Digital Health Trends Journal. 2026;2(1):23-S1.
doi:10.34172/dhtj.10

https://dhtjournal.ae

Supplementary file 1

Table S1. Characteristics of the included studies

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
Riley J. Lyons 2024 USA [1]		44 clinical vignettes/ Single-centre ophthalmology context	ChatGPT & Bing Chat → LLM, NLP, deep learning; WebMD → rule-based	Initial triage, symptom checking (ophthalmic conditions) Initial triageoklInitial triage, symptom cheking (ophthalmic condition)	Accuracy (Top 3 diagnosis): Trainees 95%, ChatGPT 93%, Bing Chat 77%, WebMD 33%; Triage urgency accuracy: Trainees 86%, ChatGPT 98%, Bing Chat 84%; No grossly inaccurate statements (ChatGPT); Bing Chat overestimates urgency	Simulated vignettes (not real patients), single-centre, limited to ophthalmology, small sample size	Moderate
Wolmer 2023 USA [2]	Observational / scenario-based evaluation	60 clinical scenarios/ Plastic surgery websites (real-world digital platforms)	Likely NLP/ML-based/Web-based chatbots embedded in websites/AI chatbots on top-ranking plastic surgery websites	Initial triage, classification accuracy, patient interaction quality	Triage performance: Emergent cases misclassified as urgent (sensitivity 20%, NPV 0.71, false negative rate 80%), moderate agreement with physician (Cohen's kappa = 0.47); Usability: Correct classifications → 60.8 vs misclassified → 49.1; Over 50% required human escalation; Reliance on templated administrative language	Limited to plastic surgery domain, scenario-based evaluation (not real patients), AI techniques not fully detailed, performance varies by platform	Low
Taylor Kring 2025 USA	Cross-sectional / scenario-based evaluation	25 patient-like symptom queries/Head and neck cancer patient	NLP/LLM-based AI/hatGPT, Google Gemini, Microsoft Copilot, Open Evidence/AI	Symptom checking, patient information/triage support	Quality (DISCERN score): Microsoft Copilot 41.40 > ChatGPT, Google Gemini, Open Evidence; Readability (SMOG score): Copilot 12.56 < others;	Limited to head and neck cancer symptom queries, scenario-based (not real patients), AI techniques not fully detailed, focus on	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
[3]		information context (digital platforms)	chatbots for patient education		Significant variability in quality and readability across platforms	readability and quality rather than clinical diagnosis accuracy	
İbrahim Sarbay 2023 Turkey [4]	Preliminary, cross-sectional, scenario-based	50 case scenarios/Emergency medicine context, simulated scenarios	Supervised machine learning, NLP /Open access NLP-based chatbot/ ChatGPT	Emergency triage prediction	Overall performance: Sensitivity 57.1%, Specificity 34.5%, PPV 38.7%, NPV 52.6%, F1 score 0.461; High acuity (ESI-1 & ESI-2): Sensitivity 76.2%, Specificity 93.1%, PPV 88.9%, NPV 84.4%, F1 score 0.821; Cohen's Kappa with EM specialists 0.341; ROC AUC 0.846 for high-acuity cases	Scenario-based (not real patients), preliminary study, single AI system, limited generalizability to other triage levels, no real-world clinical outcomes	Low
Inès Schumacher 2025 Europe [5]	Cross-sectional, scenario-based evaluation	100 hypothetical ophthalmic cases/Ophthalmic emergency department	LLM,NLP/ Web-based/clinical platform/ Customized ChatGPT-based chatbot	Emergency triage (ophthalmology)	Accuracy/Agreement: Cohen's kappa with ophthalmologists: 0.737–0.751; Fleiss' kappa overall: 0.79; No significant difference in grade distribution vs human graders (p=0.967); Bootstrap analysis confirms comparable performance	Scenario-based (not real patients), limited to ophthalmology, single AI system, potential generalizability limits	Moderate
Jonathan C. Tsui 2023 USA [6]	Cross-sectional, scenario-based evaluation	10 patient-like prompts (each submitted 3 times)/ Ophthalmology patient inquiry context	LLM/NLP (GPT-based)/ OpenAI online chatbot / ChatGPT (Feb 13 version)	Symptom triage, patient information	Precision & Suitability: 8/10 sets graded both precise and suitable; 2/10 sets imprecise and unsuitable; Fleiss' kappa inter-rater reliability: precision 0.28, suitability 0.04; No follow-up questions or source citations	Scenario-based (not real patients), small sample size, limited to ophthalmology, single version of ChatGPT, results may not generalize to other versions/platforms, scripted prompts only	Low
Yue You,	Mixed-methods	Not specified / General	AI-enabled chatbot-based	Symptom checking, self-	Existing CSC apps lack comprehensive support	Does not evaluate clinical outcomes;	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
2020, USA [7]	study: feature review, user review analysis, and interview study	consumer health context (AI-enabled CSC apps)	symptom checkers (CSC apps); Mobile & web-based apps; NLP + AI-based conversational algorithms	triage, diagnosis support	for the entire diagnostic process; Users report insufficient support for complete medical history, flexible symptom input, comprehensible questions, and broader disease coverage; Recommendations provided for improving conversational design and user experience	Limited to app functionality and user perceptions; AI techniques not fully detailed; Lack of real-patient data; Generalizability to clinical settings is limited	
Prabod Rathnayaka 2022 [8]	Pilot participatory evaluation study	Pilot study / Remote mental health monitoring context; working-age individuals	AI-enabled mental health chatbot with cognitive skills, based on Behavioural Activation (BA) therapy; Cross-platform smartphone app; AI + NLP techniques for personalized support and remote monitoring	Mental health support, behavioural activation therapy, personalised intervention, remote health monitoring	Pilot evaluation confirmed chatbot's effectiveness in providing recurrent emotional support, personalized behavioural activation assistance, and continuous remote monitoring for mental health	Small sample size; Pilot study limits generalizability; Effectiveness evaluated only in a controlled setting; No long-term clinical outcomes; Limited comparison with other therapy-based chatbots	Moderate
Anil Erkan 2024, Turkey [9]	Cross-sectional evaluation study	3 AI chatbots evaluated / Urogenital cancer treatment context	ChatGPT, Google Gemini, Microsoft Copilot; Web-based conversational agents; AI + NLP-based chatbots	Patient education, treatment decision support, cancer information provision	Quality of information: DISCERN score → ChatGPT: 41, Gemini: 42 (moderate), Copilot: 35 (low); PEMAT-P Understandability: Low across all (≈40%); PEMAT-P Actionability: Gemini moderate (60%), ChatGPT & Copilot low (40%); Readability: Coleman-Liau index = ChatGPT 16.9, Gemini 17.2, Copilot 16 → above	Does not evaluate clinical effectiveness or patient outcomes; Limited to three chatbots only; Focused only on urogenital cancers; Lacks assessment of stage-specific treatment options; Results may not generalize to other diseases or platforms	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
					college level; Overall: Limited reliability, moderate information quality, and poor readability		
Tze Chin Tan 2023, Singapore [10]	Cross-sectional survey study	200 patients (100 initial consultations, 100 follow-up visits) / Outpatient rheumatology referral center	Chatbot tailored for Autoimmune Inflammatory Rheumatic Diseases (AIIRDs); Platform not explicitly detailed; Likely AI/NLP-based for symptom screening and patient education	Symptom screening, patient education, patient engagement	Patient acceptability: High (mean scores 4.01–4.41/5); Willingness to reuse: Higher in follow-up patients (P=0.01); Comfort with chatbot diagnosis: Increased after physician consultation (P<0.001); Positive reception across sex, education level, and diagnosis category	Single-center study; Survey-based perceptions only, no clinical effectiveness measured; Platform/technical details of the chatbot not fully described; No long-term follow-up on engagement or outcomes; Generalizability to other populations limited	Low
Daniel Mendonça de Moura 2024, Brazil [11]	Cross-sectional comparative study	11 fictional pulpal and periradicular disease cases / Dental diagnostic context	AI chatbots: ChatGPT 3.5, ChatGPT 4.0, Bard, Bing; Web-based AI/NLP chatbots	Diagnosis support, treatment recommendation	Diagnostic accuracy: Bing 86.4%, ChatGPT 4.0 85.3%, ChatGPT 3.5 46.5%, Bard 28.6%; Treatment recommendation accuracy: ChatGPT 4.0 94.4%, Bing 93.2%, ChatGPT 3.5 86.3%, Bard 75%; Overall consistency rate 98.29%; Language and text order did not significantly affect accuracy; Portuguese cases prompted more requests for additional information	Small sample size of cases; Scenario-based (not real patients); Limited to dental context; Responses require critical interpretation by clinicians; Findings may not generalize to other diseases or real-world clinical settings	Low
Akanksha Singh 2023	Cross-sectional study	82,222 chat sessions / Public users in South Carolina,	AI chatbot for COVID-19 symptom checking and education; Web-	Symptom checker, patient follow-up, education, referral to care	High perceived helpfulness among high-risk users; Symptom reporting, risk assessment, and follow-	Observational, no clinical outcomes measured; Limited to one US state; Only COVID-19 context;	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
USA [12]		Prisma Health system	based platform; NLP and AI-driven conversational recommender system		up care options (telehealth, in-person visits, vaccination) increased engagement; Older age (>65), comorbidities, and recent COVID-19 contact associated with higher chatbot satisfaction	Engagement metrics based on user-reported perception, not objective clinical endpoints; No long-term follow-up	
Byeong Jin Ye 2020 South Korea [13]	Pilot study	23 healthcare providers surveyed + 6 nurses interviewed / Workers' general health examination follow-up context	AI chatbot implemented on AWS EC2, using KakaoTalk and Web Chat as user channels; Database-driven chatbot for follow-up management	Patient follow-up, occupational health management, general health exam follow-up	Effectiveness: 91.3% rated need for chatbot as very high; Usability: 47.8% rated usability not high; Overall satisfaction: 60.9%; Nurses appreciated chatbot for accessibility and supporting explanation of results	Small sample size; Pilot study limits generalizability; Evaluated mainly healthcare providers and nurses, not patients; Usability concerns reported; No clinical outcome data; Limited long-term evaluation	High
Stephen R Ali 2022 UK [14]	Early adoption / Use-case descriptive study	Not specified / Microsurgery department, free flap monitoring context	FlapBot, a chatbot to support clinical decision-making; Digital conversion of paper-based flap monitoring charts; Platform/AI techniques not fully detailed	Clinical decision support, patient follow-up, escalation of care	Facilitates early recognition and escalation in free flap monitoring; Supports decision-making and potentially improves timeliness of interventions	Small-scale descriptive study; No quantitative evaluation of clinical effectiveness; Sample size not specified; AI/technical details not provided; Findings may not generalize to other departments or settings	Moderate
Antoine Piau 2019, France [15]	Pilot feasibility study	9 unselected older patients (mean age 83) / Outpatient cancer follow-up at home	Smartphone-based semi-automated messaging Chatbot; Supports remote monitoring, collects patient-reported outcomes;	Patient follow-up, remote monitoring, chemotherapy adherence	Compliance: 86%; Questionnaire completion: 100% answers, avg 3.5 min per questionnaire; Free-text used in 58%; Detected health (e.g., fever) and adherence (e.g., blood test) issues; Feasible and	Very small sample size; Pilot study, preliminary results; No control group; Limited generalizability; AI or advanced analytics not applied; Long-	Low

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
			Platform integrated with regional cancer network; No AI/ML techniques specified		acceptable integration into workflow Chatbot integration into healthcare system is feasible and acceptable; Early identification of health/adherence issues supports timely intervention; Relies on familiar technology for seamless adoption	term effectiveness not yet evaluated	
Taylor N. Stephens 2019 USA 16	Feasibility study, pilot trial	3 adolescents (Mage = 15.2, range 9.8–18.5; 57% female) / Pediatric weight management and prediabetes care	Type: AI behavioral coaching chatbot ("Tess") Platform: SMS text messaging, Facebook Messenger AI techniques: Rule-based + NLP conversational agent with capacity for continuous learning	Treatment adherence, behavior change support, patient follow-up, wellness coaching	Clinical effectiveness: 81% of patients reported positive progress toward goals Patient satisfaction: 96% rated chatbot as useful Usability & engagement: 4,123 messages exchanged, high engagement through preferred channels Accuracy: Not directly evaluated (focus was on support/engagement, not diagnostic accuracy) Engagement metrics: Sustained interaction and message volume indicate high adherence Summary conclusion: Tess is feasible, acceptable, and beneficial as an adjunct to pediatric obesity/prediabetes care, extending therapeutic interaction outside office hours	Small sample size (n=23) Single-site pilot study Short-term follow-up Focused only on feasibility and engagement, not long-term clinical outcomes Further iterations needed to enhance user experience and scalability	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
Friederike Eva Roch 2025 Germany [17]	Comparative study (chatbot responses vs clinical guidelines)	Not patient-based; dataset = structured questionnaire + treatment algorithms / Orthopedic, trauma care context	Type: Free chatbots powered by LLMs Platforms: ChatGPT, Claude, Perplexity AI techniques: Large Language Models (machine learning, NLP)	Symptom checker, diagnosis/treatment support, guideline adherence	Clinical effectiveness / accuracy: -All chatbots met the 60% pass threshold vs guidelines - Perplexity performed best on questionnaire ($p < 0.001$) - ChatGPT scored highest on algorithm evaluation ($p = 0.023$) Patient satisfaction: Not assessed (focus was expert evaluation) Usability: Not directly measured Engagement metrics: Not applicable (evaluation study only) Summary conclusion: Chatbots provide useful recommendations broadly aligned with guidelines but miss crucial details → cannot replace professional consultation yet	- Evaluation limited to LAS (ankle sprains) only - No real patient interaction (simulated questionnaire approach) - Small scope (only 3 chatbots compared) - Did not assess patient-facing usability, satisfaction, or outcomes - Potential bias as guidelines used as sole benchmark	Moderate
Joe Hasei 2025 Japan [18]	Pilot feasibility study	5 pediatric & AYA cancer patients / Oncology care (supportive mental health in pediatric & young adult cancer)	Type: Generative AI chatbot Platform: Messaging platform (not specified, likely mobile/online) AI techniques: GPT-4 (LLM,	Patient follow-up, treatment adherence (engagement), psychological support	Clinical effectiveness: 4/5 patients reported reduced anxiety & stress Patient satisfaction: Positive — 80% shared concerns with chatbot they had not told providers Usability: Average use every 2–3 days, ~10 min	- Very small sample size (n=5) - Short duration (2 weeks) - No control group for comparison - Self-reported outcomes only	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
			NLP, generative AI)		per session; 24/7 availability valued Accuracy: Not main focus (empathetic support, not medical accuracy) Engagement metrics: Consistent engagement, disclosure of sensitive issues, increased motivation Summary conclusion: GPT-4 chatbots showed feasibility and promise as complementary psychological support tools for pediatric & AYA cancer patients. They improved motivation, reduced anxiety/stress, and filled gaps between clinic visits.	- Results not generalizable without larger trials	
anya Melnik 2023 USA 19	Observational / retrospective analysis of patient–chatbot interactions	3,248 patients (6,262 comments analyzed) / Remote patient monitoring (RPM) program at M Health Fairview, COVID-19 patients	Type: Chatbot for Remote Patient Monitoring (COVID-19) Platform: Integrated into RPM program (exact platform not specified) AI techniques: Topic modeling using LDA (Latent Dirichlet Allocation) and CorEx for	Remote monitoring, symptom tracking, diagnosis support	Clinical effectiveness: Not directly tested; showed feasibility of semi-automated curation of patient messages Patient satisfaction: Not reported Usability: Chatbot collected a large volume of meaningful patient-generated data Accuracy: Topic assignment accuracy 72.8% (LDA) and 88.2% (CorEx) Engagement metrics: 6,262 patient comments	- No evaluation of patient outcomes or satisfaction - Limited to COVID-19 context at one health system (M Health Fairview) - Retrospective design, not prospective validation - Focus on text mining, not clinical impact	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
			clinical content curation		from 3,248 individuals during COVID-19 monitoring Summary conclusion: Semi-automated curation using AI topic modeling can efficiently process patient–chatbot communications in RPM, identifying key symptom trends and correlating with real-world events (e.g., test availability).		
Niv Ben-Shabat 2022 Israel 20	Clinical vignettes study (simulation)	28 clinical vignettes, entered by 3 medical students / Simulated patient–chatbot interactions (not real patients)	Type: Chatbot-based symptom checkers (8 platforms evaluated) Platform names: Kahun, Your.MD, and 6 others (not specified in abstract) AI techniques: Rule-based / NLP (exact algorithms not detailed, but symptom-checker chatbot systems)	Symptom checker / data gathering for diagnosis support	Clinical effectiveness: Not directly assessed; performance measured in data gathering Patient satisfaction: Not measured (study used simulations) Usability: Indirectly evaluated via efficiency of data collection Accuracy: Recall rate 0.32 overall; best platform (Kahun) had 0.51 Efficiency: Overal 0.46; Kahun most efficient (0.74) Engagement metrics: Not applicable (no real patients) Summary conclusion: Current symptom checkers show limited ability to gather complete clinical data. Kahun outperformed others but	- Use of simulated vignettes, not real patients - Limited to 28 cases, which may not represent real-world variability - Only data gathering assessed (not diagnostic accuracy or clinical outcomes) - AI methods used by platforms not transparent	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
					overall performance was suboptimal.		
Xiangmin Fan, 2021 China 21	Case study using system log analysis	47,684 consultation sessions from 16,519 users / Real-world online consultations with a self-diagnosis chatbot (general public users in China)	Type: Self-diagnosis health chatbot Platform: Widely deployed chatbot in China (not named) AI techniques: Likely NLP-based conversational AI (details not specified in abstract)	Symptom checker / self-diagnosis	Clinical effectiveness: Mixed – chatbot provided diagnostic suggestions but users reported perceived inaccuracies Patient satisfaction: Issues raised: insufficient actionable info, perceived inaccuracies Usability: Dropouts during sessions, onboarding issues Accuracy: Users perceived diagnostic suggestions as sometimes inaccurate (no quantitative accuracy provided) Engagement metrics: Large dataset (16,519 users; 47,684 sessions) but dropout was common Summary conclusion: Chatbots have potential for scalable, patient-centered self-diagnosis but suffer from user trust, accuracy, and engagement problems.	- High dropout rates during chatbot use - Misuse of chatbot (fake queries by users) - Perceived low accuracy of diagnostic suggestions - Lack of actionable information - No clinical validation of chatbot's diagnostic performance	Moderate
Caretia J 2024 USA 22	Clinical vignettes study	40 clinical vignettes/ Two tertiary care institutions; evaluated by 3 fellowship-	Type: ChatGPT-3.5 chatbot Platform: OpenAI ChatGPT AI techniques: Large Language	Diagnosis support, treatment recommendations, patient education	Clinical effectiveness: 95% accuracy in first-line treatment recommendations (per NCCN), but 55% incorrect staging; neck dissection omitted in	- Inaccurate tumor staging (TNM) - Omission of critical treatment (e.g., neck dissection) - Over-treatment recommendations	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
		trained head & neck surgeons	Model (LLM), NLP, machine learning–based generative AI		50% of cases; 40% unnecessary treatments suggested Patient satisfaction: Not directly assessed (implication for patient education noted) Usability: Provides quick responses, but requires expert oversight Accuracy: High for initial treatment, poor for staging and surgical details Engagement metrics: Not reported Summary conclusion: ChatGPT shows promise for patient education and improving health literacy but is unsafe for standalone clinical decision-making due to errors in staging and treatment recommendations.	(~40% cases) - Risk of misleading patients/trainees without expert oversight - Limited to vignette study (no real patient data)	
Jonathan Shapiro 2024 Israel 23	Explorative study	Not explicitly reported (pilot/explorative nature) / Teledermatology consultations	Type: GPT-based chatbot ("Dr. DermBot") Platform: Custom prototype for teledermatology AI techniques: Generative Pre-trained Transformer (NLP, large language model); integration with image analysis	Anamnesis, diagnosis support, treatment planning, teledermatology consultations	Clinical effectiveness: Promising performance in enhancing consultation quality, diagnosis precision, and treatment personalization Patient satisfaction: Not directly measured, but accessibility for underserved populations emphasized Usability: Demonstrated potential for autonomous consultations	- Ethical and legal concerns (privacy, regulatory compliance) - Need for validation in real-world clinical trials - Risk of over-reliance without dermatologist oversight - No reported patient-level outcomes in this exploratory study	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
			AI for dermatology		Accuracy: Enhanced diagnostic accuracy when combined with teledermatology workflows Engagement metrics: Not reported Summary conclusion: AI-driven chatbots (like DermBot) can broaden access to dermatologic care, improve diagnostic precision, and support clinicians, but need validation in real-world clinical practice.		
ederico Guede-Fernández 2024 Portugal 24	ilot crossover trial	27 patients / Postoperative cardiac surgery follow-up	Type: Text-based chatbot integrated with remote patient monitoring (RPM) Platform: Mobile SMS / app-based AI techniques: Not explicitly described; functions include patient reporting and therapeutic dose adjustment guidance	atient follow-up, treatment adherence, remote monitoring for anticoagulation therapy	linical effectiveness: Improved median time in therapeutic range (TTR) during RPM periods compared with SOC Patient satisfaction: High trust and satisfaction reported by patients and clinicians Usability: Effective integration with Coaguchek© device and mobile reporting Accuracy: TTR values suggest clinically relevant improvement Engagement metrics: 27 patients actively engaged, with successful reporting and dose adjustments Summary conclusion: Portable coagulometers plus chatbot-based RPM can enhance anticoagulation	- Small sample size (27 patients) - Short-term follow-up (12 months crossover) - Cost differences noted depending on RPM timing - Generalizability limited; further larger trials needed	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
					management after cardiac surgery, improve patient experience, and offer a promising alternative to standard care.		
ohanna Habicht 2024 UK [25]	Multisite observational study	129,400 patients / NHS mental health services across England	Type: Personalized self-referral chatbot Platform: Web-based/NHS digital service AI techniques: Natural language processing (NLP) used to analyze qualitative feedback and interact with users	Patient self-referral, mental health access, treatment engagement	Clinical effectiveness: Increased patient referral volume (15% vs 6% in control) Patient satisfaction: Not quantitatively reported, but qualitative feedback indicated positive engagement Usability: High usability suggested by large-scale engagement (42,332 feedback responses analyzed) Accuracy: Not a diagnostic tool; accuracy not applicable Engagement metrics: Increased referrals among minorities (nonbinary: +179%, ethnic minorities: +29%) Summary conclusion: Personalized AI chatbot improved accessibility and equity in mental health referrals, particularly benefiting underserved and minority populations.	- Observational design, no randomized control - Effectiveness limited to referral stage, not treatment outcomes - Feedback analysis relies on self-reported qualitative data - Findings may be specific to NHS context; generalizability to other systems uncertain	Moderate
Stephanie Greer 2019 USA	andomized controlled feasibility trial	45 young adults (36 women; experimental: n=25, control:	Type: Vivibot chatbot delivering positive psychology skills	ental health support, psychosocial well-being, anxiety reduction, post-cancer recovery	Clinical effectiveness: Trend-level reduction in anxiety in experimental group vs control (effect size 0.41, P=0.09) Patient satisfaction:	- Small sample size lir - Trend-level effects; r - Short intervention du	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
[26]		n=20) / Post-cancer treatment follow-up, online via Facebook Messenger	Platform: Facebook Messenger AI techniques: Not explicitly stated as machine learning; human-centered design used; automated content delivery through chatbot interface		Rated helpful (mean 2.0/3), likely to recommend (mean 6.9/10) Usability: High engagement (average 74 minutes across 12 sessions) Accuracy: Not applicable (psychosocial intervention, not diagnostic) Engagement metrics: Greater anxiety reduction with more sessions; open-ended feedback highlighted nonjudgmental nature Summary conclusion: Vivibot is a feasible and acceptable method for delivering positive psychology interventions to young adults after cancer treatment, supporting anxiety reduction.	- Limited generalizability; only youn - No significant effects on depressio	
Matthew X Luo 2023 USA [27]	Observational / comparative study using a curated questionnaire	Pathology faculty (number not specified), research-prepared residents, unprepared residents, and AI chatbot / Genitourinary treatment	Type of chatbot: AI chatbot (OpenAI ChatGPT, January 30, 2023 release) Platform: ChatGPT AI techniques used: Large language model, natural language processing (NLP)	Diagnosis support: Evaluated answers to real-world clinical questions	Clinical effectiveness / accuracy: ChatGPT scored 4.10 vs faculty 4.75; comparable to research-prepared residents Usability / engagement: Not explicitly measured Patient satisfaction: Not applicable Summary conclusion: ChatGPT provides clinically relevant and reasonably accurate	Chatbot cannot provide references to support its answers Small sample of faculty and residents; number not fully specified Focused on a specific treatment planning context; may not generalize to broader clinical applications	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
		planning conference			answers compared with trained human faculty, but lacks reference support for integration into decision-making		
Hamza Ejaz 2024 UK 28	Observational survey study with qualitative analysis	Sample size: 21 surgeons tested Flapbot and completed surveys (from 42 who agreed) Healthcare setting: Plastic and reconstructive surgery, global survey	Type of chatbot: Flapbot Platform: Google DialogFlow AI techniques used: NLP-based conversational agent, rule-based validation	Remote monitoring: Free-flap post-operative monitoring	Clinical effectiveness / accuracy: Not directly measured, but content validity indices (I-CVI, S-CVI) indicated high relevance (I-CVI >0.78 for 9/13 items; S-CVI = 0.82) Usability: System Usability Score (SUS) = 68 (average usability) Engagement metrics: Survey responses and qualitative thematic feedback Summary conclusion: Flapbot is a valid and moderately usable tool for free-flap monitoring, with potential for broader clinical use after improvements	Sample size relatively small (21 surgeons completed survey) Survey-based study; no direct patient use or clinical outcomes measured Usability only average; improvements needed for global scalability Dependence on digital tools in clinical practice may raise concerns	Moderate
Chun-Chia Chen 2024 Taiwan 29	Case study of a telemedicine diagnostic system	Not specified / Telemedicine for wound care (pressure injuries)	Type of chatbot: ChatGPT integrated in telemedicine platform Platform: Front-end web interface with responsive design	Diagnosis support: Pressure injury classification and severity assessment Patient follow-up: Indirectly, through real-time recommendations	Clinical effectiveness / accuracy: F1 score of 0.9238 for pressure injury classification Usability / engagement metrics: Chatbot provides immediate guidance for users, supporting teleconsultation; no	Small-scale case study; actual patient usage data not detailed Focused on a single clinical domain (pressure injuries) Usability, patient satisfaction, and	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
			AI techniques used: YOLOv7 for object detection and wound classification Large Language Model (ChatGPT) for conversational interface and guidance	for medical assistance	formal usability metrics reported Summary conclusion: The system successfully integrates object detection and a generative AI chatbot to support real-time diagnosis and guidance in pressure injury management. It demonstrates high classification accuracy and can guide patients toward appropriate medical assistance.	engagement not rigorously evaluated	
Stefanie Maria Jungmann 2019 Germany 30	Comparative case study	6 users (2 psychotherapists, 2 psychology students, 2 laypersons); each evaluated 20 case vignettes /Diagnostic support for mental	Type of chatbot: Ada-Your Health Guide (health app / conversational AI) Platform: Mobile health app AI techniques used: Rule-based symptom assessment with algorithmic diagnosis	Screening and diagnosis of mental disorders in adults, adolescents, and children	Clinical effectiveness / accuracy: Moderate overall diagnostic agreement: kappa = 0.64 (adults), 0.40 (children/adolescents) Psychotherapists achieved higher agreement (kappa = 0.78 adults, 0.53 children/adolescents) Laypersons performed worst (kappa = 0.29 children/adolescents) Usability / engagement metrics: Average 34 questions per assessment, 7 minutes to complete Patient satisfaction: Not assessed	Small user sample and pilot nature Diagnostic accuracy is highly dependent on user expertise Limited applicability for pediatric/adolescent cases Evaluated using case vignettes rather than real patients	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
					Summary conclusion: Ada can support diagnostic screening in adults and potentially assist clinicians, but diagnostic accuracy is user-dependent. Improvements are needed for childhood/adolescent mental disorder screening.		
Seray Gizem Gur Ozcan 2025 Türkiye 31	Comparative evaluation study	4 AI chatbots; evaluated responses to top questions on contrast-associated acute kidney injury from Google Trends (Jan 2022–Jan 2024) / Information provision for patients regarding post-contrast AKI	Type of chatbot: AI chatbots for health information (ChatGPT, Gemini, Copilot, Perplexity) Platform: Web-based AI services AI techniques used: Large language models / generative AI	Diagnosis support: Information on diagnosis of post-contrast acute kidney injury Prevention / treatment guidance: Evaluated educational support provided by chatbots Other domains: Patient education	Clinical effectiveness / accuracy: DISCERN scores: Perplexity "good"; ChatGPT, Gemini, Copilot "average" Readability (Coleman-Liau index) >11, indicating high complexity Usability / engagement metrics: Understandability and applicability scores were low across all chatbots Likert scale ratings favorable Patient satisfaction: Not assessed directly Summary conclusion: Chatbots provide potentially useful information but content is complex and may be difficult for patients to understand; improvements needed for	Study focused on information quality, not real patient outcomes Limited generalizability due to evaluation based on Google Trends questions AI chatbot responses may vary over time (dynamic outputs) Readability too high for general population; practical patient use may be limited	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
					readability and applicability		
Marcelo Santos Coelho 2025 Brazil [32]	Pilot randomized controlled study	24 undergraduate dental students (22 completed) / Undergraduate dental education	Type of chatbot: Educational chatbot for pulpal and periapical diagnosis Platform: Telegram Messenger AI techniques used: Not specified; likely rule-based for educational delivery	Teaching diagnosis in dental health	Clinical effectiveness / knowledge acquisition: Both lecture and chatbot improved test scores significantly; no significant difference between groups Usability / engagement metrics: Chatbot rated 4.95/5 for ease of use Perceived as more fun and simpler than the lecture Patient satisfaction: Not applicable (student feedback collected instead) Summary conclusion: Chatbot is as effective as a lecture in delivering basic diagnostic content. Students found the chatbot more engaging, but interactive lectures are better for in-depth understanding.	Small sample size (pilot study) Short-term assessment; long-term knowledge retention not evaluated Chatbot does not replace faculty in discussions or complex content explanation AI capabilities not fully explored; mainly delivery platform	Moderate
Gemma Sharp 2025 Australia [33]	randomized controlled trial	60 participants (30 in chatbot group, 30 in control group)	Type of chatbot: ED ESSI (Eating Disorder Electronic Single-Session Intervention)	Patient follow-up / treatment adherence: Early intervention support and motivation for treatment	Clinical effectiveness: Significant reductions in eating disorder pathology (P=.003) Reduced psychosocial impairment (P=.008),	Sample size relatively small (pilot-scale RCT) Short-term follow-up; longer-term effectiveness unknown	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
		/ Participants on waitlists for eating disorder treatment	Platform: Web-based, rule-based chatbot AI techniques used: Rule-based; no mention of machine learning or NLP		depression (P=.002), anxiety (P=.040) Increased confidence in ability to change (P<.001; Cohen d=0.74) Patient satisfaction / usability: Chatbot rated as "excellent" on the System Usability Scale Engagement metrics: 93% of participants in chatbot group entered treatment by 3 months, vs 70% in control (P=.042) Summary conclusion: ED ESSI is an effective, accessible, and scalable early intervention for individuals awaiting eating disorder treatment; benefits sustained up to 3 months	Participants limited to those on waitlists; generalizability may be restricted Chatbot is rule-based; may not adapt to complex or unanticipated responses	
Sainan Zhang 2024 Korea [34]	Experimental/validation study	Not fully specified for all users / Household / patient self-assessment	Type of chatbot: Chat Ella, diagnostic chatbot Platform: Dialog-based interface (user-friendly interface) AI techniques used: GPT-2 large language model, transfer learning, fine-tuning, deep learning	Symptom checker: Yes – assesses user-reported symptoms to predict chronic diseases Diagnosis support: Yes – predicts 24 common chronic diseases Patient follow-up / treatment adherence: Not applicable	Clinical effectiveness: Accuracy 97.50%, AUC 99.91% Patient satisfaction / usability: 68.7% approved the system, 45.3% found it made daily consultations more convenient Summary conclusion: Chat Ella provides a highly accurate, user-friendly auxiliary diagnostic tool for chronic diseases; suitable for household use to	Limited reporting on sample size and generalizability Focused on 24 common chronic diseases only User satisfaction measured, but long-term clinical outcomes or integration with professional care not evaluated Real-world performance outside	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
					support symptom-based assessments	controlled validation not assessed	
Shameek Ghosh 2018 Australia 35	Experimental/validation study	Two sets of patient test cases (exact number not reported) / Primary care / pre-assessment context	Type of chatbot: Quro, symptom-checker and triage chatbot Platform: Natural language dialogue system AI techniques used: Natural language processing (NLP), rule-based or ML approach for condition prediction	Triage: Yes – predicts urgency and guides pre-assessment Symptom checker: Yes – predicts user conditions based on symptoms Diagnosis support: Indirect – pre-synopsis provided,	Clinical effectiveness: Precision of prediction: 0.82 Summary conclusion: Quro demonstrates that a personalized conversational chatbot can support symptom assessment and triage in primary care, enabling patient pre-assessment without cumbersome forms	Sample size and diversity of test cases not clearly defined Usability and patient satisfaction metrics not reported Real-world deployment and clinical outcomes not evaluated Limited to initial symptom assessment, not full diagnosis	Low
Gemma Sharp 2025 Australia 36	Qualitative study	17 participants: 10 adults with eating disorders + 7 psychologists; / Setting: online interviews & workshops	Conversational AI-driven chatbot prototype; Co-designed for delivering single-session interventions; Platform not specified; Designed for empathetic tone, safety, and structured therapeutic content	Mental health support, patient follow-up, treatment gap management	Positive feedback on chatbot's design, structure, and potential usability; Key improvements achieved through iterative co-design; Identified four major themes for optimization: conversational tone, safety/risk management, user journey, and structured content Co-designing with end-users and psychologists improved feasibility and acceptability; Chatbot could help reduce treatment gaps for eating	Prototype-only evaluation; No clinical efficacy or longitudinal outcome testing; Limited sample size and geographic restriction; Concerns remain about chatbot's ability to fully empathize with users; Needs further research to validate effectiveness in real treatment settings	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
					disorder patients; Final prototype well-received		
Krithi Pushpanathan 2023 Singapore 37	Comparative performance evaluation study	37 ocular symptom queries; Evaluated by 3 consultant ophthalmologists; Setting: Online testing environment	ChatGPT-3.5, ChatGPT-4.0, Google Bard; Large Language Model (LLM)-based chatbots; AI techniques: Transformer-based deep learning + natural language processing	Symptom checker & diagnosis support for ocular conditions	Accuracy: ChatGPT-4.0 scored highest (89.2% "good" vs. ChatGPT-3.5 59.5%, Bard 40.5%). Comprehensiveness: All models scored high (4.6–4.7/5). Self-awareness: Weak to moderate ability to self-check and self-correct. User engagement & satisfaction: Not directly measured but implied via comprehensiveness ratings ChatGPT-4.0 outperforms ChatGPT-3.5 and Google Bard in accuracy and comprehensiveness when answering ocular symptom queries. LLMs have potential in supporting clinical decision-making and patient self-assessment but require further clinical validation before deployment.	- Lack of real patient interaction and clinical trial validation. - Study focused on simulated scenarios only. - Self-awareness and error correction capabilities remain limited. - Reliability in diverse patient populations remains unproven.	Moderate
Shan Chen 2023 USA 38	Survey-based evaluation study	sample size not reported; Online testing environment	Large Language Model (LLM)-based chatbot; AI techniques: Transformer-based NLP; No specific platform mentioned	Diagnosis & treatment support for breast, prostate, and lung cancer	Accuracy: Assessed concordance of chatbot responses with NCCN guidelines; Performance varied depending on cancer type. Comprehensiveness: Responses were generally coherent and detailed. LLM chatbots have	- Potential misinformation risk due to incorrect or incomplete treatment recommendations. - Lack of clinical validation against real-world patient cases. - Limited to simulated guideline-	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
			(likely ChatGPT or similar)		potential to support cancer patients by providing detailed treatment-related information, but risk of misinformation remains. Chatbots cannot currently replace medical professionals for cancer treatment recommendations.	based scenarios; not tested in patient-facing settings. - Reliability and safety concerns remain before adoption in clinical oncology.	
Ridvan Guler 2024 Turkey [39]	Comparative diagnostic performance study	23 patients (9 cysts, 14 neoplasms) Healthcare setting: Dicle University Faculty of Dentistry	Type of chatbot: AI-based diagnostic chatbots Platforms tested: ChatGPT, Grok, Blackbox AI, Claude AI AI techniques: Large Language Models (LLMs), NLP-driven responses Platform integration: Web-based testing environment	Diagnosis support — preliminary diagnosis of maxillofacial pathologies (cysts and neoplasms)	Accuracy: - ChatGPT: 65.2% (15/23 correct) → Best performance overall - Blackbox AI: 52.17% - Grok: 52.17% - Claude AI: 30.43% Cyst diagnosis: Blackbox AI highest (66.6%) Neoplasm diagnosis: ChatGPT highest (71.4%) Statistical significance: No significant difference among models (p=0.125) Clinical effectiveness: Shows potential for improving preliminary diagnosis accuracy. ChatGPT outperformed the other AI chatbots in diagnosing maxillofacial pathologies, especially neoplasms. AI-driven chatbots demonstrate promising potential for assisting dentists in early diagnosis and treatment recommendations, but they cannot yet replace expert judgment.	- Small sample size (23 patients) limits generalizability. - Only four chatbot models tested; may not represent broader AI performance. - Study limited to maxillofacial pathologies; results not applicable to other dental or medical conditions. - Lack of real-world patient interaction testing; only simulated diagnostic questions were used. - No evaluation of patient satisfaction, usability, or safety risks.	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
Takanobu Hirosawa, 2023 Japan [40]	Pilot study evaluating diagnostic accuracy	30 clinical vignettes based on 10 common chief complaints / General internal medicine	Type of Chatbot: GPT-3-based medical chatbot Platform: ChatGPT-3 AI Techniques: Large Language Model (LLM) + Natural Language Processing (NLP)	Diagnosis Support	Diagnostic Accuracy: • ChatGPT-3: 93.3% (28/30 correct) across 10 differential-diagnosis lists • Physicians: 98.3% on 5 differential-diagnosis lists • For the top diagnosis: Physicians: 93.3% vs ChatGPT: 53.3% Clinical Effectiveness: Demonstrated high capability to generate comprehensive differential diagnosis lists. ChatGPT-3 showed high diagnostic accuracy in generating well-differentiated lists of possible diagnoses for common chief complaints, approaching physician-level performance. However, physicians still outperformed ChatGPT when ranking the most likely diagnosis. GPT-3 can be considered a clinical decision support tool but not a replacement for physicians.	- Small sample size (30 vignettes) → limited generalizability - Restricted to 10 common chief complaints only - Used simulated vignettes, not real patient data - Only GPT-3 tested → other LLMs not evaluated - Did not assess usability, patient acceptance, or clinical workflow integration	Low
Emilie A. C. Dronkers 2025 Europe	Multicenter retrospective evaluation study	20 clinical cases Healthcare Setting: Four tertiary laryngology centers	Type of Chatbot: AI-based medical chatbots Platforms: ChatGPT-4.0, LLaMA Chat-	Treatment Decision-Making / Clinical Decision Support	Accuracy: • ChatGPT-4.0 achieved 50% accuracy in providing partially correct treatment suggestions • LLaMA Chat-2.0	- Small sample size (20 clinical cases) - Limited to retrospective data from four centers - Only ChatGPT-4.0 and LLaMA-2.0	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
[41]		across Europe	2.0 AI Techniques: Large Language Models (LLMs) using Natural Language Processing		achieved 15% accuracy • Maximum AIPI (Artificial Intelligence Performance Instrument) score achieved in only 5% of cases Clinical Effectiveness: Both chatbots performed poorly overall; ChatGPT performed better but remained unreliable. Safety Concerns: Some potentially harmful recommendations were made, such as suggesting vocal fold medialization for patients with stridor and dyspnea. ChatGPT-4.0 outperformed LLaMA Chat-2.0 but neither chatbot provided clinically reliable treatment recommendations for BVFP. Complex treatment decision-making in rare conditions remains beyond the current capability of LLM-based chatbots. There is a need for specialized guidelines and more advanced medical AI models.	tested; no broader AI comparison - Evaluated a rare, complex condition → results may not generalize - Did not assess real-time clinical usability or patient impact	
Nathanael Rebelo 2022 Canada [42]	Development and testing study	Sample Size: Not explicitly Healthcare Setting: Cancer hospital/cent	Type of Chatbot: AI-assisted virtual assistant Platform: IBM Watson Assistant	Patient education and treatment process explanation for radiotherapy. Also indirectly supports treatment	Reported Results: • Chatbot guides patients through radiotherapy treatment process • Provides interactive responses and educational support	Sample size of testing users not detailed in abstract - No quantitative usability scores or comparative evaluation against	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
		er radiotherapy departments	AI Techniques Used: Machine learning + natural language processing (NLP) through Watson API	adherence by improving patient understanding.	• Tested by users — performance metrics rated "excellent" • Capable of acquiring user feedback for continuous improvement No clinical outcome measures reported (e.g., survival, adherence rates). The chatbot successfully provides accurate, accessible, and personalized information about radiotherapy treatment to patients, families, and the public. It improves knowledge transfer and offers a user-friendly interface for education.	standard education methods - The chatbot focuses on knowledge transfer only, not real-time clinical decision-making - Long-term patient adherence and clinical outcomes not assessed	
Stephan Rau 2024 Germany 43	Proof-of-concept experimental study	Sample Size: 50 gastrointestinal radiology cases Healthcare Setting: Radiology/imaging departments	Type of Chatbot: GPT-4-based retrieval-augmented chatbot (GIA-CB) Platform: GPT-4 + LlamaIndex framework AI Techniques Used: Large language model, zero-shot learning, knowledge retrieval from context-specific radiology documents	Diagnosis support in gastrointestinal radiology	Clinical effectiveness / accuracy: • Correct primary differential in 78% of cases (GIA-CB) vs 54% for generic GPT-4 • Primary differential included in top 3: 90% (GIA-CB) vs 74% (generic GPT-4) • Provided rationale and source excerpts for decision support Usability / engagement: Median response time 29.8 s per case; cost per case $0.15. The context-aware GPT-4 chatbot outperformed generic GPT-4 in providing accurate	- Proof-of-concept study with simulated cases, not real-time patient care - Small sample size (50 cases) - No assessment of real-world clinical workflow integration - Did not evaluate impact on actual patient outcomes or clinician decision-making	High

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
					differential diagnoses. Integration of domain-specific documents improves trustworthiness and explainability of AI decision-support in radiology.		
Peter A Giammanco 2025 USA 44	omparative evaluation study	Sample Size: 10 patient questions about clavicle fractures, evaluated across 6 AI chatbots Healthcare Setting: Orthopedic patient education context	Type of Chatbot: Generative AI chatbots (ChatGPT 4, ChatGPT 4o, Gemini 1.0, Gemini 1.5 Pro, Microsoft Copilot, Perplexity) Platform: Not specified; general AI platforms AI Techniques Used: Large language models, generative AI, natural language processing	Patient education / decision support for clavicle fracture treatment options; indirect support for treatment adherence.	Readability: Measured with Flesch-Kincaid, Gunning Fog, SMOG – no significant differences among models Quality: DISCERN criteria assessed by six orthopedists; Microsoft Copilot and Perplexity had higher scores (70.33 and 71.83) than ChatGPT 4 and Gemini 1.5 Pro. Generative AI chatbots can serve as supplementary patient education tools. Microsoft Copilot and Perplexity provided the highest educational utility for clavicle fracture information. Overall readability was good across all chatbots, and quality ratings were above average.	- Only evaluated textual responses, not real patient interaction - Small set of 10 patient questions - Did not assess actual patient comprehension or behavior change - Chatbots were used without training or customization; results may differ in real clinical deployment	Moderate
Jason S DeFrancis is 2025 USA 45	Comparative evaluation study	Sample Size: 2 AI chatbots evaluated with multiple frequent meniscal tear questions Healthcare Setting: Orthopaedic	Type of Chatbot: Generative AI chatbots (ChatGPT-4o, Gemini 2.0 Flash) Platform: General AI platforms,	iagnosis support and patient education for meniscal tears	Accuracy: ChatGPT-4o 58.22% verifiable (UpToDate only), 83.56% (UpToDate + peer-reviewed), Gemini 2.0 Flash 58.97% and 84.62% respectively. Comparison: Minimal difference between the	- Limited to textual accuracy assessment, not real patient interaction - Evaluation restricted to meniscal tear questions only - Chatbots were untrained for this specific clinical	Moderate

Author(s), year, country	Study Design	Sample Size/ Healthcare Setting	Technology detail/Platform / Type of Chatbot/VHA	Application domain	Outcomes	Limitations	Study Quality
		patient education / diagnostic context	cloud-based AI Techniques Used: Large language models, natural language processing, generative AI		two AI chatbots; accuracy improved with broader verification sources. AI chatbots can provide useful orthopaedic information, but cannot replace clinical judgment. Accuracy improves when using multiple reference sources. Chatbots may supplement patient education but have clinical limitations in orthopaedics.	context - Did not assess impact on patient outcomes or behavior	
Nadav Grinberg 2025 Israel 46	Clinical validation study (retrospective)	Sample Size: 100 oral soft tissue lesions Healthcare Setting: Oral medicine clinic	Type of Chatbot: AI chatbot (ChatGPT-4.0) Platform: OpenAI platform AI Techniques Used: Large language model, natural language processing	Diagnosis support for oral mucosal lesions	Accuracy: ChatGPT correctly suggested differential diagnoses; statistically significant correlation with specialist diagnoses ($P < 0.001$) Sensitivity: High for urgent/malignant lesions (no malignancies missed) Specificity: Lower than specialist for malignant cases ($p < 0.05$). ChatGPT-4 demonstrates consistent and reliable ability to assist in differential diagnosis of oral mucosal lesions, particularly in identifying suspicious malignant lesions. AI chatbots can serve as supporting tools in oral medicine.	- Retrospective design - Single-center study - Limited to one AI model (ChatGPT-4) - Did not assess direct clinical outcomes or patient management impact - Specialist still had higher specificity for malignant lesions	Moderate

1. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Canadian Journal of Ophthalmology. 2024;59(4):e301–e8.
2. Wolmer S, Shauly O. Evaluating Plastic Surgery Chatbot Performance: Insights into Medical Triage, Classification Accuracy, and Escalation Trends. Aesthetic Surgery Journal. 2025:sjaf123.
3. Kring T, Prasad S, Dadi S, Sokhn E, Franzmann E. A comparison of quality and readability of Artificial Intelligence chatbots in triage for head and neck cancer. American journal of otolaryngology. 2025:104710.
4. Sarbay İ, Berikol GB, Özturan İU. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): A preliminary, scenario-based cross-sectional study. Turkish Journal of Emergency Medicine. 2023;23(3):156–61.
5. Schumacher I, Ferro Desideri L, Bühler VMM, Sagurski N, Subhi Y, Bhardwaj G, et al. Performance analysis of an emergency triage system in ophthalmology using a customized CHATBOT. Digital Health. 2025;11:20552076251320298.
6. Tsui JC, Wong MB, Kim BJ, Maguire AM, Scoles D, VanderBeek BL, et al. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. Eye. 2023;37(17):3692–3.
7. You Y, Gui X, editors. Self-diagnosis through AI-enabled chatbot-based symptom checkers: user experiences and design considerations. AMIA Annual Symposium Proceedings; 2021.
8. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors. 2022;22(10):3653.
9. Tan TC, Roslan NEB, Li JW, Zou X, Chen X, Santosa A. Patient Acceptability of Symptom Screening and Patient Education Using a Chatbot for Autoimmune Inflammatory Diseases: Survey Study. JMIR Formative Research. 2023;7(1):e49239.
10. Erkan A, Koc A, Barali D, Satir A, Zengin S, Kilic M, et al. Can Patients With Urogenital Cancer Rely on Artificial Intelligence Chatbots for Treatment Decisions? Clinical Genitourinary Cancer. 2024;22(6):102206.
11. de Moura JDM, Fontana CE, da Silva Lima VHR, de Souza Alves I, de Melo Santos PA, de Almeida Rodrigues P. Comparative accuracy of artificial intelligence chatbots in pulpal and periradicular diagnosis: A cross-sectional study. Computers in Biology and Medicine. 2024;183:109332.
12. Singh A, Schooley B, Patel N. Effects of user-reported risk factors and follow-up care activities on satisfaction with a COVID-19 chatbot: Cross-sectional study. JMIR mHealth and uHealth. 2023;11(1):e43105.
13. Ye BJ, Kim JY, Suh C, Choi SP, Choi M, Kim DH, et al. Development of a chatbot program for follow-up management of workers' general health examinations in Korea: a pilot study. International Journal of Environmental Research and Public Health. 2021;18(4):2170.
14. Ali SR, Dobbs TD, Whitaker IS. Using a ChatBot to support clinical decision-making in free flap monitoring. Journal of Plastic, Reconstructive & Aesthetic Surgery. 2022;75(7):2387–440.

15. Piau A, Crissey R, Brechemier D, Balardy L, Nourhashemi F. A smartphone Chatbot application to optimize monitoring of older patients with cancer. International journal of medical informatics. 2019;128:18–23.

16. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Translational behavioral medicine. 2019;9(3):440–7.

17. Qazi F, Shaheen O, Andrabi WI, Arif M, Begum F, Mansoor M. Evaluating the Incidence of Co-Existing Injuries in Anterior Talofibular Ligament Injuries a Magnetic Resonance Imaging Study: Co-Existing Injuries in Anterior Talofibular Ligament. Pakistan Journal of Health Sciences. 2025:115–20.

18. Hasei J, Hanzawa M, Nagano A, Maeda N, Yoshida S, Endo M, et al. Empowering pediatric, adolescent, and young adult patients with cancer utilizing generative AI chatbots to reduce psychological burden and enhance treatment engagement: a pilot study. Frontiers in Digital Health. 2025;7:1543543.

19. Melnik T, Thompson JA, Vasilakes J, Annis T, Zhou S, Schutte D, et al., editors. Semi-automated Clinical Content Curation of COVID-19 Chatbot Remote Patient Monitoring Solution. AMIA Annual Symposium Proceedings; 2023.

20. Ben-Shabat N, Sharvit G, Meimis B, Joya DB, Sloma A, Kiderman D, et al. Assessing data gathering of chatbot based symptom checkers-a clinical vignettes study. International Journal of Medical Informatics. 2022;168:104897.

21. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of self-diagnosis health chatbots in real-world settings: case study. Journal of medical Internet research. 2021;23(1):e19928.

22. Washington CJ, Abouyared M, Karanth S, Braithwaite D, Birkeland A, Silverman DA, et al. The use of Chatbots in head and neck mucosal malignancy treatment recommendations. Otolaryngology–Head and Neck Surgery. 2024;171(4):1062–8.

23. Shapiro J, Lyakhovitsky A. Revolutionizing teledermatology: Exploring the integration of artificial intelligence, including Generative Pre-trained Transformer chatbots for artificial intelligence-driven anamnesis, diagnosis, and treatment plans. Clinics in Dermatology. 2024;42(5):492–7.

24. Guede-Fernández F, Silva Pinto T, Semedo H, Vital C, Coelho P, Oliosi ME, et al. Enhancing postoperative anticoagulation therapy with remote patient monitoring: A pilot crossover trial study to evaluate portable coagulometers and chatbots in cardiac surgery follow-up. Digital Health. 2024;10:20552076241269515.

25. Habicht J, Viswanathan S, Carrington B, Hauser TU, Harper R, Rollwage M. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. Nature medicine. 2024;30(2):595–602.

26. Greer S, Ramo D, Chang Y-J, Fu M, Moskowitz J, Haritatos J. Use of the chatbot “vivibot” to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR mHealth and uHealth. 2019;7(10):e15018.

27. Luo MX, Lyle A, Bennett P, Albertson D, Sirohi D, Maughan BL, et al. Artificial intelligence chatbot vs pathology faculty and residents: Real-world clinical questions from a genitourinary treatment planning conference. American Journal of Clinical Pathology. 2024;162(6):541–3.

28. Ejaz H, Ali SR, Berner JE, Dobbs TD, Whitaker IS, Collaborative F. The “Flapbot”: A Global Perspective on the Validity and Usability of a Flap Monitoring Chatbot. Journal of Reconstructive Microsurgery. 2025;41(03):227–36.

29. Chen C-C, Wei C-J, Tseng T-Y, Chiu M-C, Chang C-C. Applying Object Detection and Large Language Model to Establish a Smart Telemedicine Diagnosis System with Chatbot: A Case Study of Pressure Injuries Diagnosis System. Telemedicine and e-Health. 2024;30(6):e1705–e12.

30. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR formative research. 2019;3(4):e13863.

31. Ozcan SGG, Erkan M. Reliability and quality of information provided by artificial intelligence chatbots on post-contrast acute kidney injury: an evaluation of diagnostic, preventive, and treatment guidance. Revista da Associação Médica Brasileira. 2024;70(11):e20240891.

32. Coelho MS, Piva GB, Vasconcelos RA, Toia CC, Santos Zambon L, Brenelli S. Chatbot Versus Lecture in the Teaching of Endodontic Diagnosis for Undergraduate Students—A Pilot Study. Journal of Dental Education. 2025:e13940.

33. Sharp G, Dwyer B, Randhawa A, McGrath I, Hu H. The Effectiveness of a Chatbot Single-Session Intervention for People on Waitlists for Eating Disorder Treatment: Randomized Controlled Trial. Journal of Medical Internet Research. 2025;27:e70874.

34. Zhang S, Song J. A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model. Scientific reports. 2024;14(1):17118.

35. Ghosh S, Bhatia S, Bhatia A. Quro: facilitating user symptom check using a personalised chatbot-oriented dialogue system. Connecting the System to Enhance the Practitioner and Consumer Experience in Healthcare: IOS Press; 2018. p. 51–6.

36. Sharp G, Dwyer B, Xie J, McNaney R, Shrestha P, Prawira C, et al. Co-design of a single session intervention chatbot for people on waitlists for eating disorder treatment: a qualitative interview and workshop study. Journal of Eating Disorders. 2025;13(1):46.

37. Pushpanathan K, Lim ZW, Yew SME, Chen DZ, Lin HAHE, Goh JHL, et al. Popular large language model chatbots’ accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. Iscience. 2023;26(11).

38. Chen S, Kann BH, Foote MB, Aerts HJ, Savova GK, Mak RH, et al. Use of artificial intelligence chatbots for cancer treatment information. JAMA oncology. 2023;9(10):1459–62.

39. Guler R, Yalcin E. Performance of AI chatbots in preliminary diagnosis of maxillofacial pathologies. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research. 2025;31:e949076.

40. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. International journal of environmental research and public health. 2023;20(4):3378.

41. Dronkers EA, Geneid A, Al Yaghchi C, Lechien JR. Evaluating the potential of AI chatbots in treatment decision-making for acquired bilateral vocal fold paralysis in adults. Journal of Voice. 2024.

42. Rebelo N, Sanders L, Li K, Chow JC. Learning the treatment process in radiotherapy using an artificial intelligence–assisted chatbot: development study. JMIR Formative Research. 2022;6(12):e39443.

43. Rau S, Rau A, Nattenmüller J, Fink A, Bamberg F, Reisert M, et al. A retrieval-augmented chatbot based on GPT-4 provides appropriate differential diagnosis in gastrointestinal radiology: a proof of concept study. European radiology experimental. 2024;8(1):60.

44. Giammanco PA, Collins CE, Zimmerman J, Kricfalusi M, Rice RC, Trumbo M, et al. Evaluating the Quality and Readability of Information Provided by Generative Artificial Intelligence Chatbots on Clavicle Fracture Treatment Options. Cureus. 2025;17(1).

45. DeFrancisis JS, Richa P, Oar D, Henwood L, Buchman ZJ, Grewal G, et al. Assessing the Accuracy of Artificial Intelligence Chatbots in the Diagnosis and Management of Meniscal Tears. Cureus. 2025;17(5).

46. Grinberg N, Whitefield S, Kleinman S, Ianculovici C, Wasserman G, Peleg O. Assessing the performance of an artificial intelligence based chatbot in the differential diagnosis of oral mucosal lesions: clinical validation study. Clinical Oral Investigations. 2025;29(4):188.