Clinical Ophthalmology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Evaluating the Application of Artificial Intelligence and Ambient Listening to Generate Medical Notes in Vitreoretinal Clinic Encounters

Neeket R Patel[1], Corey R Lacher[1], Alan Y Huang[2], Anton Kolomeyer[2,3], J Clay Bavinger[4], Robert M Carroll[2,5], Benjamin J Kim[2], Jonathan C Tsui[1,2,6]

[1]Institute of Ophthalmology and Visual Science, Rutgers New Jersey Medical School, Newark, NJ, 07103, USA; [2]Scheie Eye Institute, Department of Ophthalmology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [3]NJ Retina (PRISM Vision Group), Manassas, VA, USA; [4]The Retina Group of Washington (PRISM Vision Group), New Providence, NJ, USA; [5]Eye Physicians and Surgeons, Wilmington, DE, USA; [6]Department of Ophthalmology, Veterans Affairs New Jersey Healthcare System, East Orange, NJ, USA

Correspondence: Jonathan C Tsui, Department of Ophthalmology, Veterans Affairs NJ Healthcare System, 385 Tremont Ave, East Orange, NJ, 07018, USA, Tel +1 973-676-1000 Ext 203213, Email jonathan.tsui@va.gov

Purpose: Analyze the application of large language models (LLM) to listen to and generate medical documentation in vitreoretinal clinic encounters.

Subjects: Two publicly available large language models, Google Gemini 1.0 Pro and Chat GPT 3.5.

Methods: Patient-physician dialogues simulating vitreoretinal clinic scenarios were scripted to simulate real-world encounters and recorded for standardization. Two artificial intelligence engines were given the audio files to transcribe the dialogue and produce medical documentation of the encounters. Similarity of the dialogue and LLM transcription was assessed using an online comparability tool. A panel of practicing retina specialists evaluated each generated medical note.

Main Outcome Measures: The number of discrepancies and overall similarity of LLM text compared to scripted patient-physician dialogues, and scoring on the physician documentation quality instrument-9 (PDQI-9) of each medical note by five retina specialists.

Results: On average, the documentation produced by AI engines scored 81.5% of total possible points in documentation quality. Similarity between pre-formed dialogue scripts and transcribed encounters was higher for ChatGPT (96.5%) compared to Gemini (90.6%, p<0.01). The mean total PDQI-9 score among all encounters from ChatGPT 3.5 (196.2/225, 87.2%) was significantly greater than Gemini 1.0 Pro (170.4/225, 75.7%, p=0.002).

Conclusion: The authors report the aptitude of two popular LLMs (ChatGPT 3.5 and Google Gemini 1.0 Pro) in generating medical notes based on audio recordings of scripted vitreoretinal clinical encounters using a validated medical documentation tool. Artificial intelligence can produce quality vitreoretinal clinic encounter medical notes after listening to patient-physician dialogues despite case complexity and missing encounter variables. The performance of these engines was satisfactory but sometimes included fabricated information. We demonstrate the potential utility of LLMs in reducing the documentation burden on physicians and potentially streamlining patient care.

Keywords: retina, clinical documentation, large language model, ophthalmology

Introduction

Artificial intelligence (AI) is becoming increasingly implemented in healthcare settings, offering streamlined clinical workflows and reduced administrative burden.[1] One key facet of patient care is the successful documentation of clinic encounters which often contributes to physician burnout.[2] Moreover, physician-written notes frequently contain mistakes which can lead to medical error and detrimental impact on patient care.[3] To offset the physical burden of documenting within electronic medical records (EMR), scribes and speech-to-text transcription softwares have often been employed. Many disadvantages exist including high costs, potential errors, and personnel turnover.[4–6] One opportunity for AI that remains understudied is automated medical documentation of vitreoretinal clinic encounters, or "ambient listening".

Received: 22 December 2024
Accepted: 23 May 2025
Published: 3 June 2025

Recent advances in large language models (LLMs) like ChatGPT and Google Gemini present promising alternatives for documentation by performing speech recognition and subsequently generating appropriate medical notes.[7,8] To this end, several software companies and health systems are developing tools using LLMs to create ambient AI scribes—software that will listen to clinical encounters and immediately generate relevant medical documentation.[9–11] High-volume sub-specialties like vitreoretinal surgery can benefit significantly from such tools to improve patient care and workflow.[12] Within ophthalmology, LLMs have already demonstrated aptitude in generating discharge summaries using pre-written prompts, identifying differential diagnoses, and answering board exam questions.[13–16] However, to the authors' knowledge, no studies have evaluated the use of LLMs as ophthalmic ambient scribes.

Methods

Ten vitreoretinal diagnoses were selected, ranging from common conditions, such as age-related macular degeneration and diabetic retinopathy to less common conditions such as multifocal choroiditis and central retinal artery occlusion. Physician-patient dialogues, in English, were created for each diagnosis to reflect a representative clinical encounter. Variability was introduced to reflect real-world interactions by adding extraneous verbosity, or omitting formal treatment plans, physical examinations, or both. This allowed evaluation of each AI's ability to generate complete medical notes and suggest accurate treatment plans without consistently given complete clinical information.

Two individuals (one female, one male) then role-played and recorded each of the ten patient-physician dialogues in an MPEG-4 audio format. An MPEG-4 audio file was kept as the standard for testing between the LLM models. The audio file was played aloud for the AI engine (Google Gemini 1.0 Pro and Chat GPT 3.5) to listen to using the audio function on each respective LLM engine. The LLMs each transcribed the audio in real-time and subsequently entered the transcription in the input textbox. Using an online text similarity calculator, GoTranscript.com, each AI transcription was compared to the original script to evaluate the number of discrepancies (instances of differences) and degree of similarity (extent of differences). Punctuation and capitalization differences were not considered discrepancies. Each missing or added word was counted as a discrepancy in the AI transcriptions.

After the AI engine had transcribed the conversation it "heard", we inputted the question: "Based on the previous dialogue please generate a comprehensive electronic medical record style note". This was completed for each scenario and LLM for a total of twenty notes (ten from each AI engine). Examples of the dialogue transcript, recorded text, and generated notes are provided (Supplement 1–3). Five practicing board-certified retina specialists were then given half of the encounters, in an equal distribution of ChatGPT vs Gemini, in an unmasked fashion including audio files and encounter notes. The audio files and encounter notes were provided simultaneously in a single file, with the former as a supplement to the evaluator. The retina specialists were instructed to individually grade each note based on the Physician Documentation Quality Instrument (PDQI-9) which has nine categories utilizing a Likert scale from 1 to 5 with 5 linked to an "Excellent" note, and to note any relevant findings.[17] After analysis of the first round of data, a significant difference in ChatGPT and Gemini scores was evident. A second, masked arm of the study was then implemented to determine if difference in scores or rater reliability existed based on previous knowledge of the LLMs. After a six-week washout period, the retina specialists were then given the remaining ten random audio files and generated encounter notes to grade, in a masked fashion. Scores per question for each scenario for each AI engine were compiled (Supplement 4). Each encounter was tallied for a total possible score of 9–45 points per encounter. Scores from each vitreoretinal surgeon were compiled to give a final score to each encounter.

Statistical analyses were performed using IBM Statistical Package for Social Sciences (SPSS) and Microsoft Excel. Paired two-tailed *t*-tests were utilized to compare transcription accuracy and final PDQI-9 scores. Sub-analyses were performed using paired two-tailed *t*-tests to compare average scores from each round of grading, scores from each LLM platform, and scores from each rater. Intraclass coefficient (ICC) was calculated using a two-way mixed effect, average score model.

Results

Overall, the LLMs achieved a transcription accuracy of 93.6%. On average, the documentation produced by AI engines scored 81.5% (183.3/225) of total possible points on the PDQI-9. Overall, ICC for PDQI-9 scores was 0.595 (95% CI: 0.493–0.682). Overall, physicians noted positives including documentation of informed consent for injections, ability to note worsening/improvement with correct laterality, and detailed interpretations of diagnostic testing. They also made note of pertinent drawbacks including errors and false information.

In terms of transcription accuracy, ChatGPT performed superiorly to Gemini with an average 14.4 discrepancies per encounter, while Gemini had 30.8 discrepancies per encounter ($p<0.01$). Likewise, the similarity between pre-formed scripts and transcribed encounters was higher for ChatGPT (96.5%) compared to Gemini (90.6%, $p<0.01$). ChatGPT performed with the least amount of similarity in the case of multifocal choroiditis (91%). Gemini performed with the least amount of similarity in the case of proliferative diabetic retinopathy with vitreous hemorrhage requiring injection (84%). ChatGPT performed with 100% similarity in the script with the fewest number of words (choroidal nevus) and tended to have approximately 0.02% decreasing similarity with each additional transcribed word. Gemini performed with the highest degree of similarity in the case of central retinal artery occlusion with approximately 0.03% decreasing similarity with each additional transcribed word.

We then compiled the individual PDQI-9 scores per scenario (Figure 1). The mean total PDQI-9 score among all encounters from ChatGPT 3.5 was significantly greater than Gemini 1.0 Pro (196.2/225, 87.2% and 170.4/225, 75.7%, respectively, $p=0.002$). Analysis of significant differences in score showed ChatGPT was significantly superior in 80% of scenarios ($p<0.05$). Paired *t*-test sub analyses also demonstrated ChatGPT was significantly superior compared to Gemini in all but two components of the PDQI-9 (Table 1). In the first round of unmasked grading, the calculated ICC was 0.746 (95% CI 0.653–0.821) and in the second round of masked grading, the ICC was 0.667 (95% CI 0.545–0.765). No significant difference was observed between unmasked and masked ICCs ($p=0.245$).

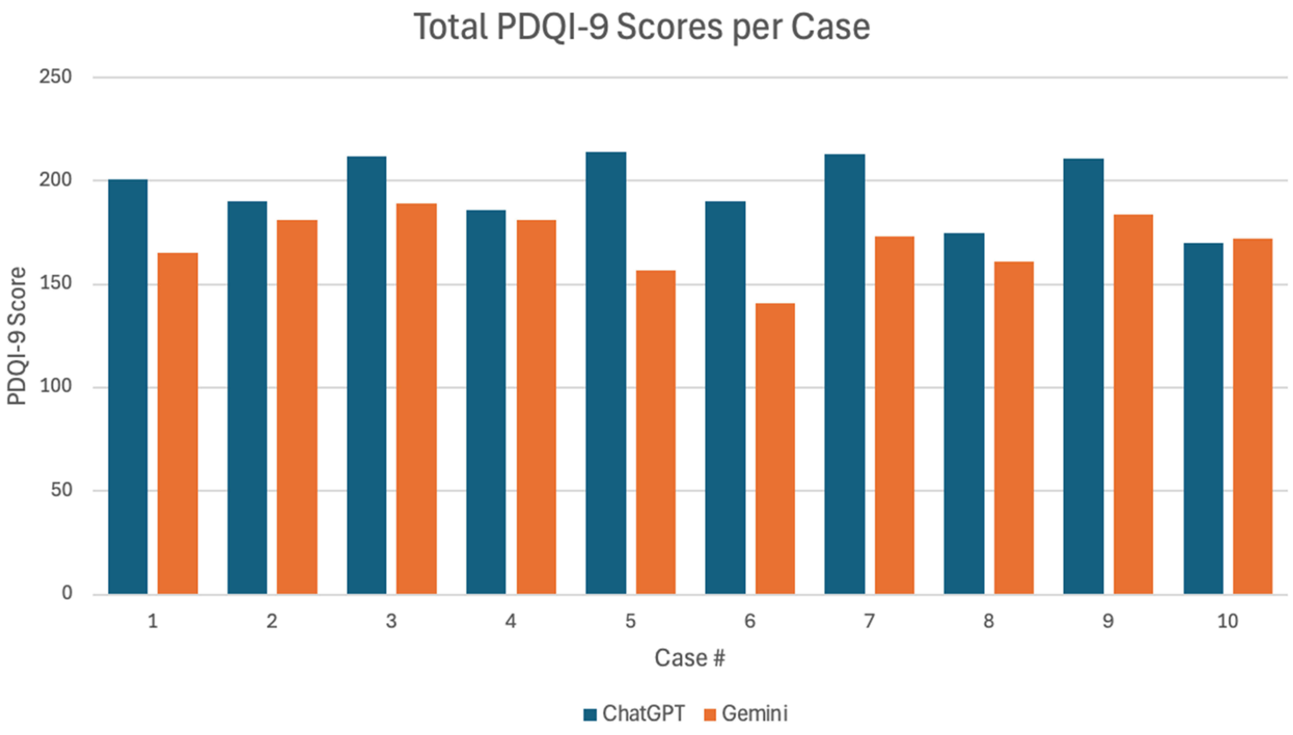

Figure 1 Total PDQI-9 score for each case demonstrated by total score per scenario for both Chat GPT 3.5 and Google Gemini 1.0 Pro.

Table 1 Paired *t*-Test Sub Analyses to Evaluate the Overall Difference Between ChatGPT 3.5 (C) and Google Gemini 1.0 Pro (G) for Each PDQI-9 Question. ChatGPT Was Significantly Superior Compared to Gemini in All but Two Components of the PDQI-9

Paired Differences	Mean	Std. Deviation	Two-sided *t*-test
Question 1C - Question 1G	3.700	2.359	<0.001*
Question 2C - Question 2G	3.600	5.190	0.056
Question 3C - Question 3G	4.000	2.261	<0.001*
Question 4C - Question 4G	3.700	3.561	0.009*
Question 5C - Question 5G	2.000	2.055	0.013*
Question 6C - Question 6G	1.900	1.729	0.007*
Question 7C - Question 7G	0.500	2.799	0.586
Question 8C - Question 8G	3.300	2.830	0.005*
Question 9C - Question 9G	3.100	2.470	0.003*

Note: *p<0.05.

Discussion

Google Gemini 1.0 Pro and ChatGPT 3.5 demonstrate LLMs can be utilized to achieve satisfactory documentation in vitreoretinal patient encounters. The first aspect assessed in this study was the ability to provide accurate transcription of encounters. Using identical recordings for evaluation, ChatGPT demonstrated superior transcription ability based on fewer instances of discrepancies and higher overall similarity. Among both AIs, reviewers pointed out grammatical errors and wrong word transcriptions; for example, "demon" was transcribed in place of "edema", but these did not seem to detract significantly from the output. Furthermore, both AIs consistently and accurately transcribed common ophthalmic abbreviations such as "OCT" and "anti-VEGF", though some abbreviated exam findings were occasionally recorded incorrectly. For instance, in the multifocal choroiditis case, Gemini incorrectly transcribed "sub-RPE deposits" as "sub-retinal pigment deposits". Despite these shortcomings, when compared to existing transcription services, a recent study found that a widely used dictation software [Dragon Medical 360 | eScription (Nuance)] transcribed medical notes with a similar 93.0% accuracy rate.[18] Thus, the authors deemed the LLMs suitable for further evaluation of medical note documentation performance using the PDQI-9.

The PDQI was originally created using psychometric measures to evaluate essential components of high-quality medical documentation for the purposes of clinical communication, and it was later refined to a simplified 9-item version.[19,20] The PDQI-9 has been used to assess both inpatient and outpatient notes in the era of electronic medical documentation.[17,19] While numerical scores have not been formally stratified to specify quality levels for various note types, for comparison, one study of primary care outpatient notes described average individual item scores above 4.0 to be "good".[17] Based on this general guideline, ChatGPT achieved a minimum score of "good" in all PDQI-9 categories except for accuracy, while Gemini only met this threshold for organization, comprehensibility, and succinctness.

Alternatively, another study found discharge and admission notes with a total PDQI-9 score of 26.2 to be considered by majority graders as "Terrible or Bad", while a score of 36.6 was described as "Good or Excellent".[19] Based on these thresholds, 80% of ChatGPT notes would be "Good or Excellent", while only 20% of Gemini notes would meet this standard. No notes were considered "Terrible or Bad" for either LLM. These results are taken in light of transcription accuracy findings; it is possible ChatGPT's superior PQDI-9 score may be related in part to superior transcription ability.

One common feature of real-world encounters is rapport-building which is not commonly documented in medical records. By including extraneous material in several cases, this study sought to test whether or not LLMs could remove non-value-added material as reflected in question 7 regarding succinctness; overall there was no difference between

LLMs, and both scored fairly well (ChatGPT 21.5, Gemini 21, p=0.586). Another feature of patient encounters is that physicians may not always voice exam findings and the design of this study included four encounters that were lacking physical exams. The PDQI-9 questions that reflected the AI's ability to anticipate the most common exam findings based on limited encounter information are reflected in questions 2 and 9 regarding accuracy and internal consistency. Reviewers found that ChatGPT had greater internal consistency compared to Gemini (23 vs 19.9 respectively, p=0.003), but no significant difference in accuracy (p=0.056). A subset of those four encounters lacked both physical exam and a plan of care in the scripts; two encounters also tested the ability to include a physical exam, given a plan of care. This additionally tested PDQI-9 question 8 which inquired whether or not the note demonstrated an understanding of patient status and ability to develop a plan of care. Reviewers found ChatGPT was superior (21.7) compared to Gemini (18.4, p=0.005) in this regard as well.

The advantages of ambient listening software compared to scribes or physicians include potential cost savings, speed, and lack of fatiguability. AI engines are continually being updated to newer generations which allows updates to pace with diagnostic advancements and ever-changing billing requirements.[21] EMRs have been shown to contain features that may improve Healthcare Effectiveness Data and Information Set (HEDIS) quality measures in outpatient settings, and further studies should evaluate whether AI features may have beneficial effects on improving quality metrics.[22]

Nonetheless, shortcomings still exist as results demonstrated that the less information given to the AI, the poorer the documentation. In fact, in order to make up for this shortcoming, AIs would also include "hallucinations" and inconsistently falsify information it believed should be included. One aspect of the note which both LLMs frequently "hallucinated" was the systemic review of systems (ROS). Despite the omission of cardiac, respiratory, gastrointestinal, etc., ROS questions in all scenarios, the LLMs incorrectly interpreted the systemic ROS to be negative in 50% of notes. This falsification could result in patient harm and physician liability, which stresses the importance of oversight.[17] When not provided enough information, reviewers also noted that the AI sometimes recommended referral to an ophthalmologist, simulating an internal medicine or optometry note.

While these LLM notes fell short, human notes have had fallacies as well. One study compared the electronic medical record notes to covertly recorded audio files from the same patient encounters and demonstrated that 90% of physician notes had at least one error; among 105 encounter notes, there were 636 errors that occurred, including 181 charted findings that did not occur and 455 findings from the encounter that were omitted.[23] In addition, a study using PDQI-9 to assess outpatient primary care notes in the US Department of Veterans Affairs showed lack of assessments, plans, verifiable observations, and information about return visits.[17]

Optimal implementation of ambient listening AIs into clinical workflow should limit high-risk text-field alterations such as medication lists and allergies from AI input. In addition, ambient listening AI distributed to various subspecialties should be trained with separate funds of knowledge to improve output especially in cases of subspecialty complexity. While AI continues to improve, physician oversight is compulsory to limit documentation errors. Future steps should address the reproducibility of encounter notes with larger samples, non-retinal ophthalmologic scenarios, differences in speaker characteristics, and with additional LLMs. Using real-world encounters and comparing medical scribe, physician, and AI notes for quality and billing compliance would be most insightful. While this study examined in-office encounters, AI's ability to document telephone calls and video encounters also warrants investigation.

Limitations of this study include using scripted dialogues to standardize testing between the two AI's which limits clinical generalizability. Bias may be present in the formation of these scripts as they were scripted to detail a given diagnosis (Supplement 1–3). Real-world variables include distance from and type of recording device, and differences in speaker gender, accents, enunciation, volumes, cadence, and tones. The effects of these variables likely impact the quality of medical documentation. This study utilized a standardized prompt to request outputs, and different prompts in larger samples may yield different results. In addition, graders were not masked as to which AI was being used in the first round of medical note evaluation, which may have permitted some recall bias. However, comparison of the intraclass coefficients between the two rounds was not significantly different, suggesting that masking had minimal impact on final PDQI-9 scores.

https://doi.org/10.2147/OPTH.S513633

Conclusions

To our knowledge, this study is the first to formally evaluate documentation produced by artificial intelligence in vitreoretinal clinic encounters. With the large rise in investigations of AI's ability to perform in multiple clinical activities, integration throughout clinical workflow can be the ultimate goal of assisting physicians in clinical practice. Artificial intelligence and the large language models utilized in this study are evolving rapidly and will undoubtedly find their niche in ophthalmology as advancements continue. This study serves as an early step in harnessing AI to decrease documentation burden and allow for optimized focus on the quality and quantity of patient care interactions.

Disclosure

Dr Anton Kolomeyer reports personal fees from Astellas (Iveric), Genentech, Regeneron, Alimera Sciences, Apellis, Biogen, Allergen, Oculis, Vial, and Retina Labs, outside the submitted work. Dr Benjamin Kim reports grants from Research to Prevent Blindness and Paul and Evanina Mackall Foundation, during the conduct of the study. The authors report no other conflicts of interest in this work.

References

1. Vo V, Chen G, Aquino YSJ, et al. Multi-stakeholder preferences for the use of artificial intelligence in healthcare: a systematic review and thematic analysis. *Soc Sci Med*. 2023;338:116357. doi:10.1016/j.socscimed.2023.116357
2. Kroth PJ, Morioka-Douglas N, Veres S, et al. Association of electronic health record design and use factors with clinician stress and burnout. *JAMA Network Open*. 2019;2:e199609. doi:10.1001/jamanetworkopen.2019.9609
3. Bell SK, Delbanco T, Elmore JG, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA Network Open*. 2020;3:e205867. doi:10.1001/jamanetworkopen.2020.5867
4. Florig ST, Corby S, Rosson NT, et al. Chart completion time of attending physicians while using medical scribes. *AMIA Annu Symp Proc*. 2021;2021:457–465.
5. Pranaat R, Mohan V, O'Reilly M, et al. Use of simulation based on an electronic health records environment to evaluate the structure and accuracy of notes generated by medical scribes: proof-of-concept study. *JMIR Med Inform*. 2017;5:e7883. doi:10.2196/medinform.7883
6. Goss FR, Blackley SV, Ortega CA, et al. A clinician survey of using speech recognition for clinical documentation in the electronic health record. *Int J Med Inform*. 2019;130:103938. doi:10.1016/j.ijmedinf.2019.07.017
7. Garg RK, Urs VL, Agarwal AA, et al. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. *Health Promot Perspect*. 2023;13:183–191. doi:10.34172/hpp.2023.22
8. Joseph T, Sanghavi N, Kanyal S, et al. Comparative analysis of ChatGPT and Google Gemini in the creation of patient education materials for acute appendicitis, cholecystitis, and hydrocele. *Indian J Surg*. 2024:1–6.
9. Crampton NH. Ambient virtual scribes: mutuo health's autoScribe as a case study of artificial intelligence-based technology. *Healthc Manage Forum*. 2020;33:34–38. doi:10.1177/0840470419872775
10. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal*. 2024;5:CAT.23.0404. doi:10.1056/CAT.23.0404
11. Dowling R. Ambient clinical documentation shows promise for physicians. *Urol Times*. 2023;51.
12. Srivastava O, Tennant M, Grewal P, et al. Artificial intelligence and machine learning in ophthalmology: a review. *Indian J Ophthalmol*. 2023;71:11. doi:10.4103/ijo.IJO_1569_22
13. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Sem Ophthalmo*. 2023;38:503–507. doi:10.1080/08820538.2023.2209166
14. Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel pro differential diagnosis generator. *JFO Open Ophthalmol*. 2023;1:100005. doi:10.1016/j.jfop.2023.100005
15. Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci*. 2023;3:100324. doi:10.1016/j.xops.2023.100324
16. Tsui JC, Wong MB, Kim BJ, et al. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. *Eye*. 2023;37:3692–3693. doi:10.1038/s41433-023-02556-2
17. Weiner M, Flanagan ME, Ernst K, et al. Accuracy, thoroughness, and quality of outpatient primary care documentation in the U.S. department of veterans affairs. *BMC Prim Care*. 2024;25:262. doi:10.1186/s12875-024-02501-6
18. Zhou L, Blackley SV, Kowalski L, et al. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. *JAMA Network Open*. 2018;1:e180530. doi:10.1001/jamanetworkopen.2018.0530
19. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). *Appl Clin Inform*. 2012;3:164–174. doi:10.4338/ACI-2011-11-RA-0070
20. Stetson PD, Morrison FP, Bakken S, et al. Preliminary development of the physician documentation quality instrument. *J Am Med Inform Assoc*. 2008;15:534–541. doi:10.1197/jamia.M2404
21. Koteluk O, Wartecki A, Mazurek S, et al. How do machines learn? Artificial intelligence as a new era in medicine. *J Pers Med*. 2021;11:32. doi:10.3390/jpm11010032
22. Poon EG, Wright A, Simon SR, et al. Relationship between use of electronic health record features and health care quality: results of a statewide survey. *Med Care*. 2010;48:203–209. doi:10.1097/MLR.0b013e3181c16203
23. Weiner SJ, Wang S, Kelly B, et al. How accurate is the medical record? A comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. *J Am Med Inform Assoc*. 2020;27:770–775. doi:10.1093/jamia/ocaa027