Article:
Alabed, S., Anderson, A., Maiter, A. et al. (2026) Large language models for simplifying radiology reports: a systematic review and meta-analysis of patient, public, and clinician evaluations. The Lancet Digital Health. 100960. ISSN: 2589-7500

https://doi.org/10.1016/j.landig.2025.100960

Large language models for simplifying radiology reports: a systematic review and meta-analysis of patient, public, and clinician evaluations

Samer Alabed, Abigail Anderson, Ahmed Maiter, Anthony Hughes, Niamh McAnenly, Mahan Salehi, Michael Sharkey, Krit Dwivedi, Alireza Hokmabadi, Fares Alahdab, Mark Stevenson, Ning Ma, Robert Gaizauskas, Tim J Chico, Andy J Swift, Junyi Jessy Li, Jens Kleesiek, Curtis Langlotz

Summary

Background Radiology reports are typically written in language that is difficult for patients to understand. Large language models (LLMs) excel at simplifying text. We aimed to evaluate the ability of LLMs to improve the understanding of radiology reports.

Methods In this systematic review and meta-analysis, we searched CENTRAL, MEDLINE, and Embase from inception to Nov 11, 2025, without restrictions on language. Full-text articles and preprints were considered for inclusion. Eligible studies applied LLMs to simplify radiology reports and had these reports assessed by members of the public or medical professionals. We excluded studies that focused solely on dialogues with interactive chatbots, preimaging leaflets, educational materials, appointment letters, or summarising findings without simplifying them for patients. Search results were screened independently by two authors and full-text review and data extraction were done by three authors; disagreements were resolved by consensus. The main outcomes were patient, public, and clinician evaluations (Likert scores) and text readability metrics. We assessed study quality with the MAIC-10 tool. This study was registered with PROSPERO (CRD420251027489).

Findings We identified 2385 records, of which 38 studies were eligible. These 38 studies generated 12 922 simplified reports, assessed by 508 evaluators (387 lay people and 121 clinicians). 35 (92%) of 38 studies used OpenAI GPT models and 29 (76%) produced simplified reports in English. Patients perceived LLM-rewritten reports as significantly more understandable than radiologist reports (mean Likert score 4·04 [SD 1·20] for simplified reports *vs* 2·16 [SD 0·94] for original reports; mean difference 2·00 [95% CI 1·54–2·46]). Clinicians rated LLM-rewritten reports highly for accuracy (mean 4·45 [95% CI 4·27–4·63]; 27 studies) and completeness (mean 4·53 [95% CI 4·30–4·76]; 14 studies). Readability was improved across imaging modalities, with lower Flesch–Kincaid Grade Level for LLM-rewritten reports, including a mean difference of −6·20 (95% CI −6·91 to −5·48) for CT, −5·07 (−5·99 to −4·15) for x-ray, and −5·0 (−6·0 to −4·0) for MRI. The error rate in LLM-rewritten reports was 7·2% (95% CI 5·1%–10·0%; 13 studies) and 0·9% (95% CI 0·6–1·5%; 2 studies) for clinically significant errors.

Interpretation LLM-simplified radiology reports improved patient-perceived understanding and readability and were rated by clinicians as largely accurate and complete, although a small proportion contained clinically significant errors. LLM-based simplification shows promise for making radiology communication more patient-centred, but further evaluation of its effect on patient outcomes and clinical workflows is required.

Funding National Institute for Health and Care Research Sheffield Biomedical Research Centre.

Introduction

When a patient has medical imaging (eg, x-rays, CT, or MRI scans), the images are reviewed by a radiologist who provides a written radiology report. These reports are intended for the requesting clinician and are written in highly technical medical and anatomical detail. Patients are now increasingly accessing and reading such reports.[1,2] This access can promote agency, support shared decision making, and reinforce autonomy, reflecting a broader shift towards patient-centred health care, accelerated by legislation and policies. In the USA, the 21st Century Cures Act mandates immediate release of medical records to patients, whereas in the UK and EU, patients have a legal right to access their records under the General Data Protection Regulation. The digital transformation of health-care systems has removed barriers to accessing medical data. For example, patients in the UK can read their radiology reports and other test results on the NHS App.[3,4]

However, because radiology reports are produced by clinicians for clinicians, the technical terminology often poses problems when patients read them. Medical jargon and unfamiliarity with reporting styles can cause

Lancet Digit Health 2026

Published Online https://doi.org/10.1016/j.landig.2025.100960

See Comment https://doi.org/10.1016/j.landig.2026.100979

School of Medicine and Population Health, Institute for In Silico Medicine, National Institute for Health and Care Research, University of Sheffield, Sheffield, UK (S Alabed PhD, A Anderson, A Maiter FRCR, N McAnenly, M Salehi MSc, M Sharkey MSc, K Dwivedi PhD, A Hokmabadi PhD, Prof T J Chico PhD, Prof A J Swift PhD); **Department of Clinical Radiology, Sheffield Teaching Hospitals, Sheffield, UK** (S Alabed PhD, A Maiter FRCR, M Salehi MSc, K Dwivedi PhD, Prof A J Swift PhD); **School of Computer Science, University of Sheffield, Sheffield, UK** (A Hughes MSc, M Stevenson PhD, N Ma PhD, Prof R Gaizauskas PhD); **Department of Biomedical Informatics, Biostatistics, and Medical Epidemiology and Department of Cardiology, University of Missouri, Columbia, MO, USA** (F Alahdab MD); **Department of Linguistics, University of Texas at Austin, Austin, TX, USA** (J J Li PhD); **Institute for AI in Medicine, University Medicine Essen, Essen, Germany** (Prof J Kleesiek PhD); **Department of Radiology, Department of Medicine, and Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA** (Prof C Langlotz PhD)

Correspondence to: Samer Alabed, School of Medicine and Population Health, Institute for In Silico Medicine, National Institute for Health and

Care Research, University of Sheffield, Sheffield S10 2RX, UK
s.alabed@nhs.net

See Online for appendix

Research in context

Evidence before this study

Patients' access to radiology reports has expanded rapidly, driven by patient-centred care initiatives and policies mandating transparency of medical records. However, most radiology reports are written well above average literacy levels and remain difficult for patients to understand. In parallel, artificial intelligence tools such as large language models (LLMs), trained on vast amounts of text, offer the potential to improve report readability. Despite this promise, evidence from patients, the public, and clinicians is required to determine whether LLM-generated simplifications are comprehensible and accurate. We searched MEDLINE, Embase, CENTRAL, medRxiv, bioRxiv, and arXiv from inception to Nov 11, 2025, with terms related to LLMs, radiology, patient preferences, and reports (appendix p 1). Reference lists were also screened. No restrictions on language or publication type were applied. Eligible studies applied automated methods, including LLMs, to simplify radiology reports and used patient, public, or professional assessors to evaluate comprehension, accuracy, or usability. 38 studies, representing 12 922 simplified reports assessed by 508 raters, were identified, most with small, single-centre samples and variable methods. The meta-analyses suggested that LLM-simplified reports were accurate and complete and improved patient-perceived understanding and readability metrics; however, heterogeneity between studies was high.

Added value of this study

This is the first systematic review and meta-analysis to synthesise evaluations of LLM-rewritten radiology reports by both patients and clinicians. Pooled analyses showed that simplified reports were rated substantially more understandable by patients and lay assessors, and clinicians judged them accurate and complete. We also identified limitations, including occasional clinically significant errors, report lengthening, and uncertainties around trust, governance, and workflow integration.

Implications of all the available evidence

The combined evidence indicates that LLMs can make radiology reports more accessible to patients without compromising clinical accuracy. However, safe deployment requires human oversight, co-design with patients, and careful integration into workflows. Future research should test LLM-assisted reporting in real-world clinical settings, establish standard evaluation metrics, and explore adaptive formats that balance clarity with brevity. If developed responsibly, LLM-generated simplifications could become a mainstay of patient-centred communication in radiology.

confusion,[5,6] anxiety,[7,8] and reduced satisfaction with care.[9] Patients misunderstanding radiology reports can prompt further appointments, investigations, or hospital admissions and therefore can be harmful for both patients and health-care systems.[10,11] Furthermore, quality of care could be diminished, as clinician time and attention during an appointment might be diverted away from addressing pertinent medical issues towards explaining minor incidental findings of no relevance.[12] Enabling access to imaging reports could also perpetuate health-care inequalities. In the UK, 40% of adults struggle with health content, 2% have poor English proficiency, and the average reading age is approximately 9 years.[13,14] Therefore, patients with poor literacy or disability might not benefit equally from access to their medical records.

Radiologists and other reporters might have additional concerns. The knowledge that reports will be read by patients could adversely affect the accuracy and quality of reporting. The Royal College of Radiologists notes that radiology departments do not have the capacity to deal with large volumes of queries from patients[15] and a radiologist might choose to omit findings to reduce queries, even if such findings could be relevant in the future.[6] Alternatively, a radiologist might be concerned about the emotional effect of their report on a patient and attempt to lessen their description of findings in a way that impairs accurate clinical communication, such as implying cancer by recommending referral to a multidisciplinary team meeting rather than explicitly stating it.

Since patients can access radiology reports, there is a clear need for versions comprehensible to the patient while remaining accurate. In most settings, workload pressure makes it impractical for radiologists to manually produce patient-friendly versions alongside usual clinical reports; rather, automated solutions to this problem are required. Large language models (LLMs) are artificial intelligence (AI) systems that generate human-like text from vast pre-training datasets. In radiology, LLMs are increasingly explored to draft or simplify reports to improve readability in a cost-effective manner.[16–23] Although more than 1000 AI-enabled radiology tools have US Food and Drug Administration approval, none include generative LLMs for clinical reporting.[24,25] LLMs are prone to errors which, although improving, remain a crucial barrier to safe use in high-stakes clinical contexts. These limitations necessitate ongoing clinician validation and raise questions about patient trust and acceptance.[26–29]

The aim of this study is to evaluate how patients, the public, and medical professionals rate the quality of LLM-rewritten radiology reports. The findings aim to guide the development of patient-centred reporting tools that enhance communication, support equitable health care, and align with the needs and expectations of patients as the ultimate end users of medical imaging services.

Methods

Search strategy and selection criteria

In this systematic review and meta-analysis, we searched MEDLINE, Embase, and CENTRAL from database inception to Nov 11, 2025, with the key terms (and their variations) LLM, radiology, patient preferences, and report for full-length articles. The detailed search strategy is provided in the appendix (p 1). In addition, preprint and grey literature articles were searched in medRxiv, bioRxiv, and arXiv with broad terms for radiology and imaging reports (appendix p 1). Reference lists of included studies were also screened. No restrictions on language were applied. Full-text articles and preprints were considered for inclusion.

Studies were eligible for inclusion if they applied automated methods (including LLMs) to rewrite radiology reports (from any imaging modality) to improve patient understanding and used patient, public, or medical assessors to provide qualitative or quantitative assessments of rewritten reports, either with or without subjecting the original medical report to the same assessment. Studies were excluded if they focused solely on dialogues with interactive chatbots, preimaging leaflets, educational materials, appointment letters, or summarising findings without simplifying them for patients. All stages of screening, full-text assessment, data extraction, and quality assessment were done independently, in duplicate, by at least two authors. Search results were screened by two authors (SA and AA) by title, abstract, and keywords. Full-text review was predefined and done by three authors (SA, AA, and NM). Rayyan was used for screening and assessment. Disagreements over inclusion or data extraction were resolved by consensus.

This systematic review and meta-analysis adhered to the PRISMA guidelines (appendix pp 20–21)[30] and was prospectively registered with the PROSPERO International Prospective Register of Systematic Reviews (CRD420251027489).

Data analysis

Data extraction was predefined and done independently by three authors (SA, AA, and NM). Duplicate records were checked through Rayyan and Ovid.

Primary outcomes were patient or lay person self-reported understanding of LLM-rewritten radiology reports and clinician-assessed quality of LLM-rewritten reports, both measured with Likert scale scores and summarised as mean differences with 95% CIs. Secondary outcomes were objective readability metrics (eg, Flesch Reading Ease Score [FRES], Flesch–Kincaid Grade Level [FKGL], Automated Readability Index [ARI]), word counts, error rates of LLM-rewritten reports, and comparison of LLM prompting methods. We did sensitivity analyses to evaluate robustness of findings across LLM types (GPT *vs* other LLMs), LLM version (GPT-4 *vs* GPT-3.5), and clinician assessor type (radiologist *vs* non-radiologist).

Extracted information was study characteristics (ie, authors, journal, year of publication, publication type, funding and conflicts of interests, country of origin and language or reports, number of reports, study design, and population details such as demographics, medical history, education level, and language of participants), imaging modality (eg, CT or MRI), body part imaged (eg head or chest), LLM details (eg, dataset, model, prompts, parameters, and hardware used), report assessment methods (eg, type of assessment and number and type of assessors), and outcome metrics (ie, Likert scores, readability scores, and word count). Methodological quality and risk of bias were assessed by two authors (SA and AA) with the MAIC-10 tool.[31]

All meta-analyses were done with R (version 4.4.3) and used a random effects model (metacont function in the meta package for R). Further data analysis details are provided in the appendix (p 2).

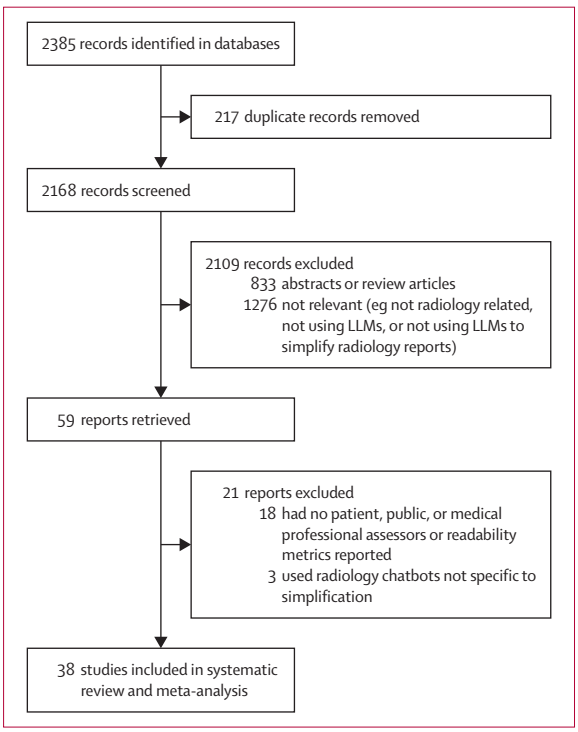


Figure 1: **Flow diagram of studies**
LLM=large language model.

For more details on **Rayyan** see www.rayyan.ai

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Our search identified 2385 records, of which 38 studies were included (figure 1). These studies were published between 2022 and 2025 and encompassed 12 922 simplified reports evaluated by 508 assessors, of which 387 were lay people (median 17 people [IQR 6–30]) and 121 were

	Language	Imaging modalities included	Radiology subspecialty	Number of reports assessed	Assessors	Large language models used
Amin et al (2023)[32]	English	X-ray, CT, MRI, and ultrasound	Mixed	150	2 medical professionals	GPT 4, Bard, and Copilot
Bai et al (2025)[33]	English	MRI	Mixed	6174	2 medical professionals	GPT-O1 and Deepseek-R1
Berigan et al (2024)[34]	English	X-ray, CT, MRI, ultrasound, and mammography	NR	27	22 lay people	Bard
Berzolla et al (2025)[35]	English	MRI	Musculoskeletal	32	32 lay people and 4 medical professionals	GPT 4
Bozer and Pekçevik (2025)[36]	English	CT and MRI	Mixed	100	2 medical professionals	GPT 3.5, Gemini 2.5 Flash, and Copilot
Butler et al (2025)[37]	English	X-ray, CT, and MRI	Hand musculoskeletal	300	2 medical professionals	GPT 3.5
Butler et al (2024a)[38]	English	X-ray, CT, and MRI	Knee musculoskeletal	300	2 medical professionals	GPT 3.5
Butler et al (2024b)[39]	English	X-ray, CT, and MRI	Foot musculoskeletal	300	2 medical professionals	GPT 3.5
Çamur et al (2024)[40]	Turkish	CT	Mixed	50	3 medical professionals	GPT 4, GPT 3.5, Claude Opus 3, and Gemini 1.5
Can et al (2025)[41]	German	Interventional radiology	Interventional radiology	109	3 medical professionals	GPT 4, GPT 3.5, Claude Opus 3, Gemini 1.5, Mistral 7b, and Mixtral 8 7b
Cesur and Çamur (2024)[42]	Turkish	MRI	Neuroradiology, musculoskeletal, and gastrointestinal	50	3 medical professionals	GPT 4, Claude Opus 3, Gemini, and Perplexity
Chung et al (2023)[43]	English	MRI	Oncology	5	12 medical professionals	GPT 3.5
Doshi et al (2024)[44]	English	X-ray, CT, MRI, ultrasound, and mammography	Mixed	750	NA	GPT 4, GPT 3.5, Bard, and Copilot
Gupta et al (2025)[45]	English	CT	Oncology	50	100 lay people and 3 medical professionals	GPT 4o, Gemini, Claude Opus 3, Llama-3.1 8B, and Phi-3.5-mini
Güneş et al (2024)[46]	Turkish	Ultrasound	Mixed	50	3 medical professionals	GPT 4, Claude Opus, Gemini 1.5, and Perplexity
Jeblick et al (2024)[47]	English	CT and MRI	Neuroradiology, musculoskeletal, and oncology	3	15 medical professionals	GPT 3.5
Kuckelmann et al (2024)[48]	English	MRI	Musculoskeletal	60	3 medical professionals	GPT 4
Li et al (2023)[49]	English	X-ray, CT, MRI, and ultrasound	Mixed	400	NA	GPT 3
Li et al (2025)[50]	English, Spanish, Korean, Chinese, Swahili	CT, ultrasound, and fluoroscopy	Interventional radiology	200	26 lay people and 8 medical professionals	GPT 4
Lyu et al (2023)[51]	English	CT and MRI	Neuroradiology and chest	138	2 medical professionals	GPT 4
Maroncelli et al (2024)[52]	Italian	Mammography, ultrasound, and MRI	Breast	21	5 lay people and 5 medical professionals	GPT 4o
Park et al (2024)[53]	English	MRI	Spinal	685	2 lay people and 2 medical professionals	GPT 3.5
Pisarcik et al (2025)[54]	German	Mammography and ultrasound	Breast	27	40 lay people and 2 medical professionals	GPT 4, GPT 4o, and Gemini 1.5
Prucker et al (2025)[55]	English	CT	Chest	50	3 medical professionals	GPT 4, GPT 3.5, Claude Opus 3, Gemini 1.5, Llama 3 70B, Mistral 7b, and Mixtral 8 7b
Rogasch et al (2023)[56]	English	PET-CT	Oncology	5	3 medical professionals	GPT 4
Salam et al (2024)[57]	English	MRI	Cardiac	20	13 lay people and 2 medical professionals	GPT 4
Sarangi et al (2023)[58]	English, Hindi	CT and MRI	Mixed	9	8 medical professionals	GPT 3.5
Schmidt et al (2024)[59]	German	MRI	Knee musculoskeletal	3	20 lay people and 4 medical professionals	GPT 3.5
Stephan et al (2025)[60]	German	X-Ray	Dentistry	100	300 lay people	GPT 3
Sterling et al (2024)[61]	English	X-ray, CT, MRI, and ultrasound	Mixed	1982	8 medical professionals	GPT 3.5
Sudarshan et al (2024)[62,*]	English	CT, MRI, and ultrasound	Mixed	16	NA	GPT 4o
Sunshine et al (2025)[63]	English	CT and MRI	Neuroradiology	30	4 lay people and 4 medical professionals	GPT 4

(Table 1 continues on next page)

	Language	Imaging modalities included	Radiology subspecialty	Number of reports assessed	Assessors	Large language models used
(Continued from previous page)						
Tang et al (2024)[64]	English	CT	Neuroradiology	100	1 medical professionals	Claude 1.3
Tariq et al (2025)[65]	English	CT	Chest	23	3 lay people and 3 medical professionals	T5 fine-tuned
Tepe and Emekli (2024)[66]	English	CT and MRI	Mixed	30	2 medical professionals	GPT 4, Bard, and Copilot
Tripathi et al (2025)[67]	English	X-ray	Chest	500	3 medical professionals	GPT 4
van Driel et al (2025)[68]	Dutch	CT and MRI	Gastrointestinal	10	12 lay people and 2 medical professionals	GPT 4
Yang et al (2024)[69]	English	X-ray	NR	63	8 lay people and 1 medical professional	GPT 3.5

NA=not applicable. NR=not reported. *This paper is a preprint.

Table 1: **Study characteristics table of the included studies**

medical professionals (median 3 assessors [IQR 2–4]). Study characteristics are presented in table 1.

Across eight (21%) of 38 studies, 330 patients (median 29 patients [IQR 24–37]) were enrolled either at the time of ordering or attending imaging examinations.[34,35,45,54,59,60,63,68] Recruitment was random in one study,[60] consecutive in three studies,[35,45,59] and based on convenience sampling in four studies.[34,54,63,68] In addition, two studies[34,45] randomly assigned patients to receive either LLM-generated reports or standard reports.

Across six (16%) of 38 studies, 57 lay participants (median 22 participants [IQR 20–53]) were recruited through direct invitation in two studies[57,69] and via Amazon Mechanical Turk in one study,[50] whereas in three studies the recruitment approach was not specified.[52,53,65]

The demographics of the patient and public assessors were inconsistently reported (appendix p 3). Eight of 38 studies (21%; 248 participants) reported age data with a pooled mean age of 50 years (SD 18). Sex was reported in eight studies (234 participants), with 131 (56%) participants being female (range 31–100%) and 103 (44%) being male (range 0–69%). Education level was reported in ten studies (26%; 250 participants), and 124 (50%) of 250 were school graduates, 96 (38%) were undergraduates, and 31 (12%) were postgraduates. Ethnicity was reported in only two studies. Four studies did not report any assessors' demographics.

The simplified reports covered six different imaging modalities, with the most common being MRI (25 [66%] of 38) and CT (22 [58%] of 38). Multiple modalities were included in 19 (50%) of 38 studies and multiple imaging specialties in 10 (26%) studies. The most common imaging specialty was musculoskeletal imaging (8 [21%] of 38). Reports were mostly generated in English (29 [76%] of 38). Descriptive information is presented in figure 2.

A single LLM was used in 26 (68%) of 38 studies, whereas 12 (32%) studies compared multiple LLMs. GPT (OpenAI, San Francisco, CA, USA) was used in 35 (92%) studies, with GPT-4 being the most common model (18 [47%] of 38; figure 2).

The included studies reported patient-related and public-related metrics (ie, understanding, trust, empathy, and likability) and medical professional-related metrics (ie, accuracy, completeness, releasability, and harm), most commonly assessed on Likert scales ranging from 1 (strongly disagree or very poor) to 5 (strongly agree or very good).

Quality assessment and the completed MAIC-10 tool for all included studies are available in the appendix (pp 4, 7–8).

The pooled mean Likert score for perceived understanding for original radiologist-generated reports was 2·16 (SD 1·20) across 11 studies (representing 1048 reports and 277 assessors). In comparison, simplified reports generated by LLMs had a pooled mean score of 4·04 (SD 0·94) across 14 studies (representing 1111 reports and 387 assessors). Relative to the original radiologist reports, these data represent an 87% improvement in perceived understanding in patient or lay participant ratings of simplified reports. In studies in which assessors evaluated both the original and LLM-rewritten report, the pooled mean difference (MD) in Likert scores was 2·00 (95% CI 1·54–2·46) across ten studies (representing 1018 reports and 268 assessors), showing that the LLM-rewritten reports were substantially more understandable to lay readers and patients (figure 3).

Patient satisfaction with the radiologist-written report was reported by Sunshine and colleagues,[63] with a mean score of 1·61 (SD 1·21) across 30 reports and four patients and had an MD of 3·17 (95% CI 1·78–4·56). In comparison, patient satisfaction in LLM-rewritten reports was assessed in three studies[60,63,68] with a pooled Likert score of 3·81 (SD 0·82) across 140 reports and 316 patients. Three studies evaluated how well LLM-rewritten radiology reports conveyed empathy to patients.[54,55,60] The pooled Likert score for empathy was 3·61 (SD 0·79) across 177 reports and 143 patients, indicating a moderately positive perception. No study assessed patients' perception of empathy in the original radiologist-generated reports.

Patient trust in simplified reports was reported in one study of 32 patients with a mean Likert score of 4·09

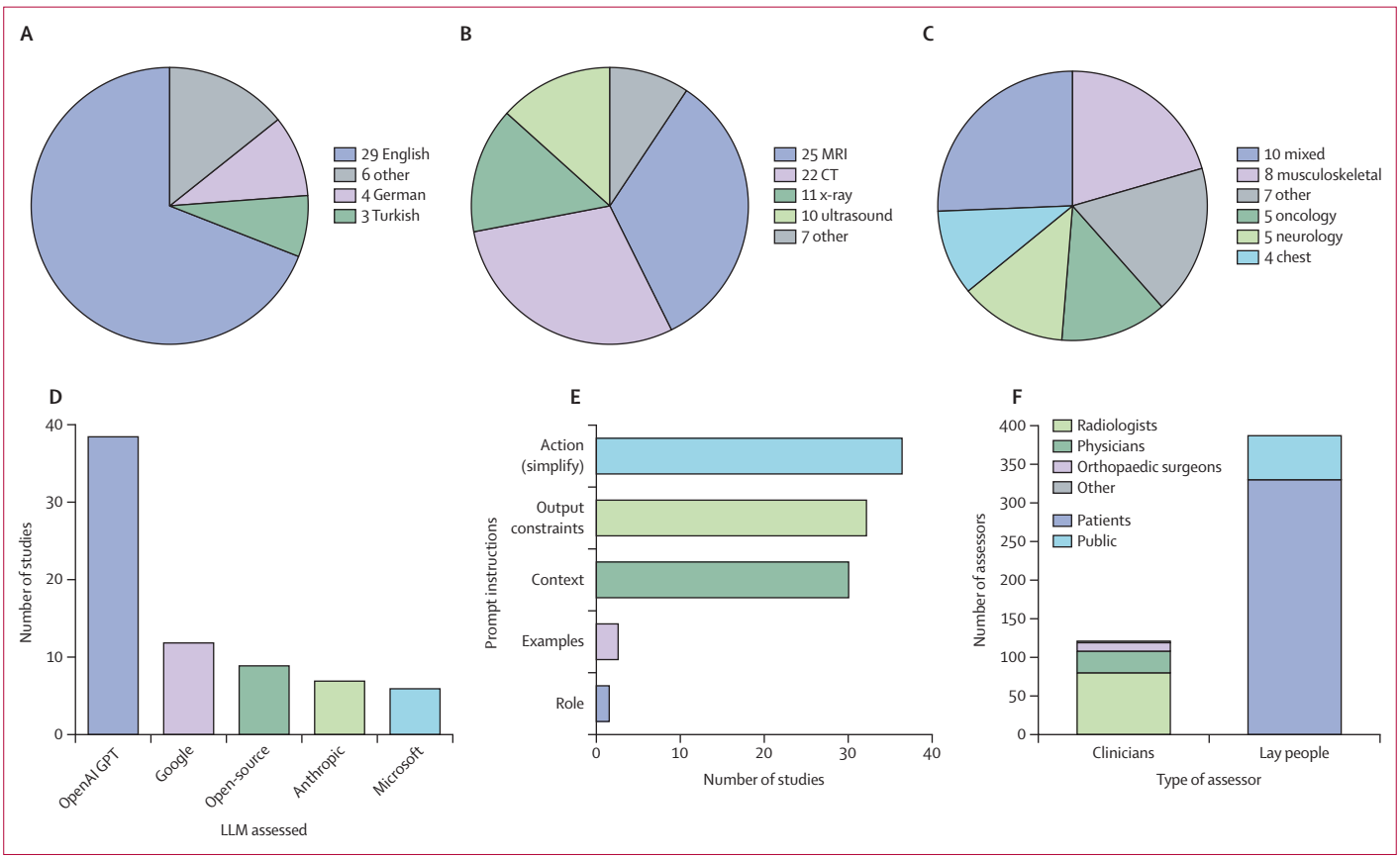


Figure 2: **Overview of included studies evaluating LLM-rewritten radiology reports**
(A) Language of simplified reports across studies (NB, some studies had reports in multiple languages). (B) Imaging modality. (C) Imaging specialty. (D) LLM vendors evaluated. (E) Descriptive information for the prompt anatomy used in report simplification. (F) Assessors of simplified reports. LLM=large language model.

(SD 0·68) for the simplified report and 4·48 (0·74) for the radiologist report.[35]

In 32 (84%) of 38 studies, 121 medical evaluators—including 80 (66%) radiologists, 28 (23%) physicians, and 11 (9%) orthopaedic surgeons—participated in the evaluation of the LLM-rewritten reports.

Pooled analysis of study outcomes showed high ratings for most measures (figure 4). Accuracy was rated at a pooled mean of 4·45 (95% CI 4·27–4·63) across 27 studies (representing 11 400 reports and 108 raters). Completeness had a pooled mean of 4·53 (95% CI 4·30–4·76) across 14 studies (1113 reports and 68 raters). Simplicity had a pooled mean score of 4·32 (95% CI 3·97–4·67) across eight studies (976 reports and 27 raters). Ratings for suitability to release the LLM-generated reports to patients (ie, the releasability) and for agreement that there was no potential for harm were lower than other scores, with a pooled mean of 3·93 (95% CI 3·10–4·77) for releasability across three studies (165 reports and 16 raters) and 3·79 (95% CI 3·10–4·51) for no potential for harm across six studies (122 reports and 42 raters). The sensitivity analysis restricted to studies using OpenAI models showed that GPT-4 achieved a higher mean accuracy rating of 4·77 (95% CI 4·60–4·94) across 10 studies (1199 reports and 28 raters) than GPT-3.5, with a mean accuracy rating of 4·09 (95% CI 3·78–4·40; p=0·0002) across 12 studies (3786 reports and 62 raters; appendix p 9). Mean accuracy ratings were similar for radiologists (4·51 [95% CI 4·29–4·72] in 17 studies, 1798 reports, and 66 raters) and non-radiologists (4·41 [95% CI 4·07–4·74] in nine studies, 9599 reports, and 38 raters; p=0·62; appendix p 10). Assessor agreement was substantial in six studies and fair or moderate in two studies (appendix p 4).

17 (45%) of 38 studies assessed error rates, including inaccurate, missing, or fabricated information. The pooled error rate for any error was 7·2% (95% CI 5·1%–10·0%) across 13 studies (representing 7774 reports and 353 incorrect reports; appendix p 11). For clinically significant errors, the pooled error rate was 0·9% (95% CI 0·6%–1·5%) across two studies (4037 reports and 38 incorrect reports).

19 (50%) of 38 studies reported readability scores and simplified reports showed substantial improvements across all metrics. A meta-analysis of mean difference in FKGL scores by imaging modality showed consistent improvements in simplified reports across all groups

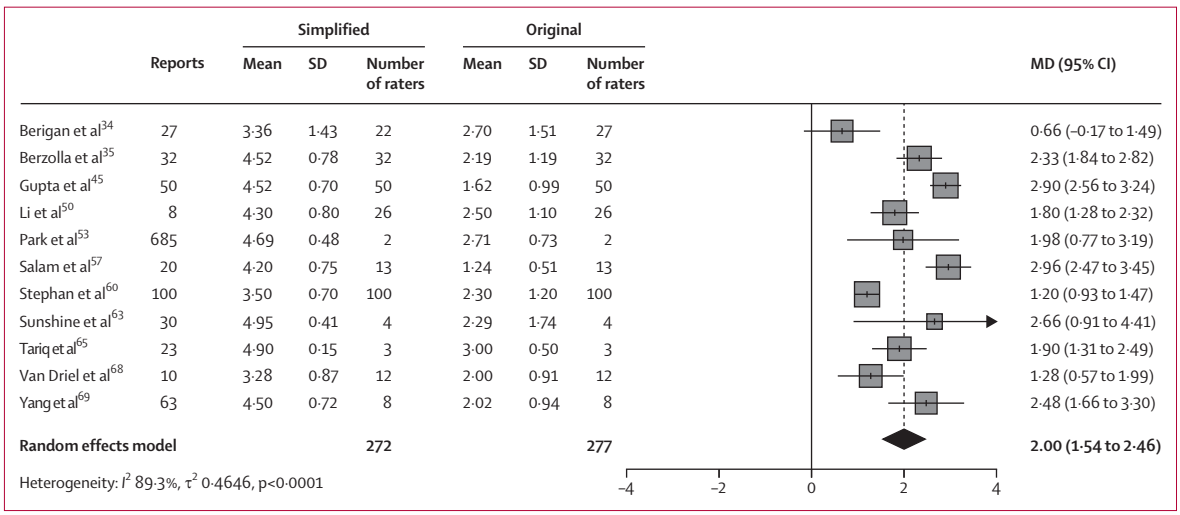


Figure 3: **Pooled mean differences in Likert scores for perceived understanding**
Positive values indicate a preference for simplified reports. MD=mean difference.

(appendix p 12). For x-ray studies, the pooled MD in FKGL score was –5·07 (95% CI –5·99 to –4·15) across five studies and 1100 reports. For CT studies, the MD was –6·20 (–6·91 to –5·48) across seven studies and 1400 reports, and for MRI, the MD was –5·0 (–6·0 to –4·0) across seven studies and 6729 reports. In studies simplifying multimodality reports, the pooled MD was –4·96 (–6·57 to –3·36) across four studies and 542 reports.

The FKGL, FRES, and ARI readability scores for both the original and simplified x-ray, CT, and MRI reports are summarised in the appendix (p 13). For CT imaging, the original reports mean FKGL score was 13·4 (95% CI 12·4–14·4) and the simplified reports had a score of 7·3 (6·7–7·9), representing a 47% reduction; the original reports mean ARI score was 20·0 (19·4–20·6) and the simplified reports was 9·1 (6·3–12·0), a 57% reduction; and the original reports mean FRES score was 28·0 (25·5–30·5) and the simplified reports were 73·6 (66·7–80·4), a 163% increase. For x-ray reports, the original reports mean FKGL score was 12·4 (11·1-13·6) and the simplified reports score was 7·3 (6·4–8·2), a 42% reduction in FKGL; the original reports mean ARI score was 13·3 (13·2–13·5) and the simplified reports score was 8·3 (3·5–13·0), a 53% reduction in ARI; and the original reports FRES score was 37·1 (35·1–39·1) and the simplified reports score was 77·8 (73·7–82·0), a 111% increase in FRES (appendix p 13). For MRI, the original reports mean FKGL score was 12·5 (11·3–13·8) and the simplified reports score was 7·5 (6·4–8·7), a 21% decrease in FKGL; the original reports mean ARI score was 13·5 (6·8–20·1) and the simplified reports score was 6·1 (3·9–8·2), a 54% decrease in ARI score; and the original reports mean FRES score was 26·3 (21·2–31·4) and the simplified reports score was 67·7 (55·9–79·6), a 72% increase in FRES (appendix p 13). These changes correspond to a shift from university-level to school-level (ages 11–13 years) text and from highly technical language to a general-audience style for CT and x-ray.

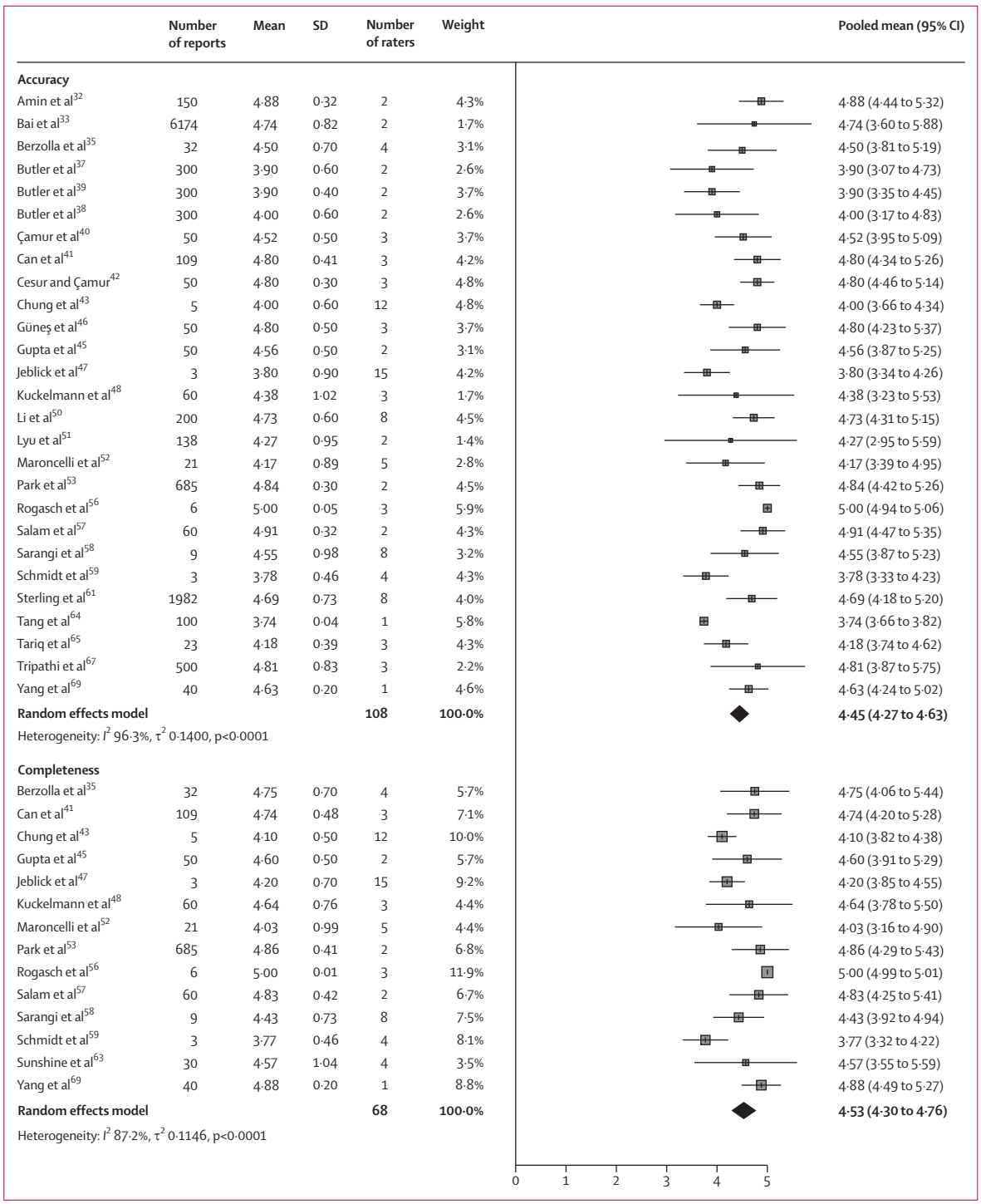
Further analyses are found in the appendix on report lengths, LLM model comparisons, and prompting strategies (appendix pp 14–19).

Discussion

This systematic review and meta-analysis is the first to evaluate LLMs used for improving patient understanding of radiology reports. 38 studies, published since 2022, were included, encompassing multiple imaging modalities, subspecialties, and LLMs. Lay people rated LLM-rewritten reports as 87% more understandable than original radiologist-authored reports. Clinicians also rated them highly for accuracy, completeness, and simplicity. The findings indicate that LLM-simplified reports can improve patient perceived comprehension without sacrificing correctness (figure 5).

Clinicians expressed confidence in the accuracy and completeness of LLM-rewritten reports. Such accuracy is crucial, since legislation such as the 21st Century Cures Act gives patients direct access to radiology reports often written above average literacy levels.[70] However, this confidence was tempered by caution, consistent with experiences from other LLM-generated explanations.[71] Releasability and safety were rated lower, reflecting concerns about unsupervised release because of errors, oversimplification, and patient anxiety when accessed without context.

LLM-rewritten reports had a low overall error rate; however, approximately 1 in 100 reports contained clinically significant errors, potentially altering diagnosis or severity. A human-in-the-loop model, in which clinicians review and authorise LLM-generated drafts before patient access, offers a potential solution.[20,22,72] However, the added workload might render this approach impractical in most



(Figure 4 continues on next page)

radiology settings.[73] The best method for verifying and releasing LLM-reports remains unresolved and no study in this review evaluated real-world use. Responsibility for release also remains unclear. Radiologists are best positioned to check technical accuracy, whereas referring physicians can contextualise results within broader patient care.

The timing of release adds complexity. Instant access risks alarming patients before their clinician has reviewed the findings (eg, reading about fetal death before the

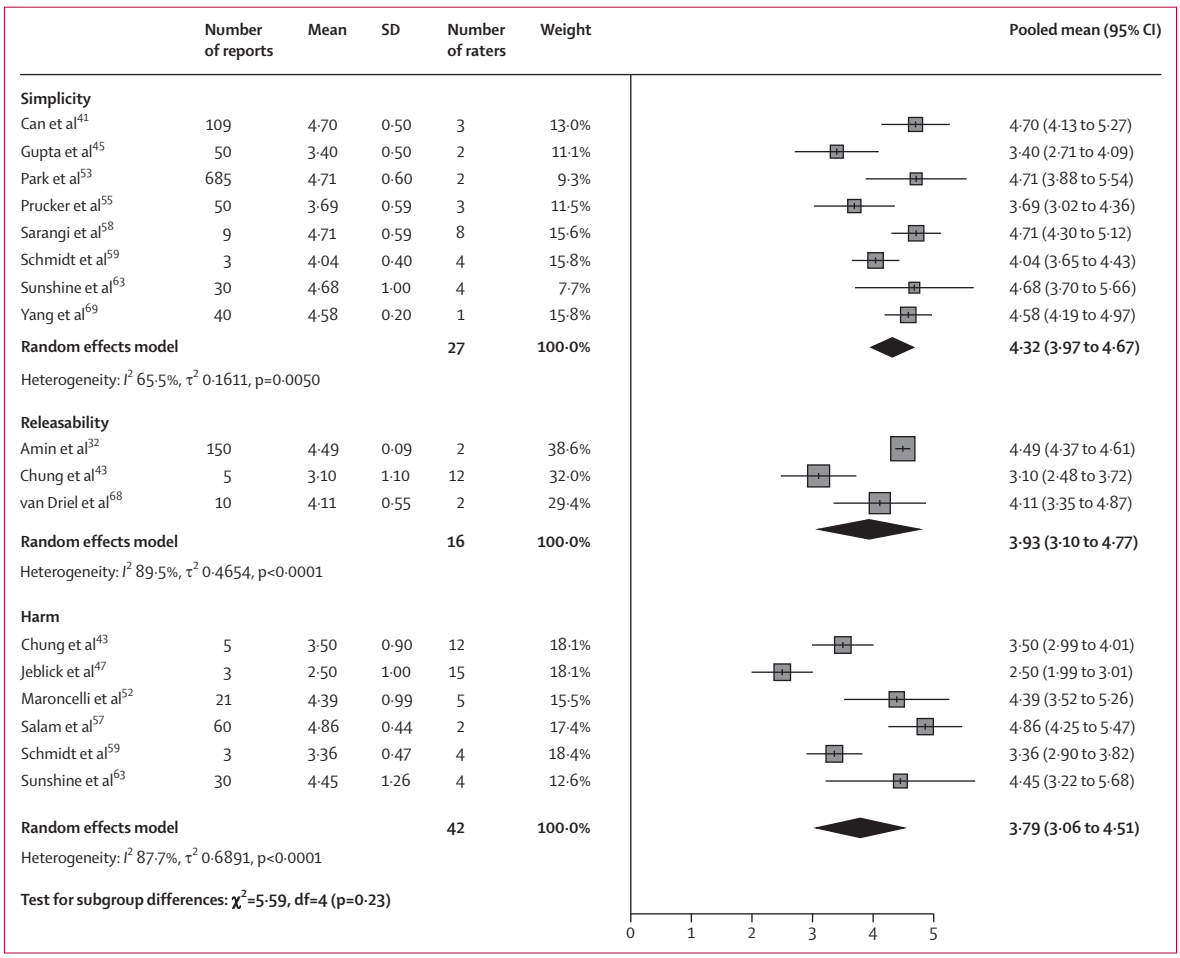


Figure 4: **Pooled means in Likert scores for medical professionals' assessment of large language model simplified reports**

clinician can compassionately explain results[74]), whereas delayed access could undermine engagement since patients prefer to know their results before consultations and prepare for discussions with their health-care practitioner.[75] Governance concerns, liability, disclaimers, and quality assurance remain unresolved, and dissemination methods should ensure equitable access for less digitally literate populations.[76] Inconsistency in outputs and no standard templates further threaten trust and reproducibility.[20,77]

Readability remains a central barrier to patient-centred radiology. LLMs consistently improved readability metrics and shifted reports from university-level language to one with a reading age of 11–13 years, aligning with health literacy recommendations.[78] These improvements should, however, be interpreted with caution. Readability formulas rely on surface-level features of text such as counts of sentences, words, and syllables, which are not consistently defined and vary across tools. Radiology reports in particular are full of abbreviations, numbers, and unusual punctuation, which makes boundary detection difficult and the metrics unstable. Improved readability can also come at the cost of reduced accuracy.[79,80] Moreover, traditional readability scores are decades old, can be gamed, and do not necessarily measure comprehension.[81]

Notably, improved readability was often accompanied by longer reports (appendix p 14). This paradox stems from how LLMs simplify: they not only substitute with simpler words but also add definitions, explanations, and analogies to unpack complex terminology. Such expansion might enhance comprehension but also risks information overload and increases burden for clinicians reviewing them. Thus, readability should not be equated with usability.

Although the meta-analysis showed that patients and lay assessors preferred simplified reports, none of the included studies evaluated patients independently applying LLMs to their own radiology reports or explored what patients specifically want to see in these simplified reports. Co-design studies highlighted a shared set of expectations.[82–85] Patients prefer reports in clear, lay-friendly language with concise explanations matched to their literacy level. They want unambiguous statements about typical or atypical results, glossaries for technical terms, and structured sections

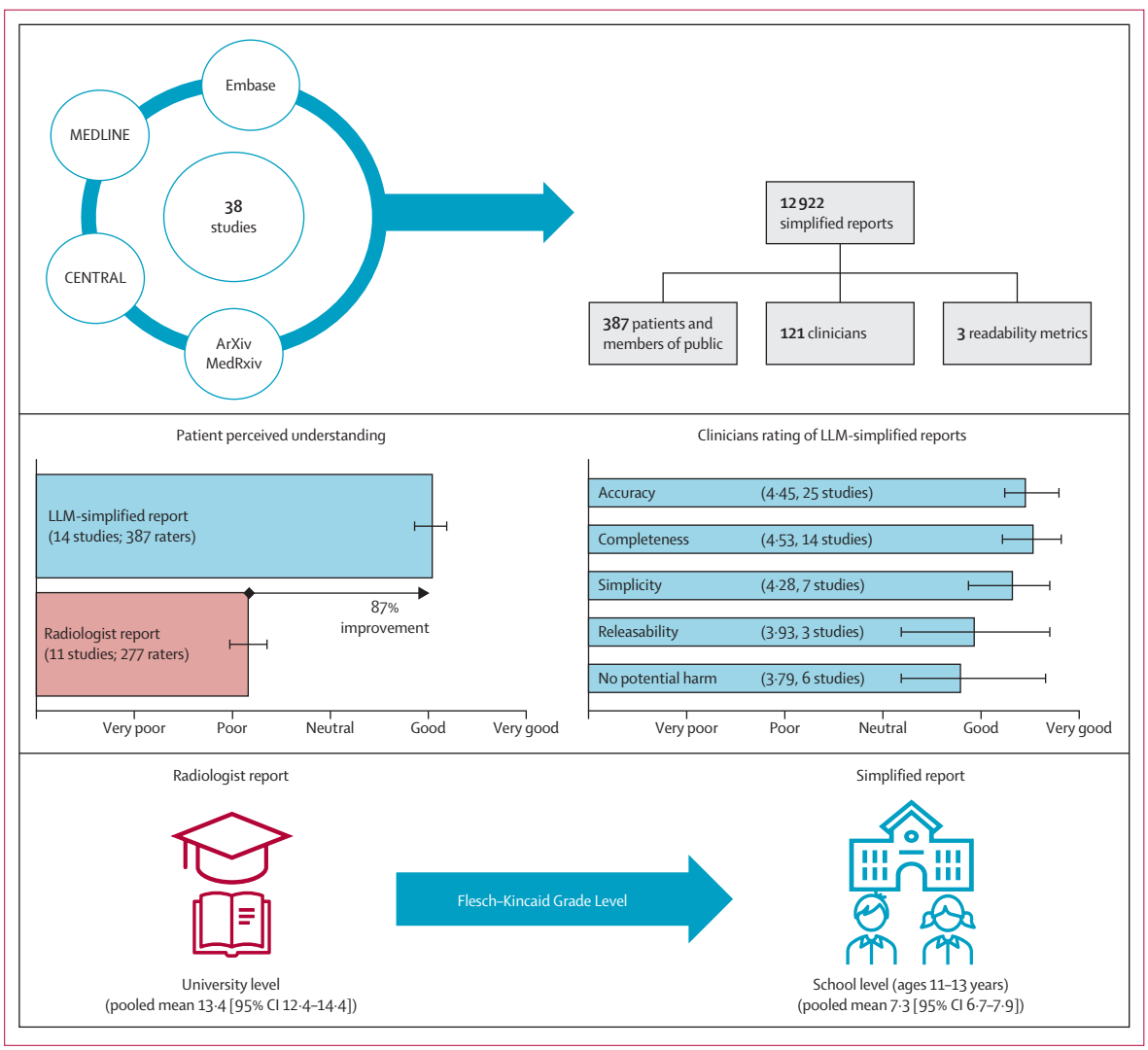


Figure 5: **Overview of the findings from this meta-analysis**
Readability outcomes (Flesch–Kincaid Grade Level) are shown for CT reports. LLM=large language model.

about what the results mean for them and what happens next. Emotional sensitivity is valued when conveying bad news or unexpected findings, alongside actionable advice on urgency, prognosis, and questions to prepare for consultations.[82–84] Visual aids, annotated images, and links to trusted resources help patients' understanding and reduce their anxiety.[82,83]

Our synthesis found moderately positive experiences with empathy shown in the LLM-rewritten reports. However, trust is more nuanced: in one study, patients rated radiologist reports slightly higher than LLM-rewritten reports for trust,[35] possibly reflecting patient concerns about depersonalisation, standardised phrasing, bias, or data privacy. Patients supported AI-generated explanations, but only if clinician-supervised to ensure accuracy and transparency.[73,86] Patients also want to see both the original and simplified report side by side, for cross-reference and second opinions.[82,85] This approach has been adopted in research and commercial solutions, which present the original report with hyperlinks that patients can explore.[34,87]

The evidence base for the use of LLMs to simplify radiology reports is growing but currently remains small. Most studies were small, single-centre studies. Patient participants tended to be younger, English-speaking, and with higher levels of education, limiting generalisability of the study findings. Full masking of assessors to the report origin is difficult, as LLM outputs have a distinctive style. Studies assessed self-reported (ie, perceived) understanding of radiology reports. However, this perceived understanding might not reflect actual understanding and correlation between the two is not always reliable.[88] Thus, more accurate ways of assessing comprehension of radiology reports should be investigated, such as tests involving

question answering, summarisation, inference, or connection to other knowledge. Reproducibility was constrained by the absence of shared datasets or code and variable LLM prompting methods, which reduces the robustness and transparency of the evidence base. Included studies focused patient and clinician assessments primarily on CT and MRI, limiting generalisability to other imaging modalities. Crucially, no studies incorporated patient perspectives in the design of simplified reports, despite their centrality to patient-centred innovation.

Heterogeneity across the meta-analyses was very high, probably reflecting differences in imaging modalities, clinical specialties, report complexity, disease severity, institutional practices, reporter expertise, assessor experience, rating frameworks, and LLM versions. These differences could not be adjusted for in sensitivity analyses or meta-regression and could reduce the accuracy of pooled estimates. Therefore, the precision and generalisability of pooled effect estimates are reduced, and the summary effects should be interpreted with appropriate caution. Nevertheless, the consistent direction of effect across studies supports the reliability of the overall conclusions.

Future research should prioritise co-design with patients and clinicians and establish standard evaluation metrics and prompting strategies. Prospective implementation studies are needed to assess acceptability, accessibility, equity, safety, workflow effects, and patient-level and system-level outcomes and should incorporate objective comprehension testing (eg, structured questionnaires or assessments) to determine whether improved readability translates into better understanding. To accommodate diverse preferences and literacy levels, future systems might need to offer adaptive formats that balance brevity with detail, such as concise summaries alongside detailed explanations.[70,89] Further development should explore embedding visual or explanatory aids to ensure reports are not only more readable but genuinely usable for patients.[90–92]

In conclusion, LLM simplification of radiology reports improves patient perceived understanding while maintaining clinical accuracy, positioning it as a powerful tool to make radiology reports more patient centred. By adopting a careful, evidence-based approach, LLM-rewritten reports could evolve from a technical novelty into a cornerstone of patient communication.

Contributors

SA conceived the idea and need for the systematic review and contributed to the study conception and design. AA and SA registered the protocol on PROSPERO. SA created the search strategy and did the literature search. SA, AA, and NMcA did the screening and eligibility assessments independently. SA, AA, and NMcA evaluated the included studies and collected relevant data independently. SA, AA, and NMcA verified the data and all authors had access to the data. SA did the meta-analyses. SA, AM, and AA drafted the manuscript, figures, and tables. AHu, MSa, MSh, KD, AHo, FA, MSt, NM, RG, TJC, AJS, JJL, JK, and CL reviewed the draft and provided comments and suggestions. All authors contributed to the interpretation of data, took part in the critical review and editing of the manuscript, and have read and approved the final manuscript.

Declaration of interests

We declare no competing interests.

Data sharing

Search terms, study characteristics (data, model, and performance metrics), quality assessment criteria and results, risk of bias assessment criteria and results, and meta-analysis findings are available in the appendix.

Acknowledgments

This study was supported by the UK National Institute for Health and Care Research (NIHR) Sheffield Biomedical Research Centre (NIHR203321 and NIHR304150). CPL is supported, in part, by the Medical Imaging and Data Resource Center, funded by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (under contract 75N92020D00021) and through The Advanced Research Projects Agency for Health. AJS is supported by the British Heart Foundation Senior Clinical Research Fellowship (FS/SCRF/24/32034). MSh is supported by the Wellcome Trust doctoral fellowship grant (223521/Z/21/Z). During the preparation of this work the authors used ChatGPT Business (GPT-5, OpenAI; data not used for training) on Sept 1, 2025, to generate a draft for the research in context panel based on the abstract (appendix p 25). The authors reviewed and substantially edited the content and take full responsibility for the content of the publication.

References

1 Amin KS, Davis MA, Naderi A, Forman HP. Increasing patient viewership of complex imaging reports: the paradox of the Cures Act. *Clin Imaging* 2025; **119:** 110398.

2 Miles RC, Hippe DS, Elmore JG, Wang CL, Payne TH, Lee CI. Patient access to online radiology reports: frequency and sociodemographic characteristics associated with use. *Acad Radiol* 2016; **23:** 1162–69.

3 Triggle N. Patients to get full access to record on NHS App, BBC News. Oct 21, 2024. https://www.bbc.co.uk/news/articles/cz7j73vx9v3o (accessed Nov 23, 2024).

4 Mehan WA Jr, Brink JA, Hirsch JA. 21st Century Cures Act: patient-facing implications of information blocking. *J Am Coll Radiol* 2021; **18:** 1012–16.

5 Mervak BM, Davenport MS, Flynt KA, Kazerooni EA, Weadock WJ. What the patient wants: an analysis of radiology-related inquiries from a web-based patient portal. *J Am Coll Radiol* 2016; **13:** 1311–18.

6 Bruno MA, Petscavage-Thomas JM, Mohr MJ, Bell SK, Brown SD. The "open letter": radiologists' reports in the era of patient web portals. *J Am Coll Radiol* 2014; **11:** 863–67.

7 Rogers C, Willis S, Gillard S, Chudleigh J. Patient experience of imaging reports: a systematic literature review. *Ultrasound* 2023; **31:** 164–75.

8 Tapuria A, Porat T, Kalra D, Dsouza G, Xiaohui S, Curcin V. Impact of patient access to their electronic health record: systematic review. *Inform Health Soc Care* 2021; **46:** 192–204.

9 Mangano MD, Bennett SE, Gunn AJ, Sahani DV, Choy G. Creating a patient-centered radiology practice through the establishment of a diagnostic radiology consultation clinic. *AJR Am J Roentgenol* 2015; **205:** 95–99.

10 Johnson AJ, Easterling D, Nelson R, Chen MY, Frankel RM. Access to radiologic reports via a patient portal: clinical simulations to investigate patient preferences. *J Am Coll Radiol* 2012; **9:** 256–63.

11 Davis Giardina T, Menon S, Parrish DE, Sittig DF, Singh H. Patient access to medical records and healthcare outcomes: a systematic review. *J Am Med Inform Assoc* 2014; **21:** 737–41.

12 Martin-Carreras T, Cook TS, Kahn CE Jr. Readability of radiology reports: implications for patient-centered care. *Clin Imaging* 2019; **54:** 116–20.

13 UK Office for National Statistics. Language. 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language (accessed Jan 8, 2025).

14 The Reading Agency. The state of the nation's adult reading: 2024. Focus on... reading, skills development and career opportunities. 2024. https://readingagency.org.uk/wp-content/uploads/2024/08/State-of-the-Nations-Adult-Reading-2024-Focus-on-Skills-Engagement-and-Career-Opportunities.pdf (accessed Sept 1, 2025).

15 The Royal College of Radiologists. RCR/SoR position statement on access to imaging reports. https://www.rcr.ac.uk/news-policy/latest-updates/rcrsor-position-statement-on-access-to-imaging-reports/ (accessed Jan 5, 2026).

16 Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models in simplifying radiological reports: systematic review. *bioRxiv* 2024; published online Jan 9. https://doi.org/10.1101/2024.01.05.24300884 (preprint).

17 Casey A, Davidson E, Poon M, et al. A systematic review of natural language processing applied to radiology reports. *BMC Med Inform Decis Mak* 2021; **21:** 179.

18 Davidson EM, Poon MTC, Casey A, et al. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. *BMC Med Imaging* 2021; **21:** 142.

19 Liu C, Tian Y, Song Y. A systematic review of deep learning-based research on radiology report generation. *arXiv* 2023; published online Nov 23. http://arxiv.org/abs/2311.14199 (preprint).

20 Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: a systematic review of performance, pitfalls, and future perspectives. *Diagn Interv Imaging* 2024; **105:** 251–65.

21 Gorenstein L, Konen E, Green M, Klang E. Bidirectional encoder representations from transformers in radiology: a systematic review of natural language processing applications. *J Am Coll Radiol* 2024; **21:** 914–41.

22 Temperley HC, O'Sullivan NJ, Mac Curtain BM, et al. Current applications and future potential of ChatGPT in radiology: a systematic review. *J Med Imaging Radiat Oncol* 2024; **68:** 257–64.

23 Zhang H, Yu PS, Zhang J. A systematic survey of text summarization: from statistical methods to large language models. *ACM Computing Surveys* 2024; **57:** 277.

24 Weissman GE, Mankowitz T, Kanter GP. Unregulated large language models produce medical device-like output. *NPJ Digit Med* 2025; **8:** 148.

25 Smith MS. AI isn't replacing radiologists. Instead, they're using it to tackle time-sucking administrative tasks. Business Insider. June 5, 2025. https://www.businessinsider.com/radiology-embraces-generative-ai-to-streamline-productivity-2025-6 (accessed Sept 9, 2025).

26 Karran EL, Medalian Y, Hillier SL, Moseley GL. The impact of choosing words carefully: an online investigation into imaging reporting strategies and best practice care for low back pain. *PeerJ* 2017; **5:** e4151.

27 Bossen JKJ, Hageman MGJS, King JD, Ring DC. Does rewording MRI reports improve patient understanding and emotional response to a clinical report? *Clin Orthop Relat Res* 2013; **471:** 3637–44.

28 Farmer C, Bourne A, Haas R, Wallis J, O'Connor D, Buchbinder R. Can modifications to how medical imaging findings are reported improve quality of care? A systematic review. *Clin Radiol* 2022; **77:** 428–35.

29 Witherow JL, Jenkins HJ, Elliott JM, et al. Characteristics and effectiveness of interventions that target the reporting, communication, or clinical interpretation of lumbar imaging findings: a systematic review. *AJNR Am J Neuroradiol* 2022; **43:** 493–500.

30 Moher D, Liberati A, Tetzlaff J, Altman DG, and the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009; **6:** e1000097.

31 Cerdá-Alberich L, Solana J, Mallol P, et al. MAIC-10 brief quality checklist for publications using artificial intelligence and medical images. *Insights Imaging* 2023; **14:** 11.

32 Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. *Radiology* 2023; **309:** e232561.

33 Bai X, Feng M, Ma W, Liao Y. Application of artificial intelligence chatbots in interpreting magnetic resonance imaging reports: a comparative study. *Sci Rep* 2025; **15:** 31266.

34 Berigan K, Short R, Reisman D, et al. The impact of large language model-generated radiology report summaries on patient comprehension: a randomized controlled trial. *J Am Coll Radiol* 2024; **21:** 1898–903.

35 Berzolla E, Gosnell GG, Chen L, Vonck C, Alaia E, Meislin R. Artificial intelligence large language models improve patient comprehension of radiologist magnetic resonance imaging reports. *Arthroscopy* 2025; **41:** 4607–14.

36 Bozer A, Pekçevik Y. Comparative evaluation of large language models in explaining radiology reports: expert assessment of readability, understandability, and communication features. *Insights Imaging* 2025; **16:** 232.

37 Butler JJ, Acosta E, Kuna MC, et al. Decoding radiology reports: artificial intelligence-large language models can improve the readability of hand and wrist orthopedic radiology reports. *Hand* 2025; **20:** 1144–52.

38 Butler JJ, Puleo J, Harrington MC, et al. From technical to understandable: artificial intelligence large language models improve the readability of knee radiology reports. *Knee Surg Sports Traumatol Arthrosc* 2024; **32:** 1077–86.

39 Butler JJ, Harrington MC, Tong Y, et al. From jargon to clarity: improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. *Foot Ankle Surg* 2024; **30:** 331–37.

40 Çamur E, Cesur T, Güneş YC. A comparative study: performance of large language models in simplifying Turkish computed tomography reports. *J Istanb Univ Fac Med* 2024; **87:** 321–26.

41 Can E, Uller W, Vogt K, et al. Large language models for simplified interventional radiology reports: a comparative analysis. *Acad Radiol* 2025; **32:** 888–98.

42 Cesur YCG, Çamur E. Use of large language models in radiological reports: a study on simplifying Turkish MRI findings. *Ann Clin Anal Med* 2024; **15:** 586–90.

43 Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. *Digit Health* 2023; **9:** 20552076231221620.

44 Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis. *Radiology* 2024; **310:** e231593.

45 Gupta A, Singh S, Malhotra H, et al. Provision of radiology reports simplified with large language models to patients with cancer: impact on patient satisfaction. *JCO Clin Cancer Inform* 2025; **9:** e2400166.

46 Güneş YC, Cesur T, Çamur E. Comparative analysis of large language models in simplifying Turkish ultrasound reports to enhance patient understanding. *Eur J Ther* 2024; **30:** 714–23.

47 Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol* 2024; **34:** 2817–25.

48 Kuckelman IJ, Wetley K, Yi PH, Ross AB. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. *Skeletal Radiol* 2024; **53:** 1621–24.

49 Li H, Moon JT, Iyer D, et al. Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clin Imaging* 2023; **101:** 137–41.

50 Li HH, Moon JT, Kumar S, et al. Evaluation of multilingual simplifications of IR procedural reports using GPT-4. *J Vasc Interv Radiol* 2025; **36:** 696–703.e1.

51 Lyu Q, Tan J, Zapadka ME, 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* 2023; **6:** 9.

52 Maroncelli R, Rizzo V, Pasculli M, et al. Probing clarity: AI-generated simplified breast imaging reports for enhanced patient comprehension powered by ChatGPT-4o. *Eur Radiol Exp* 2024; **8:** 124.

53 Park J, Oh K, Han K, Lee YH. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. *Sci Rep* 2024; **14:** 13218.

54 Pisarcik D, Kissling M, Heimer J, et al. Artificial intelligence language models to translate professional radiology mammography reports into plain language—impact on interpretability and perception by patients. *Acad Radiol* 2025; **32:** 4988–96.

55 Prucker P, Busch F, Dorfner F, et al. Performance of open-source and proprietary large language models in generating patient-friendly radiology chest CT reports. *Clin Imaging* 2025; **125:** 110557.

56 Rogasch JMM, Metzger G, Preisler M, et al. ChatGPT: can you prepare my patients for [^{18}F]FDG PET/CT and explain my reports? *J Nucl Med* 2023; **64:** 1876–79.

57 Salam B, Kravchenko D, Nowak S, et al. Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand. *J Cardiovasc Magn Reson* 2024; **26:** 101035.

58 Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. *Cureus* 2023; **15:** e50881.

59 Schmidt S, Zimmerer A, Cucos T, Feucht M, Navas L. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. *Arch Orthop Trauma Surg* 2024; **144:** 611–18.

60 Stephan D, Bertsch AS, Schumacher S, et al. Improving patient communication by simplifying AI-generated dental radiology reports with ChatGPT: comparative study. *J Med Internet Res* 2025; **27:** e73337.

61 Sterling NW, Brann F, Frisch SO, Schrager JD. Patient-readable radiology report summaries generated via large language model: safety and quality. *J Patient Exp* 2024; **11:** 23743735241259477.

62 Sudarshan M, Shih S, Yee E, et al. Agentic LLM workflows for generating patient-friendly medical reports. *arXiv* 2024; published online Aug 2. http://arxiv.org/abs/2408.01112 (preprint).

63 Sunshine A, Honce GH, Callen AL, et al. Evaluating the quality and understandability of radiology report summaries generated by ChatGPT: survey study. *JMIR Form Res* 2025; **9**; e76097–e76097.

64 Tang CC, Nagesh S, Fussell DA, et al. Generating colloquial radiology reports with large language models. *J Am Med Inform Assoc* 2024; **31:** 2660–67.

65 Tariq A, Trivedi S, Urooj A, et al. Patient-centric summarization of radiology findings using two-step training of large language models. *ACM Trans Comput Healthc* 2025; **6:** 21.

66 Tepe M, Emekli E. Decoding medical jargon: the use of AI language models (ChatGPT-4, BARD, Microsoft Copilot) in radiology reports. *Patient Educ Couns* 2024; **126:** 108307.

67 Tripathi S, Mutter L, Muppuri M, et al. PRECISE framework: enhanced radiology reporting with GPT for improved readability, reliability, and patient-centered care. *Eur J Radiol* 2025; **187:** 112124.

68 van Driel MHE, Blok N, van den Brand JAJG, et al. Leveraging GPT-4 enables patient comprehension of radiology reports. *Eur J Radiol* 2025; **187:** 112111.

69 Yang Z, Cherian S, Vucetic S. Two-pronged human evaluation of ChatGPT self-correction in radiology report simplification. *Find ACL* 2024; **2024:** 4701–14.

70 Amin K, Khosla P, Doshi R, Chheang S, Forman HP. Artificial intelligence to improve patient understanding of radiology reports. *Yale J Biol Med* 2023; **96:** 407–17.

71 Shah SJ, Nair A, Murtagh K, et al. Clinician perspectives on AI-generated drafts of patient test result explanations. *JAMA Netw Open* 2025; **8:** e2528794.

72 Artsi Y, Klang E, Collins JD, et al. Large language models in radiology reporting—a systematic review of performance, limitations, and clinical implications. *Intell Based Med* 2025; **12:** 100287.

73 Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. *J Med Internet Res* 2025; **27:** e63649.

74 Rotholz S, Lin C-T. "I don't think it should take you three days to tell me my baby is dead." A case of fetal demise: unintended consequences of immediate release of information. *J Am Med Inform Assoc* 2023; **30:** 1301–04.

75 Steitz BD, Turer RW, Lin C-T, et al. Perspectives of patients about immediate access to test results through an online patient portal. *JAMA Netw Open* 2023; **6:** e233572.

76 Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* 2025; **333:** 319–28.

77 Reichenpfader D, Müller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. *NPJ Digit Med* 2024; **7:** 222.

78 Lopez C, Kim B, Sacks K. Health literacy in the United States: enhancing assessments and reducing disparities. *SSRN* 2022; published online Aug 22. https://doi.org/10.2139/ssrn.4182046 (preprint).

79 Lee H-S, Song S-H, Park C, et al. The ethics of simplification: balancing patient autonomy, comprehension, and accuracy in AI-generated radiology reports. *BMC Med Ethics* 2025; **26:** 136.

80 Lee H-S, Kim S, Kim S, et al. Readability versus accuracy in LLM-transformed radiology reports: stakeholder preferences across reading grade levels. *Radiol Med* 2025; published online Sept 29. https://doi.org/10.1007/s11547-025-02098-5.

81 Tanprasert T, Kauchak D. Flesch–Kincaid is not a text simplification evaluation metric. In: Proceedings of the first workshop on natural language generation, evaluation, and metrics (GEM). Association for Computational Linguistics, 2021: 1–14.

82 Perlis N, Finelli A, Lovas M, et al. Creating patient-centered radiology reports to empower patients undergoing prostate magnetic resonance imaging. *Can Urol Assoc J* 2021; **15:** 108–13.

83 Alarifi M, Patrick T, Jabour A, Wu M, Luo J. Designing a consumer-friendly radiology report using a patient-centered approach. *J Digit Imaging* 2021; **34:** 705–16.

84 Zhang Z, Citardi D, Wang D, Genc Y, Shan J, Fan X. Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data. *Health Informatics J* 2021; **27:** 14604582211011215.

85 Lockwood P, Mitchell M. A co-designed patient reported experience measure for understanding the patient's and public experience of receiving x-ray results. *Radiography* 2025; **31:** 102990.

86 Grover V, Balusamy BMKN, Anand V, Milanova M. Approaches to human-centered AI in healthcare, 2024, *IGI Global*.

87 Herwald SE, Shah P, Johnston A, Olsen C, Delbrouck J-B, Langlotz CP. RadGPT: a system based on a large language model that generates sets of patient-centered materials to explain radiology report information. *J Am Coll Radiol* 2025; **22:** 1050–59.

88 Prinz A, Golke S, Wittwer J. How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educ Res Rev* 2020; **31:** 100358.

89 Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. *Front Med* 2024; **11:** 1477898.

90 Rockall AG, Justich C, Helbich T, Vilgrain V. Patient communication in radiology: moving up the agenda. *Eur J Radiol* 2022; **155:** 110464.

91 Schreyer AG, Schneider K, Dendl LM, et al. Patient centered radiology—an introduction in form of a narrative review. *Rofo* 2022; **194:** 873–81 (in German).

92 Mityul MI, Gilcrease-Garcia B, Mangano MD, Demertzis JL, Gunn AJ. Radiology reporting: current practices and an introduction to patient-centered opportunities for improvement. *AJR Am J Roentgenol* 2018; **210:** 376–85.