges in feeling understood and supported by healthcare providers. They also struggle to access sufficient information for managing their health conditions, obtaining adequate social support, and engaging actively in their healthcare. Furthermore, individuals with a high symptom burden or those reporting a large number of symptoms to their general practitioner (GP) tend to experience difficulties in receiving adequate information and engaging with healthcare providers (3).

Consequences of Inadequate Health Literacy Support

Low health literacy is a widespread and persistent challenge with significant implications at both the individual and systemic levels. At the level of healthcare systems, it contributes to increased service utilization and inefficient allocation of resources. Individuals with inadequate or marginal health literacy often struggle to navigate healthcare services, which can lead to delayed access to appropriate care, higher number of revisits to emergency departments, and longer hospital stays (7). These inefficiencies not only burden the healthcare system economically but may also indicate missed opportunities for timely and effective interventions.

At the individual level, the impact of low health literacy is particularly notable. Research consistently links limited health literacy to poorer health outcomes. For example, individuals with difficulties understanding health-related information are more likely to substitute routine primary care with emergency department visits (8). This may reflect challenges in identifying when and where to seek appropriate care, as well as difficulties in understanding communication from healthcare providers. In such cases, patients may default to emergency services, resulting in fragmented and potentially inappropriate care.

Medication safety is another area where limited health literacy poses significant risks. Studies have shown that many patients struggle to interpret prescription labels correctly, which can lead to medication errors, adverse drug reactions, and poor disease management (9), a factor that may contribute to increased morbidity and mortality, as also shown by Baker et al. (10). In surgical settings, low health literacy has similarly been associated with poor adherence to preoperative instructions, which may compromise patient safety (11).

Beyond access to and use of care, low health literacy also influences the quality of care patients receive. Individuals with limited health literacy often experience difficulties understanding health insurance coverage, locating appropriate services, or navigating administrative processes within healthcare systems (12). These challenges can result in delayed or avoided care, further exacerbating health inequalities. Effective communication between patients and providers is also hindered by low health literacy. Patients with limited comprehension frequently report poorer experiences with healthcare communication, including confusion regarding diagnoses, treatment plans, and follow-up procedures (12). Koh et al. describe this dynamic as a "cycle of crisis care," in which patients repeatedly access healthcare services without fully understanding the care being provided, leading to recurring health issues (12). Factors such as complex forms, use of medical jargon, and unclear discharge instructions further contribute to this cycle.

Worksheet 2: Medical Documentation in the Danish Health Care System

Medical Documentation in Denmark

Medical documentation plays a central role in the Danish healthcare system by ensuring continuity of care, supporting clinical decisions, and fulfilling legal requirements. Healthcare professionals in Denmark follow specific documentation guidelines established by the individual regions and the Danish Patient Safety Authority (13). These guidelines emphasize the importance of accuracy, timelines, and structured records to ensure patient safety. Documentation serves several purposes: it helps maintain continuity of care across healthcare providers and sectors, supports decision-making in clinical settings, and meets legal and ethical standards.

While the primary function of clinical documentation is to serve as a professional tool for healthcare providers, it also plays a secondary role in supporting patient involvement. According to §3 of the Danish Executive Order on Medical Recordkeeping (*Journalføringsbekendtgørelsen*), the patient record is intended to ensure safe and effective care through necessary clinical notes. However, it may also contribute to patient engagement by enabling individuals to participate more actively in their treatment and to safeguard their own interests (13).

eHealth in Denmark

Denmark has a well-established electronic health record (EHR) system, with *sundhed.dk* serving as the national eHealth portal. The platform provides both citizens and healthcare professionals with access to patient data created at general practices, specialist clinics, and hospital systems. The central aim of this initiative was to empower patients by providing them with digital access to their health data, including medical records and laboratory results, thus enhancing transparency, patient engagement, and continuity of care. In 2019, the *MinSundhed* app was introduced as a complementary mobile platform, offering access to many of the portal's features along with additional functionalities such as emergency contact tools (14).

The data accessed through *sundhed.dk* is created at various healthcare settings and transferred to the portal, where it becomes available to both patients and healthcare providers (15). To protect patient privacy, data is subject to a two-week delay, and healthcare providers can only access records for patients under their care. Hospitals are legally mandated to submit care summaries, and prescription data is available via the patient portal, allowing both citizens and general practitioners to access critical medical information (15).

Currently, patients do not have access to their general practitioners' notes via *Sundhed.dk*. However, a joint digitalization strategy, developed by the Danish government, Danish Regions, and Local Government Denmark (KL), aims to expand the portal's functionalities. The strategy's goal is to enable patients to access information about their entire care trajectory across various sectors, thus fostering a more cohesive and transparent healthcare experience (16). As digital access becomes more widespread increasing numbers of patients are engaging directly with their own health data. This development represents a significant step towards greater transparency and patient involvement. However, it also brings to light a number of challenges related to how clinical information is presented and understood by non-professional users.

Challenges in Medical Documentation

The growing accessibility of electronic health records in Denmark, particularly through platforms like *sundhed.dk* and *MinSundhed*, has reinforced a critical tension between traditional documentation practices and patient comprehension. While these digital tools are intended to enhance patient autonomy and involvement in healthcare decision-making, they also present a range of consequences at both the individual and societal levels. Medical records are primarily designed as clinical tools for healthcare professionals, yet patients are now secondary users of this information. As a result, the way documentation is written has significant implications for how patients engage with and understand their health data.

A British study suggests that when patients are exposed to their medical records without adequate preparation or context, can lead to confusion and elevated anxiety. For instance, patients may encounter unfamiliar or complex medical terminology that they either misinterpret or find difficult to comprehend. This can strain the doctor-patient relationship, as patients may feel overwhelmed or uncertain about the significance of their health conditions. In parallel, healthcare professionals report an increased burden, as additional time is required to explain or clarify information, often outside of scheduled consultations (17).

Furthermore, studies have shown that the use of specific medical terminology can inadvertently alienate patients. Common phrases such as "patient claims" or "patient denies" are often perceived as dismissive, while more complex diagnostic terms can lead to misinterpretations. For example, patients have misinterpreted the ICD-10 diagnosis "Dizziness and Giddiness," with the term "giddy" being perceived as trivializing their condition, while others felt that the term "pseudo-claudication" undermined the validity of their symptoms (18).

These findings emphasize the need for clear, patient-friendly language in documentation. Efforts to simplify explanations, clarify diagnostic terms, and provide accessible digital resources, such as patient-friendly summaries within EHRs, could improve comprehension and engagement. Further research is needed to assess the most effective methods for making medical records more understandable without compromising clinical accuracy.

Worksheet 3: Generative Artificial Intelligence

Generative AI

Generative artificial intelligence (generative AI) refers to AI systems capable of producing new and meaningful content, such as text, images, or audio, by learning patterns from existing data. Unlike traditional AI models focused on decision-making, generative AI uses modeling techniques that infer complex data distributions to generate original outputs that often resemble human-created content. A generative AI model is a specific type of machine learning architecture designed to create novel data instances based on observed patterns and relationships in training data. While such models play a central role in generative AI, they are inherently incomplete and typically require further fine-tuning and integration into specific systems and applications to perform targeted tasks effectively (19).

Large Language Models (LLMs)

A prominent example of generative AI is Large Language Models (LLMs), which are a class of advanced natural language processing (NLP) models within the broader category of generative AI. These models, including examples such as GPT-4 and BERT, are built upon deep learning architectures, most notably transformer models, which enable them to process and generate human-like text with a high degree of fluency and contextual awareness. Trained on massive datasets sourced primarily from the internet, LLMs are capable of performing a wide range of language-based tasks, such as text summarization, content generation, machine translation, and conversational interaction. As generative AI systems, LLMs do not merely analyze language, they also generate coherent, contextually relevant outputs, making them suitable for applications that require dynamic text generation and interaction (19,20).

Prompting

Prompting involves carefully designing textual instructions ("prompts") to guide large language models (LLMs) toward producing specific outputs. The wording and format significantly influence the effectiveness and accuracy of the model’s responses. Poorly designed prompts may result in inaccuracies or unintended outputs known as hallucinations (21).

Prompt learning leverages the inherent knowledge of LLMs to address specific tasks efficiently, without extensive fine-tuning. Recent advances have introduced more sophisticated, data-driven methods, such as reinforcement learning, to optimize prompts (19).

Strategically engineered prompts can mitigate common issues with LLMs, including hallucinations and biases, especially crucial in clinical contexts. Effective prompts emphasize specificity and inclusiveness, enhancing reliability and fairness in healthcare applications (20).

Prompt engineering primarily focuses on clearly structuring and phrasing instructions to reduce ambiguity and factual errors. Typically, prompts consist of three elements: an instruction (defining the task), an output indicator (specifying the desired response format), and context (supplementary details guiding model output). Not all prompts necessarily include every element, depending on the task requirements (20).

LLMs demonstrate unique strengths in zero-shot, one-shot, and few-shot learning. Zero-shot involves task completion based solely on instructions without prior examples, one-shot provides a single illustrative example, and few-shot supplies multiple examples to refine the model's understanding and output (19,20).

ChatGPT

ChatGPT is a conversational AI system based on LLMs from the GPT (Generative Pre-trained Transformer) family, specifically GPT-4. Developed by OpenAI, ChatGPT is designed to process and generate human-like text, supporting applications in various domains, including customer service, content creation, education, and programming assistance (22). One of ChatGPT's key features is its multimodal capability, meaning it can process both text and image inputs while generating text-based responses. The model has been extensively tested on professional and academic benchmarks, demonstrating human-level performance in certain tasks, such as passing a simulated bar exam with scores in the top 10% of test-takers. Compared to earlier models like GPT-3.5, GPT-4 exhibits improved factual accuracy, adherence to user intent, and stronger performance across multiple languages (22). ChatGPT utilizes deep learning techniques, including transformers and reinforcement learning from human feedback (RLHF), to refine responses and align them with user expectations. Additionally, its ability to analyze and explain complex topics makes it valuable for research and decision support in various industries (22).

In recent years, several LLMs have been developed specifically for the biomedical field. However, ChatGPT has emerged as a significant disruptor in the medical literature on LLMs. According to a systematic review aiming to synthesize the applications and limitations of LLMs in patient care, nearly 80% of the models examined were based on GPT-3.5 and GPT-4 (23).

Large Language Models in Patient Communication

LLMs are emerging as powerful tools in medicine, enhancing access to care, supporting diagnostic reasoning, and aiding treatment planning. Additional applications include improved surgical planning, greater accuracy in medical imaging, and more effective physician-patient communication (20).

Communication is a cornerstone of patient-centered care, encompassing both direct interactions between providers and patients and indirect channels such as electronic health records (EHRs), patient feedback systems, and automated messaging (20).

LLMs can transform unstructured clinical notes into structured formats. They also facilitate multilingual communication by providing fast, reliable translations, thereby bridging language gaps between patients and providers (20).

Empirical studies show that ChatGPT-generated clinical letters, such as those related to skin cancer, are rated highly for factual accuracy and human-like tone (24). Similarly, simplified radiology reports produced by ChatGPT received favorable ratings from radiologists regarding completeness and their potential to improve patient-centered care (24).

Beyond readability and accuracy, ChatGPT has demonstrated the ability to organize medical notes for ICU patients—even when facing abbreviations or missing context—by structuring information according to categories such as treatment status, lab values, respiratory function, and hemodynamic parameters (24).

Collectively, these capabilities highlight the significant potential for ChatGPT and similar models to improve the efficiency, accuracy, and accessibility of patient communication and clinical documentation.

Limitations and Hallucinations

LLMs offer significant potential for patient care, yet several critical limitations persist. These include design issues such as insufficient medical optimization, lack of transparency in training data, and restricted data accessibility. Furthermore, LLMs can produce non-reproducible, incomplete, or factually inaccurate outputs, raising safety and bias concerns, particularly in complex or ambiguous scenarios where errors may have serious consequences. One of the most prominent challenges is the phenomenon of hallucinations, where LLMs generate outputs that appear evidence-based but are in fact inaccurate or entirely fabricated. Such hallucinations can result in misdiagnosis or the presentation of false information as fact (23).

These risks underscore the need for ongoing human oversight, transparent data sources, and continuous safety evaluation before integrating LLMs into clinical care. To mitigate issues like hallucinations and bias, users must employ strong prompt engineering skills, emphasizing specificity and clarity in their inputs. Without narrowly defined prompts, LLMs are more likely to hallucinate or misapply valid information, while biases towards specific populations may persist. Thus, ensuring prompt precision and maintaining a critical approach are vital for the reliable clinical use of LLMs (20).

Regulation and Data Protection in the Use of Generative AI in Healthcare

In August 2024, the EU introduced the AI Act, the first binding regulatory framework for AI, categorizing systems by risk level. Clinical decision-making systems are deemed high-risk, requiring strict compliance with accuracy, bias mitigation, oversight, and transparency rules. General-purpose models like ChatGPT must meet transparency and copyright obligations (25).

Healthcare-oriented generative AI tools are generally not high-risk unless they influence clinical decisions, triggering stricter regulations (25). The Act's implementation timeline requires prohibited systems phased out by February 2025, transparency compliance by August 2025, high-risk systems by August 2026, and full compliance under existing product legislation by August 2027 (25).

In Denmark, The Danish Data Protection Agency emphasizes strict legality in handling health data, requiring clear consent or legal basis and proportional processing aligned with AI’s purpose (26).

The European Data Protection Board addresses issues around AI-generated data anonymity, legitimate interest bases, and legality concerning previously unlawfully processed data (27). Additionally, commercial platforms like GPT-4 raise privacy concerns due to external data handling; local deployments, though costly, offer enhanced security (28).

Worksheet 4: Search Protocol

In the early stages of the project, an initial exploratory search was conducted in Primo (the search database of Aalborg University Library), Google, and Google Scholar to define the research problem and formulate a relevant research question aimed at identifying gaps in the existing knowledge (see Table 1). These three platforms were selected because they enable broad searches while also retrieving relevant sources from academic databases, journals, and books. Additionally, chain searching and citation tracking were employed to identify related articles.

Table 1: Topic Description

Title:	The Role of ChatGPT in Enhancing Patient Record Comprehension Across Varying Levels of Health Literacy While Ensuring Clinical Accuracy
Initiating research question:	How can **ChatGPT** support the **understanding** of **patient records** by people with varying levels of **health literacy** without compromising clinical precision/accuracy?

Subsequently, a structured literature search was performed in PubMed and Embase to establish the state of the art based on the formulated research question. The search was designed using a three-block approach, with each block reflecting a key element of the research problem: health understanding (Block A), generative AI (Block B), and medical records (Block C). To cover these themes comprehensively, PubMed was included for its extensive access to biomedical literature, including studies on AI in healthcare (29). Embase, produced by Elsevier, includes additional journals and offers wider coverage of AI-related healthcare studies, thereby complementing PubMed and ensuring a more comprehensive search (30,31). A description of the databases can be found in Table 2, while the overarching headings for each search block are presented in Table 3.

Table 2: Description of Search Databases

Database	Reasons for choosing the database
PubMed	PubMed, maintained by the National Library of Medicine, is a free resource that comprises over 37 million citations for biomedical literature from MEDLINE, life science journals, and online books (29). PubMed's extensive indexing of biomedical literature,

	including studies on AI applications in healthcare, makes it indispensable for capturing core medical research relevant to AI-driven documentation.
Embase	Embase is a comprehensive biomedical and pharmacological database produced by Elsevier (30), containing over 32 million records from more than 8.400 currently published journals dating from 1947 to the present (31). Embase's additional coverage of AI applications in healthcare ensures a more comprehensive dataset, capturing studies that may not have been indexed in PubMed thereby supporting the width of the systematic literature search.

Table 3: Search blocks

Block A	**Block B**	**Block C**
People with varying levels of health literacy	Generative AI	Patient records

PubMed Search

To ensure comprehensive retrieval of relevant literature in PubMed, the search included both Medical Subject Headings (MeSH) and free-text terms in the title and abstract fields (32). MeSH terms were included to capture studies indexed under standardized subject headings, ensuring that relevant articles were retrieved even if different terminology was used in their titles or abstracts. Title/Abstract searches were used to identify articles that discussed the topic but were not necessarily indexed under the corresponding MeSH terms. Truncation (e.g., *) was applied to capture variations of a word (e.g., "comprehen*" to include "comprehension" and "comprehending").

The three blocks were combined using the boolean operator AND, ensuring that retrieved articles addressed all three conceptual areas. Within each block, terms were combined using OR, allowing for broad coverage of relevant studies (see Table 4). For an overview of the number of hits generated by each block individually, as well as all possible combinations of the three blocks in the PubMed search, see Tables 5 and 6.

Table 4: PudMed Search Strings for Each Block

Database: PubMed				
Search conducted: 05.03.2025				
Block A		**Block B**		**Block C**
"Health Literacy"[MeSH Terms] OR "Health Literacy"[Title/Abstract] OR "eHealth literacy"[Title/Abstract] OR "Comprehension"[MeSH Terms] OR "comprehen*"[Title/Abstract] OR "understand*"[Title/Abstract] OR "understood"[Title/Abstract] OR "interpret*"[Title/Abstract] OR "perception"[Title/Abstract] OR "perceive"[Title/Abstract]	**A** **N** **D**	"Natural Language Processing"[MeSH Terms] OR "natural language process*"[Title/Abstract] OR "large language model"[Title/Abstract] OR "LLM"[Title/Abstract] OR "Generative Artificial Intelligence"[MeSH Terms] OR "Generative Artificial Intelligence"[Title/Abstract] OR "generative AI"[Title/Abstract] OR “generative model"[Title/Abstract] OR	**A** **N** **D**	"Medical Records"[MeSH Terms:noexp] OR "Electronic Health Records"[MeSH Terms] OR "health records, personal"[MeSH Terms] OR "medical record*"[Title/Abstract] OR "electronic health record*"[Title/Abstract] OR "EHR"[Title/Abstract] OR "patient record*"[Title/Abstract] OR "health record*"[Title/Abstract] OR "clinical record*"[Title/Abstract] OR

OR "insight*"[Title/Abstract] OR "apprehen*"[Title/Abstract] OR "readab*"[Title/Abstract]		"Generative Pretrained Transformer"[Title/Abstract] OR "GPT"[Title/Abstract] OR "ChatGPT"[Title/Abstract] OR “Chat-GPT"[Title/Abstract] OR “Chat GPT”[Title/Abstract] OR "OpenAI"[Title/Abstract] OR "chatbot*"[Title/Abstract] OR “Mistral”[Title/Abstract] OR “LLaMA”[Title/Abstract] OR “Grok”[Title/Abstract] OR “Gemini”[Title/Abstract] OR “Claude”[Title/Abstract] OR “Copilot”[Title/Abstract]		"patient documentation"[Title/Abstract] OR "medical documentation"[Title/Abstract] OR "digital health record*"[Title/Abstract] OR "computerized medical record*"[Title/Abstract] OR "electronic patient record*"[Title/Abstract] OR "patient file*"[Title/Abstract] OR "healthcare record*"[Title/Abstract] OR "clinical note*"[Title/Abstract] OR "soap note*"[Title/Abstract] OR "open note*"[Title/Abstract] OR "patient health information"[Title/Abstract] OR

				"medical data"[Title/Abstract] OR “patient discharge summar*”[Title/Abstract]

Table 5: PubMed Number of Hits for Each Block

Block	**Number of hits**
A	4,002,096
B	36,775
C	318,215

Table 6: PubMed Number of Hits for All Possible Block Combinations

Block combination	**Number of hits**
A AND **B**	9,623
A AND **C**	38,492
B AND **C**	4,546
A AND **B** AND **C**	**1,163**

Embase Search

In the Embase search, Emtree terms were utilized across all three blocks to enhance the precision of the search. Additionally, the search was expanded by incorporating keywords from free-text fields, specifically in the title, abstract, and keywords (denoted as "ti, ab, kw"), to capture relevant studies not indexed under controlled vocabulary. Truncation and Boolean operators were also used during the Embase search (see Table 7). To see the number of hits for each block as well as for each possible combination of blocks, see Tables 8 and 9.

Table 7: Embase Search Strings for Each Block

Database: Embase				
Search conducted: 05.03.2025				
Block A	**A**	**Block B**	**A**	**Block C**
“health literacy”/exp OR	**N** **D**	"natural language processing"/exp	**N** **D**	“medical record”/exp OR

“comprehension”/exp OR “health literacy":ti,ab,kw OR "eHealth literacy":ti,ab,kw OR "comprehen*":ti,ab,kw OR "understand*":ti,ab,kw OR "understood":ti,ab,kw OR "interpret*":ti,ab,kw OR "perception":ti,ab,kw OR "perceive":ti,ab,kw OR "insight*":ti,ab,kw OR "apprehen*":ti,ab,kw OR "readab*":ti,ab,kw		OR "generative artificial intelligence"/exp OR "natural language process*":ti,ab,kw OR "large language model":ti,ab,kw OR "LLM":ti,ab,kw OR "generative artificial intelligence":ti,ab,kw OR "generative AI":ti,ab,kw OR "generative A.I.":ti,ab,kw OR "genAI":ti,ab,kw OR “gen-AI":ti,ab,kw OR "generative model":ti,ab,kw OR "generative pretrained transformer":ti,ab,kw OR "GPT":ti,ab,kw OR "chatbot*":ti,ab,kw OR		"medical record*":ti,ab,kw OR "electronic health record*":ti,ab,kw OR "EHR":ti,ab,kw OR "patient record*":ti,ab,kw OR “personal health record*”:ti,ab,kw OR "health record*":ti,ab,kw OR "clinical record*":ti,ab,kw OR “hospital record*”:ti,ab,kw OR "patient documentation":ti,ab,kw OR "medical documentation":ti,ab,kw OR "digital health record*":ti,ab,kw OR "computerized medicalrecord*":ti,ab,kw OR "electronic patient record*":ti,ab,kw OR

		"ChatGPT":ti,ab,kw OR "Chat-GPT":ti,ab,kw OR "Chat GPT":ti,ab,kw OR "OpenAI":ti,ab,kw OR "Mistral":ti,ab,kw OR "LLaMA":ti,ab,kw OR "Grok":ti,ab,kw OR "Gemini":ti,ab,kw OR "Claude":ti,ab,kw OR "Copilot":ti,ab,kw		"patient file*":ti,ab,kw OR "healthcare record*":ti,ab,kw OR "clinical note*":ti,ab,kw OR "soap note*":ti,ab,kw OR "open note*":ti,ab,kw OR "patient health information":ti,ab,kw OR "medical data":ti,ab,kw OR “patient discharge summar*”:ti,ab,kw

Table 8: Embase Number of Hits for Each Block

Block	Number of hits
A	4,878,990
B	47,834
C	712,909

Table 9: Embase Number of Hits for All Possible Block Combinations

Block combination	Number of hits
A AND **B**	11,745
A AND **C**	90,306
B AND **C**	6,950
A AND **B** AND **C**	**1,742**

Selection of studies

Language and publication year filters were applied directly within the databases prior to export, in accordance with the study's inclusion and exclusion criteria (Table 10). Rayyan was then used as a tool for screening the titles and abstracts of the remaining articles. Each researcher independently screened the articles, after which the inclusion and exclusion decisions were compared via Rayyan to ensure consistency. Following the initial screening, Rayyan was further employed to select studies after a full-text reading (33).

In addition to the database searches, a backward citation search of the included studies was conducted to identify additional relevant articles. Three records were identified through this process. All were assessed in full text, with two included and one excluded due to its focus on imaging-based AI rather than text generation.

A total of 10 studies were ultimately included in the review. Eight of these were identified through systematic database searches, and two were identified through citation searching (snowballing) of the included studies. An illustration of the selection process is provided in the PRISMA Flowchart (Figure 1).

Table 10: Inclusion and Exclusion Criteria

Inclusion criteria	
Criterion	**Justification for inclusion**
Date of publication: 2022-2025	Only studies published from 2022 onwards were included, as ChatGPT was released in November 2022. This ensures that findings are directly relevant and comparable to current models.
Language: Danish, English, Norwegian, Swedish	To ensure both accurate interpretation and accessibility for the research team. Limiting to these languages avoids misinterpretation due to translation and supports analytical clarity.
Adults (18+)	The focus was on adult patients, as medical communication needs and record structure differ significantly from those of pediatric populations.
Technology focus:	These criteria ensure that the included studies are relevant to the scope of this study, which

• Text-generative AI (e.g., large language models) • Applied to medical records or electronic health records • Used for summarization, rephrasing, or translation	examines how generative AI supports patient understanding of medical records.
Exclusion criteria	
Exclusion criteria	**Justification for exclusion**
Mental health disorders	Mental health documentation differs significantly in structure and terminology, often using fewer specialized clinical terms. Therefore, such studies were excluded for consistency.
Generative AI limited to non-text modalities	Studies focusing solely on generative AI models that produce non-text outputs (such as images, video, or audio) were excluded, as this study focused specifically on natural language generation and text-based applications.
Publication types: • Editorials • Conference abstracts • Preprints	To maintain scientific rigor and reliability, non-peer-reviewed and non-research-based publications were excluded.

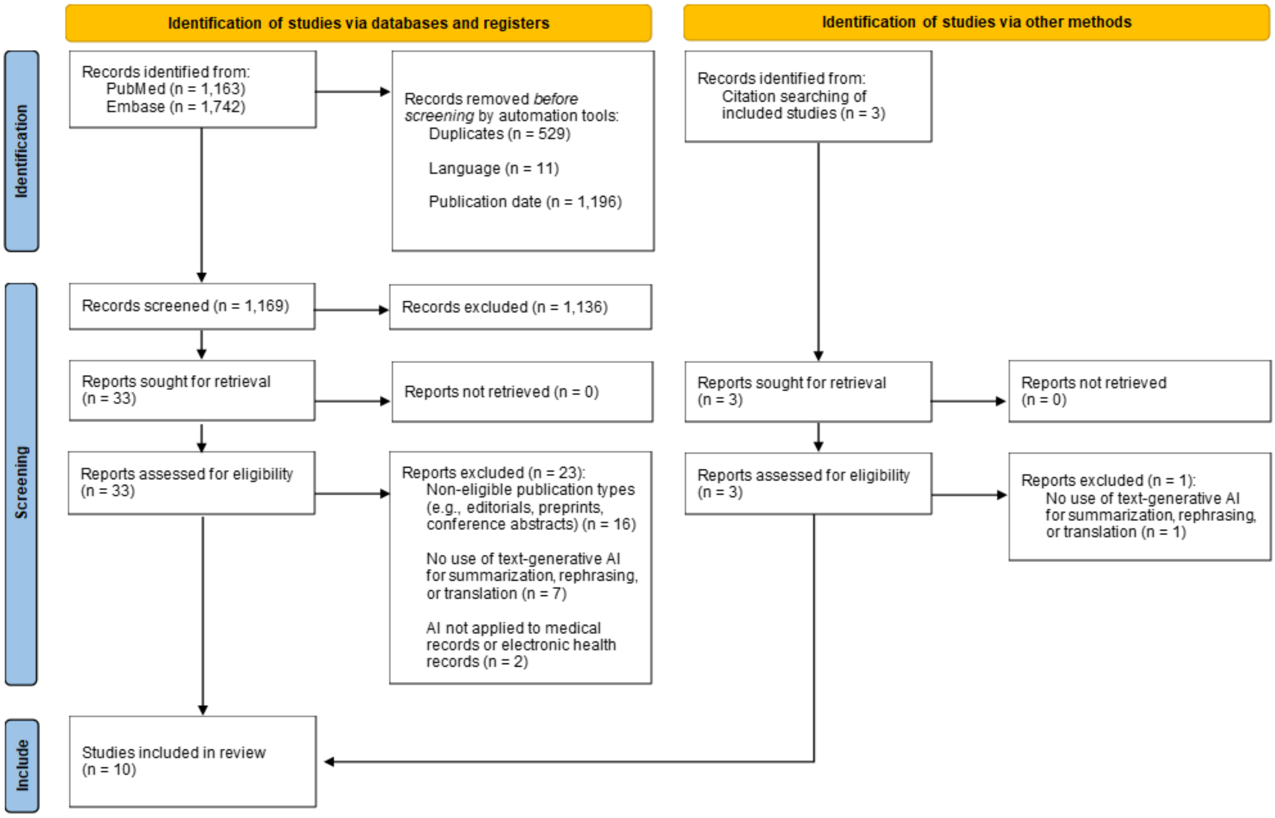

Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Include
Records identified from:
PubMed (n = 1,163)
Embase (n = 1,742)
Records removed *before screening* by automation tools:
Duplicates (n = 529)
Language (n = 11)
Publication date (n = 1,196)
Records identified from:
Citation searching of included studies (n = 3)
Records screened (n = 1,169)
Records excluded (n = 1,136)
Reports sought for retrieval (n = 33)
Reports not retrieved (n = 0)
Reports sought for retrieval (n = 3)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 33)
Reports excluded (n = 23):
Non-eligible publication types (e.g., editorials, preprints, conference abstracts) (n = 16)
No use of text-generative AI for summarization, rephrasing, or translation (n = 7)
AI not applied to medical records or electronic health records (n = 2)
Reports assessed for eligibility (n = 3)
Reports excluded (n = 1):
No use of text-generative AI for summarization, rephrasing, or translation (n = 1)
Studies included in review (n = 10)

Figure 1: PRISMA Flow-chart

References

1. Office of Disease Prevention and Health Promotion. U.S. Department of Health and Human Services. [cited 2025 May 30]. Health Literacy in Healthy People 2030. Available from: https://odphp.health.gov/healthypeople/priority-areas/health-literacy-healthy-people-2030

2. Nutbeam D, Muscat DM. Health Promotion Glossary 2021. Health Promot Int [Internet]. 2021 Dec 1;36(6):1578–98. Available from: https://doi.org/10.1093/heapro/daaa157

3. Sætre LMS, Jarbøl DE, Raasthøj IP, Seldorf SA, Rasmussen S, Balasubramaniam K. Examining health literacy in the Danish general population: a cross-sectional study on the associations between individual factors and healthcare-seeking behaviour. Eur J Public Health [Internet]. 2024 Dec 1;34(6):1125–33. Available from: https://doi.org/10.1093/eurpub/ckae150

4. Moreira L. Health literacy for people-centred care: Where do OECD countries stand? [Internet]. Paris; 2018 Dec [cited 2025 May 30]. Available from: https://doi.org/10.1787/d8494d3a-en

5. Sørensen K, Pelikan JM, Röthlin F, Ganahl K, Slonska Z, Doyle G, et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). The European Journal of Public Health [Internet]. 2015 Dec [cited 2025 May 30];25(6):1053–8. Available from: https://doi.org/10.1093/eurpub/ckv043

6. Svendsen IW, Damgaard MB, Bak CK, Bøggild H, Torp-Pedersen C, Svendsen MT, et al. Employment Status and Health Literacy in Denmark: A Population-Based Study. Int J Public Health [Internet]. 2021;Volume 66-2021. Available from: https://www.ssph-journal.org/journals/international-journal-of-public-health/articles/10.3389/ijph.2021.598083

7. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. BMC Health Serv Res [Internet]. 2022;22(1):1148. Available from: https://doi.org/10.1186/s12913-022-08527-9

8. Baker DW, Gazmararian JA, Williams M V, Scott T, Parker RM, Green D, et al. Health literacy and use of outpatient physician services by medicare managed care enrollees. J Gen Intern Med [Internet]. 2004;19(3):215–20. Available from: https://doi.org/10.1111/j.1525-1497.2004.21130.x

9. Davis TC, Wolf MS, Bass PF, Thompson JA, Tilson HH, Neuberger M, et al. Literacy and Misunderstanding Prescription Drug Labels. Ann Intern Med [Internet]. 2006 Dec 19;145(12):887–94. Available from: https://doi.org/10.7326/0003-4819-145-12-200612190-00144

10. Baker DW, Wolf MS, Feinglass J, Thompson JA. Health Literacy, Cognitive Abilities, and Mortality Among Elderly Persons. J Gen Intern Med [Internet]. 2008 Mar 11;23(6):723–6. Available from: https://doi.org/10.1007/s11606-008-0566-4

11. Chew LD, Bradley KA, Flum DR, Cornia PB, Koepsell TD. The impact of low health literacy on surgical practice. The American Journal of Surgery [Internet]. 2004;188(3):250–3. Available from: https://www.sciencedirect.com/science/article/pii/S0002961004002132

12. Koh HK, Berwick DM, Clancy CM, Baur C, Brach C, Harris LM, et al. New Federal Policy Initiatives To Boost Health Literacy Can Help The Nation Move Beyond The Cycle Of Costly 'Crisis Care.' Health Aff [Internet]. 2012 Feb 1;31(2):434–43. Available from: https://doi.org/10.1377/hlthaff.2011.1169

13. Styrelsen for Patientsikkerhed. Retsinformation. 2024 [cited 2025 May 30]. Bekendtgørelse om autoriserede sundhedspersoners patientjournaler (journalføring, opbevaring, videregivelse, overdragelse m.v.). Available from: https://www.retsinformation.dk/eli/lta/2024/713

14. sundhed.dk. sundhed.dk. 2024 [cited 2025 May 31]. Om sundhed.dk. Available from: https://www.sundhed.dk/borger/service/om-sundheddk/om-organisationen/hvem-er-sundheddk/historien-om-sundheddk/

15. Nøhr C, Parv L, Kink P, Cummings E, Almond H, Nørgaard JR, et al. Nationwide citizen access to their health data: analysing and comparing experiences in Denmark, Estonia and Australia. BMC Health Serv Res [Internet]. 2017 Aug 7 [cited 2025 May 31];17(1):534. Available from: https://doi.org/10.1186/s12913-017-2482-y

16. Ministry of Health, Ministry of Finance, Danish Regions, KL. A Coherent and Trustworthy Health Network for All: Digital Health Strategy 2018-2022 [Internet]. København S; 2018 Jan [cited 2025 May 31]. Available from: https://sundhedsdatastyrelsen.dk/om-os/strategi-og-grundlag/strategi-for-digital-sundhed/strategi-for-digital-sundhed-2018-2024

17. Turner A, Morris R, McDonagh L, Hamilton F, Blake S, Farr M, et al. Unintended consequences of patient online access to health records: a qualitative study in UK primary care. British Journal of General Practice [Internet]. 2023 Jan 1;73(726):e67. Available from: http://bjgp.org/content/73/726/e67.abstract

18. Fernández L, Fossa A, Dong Z, Delbanco T, Elmore J, Fitzgerald P, et al. Words Matter: What Do Patients Find Judgmental or Offensive in Outpatient Notes? J Gen Intern Med [Internet]. 2021 Feb 2;36(9):2571–8. Available from: https://doi.org/10.1007/s11606-020-06432-7

19. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. Business & Information Systems Engineering [Internet]. 2023 Sep 12;66(1):111–26. Available from: https://doi.org/10.1007/s12599-023-00834-7

20. Shah K, Xu AY, Sharma Y, Daher M, McDonald C, Diebo BG, et al. Large Language Model Prompting Techniques for Advancement in Clinical Medicine. J Clin Med [Internet]. 2024 Aug 28 [cited 2025 May 31];13(17):5101. Available from: https://doi.org/10.3390/jcm13175101

21. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J [Internet]. 2023 Sep 30 [cited 2025 May 31];41(3):209–16. Available from: https://doi.org/10.3857/roj.2023.00633

22. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. 2023 Mar 15 [cited 2025 May 31];1–100. Available from: https://doi.org/10.48550/arXiv.2303.08774

23. Busch F, Hoffmann L, Rueger C, van Dijk EHC, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Communications Medicine [Internet]. 2025 Jan 21;5(1):26. Available from: https://doi.org/10.1038/s43856-024-00717-2

24. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res [Internet]. 2023 Jun 28 [cited 2025 May 31];25:e48568. Available from: https://doi.org/10.2196/48568

25. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council [Internet]. Luxembourg; 2024 Jul [cited 2025 May 31]. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

26. Datatilsynet. Offentlige myndigheders brug af kunstig intelligens - Inden I går i gang [Internet]. Valby; 2023 Oct [cited 2025 May 31]. Available from: https://www.datatilsynet.dk/presse-og-nyheder/nyhedsarkiv/2023/okt/ny-vejledning-om-offentlige-myndigheders-brug-af-ai-og-kortlaegning-af-ai-paa-tvaers-af-den-offentlige-sektor

27. European Data Protection Board. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models [Internet]. 2024 Dec [cited 2025 May 31]. Available from: https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en

28. Zhang J, Sun K, Jagadeesh A, Falakaflaki P, Kayayan E, Tao G, et al. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. Journal of the American Medical Informatics Association [Internet]. 2024 Sep 1 [cited 2025 May 31];31(9):1884–91. Available from: https://doi.org/10.1093/jamia/ocae184

29. National Library of Medicine. National Library of Medicine. 2025 [cited 2025 May 31]. PubMed Overview. Available from: https://pubmed.ncbi.nlm.nih.gov/about/

30. Elsevier. Elsevier. 2025 [cited 2025 May 31]. Embase is the medical research database for high-quality, comprehensive evidence. Available from: https://www.elsevier.com/products/embase

31. Elsevier. Elsevier. 2025 [cited 2025 May 31]. Embase content is updated daily and expanding globally. Available from: https://www.elsevier.com/products/embase/content

32. DeMars MM, Perruso C. MeSH and text-word search strategies: precision, recall, and their implications for library instruction. Journal of the Medical Library Association. 2022 Jan 1;110(1).

33. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev [Internet]. 2016 Dec 5 [cited 2025 May 31];5(1):210. Available from: https://doi.org/10.1186/s13643-016-0384-4.