BMJ Open

How have generic large language models progressed in their ability to write clinic letters and provide accurate management plans in the virtual fracture clinic?

Amy Smith,[1] James Brock,[1] Harri Jones,[2] Francesca Solari ,[3] Rana Anss,[2] Charles Kimberley,[1] Claire Joyner,[4] Tariq Yasin,[1] Omar Basbous,[1] Arwel Tomos Poacher [2,5]

To cite: Smith A, Brock J, Jones H, *et al*. How have generic large language models progressed in their ability to write clinic letters and provide accurate management plans in the virtual fracture clinic? *BMJ Open* 2025;**15**:e104612. doi:10.1136/bmjopen-2025-104612

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2025-104612).

AS and JB are joint first authors.

Received 04 May 2025
Accepted 04 November 2025

For numbered affiliations see end of article.

Correspondence to
Dr Arwel Tomos Poacher;
drarwelpoacher@gmail.com

ABSTRACT

Objective To explore whether large language models (LLMs), Generative Pre-trained Transformer (GPT)-3, GPT-3.5 and GPT-4 can autonomously manage a virtual fracture clinic (VFC) as a marker of their efficacy in an emergency department and with simple orthopaedic trauma.

Setting and participants Simulated UK VFC workflow.

Design 11 clinical scenarios were generated, and GPT-4, GPT-3.5 and GPT-3 were prompted to write clinic letters and management plans.

Main outcome measures The Readable Tool was used to assess the clarity of letters. Six independent orthopaedic surgeons then evaluated the accuracy of letters and management plans.

Results Readability was compared using the Flesch-Kincaid grade level: GPT-4: 9.11 (SD 0.98); GPT-3.5: 8.77; GPT-3: 8.47, and the Flesch readability ease: GPT-4: 56.3; GPT-3.5: 58.2; GPT-3: 59.3. Surgeon-rated accuracy comparisons indicated that GPT-4 exhibited the highest accuracy for management plans (9.08/10 (95% CI 8.25 to 9.9)). This represents a statistically significant progression in the capacity of a LLM to provide accurate management plans compared with GPT-3 at 6.84 (95% CI 5.41 to 8.27) and GPT-3.5 at 7.63 (95% CI 7.23 to 8.13) ($p<0.0001$).

Conclusions LLMs can produce high-quality, readable clinical letters for common VFC presentations, and GPT-4 can generate management plans to aid clinicians in their administration. With clinician oversight, appropriately trained LLMs could meaningfully reduce routine administrative work. However, while the results of this study are promising, further evaluation of LLMs is required before they can be deemed safe for managing simple orthopaedic scenarios.

INTRODUCTION

Due to the increasing demand for trauma services following emergency department attendance, virtual fracture clinics (VFCs) have been widely adopted to comply with the British Orthopaedic Association Standards for Trauma and Orthopaedics (BOAST) Guidance. The BOAST guidelines specify that clinic reviews occur within 72 hours of a traumatic orthopaedic injury, further stating that adequate clinic letters should communicate the diagnosis, investigations and management plan.[1]

The Glasgow Fracture Pathway was established to use VFC in redesigning the management pathways for fractures managed non-operatively.[2] This model leverages a consultant-led review of clinical notes and radiology before nursing and administrative staff contact the patient and communicate the decision to their general practitioner (GP) via a letter. Approximately 4.6% of all emergency department attendances are trauma-related, demonstrating the demand for specialist orthopaedic review and the resultant administrative burden.[3]

Artificial Intelligence (AI) in the form of large language models (LLMs) is being rapidly

STRENGTHS AND LIMITATIONS

⇒ This is the first study to evaluate the progression of large language models (Generative Pre-trained Transformer (GPT)-3, GPT-3.5 and GPT-4) over time in generating clinic letters and management plans for virtual fracture clinic (VFC) presentations.

⇒ The study incorporated common VFC presentations, covering both operative and non-operative cases, to reflect typical practice.

⇒ Both objective readability indices and clinician-rated accuracy were applied, providing a comprehensive methodological assessment.

⇒ The prompts used can lead to reliance on documentation and radiographic reporting; the information used is limited by its use of simulated patient scenarios, as this reduces the amount of variation seen in real-life clinical practice.

Table 1 Summary of letter and management prompts

Fracture prompt	Management prompt
Distal radius (undisplaced extra-articular)	Non-operative
Clavicle (undisplaced midshaft)	Non-operative
Proximal humerus (4-part)	Open reduction and internal fixation
Olecranon (simple transverse)	Open reduction and internal fixation
Midshaft ulna (nightstick)	Non-operative
Distal biceps tendon rupture	Awaiting ultrasound
Buckle fracture	Discharge
Quadriceps tendon rupture	Awaiting ultrasound
Knee medial collateral ligament injury	Non-operative
Ankle (Weber A)	Non-operative
Ankle (bimalleolar)	Open reduction and internal fixation

applied to orthopaedic practice via research, patient information provision and clinical letter production.[4 5] Chat-Generated Pre-trained Transformer (ChatGPT) is a supervised learning model reinforced by human feedback, designed to generate text responses to human prompts.[6] Initial versions of ChatGPT were found to lack readability and quality of information compared with common patient search queries.[7] However, ChatGPT 4.0 has demonstrated advancements in orthopaedic literacy, with a recent study showing a pass mark of 63.4% on the Orthopaedic In-Training Exam.[8] As such, ChatGPT shows promise in orthopaedic outpatient settings, with the ability to generate detailed clinic letters, provide patient information related to injuries, and may soon be able to aid in management planning.

This study aimed to evaluate GPT-4's advances in letter and management plan generation for common VFC referrals.

METHODS

The AI software GPT-3.0, GPT-3.5 and GPT-4, a LLM produced by OpenAI accessed 23 November 2024 via ChatGPT web interface, was selected to generate attendance letters and management plans for 11 common VFC referrals outlined in table 1. The 11 clinical scenarios were selected to provide a representative sample of the presentations to VFC, with a combination of awaiting further investigation, non-operative and operative management, and paediatric and adult cases. The case complexity, relevance to guidance, radiological findings and comorbidities are summarised in online supplemental material.

ChatGPT was given the prompts 'write a letter to the patient and their GP about the following' for a clinic letter for the following patient seen in an orthopaedic clinic, based on the information provided. An example of a prompt provided for non-operative management of a distal radius fracture is outlined in table 2. Further details of the remaining prompts are available within the published online supplemental dataset.[9]

Table 2 Example prompt provided to ChatGPT for the management of a distal radius fracture

Demographics	60-year-old female, retired, right-hand dominant
Mechanism	Fall on outstretched right hand, following a trip over loose paving
Past medical historty	Hypertension, provoked deep vein thrombosis 10 years ago
Imaging	Transverse metaphyseal, non-displaced, non-comminuted extra-articular fracture of the right distal radius with no shortening Dorsal angle <5° Universal classification type I

ChatGPT, Chat-Generated Pre-trained Transformer.

Following the prompt, details regarding the patient's age, mechanism of injury, hand dominance, occupation and past medical history were documented. A comprehensive X-ray report was also recorded, outlining the fracture pattern. Lastly, an initial management plan was provided, including cast immobilisation and a review in 2 weeks, along with inclusion on the trauma list for open reduction and internal fixation or referral to physiotherapy, when formulating clinic letters.

Readability was assessed using the Flesch-Kincaid Grade, the Gunning Fog Index and the SMOG (Simple Measure of Gobbledygook) Index. Each system calculates the reading age based on the average number of syllables per word and the number of words per sentence. The lower the score, the easier the text is to read; Flesch-Kincaid is measured out of 18, while Gunning Fog Index is measured out of 17.[10 11]

Accuracy was assessed by six UK-based independent senior orthopaedic surgeons using a Likert scale from 0 to 10, where 0 represented a completely inaccurate letter or management plan, and 10 indicated full accuracy. Each assessor was blinded to the scores of the other raters and unaware of whether the letter was authored by a human, GPT-3, GPT-3.5 or GPT-4. The reviewers also qualitatively evaluated the letters and management plans for information quality, tone and readability. Outputs were further analysed for overall tone, along with any omissions or insertions noted by the evaluators and reviewed by independent authors, JB and ATP.

An additional analysis evaluated the ability of GPT-4, GPT-3.5 and GPT-3 to develop suitable management plans for each case. For each case, the following prompt was used: 'write an appropriate management plan for the following patient seen in an orthopaedic clinic based on the information provided'. Six independent senior orthopaedic surgeons assessed the management plans for accuracy against current best practices and published guidance, which is detailed in online supplemental table 1 as 'gold standard', using the same Likert scale.

The first response for each prompt was obtained to minimise clinician selection bias and then manually reviewed by six independent, blinded orthopaedic clinicians. To evaluate the consistency of clinician scoring across all management scenarios, an intraclass correlation coefficient (ICC) was calculated using a two-way random-effects model for absolute agreement (ICC (2,k)), suitable for continuous ratings from six independent assessors. ICC values were interpreted based on the classification by Koo and Li,[12] where values <0.5 indicate poor, 0.5–0.75 moderate, 0.75–0.9 good and >0.9 excellent reliability. Statistical analysis was conducted using SPSS V.28 (IBM Corp, Armonk, NY).

Patient and public involvement

None. Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

RESULTS

GPT-4, GPT-3.5 and GPT-3 generated complete clinical letters with a single prompt (see online supplemental appendix 1 for all responses). Without any additional specifications other than those mentioned above, the letters contained blanks to fill in the patient's name and the clinician responsible for drafting the letter.

For readability, all letters were of sufficient quality and were generally rated as accurate. The Flesch-Kincaid Grade Level is a test that measures how difficult a text is to read by assigning it a US school grade level. There was only a slight variation between the LLMs. The mean score for GPT-4 was 9.11, for GPT-3.5 it was 8.77 and for GPT-3 it was 8.47. This metric indicates the approximate US school grade level required to understand the letter, which in both prompts corresponds to the reading level expected for children aged 14–15.

Flesch readability ease, a measure of how easy a piece of text is for GPT-4 to read, was 56.3. For GPT-3.5, the score was 58.2, and for GPT-3, it was 59.3. The SMOG Index, a measure of readability estimating the years of education required to understand a piece of writing, was similar but increased with successive generations of LLMs for GPT-4, GPT-3.5 and GPT-3. The mean index scores for these models were 12.1, 11.6 and 11.4, respectively. Comparison of scores using paired Student's t-test showed small and non-statistically significant differences for any readability metric assessed (table 3).

The written content quality from GPT-3.5 and GPT-3 was inconsistent. In some cases, the letters summarised all content well, with good inference of some relevant information. For example, it was inferred that some occupations were relevant because their injuries could impact their work. Subjectively, this inconsistency appeared reduced in GPT-4.

Clinician-rated accuracy comparisons across 4.0, 3.5 and 3.0 revealed that GPT-4 exhibited the highest accuracy for management plans (9.08/10 (95% CI 8.25 to 9.9)). This represents a statistically significant progression of the ability of a LLM to provide accurate management plans from GPT-3 6.84 (95% CI 5.41 to 8.27), to GPT-3.5 7.63 (95% CI 7.23 to 8.13) to GPT4 (p<0.0001). The accuracy results are summarised in figure 1. Agreement between raters across all 11 clinical scenarios was excellent, with an ICC (2,k) 0.91 (95% CI 0.86 to 0.95, p<0.0001). This indicates high consistency among the six independent assessors when evaluating the accuracy of management plans and clinic letters.

DISCUSSION

Overall, the quality of information provided by the ChatGPT web interface in VFC letters to GPs was high and demonstrated an advanced reading level. The information was accurate in nearly all cases when used to develop management plans for common VFC presentations. GPT-4 produced more detailed, appropriate and less generalised management plans; however, real-world VFC letters often include contextual data relevant to the patient and their presentation that were not tested, such as allergies and additional laboratory results, including

Table 3 Summary of response readability

Metric	Mean GPT-4 response (95% CI)	Mean GPT-3.5 response (95% CI)	Mean GPT-3 response (95% CI)	P value (significance)
Flesch-Kincaid grade level	9.11 (8.79 to 9.43)	8.77 (8.15 to 9.39)	8.47 (7.81 to 9.13)	0.242 (NS)
Flesch readability ease	56.3 (53.96 to 58.64)	58.2 (55.51 to 60.89)	59.3 (54.61 to 63.99)	0.331 (NS)
SMOG Index	12.1 (11.67 to 12.53)	11.6 (11.09 to 12.11)	11.4 (10.72 to 12.08)	0.507 (NS)
General public reach (%)	80.6 (77.91 to 83.29)	80.3 (77.48 to 83.12)	81.2 (77.32 to 85.08)	0.700 (NS)

Readability of letters produced by GPT-3.5, GPT-3 and GPT-4, compared using Flesch-Kincaid grade level, Flesch readability ease, SMOG Index and reach.
GPT, Generative Pre-trained Transformer; NS, not significant; SMOG, Simple Measure of Gobbledygook.

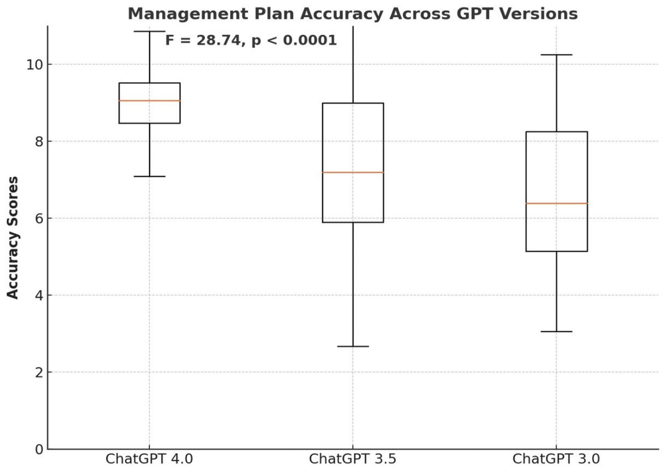

Figure 1 Box and whisker plot of management plan accuracy across GPT versions. Accuracy scores for ChatGPT and GPT-3 generate (A) letters and (B) management plans, independently scored by six senior orthopaedic clinicians. Grey lines show paired prompts. Compared using a paired t-test. ChatGPT, Chat-Generated Pre-trained Transformer; GPT, Generative Pre-trained Transformer.

blood tests, which may influence management plans. These results show that LLMs can summarise clinician notes into readable clinic letters, demonstrating progress in LLM technology, which can, in most cases, accurately manage patients within a simulated VFC setting, based on our UK study. Nevertheless, we must recognise that this was a limited simulated sample. Despite this limitation, there is a clear trend of improvement in management plans with each subsequent generation. With clinician oversight, LLMs could be used to help reduce the administrative workload involved in a VFC environment. Inter-rater reliability was excellent (ICC=0.91), confirming strong agreement among assessors and supporting the reliability of the scoring process. This indicates that the differences in accuracy between model versions were genuine and not due to variability among raters.

Our study examined the progression in the ability of LLMs to generate management plans for common orthopaedic injuries. This builds on previous research exploring ChatGPT's diagnostic capabilities across various specialities.[13–15] Earlier studies on management plan generation in response to cardiovascular symptoms produced plans and treatment protocols consistent with current literature and medical expert opinions.[16] Our data indicated that, although management plans were often appropriate in response to the prompts, they frequently lacked crucial details such as the duration or method of immobilisation. Nonetheless, GPT-4 showed clear progress; although verification is still necessary, the LLM has clearly demonstrated a remarkable capacity to handle simple orthopaedic cases. This advancement suggests that a suitably trained, Health Insurance Portability and Accountability Act-compliant AI LLM could help share the administrative and decision-making workload faced by a consultant in the VFC.

The current literature on AI applications is rapidly evolving. LLMs have demonstrated advantages in streamlining the production of radiology reports and generating outpatient clinic letters following initial consultations in plastic surgery.[17 18] A survey of General Practice revealed that 20% of practitioners used AI tools in clinical practice, with the most common application being the creation of documentation after patient appointments.[19] Recent advances in machine learning and the development of GPT-4.0 have addressed earlier concerns regarding the readability and relevance of information.[20] Data from this study support the prevailing view that clinic letters produced by LLMs maintain a high standard of readability and contain relevant clinical information.

A closely related application of AI leverage is patient information provision, whereby the ChatGPT web interface is used to generate information about conditions and surgical procedures. GPT has demonstrated the ability to produce readable, high-quality information regarding carpal tunnel release surgery and aesthetic plastic surgery by either creating patient information leaflets or integrating risk profiles into clinic letters.[5 21] GPT could also generate patient information leaflets alongside clinic letters to help improve patient understanding.

Research into the use of AI and its role in assisting clinicians in the maxillofacial trauma triage setting has shown the potential of LLMs as valuable tools for supporting clinical decision-making and providing recommendations for multidisciplinary assessment and treatment.[22] However, ongoing supervision and monitoring of LLMs remain necessary at this stage, and further large-scale studies are required to evaluate their efficacy and safety. Recent advances in AI and research into its wider application in radiology for fracture detection highlight its growing role in assisting clinicians with diagnosis and surgical planning. The development of AI models such as deep learning networks has demonstrated the ability to match the accuracy, sensitivity and specificity of reports produced by human radiologists and orthopaedic clinicians,[23] successfully detecting and correctly classifying fractures across various skeletal joints. With the rise of AI and the increasing number of systems trialled and implemented in clinical environments, research emphasises the need for establishing a reporting guideline for early-stage live clinical evaluation of these technologies.[24] DECIDE-AI (Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence) is a new, stage-specific reporting guideline employing a checklist aimed at improving the reporting of proof of clinical utility, human factors, safety and preparedness for large-scale, definitive trials.[25]

Despite their benefits in reducing administrative burden, these tools raise concerns regarding data access, search limitations and ethical considerations. We have previously highlighted the risks of data breaches and the consequences of using patient-identifiable data with LLMs.[4] Therefore, it is crucial to use established AI systems in a non-identifiable manner or to develop a

General Data Protection Regulation compliant, medically trained LLM for clinician use.[26] Implementing in-house versions of open LLMs, rather than closed models, offers significant benefits in data security, as institutions can host LLMs locally, reducing risks related to data privacy and security.[27] From a medicolegal perspective, it is advisable that clinicians use AI tools to support, rather than replace, professional judgement.[28] AI model recommendations can often be difficult to interpret because their internal workings resemble a black box, raising concerns about accountability in cases of misdiagnosis. To address these issues, regulatory frameworks must be established to ensure LLMS meet legal and ethical standards.[22 23] Although GPT performed well in generating clinic letters, it is not a regulated medical device and does not comply with GDPR for accessing patient data. Its use in clinical decision-making remains experimental and primarily demonstrates AI's potential to reduce professionals' workload and improve overall service quality.[29]

The limitations of this study relate to the prompts used, which restrict the quality of the information generated and may lead to dependence on documentation and reporting of radiographic findings. While our letter prompts offer a representative sample of VFC presentations, they are often not tailored to specific diagnoses or scenarios with clinical uncertainty that might be encountered in an in-person setting. However, the very nature of a VFC exposes clinicians to the risk of decision-making based on imaging and documentation in the emergency department. Although Flesch-Kincaid and Gunning Fog indices provide insights into readability and linguistic complexity, they do not assess clinical accuracy. Therefore, evaluation by independent clinicians was essential to determine the clinical utility of each model's output. In future applications, alongside reviews of newer generations of LLMs, we will ensure the use of a SMART prompt structure (Seeker, Mission, AI role, Register, Targeted question) to guarantee that the outputs from the LLMs are clinically relevant and to improve their clarity and completeness. Previous research within head and neck surgery has shown that employing SMART prompt structures significantly enhances the quality of AI chatbot responses, leading to more accurate, complete and relevant information.[30]

This study used a representative sample of prompts presented to multiple LLMs. Senior clinicians evaluated quality in a blinded manner. Future research should monitor AI progress as it engages in machine learning and assess its performance as these systems evolve. Additional integration of AI into VFC should be carried out through a pilot study to ensure safe information delivery and ultimately incorporate patient feedback to ensure satisfaction.

CONCLUSIONS

LLMs are valuable tools for creating high-quality, readable clinical letters for common VFC cases. When properly trained and supervised by experienced clinicians, AI software can help decrease the administrative workload. Further advances in machine learning are needed before these models can generate management plans without supervision; however, our findings emphasise the potential of this technology to reduce the administrative and decision-making burdens on clinicians, particularly in straightforward cases encountered in the VFC.

Author affiliations

[1]University Hospital of Wales, Cardiff, UK
[2]Trauma Unit, Cardiff and Vale University Health Board, Cardiff, UK
[3]Trauma and Orthopaedics, Ysbyty Gwynedd, Bangor, UK
[4]Royal Glamorgan Hospital, Llantrisant, UK
[5]Department of Biomedical Sciences, Cardiff University, Cardiff, UK

Acknowledgements The authors would like to thank the orthopaedic surgeons who generously reviewed and provided feedback on the hypothetical patient cases used in this study. Their clinical insight was invaluable to the development and interpretation of the dataset.

Contributors AS, JB, HJ and ATP contributed to the study's conceptualisation and design, data collection and analysis, drafting of the first manuscript, and subsequent edits and reviews. RA, CK, CJ, FS, TY and OB contributed to data collection and analysis, as well as editing and reviewing the manuscript. ATP is the guarantor of the work and takes full responsibility for the integrity of the study and the manuscript. This study targeted AI efficacy in routine clinical practice. The methods section describes the details of using AI to generate data for analysis. Once the clinic letters were generated as described in the methodology, AI was not used for any aspect of the analysis, write-up or manuscript preparation; minor grammar and spelling checks were performed by Grammarly for accessibility purposes only.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in the Zenodo repository (DOI: 10.5281/zenodo.16962540)

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Francesca Solari https://orcid.org/0000-0002-0853-1485
Arwel Tomos Poacher https://orcid.org/0000-0002-4200-4929

REFERENCES

1 British Orthopaedic Association. BOA statement on virtual fracture clinics. London BOA; 2015. Available: https://www.boa.ac.uk/resource/boa-statement-on-virtual-fracture-clinics.html [accessed 26 Jan 2025]

2 Khan SA, Asokan A, Handford C, *et al*. How useful are virtual fracture clinics?: a systematic review. *Bone Jt Open* 2020;1:683–90.
3 Rhind JH, Ramhamadany E, Collins R, *et al*. An analysis of virtual fracture clinics in orthopaedic trauma in the UK during the coronavirus crisis. *EFORT Open Rev* 2020;5:442–8.
4 Caterson J, Ambler O, Cereceda-Monteoliva N, *et al*. Application of generative language models to orthopaedic practice. *BMJ Open* 2024;14.
5 Brock J, Roberts R, Horner M, *et al*. Artificial Intelligence as a Consent Aid for Carpal Tunnel Release. *Cureus* 2024;16.
6 OpenAI. Introducing ChatGPT. OpenAI Blog. Available: https://openai.com/blog/chatgpt [Accessed 26 Jan 2025].
7 Dubin JA, Bains SS, Chen Z, *et al*. Using a Google Web Search Analysis to Assess the Utility of ChatGPT in Total Joint Arthroplasty. *J Arthroplasty* 2023;38:1195–202.
8 Hofmann HL, Guerra GA, Le JL, *et al*. The Rapid Development of Artificial Intelligence: GPT-4's Performance on Orthopedic Surgery Board Questions. *Orthopedics* 2024;47:e85–9.
9 Smith A, Brock J, Jones H, *et al*. Data from: How have generic large language models progressed in their ability to write clinic letters and provide accurate management plans in the virtual fracture clinic. *Zenodo* 2025.
10 Solnyshkina M, Zamaletdinov R, Gorodetskaya L, *et al*. Evaluating text complexity and Flesch–Kincaid grade level. *J Soc Stud Educ Res* 2017;238–48.
11 Świeczkowski D, Kułacz S. The use of the Gunning Fog Index to evaluate the readability of Polish and English drug leaflets. *Cardiol J* 2021;28:627–31.
12 Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* 2016;15:155–63.
13 Hoppe JM, Auer MK, Strüven A, *et al*. ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis. *J Med Internet Res* 2024;26.
14 Stoneham S, Livesey A, Cooper H, *et al*. ChatGPT versus clinician: challenging the diagnostic capabilities of artificial intelligence in dermatology. *Clin Exp Dermatol* 2024;49:707–10.
15 Huang J, Yang DM, Rong R, *et al*. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *Npj Digit Med* 2024;7:106.
16 Rizwan A, Sadiq T. The Use of AI in Diagnosing Diseases and Providing Management Plans: A Consultation on Cardiovascular Disorders With ChatGPT. *Cureus* 2023;15.
17 Soleimani M, Seyyedi N, Ayyoubzadeh SM, *et al*. Practical Evaluation of ChatGPT Performance for Radiology Report Generation. *Acad Radiol* 2024;31:4823–32.
18 Ali SR, Dobbs TD, Hutchings HA, *et al*. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 2023;5:e179–81.
19 Blease CR, Locher C, Gaab J, *et al*. Generative artificial intelligence in primary care: an online survey of UK general practitioners. *BMJ Health Care Inform* 2024;31.
20 Ulusoy I, Yılmaz M, Kıvrak A. How Efficient Is ChatGPT in Accessing Accurate and Quality Health-Related Information? *Cureus* 2023;15.
21 Roberts RHR, Ali SR, Dobbs TD, *et al*. Can Large Language Models Generate Outpatient Clinic Letters at First Consultation That Incorporate Complication Profiles From UK and USA Aesthetic Plastic Surgery Associations? *Aesthet Surg J Open Forum* 2024;6.
22 Frosolini A, Catarzi L, Benedetti S, *et al*. n.d. The Role of Large Language Models (LLMs) in Providing Triage for Maxillofacial Trauma Cases: A Preliminary Study. *Diagnostics (Basel)*14:839.
23 Kutbi M. Artificial Intelligence-Based Applications for Bone Fracture Detection Using Medical Images: A Systematic Review. *Diagnostics (Basel)* 2024;14:1879.
24 Vasey B, Novak A, Ather S, *et al*. DECIDE-AI: a new reporting guideline and its relevance to artificial intelligence studies in radiology. *Clin Radiol* 2023;78:130–6.
25 Vasey B, Nagendran M, Campbell B, *et al*. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377.
26 General Data Protection Regulation (GDPR). Article 22: automated individual decision-making, including profiling. GDPR.eu, Available: https://gdpr.eu/article-22-automated-individual-decision-making/ [Accessed 26 Jan 2025].
27 Dennstädt F, Hastings J, Putora PM, *et al*. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ Digit Med* 2025;8:143.
28 Kapsali MZ, Livanis E, Tsalikidis C, *et al*. Ethical Concerns About ChatGPT in Healthcare: A Useful Tool or the Tombstone of Original and Reflective Thinking? *Cureus* 2024;16.
29 Haupt CE, Marks M. AI-Generated Medical Advice—GPT and Beyond. *JAMA* 2023;329:1349.
30 Vaira LA, Lechien JR, Abbate V, *et al*. Enhancing AI Chatbot Responses in Health Care: The SMART Prompt Structure in Head and Neck Surgery. *OTO Open* 2025;9.