Annals of Internal Medicine ORIGINAL RESEARCH

Comparison of Initial Artificial Intelligence (AI) and Final Physician Recommendations in AI-Assisted Virtual Urgent Care Visits

Dan Zeltzer, PhD; Zehavi Kugler, MD; Lior Hayat, MD; Tamar Brufman, MD; Ran Ilan Ber, PhD; Keren Leibovich, PhD; Tom Beer, MSc; Ilan Frank, MSc; Ran Shaul, BASc; Caroline Goldzweig, MD, MSHS; and Joshua Pevnick, MD, MSHS

Background: Whether artificial intelligence (AI) assistance is associated with quality of care is uncertain.

Objective: To compare initial AI recommendations with final recommendations of physicians who had access to the AI recommendations and may or may not have viewed them.

Design: Retrospective cohort study.

Setting: Cedars-Sinai Connect, an AI-assisted virtual urgent care clinic with intake questions via structured chat. When confidence is sufficient, AI presents diagnosis and management recommendations (prescriptions, laboratory tests, and referrals).

Patients: 461 physician-managed visits with AI recommendations of sufficient confidence and complete medical records for adults with respiratory, urinary, vaginal, eye, or dental symptoms from 12 June to 14 July 2024.

Measurements: Concordance of diagnosis and management recommendations of initial AI recommendations and final physician recommendations. Physician adjudicators scored all nonconcordant and a sample of concordant recommendations as optimal, reasonable, inadequate, or potentially harmful.

Results: Initial AI and final physician recommendations were concordant for 262 visits (56.8%). Among the 461 weighted visits, AI recommendations were more frequently rated as optimal (77.1% [95% CI, 72.7% to 80.9%]) compared with treating physician decisions (67.1% [CI, 62.9% to 71.1%]). Quality scores were equal in 67.9% (CI, 64.8% to 70.9%) of cases, better for AI in 20.8% (CI, 17.8% to 24.0%), and better for treating physicians in 11.3% (CI, 9.0% to 14.2%), respectively.

Limitations: Single-center retrospective study. Adjudicators were not blinded to the source of recommendations. It is unknown whether physicians viewed AI recommendations.

Conclusion: When AI and physician recommendations differed, AI recommendations were more often rated better quality. Findings suggest that AI performed better in identifying critical red flags and supporting guideline-adherent care, whereas physicians were better at adapting recommendations to changing information during consultations. Thus, AI may have a role in assisting physician decision making in virtual urgent care.

Primary Funding Source: K Health.

Ann Intern Med. doi:10.7326/ANNALS-24-03283
For author, article, and disclosure information, see end of text.
This article was published at Annals.org on 4 April 2025.

Artificial intelligence (AI) has shown promise in various health care domains, including diagnostic radiology (1, 2), cardiology (3, 4), pathology (5), and risk prediction and screening (6, 7). However, systematic reviews highlight the limited number and quality of studies evaluating AI in real-world clinical practice, particularly in primary care (8–11). Moreover, most existing studies focus on AI performance on narrow tasks, such as image interpretation or risk prediction for specific outcomes, rather than on AI's potential to support diagnosis and management. This study aimed to address these gaps by comparing AI system diagnosis and management recommendations with treating physician decisions during real-world, acute, virtual urgent care visits. Eligible visits were those with common symptoms for which the AI system has shown high diagnostic accuracy (12).

Methods

Clinical Context

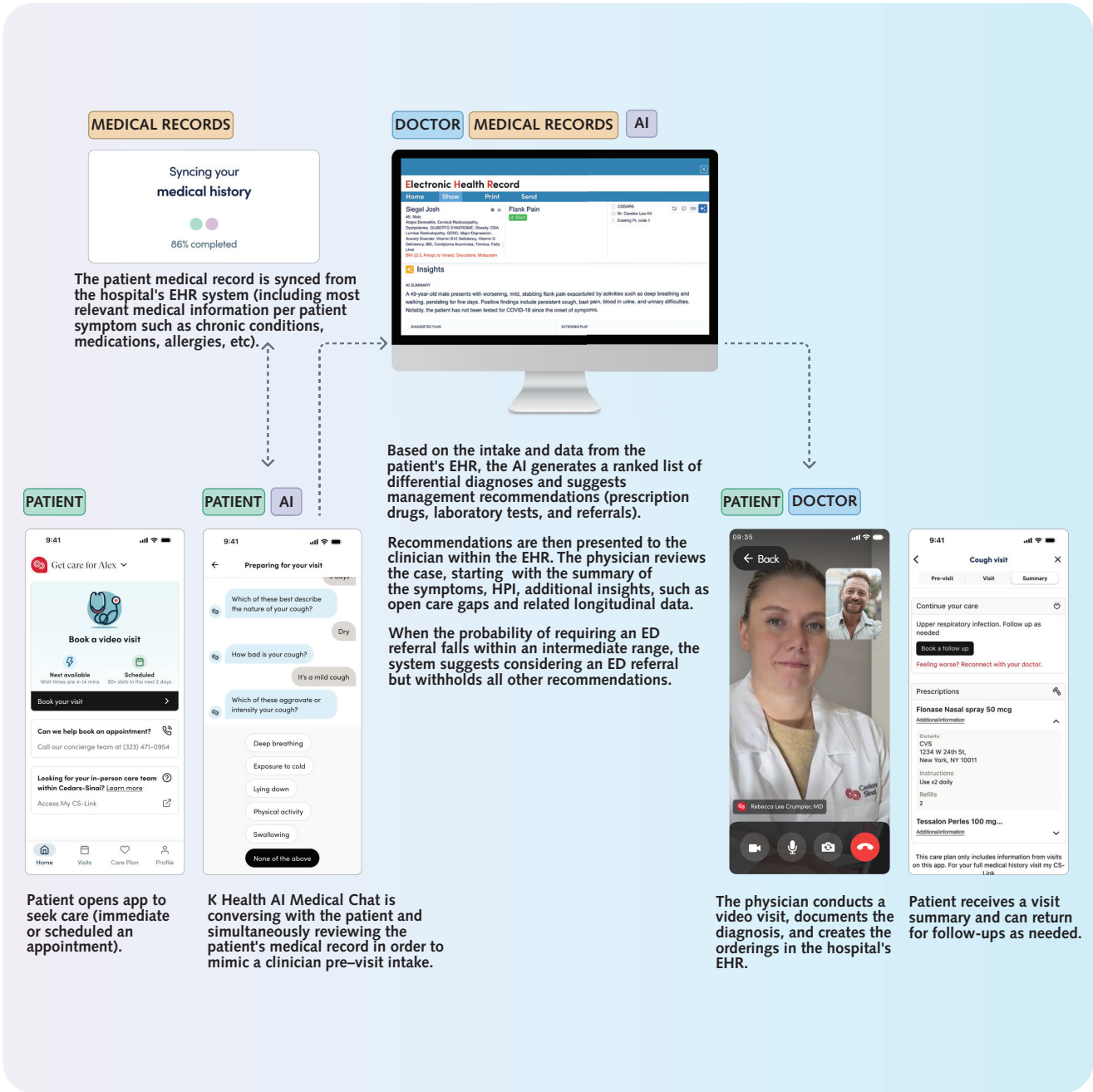
This retrospective cohort study was conducted using data from Cedars-Sinai Connect (CS Connect), a virtual primary and urgent care clinic in which physicians are presented with AI-based intake, diagnosis, and management assistance. The clinical workflow and the interfaces between the patient, AI, and physician are illustrated in Figure 1. Patients access the service via a mobile application, initiating visits by entering their medical concerns and, for first-time users, providing demographic information. An expert AI model conducts a structured dynamic interview, gathering symptom information and medical history. On average, 25 questions are answered over a period of 5 minutes. An algorithm, using this information and data from the patient's electronic health record (EHR), provides initial information about conditions with related symptoms to patients who

See also:

Editorial comment

Web-Only
Supplement

Figure 1. Clinical workflow and interfaces between patient, AI, and physician.



This figure illustrates the clinical workflow of CS Connect, a virtual primary and urgent care clinic, where AI-driven intake, diagnosis, and management assistance are integrated with physician decision making. Patients initiate care through a mobile application by describing their symptoms and medical history. The AI model dynamically interviews patients, gathering an average of 25 responses over a period of 5 minutes, and incorporates data from the EHR to generate a ranked list of differential diagnoses and management recommendations (prescriptions, laboratory tests, referrals). These recommendations are available to the physician via the EHR during a video consultation. The physician can review the AI-provided information, make final diagnoses, and determine the appropriate treatment plan. In cases where AI predictions have low confidence (defined based on intermediate probability of an ED referral), recommendations are withheld. After the consultation, patients receive a visit summary and follow-up instructions. AI = artificial intelligence; ED = emergency department; EHR = electronic medical record; HPI = history of present illness; Pt = patient. (The following images were reproduced with permission from iStock.com and Dr. Stephanie Foley: the computer, the patient, and the physician).

can then initiate a video visit with a physician. The algorithm also suggests management recommendations to treating physicians, including potential medication prescriptions, laboratory test orders, and referrals to appropriate care settings. These AI recommendations can be viewed through the EHR system during a video consultation before the physician makes their final diagnosis and treatment decisions.

The AI system was developed by K Health, a technology company, and is in use in the CS Connect clinic and other clinics in the United States operated by or jointly with K Health. The AI system is an ensemble of discriminative machine-learning models, including boosting algorithms and neural networks, trained on real-world EHR data from a large U.S.-based telehealth clinic and augmented with rule-based logic based on guidelines and medical knowledge. These models, specialized for different medical domains and data types, were developed using visit data from adult patients with acute conditions, incorporating reported symptoms, medical history, and relevant imaging. The AI system uses a selective recommendation protocol based on confidence calibration (13, 14). When the probability of requiring an emergency department (ED) referral falls within an intermediate range, the system only suggests considering ED referral and withholds all other recommendations, preserving independent physician judgment in cases of indeterminate acuity, where AI predictions may be less beneficial (henceforth "low confidence"). This design is more applicable to real-world clinical settings with several levels of complexity. The model architecture and development are discussed in more detail in the Supplement 1 (available at Annals.org). Regardless of AI recommendation availability or confidence, all patients proceed to a video visit with a physician, ensuring further evaluation and appropriate care.

Although physicians had access to AI recommendations, the CS Connect user interface for the period of analysis required them to actively scroll down to view these suggestions. It is therefore uncertain whether and how frequently physicians viewed these recommendations or incorporated them into their clinical decision making. Consequently, this study compares initial AI recommendations to the decisions made by physicians who had access to the AI recommendations but may or may not have viewed them.

Ethical Approval

The Cedars-Sinai Internal Review Board (reference STUDY00003707) reviewed and approved this study protocol. Patient consent was waived.

Study Sample

Based on an a priori decision, the study population consisted of all visits between 12 June and 14 July 2024, with chief symptoms for which prior analysis has shown high AI accuracy: respiratory, urinary, vaginal, eye, and dental (12). Cases were excluded if they had incomplete video recordings or missing critical information, if the AI withheld recommendations due to low confidence, or if the cases were managed by a nonphysician clinician. The remaining cases were automatically classified, based on predefined rules and structured EHR data, as *concordant* when all physician diagnosis and management decisions aligned with the AI recommendation and as *nonconcordant* when any of them did not align. Concordance was defined as agreement on the following criteria: having the same diagnosis group (for example, viral respiratory conditions), having the same prescription medication regimen (including class, type, and dose), having the same type of laboratory tests, recommending nonurgent referrals (that is, for in-person or specialist care), and recommending urgent referrals (that is, to an urgent care center or ED). Concordance classification required agreement across all of these criteria. If the AI recommendations and physician decisions differed on any of these diagnosis or management criteria, the case was classified as nonconcordant.

The sample of cases for manual physician adjudication consisted of all nonconcordant cases and a randomly selected subset of concordant cases. This sampling approach was intentionally unbalanced to optimize limited physician adjudication resources, as concordance across all diagnosis and management criteria between AI and physician indicates alignment in decision making, requiring fewer cases to confirm quality.

Adjudication

Four expert physicians specializing in family, internal, and emergency medicine, each with at least 10 years of experience, reviewed the cases using a 2-step process. The initial review involved 2 adjudicators independently examining each case. The review included the patient's intake questionnaire responses, the video encounter transcript, and the physician's diagnosis and management decisions. Approximately 20 minutes were spent per case. Adjudicators were blinded to the identities of the original patient and physician. Scores were provided for both the physician's and AI's diagnosis and management decisions using a 4-point scale: optimal management (appropriate diagnosis and guideline-adherent management), reasonable (for example, recommending treatment of likely viral pharyngitis, contrary to guidelines), inadequate (for example, prescribing antibiotics for a viral condition), and potentially harmful (for example, failure to refer a seemingly urgent case to the ED). Deviations from guidelines were evaluated in context (for example, considering a patient's reported travel plans when evaluating antibiotic prescriptions for borderline sinus infections). See Supplement 1 for detailed adjudication instructions.

In cases where the scores of the 2 adjudicators for either the AI or the physician differed by more than

1 level (namely, optimal vs. inadequate or potentially harmful, or reasonable vs. potentially harmful), a third adjudicator was involved. The 3 adjudicators then convened to discuss the case. Consensus was sought but not forced, with each adjudicator recording their final score. An adjudicator summarized the key reasons in cases where the AI and physician scores differed.

Statistical Analysis

A 4×4 contingency table was constructed comparing adjudicator-assigned scores between AI and physician decisions. Marginal distributions of scores for AI and physician decisions were computed, as well as the proportions of cases in which the AI score was the same, higher, or lower than the physician score. These statistics were calculated for the sample of all cases and as auxiliary analyses, separately for each group of acute symptoms.

All analyses incorporated 2 weighting factors. First, to account for stratified sampling, weights were applied to each of the 2 strata (concordant and nonconcordant cases) to reflect the ratio of the sample size to the population size within each stratum. Second, to account for the variable number of adjudicators per case, case-level weights were applied: one half for each score in cases reviewed by 2 adjudicators and one third for each score in cases reviewed by 3 adjudicators. When summing observations, the final weight for each adjudicator score was calculated as the product of these 2 weights. When these combined weights were applied, the weighted sum of adjudicator scores equaled the total study population for both AI and physician assessments.

Confidence intervals for proportions were calculated using the logit-transformed binomial method via the "svyciprop" function from the "survey" package in R with the "logit" option (15). This approach fits a logistic regression model to estimate the log-odds of the proportion, computes a Wald-type confidence interval on the log-odds scale, and back-transforms it to the probability scale using the inverse logit function. The logit transform is particularly advantageous for small proportions near 0, as it ensures that the confidence interval remains within the (0 to 1) range, avoiding negative proportions. This method also accounts for the effect of the complex sampling design, including stratified sampling and clustering at the case level, on estimation uncertainty (16).

Interrater reliability was calculated using Cohen's Kappa to measure agreement beyond chance between adjudicators (17). Kappa values were computed for each pair of adjudicators using 4×4 contingency tables of their scores for both AI and physicians in jointly reviewed cases. The "cohen.kappa" function from the "psych" package in R was used to calculate Kappa for each pair (18). A weighted mean Kappa was derived to summarize agreement across all adjudicator pairs, with weights proportional to the number of cases reviewed per pair.

Role of the Funding Source

The study was sponsored by K Health. Cedars-Sinai is involved in a joint venture for the CS Connect project. The sponsor had a role in extracting and analyzing data from EHRs, conducting case adjudication, and providing logistic and administrative support and supervision, as performed by its employees (Z.K., L.H., T. Brufman, T. Beer, R.I., K.L., I.F., and R.S.). D.Z. (Tel Aviv University) and J.P. (Cedars-Sinai) were responsible for the study design, with D.Z. supervising the statistical analysis. Data interpretation and manuscript preparation were led by D.Z., J.P., and C.G. (Cedars-Sinai), who retained full independence in interpreting the results and drafting the manuscript. The sponsor reviewed the manuscript draft to ensure the exclusion of confidential information, but did not influence the interpretation of the data, the conclusions drawn, or the decision to submit the manuscript for publication.

RESULTS

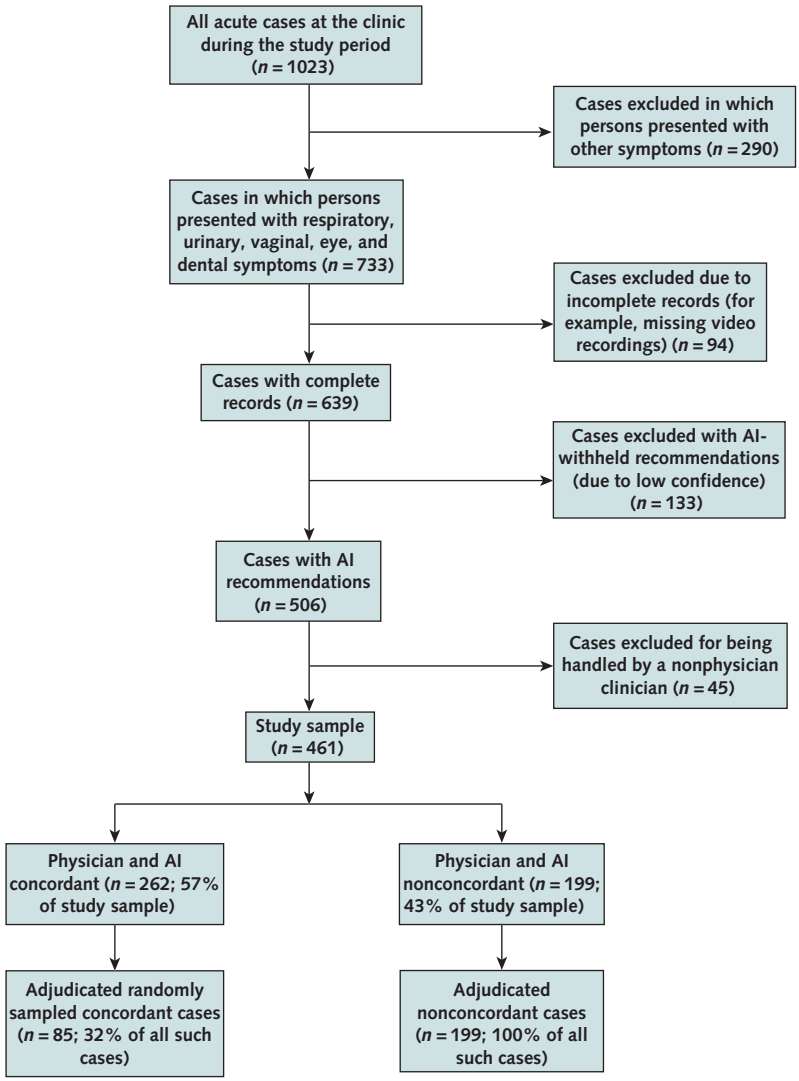
Sample Characteristics

During the study period, 1023 visits were made to the virtual clinic by adults with acute symptoms (Figure 2). Of these, 733 met the inclusion criteria. After excluding 94 cases due to incomplete records (generally due to technical issues), 133 cases where the AI withheld recommendations, and 45 cases that were managed by a nonphysician advanced practice clinician, the final sample consisted of 461 cases. Table 1 summarizes the descriptive statistics of this sample. The mean age was 45.3 years, and 70.2% of patients were female. Common background conditions included anxiety/depression (44.7%) and dyslipidemia (36.2%), with an average of 2.1 chronic conditions per patient. Acute symptoms were predominantly respiratory (65.3%), followed by urinary (20.4%), vaginal (6.9%), eye (6.5%), and dental (0.9%). Adjudicated cases involved visits with 18 physicians. Physician specialties included family medicine (10 physicians), internal medicine (4 physicians), and emergency medicine (4 physicians). All physicians had a minimum of 2 years of postresidency clinical experience.

Physician decisions were classified as concordant with AI recommendations in 262 cases (56.8%), 85 of which were randomly selected for adjudication. All 199 nonconcordant cases (43.2%) were adjudicated. Stratum sampling weights were 3.08 for concordant cases and 1.0 for nonconcordant cases, reflecting their respective sampling fractions. A third adjudicator was required in 116 cases, comprising 2 concordant and 114 nonconcordant cases.

The adjudication process yielded 1368 scores, corresponding to 684 paired scores (1 for the AI and 1 for the physician) across 284 adjudicated cases. For nonconcordant cases, 512 paired scores were generated: 170 from 85 cases reviewed by 2 adjudicators and 342 from 114 cases with 3 adjudicators. For concordant cases, 172 paired scores were generated: 166 from 83 cases with 2 adjudicators and 6 from 2 cases with

Figure 2 . Sample construction.



The diagram shows the sample inclusion and exclusion criteria. The 94 excluded incomplete records are cases missing critical information due to technical factors: 42 records with missing intake data and 36 with missing video recordings; 13 cases with partial intake data due to errors in either the mobile application or questionnaire module collecting patient responses; and 3 cases where adjudicators could not access the patient electronic health record when performing the adjudication. By its design, the AI model withheld recommendations due to low confidence in 133 cases (see Clinical Context section for details). Concordant cases refer to cases where physicians' diagnoses and management decisions were congruent with the AI recommendations. Nonconcordant cases refer to cases where they differed. Adjudicated samples include a random subsample of concordant cases ($n = 85$ adjudicated cases of 262 cases) and all nonconcordant cases ($n = 199$). AI = artificial intelligence.

3 adjudicators. The mean pairwise Cohen's Kappa for interrater reliability of final scores was 0.756 (**Table 1 in Supplement 2**, available at Annals.org). Cross-tabulations of adjudicator scores for all pairs of adjudicators are presented in **Tables 2** to **6** in **Supplement 2**.

Joint Distribution of AI and Physician Scores

Figure 3 summarizes adjudicators' scores of the diagnostic and management decisions made by AI and physicians in a weighted sample of $\tilde{n} = 461$ cases (we use $\tilde{n}$ to denote weighted case counts, which may be noninteger). **Panel A** of **Figure 3** shows the cross-tabulation of scores assigned to AI recommendations (rows) and physician decisions (columns). **Panel B** of **Figure 3** summarizes the marginal distributions of scores for both AI and physicians. **Table 7** in **Supplement 2** presents the marginal distribution of physician and AI scores separately for each adjudicator.

In the cross-tabulation (**Figure 3,** *A*), the most frequent cell was optimal scores for both AI and physicians, occurring in 58.3% of cases (95% CI, 54.3% to 62.3%; $\tilde{n} = 268.9$). Overall, adjudicators rated AI and physician scores as equal in 67.9% of cases (CI, 64.8% to 70.9%; $\tilde{n} = 313.0$). The AI scores were higher than

Table 1. Sample Summary Statistics) ($n = 461$)*

Visit Characteristics	Summary Statistics
Demographic characteristics	
Mean age, y	45.3
Female, *n (%)*	322 (69.8)
Race, *n (%)*	
White	274 (59.4)
Black	55 (11.9)
Asian	51 (11.1)
Other	60 (13.0)
Unknown	21 (4.6)
Ethnicity, *n (%)*	
Hispanic of Latino	77 (16.7)
Not Hispanic of Latino	344 (74.6)
Unknown	40 (8.7)
Background conditions	
Anxiety/depression, *n (%)*	206 (44.7)
Dyslipidemia, *n (%)*	167 (36.2)
Obesity, *n (%)*	136 (29.5)
Hypertension, *n (%)*	101 (21.9)
Asthma, *n (%)*	67 (14.5)
Mean chronic conditions, *n*	2.1
Acute symptom, *n (%)*	
Respiratory	301 (65.3)
Urinary	94 (20.4)
Vaginal	32 (6.9)
Eye	30 (6.5)
Dental	4 (0.9)
Mean visit duration, *min*	17.7

* This table summarizes the demographic and clinical characteristics of the study sample. Presenting symptoms were classified into the following categories: respiratory, urinary, vaginal, eye, and dental. Respiratory symptoms included cough, sore throat, COVID-19, nasal congestion, sinus infection, upper respiratory infection, strep throat, and runny nose. Urinary symptoms included bladder infection, urinary tract infection, burning or painful urination, and urinary urgency. Vaginal symptoms included vaginal yeast infection, itch, or unusual discharge. Eye symptoms included eye infection, redness, or discharge. Dental symptoms include dental pain. Percentages may not sum to 100 due to rounding.

the physician scores in 20.8% of cases (CI, 17.8% to 24.0%; $\tilde{n} = 95.7$) and lower in 11.3% of cases (CI, 9.0% to 14.2%; $\tilde{n} = 52.2$). Cross-tabulations and marginal distributions for subsets defined by patient symptoms (except for dental, which were present in only 4 cases) are presented in **Figures 1** to **4** in **Supplement 2**.

Marginal distributions (**Figure 3,** *B*) indicate that adjudicators rated AI recommendations as optimal, the highest score, in 77.1% of cases (CI, 72.7% to 80.9%; $\tilde{n} = 355.3$), compared with 67.1% of cases for physician decisions (CI, 62.9% to 71.1%; $\tilde{n} = 309.5$). The AI recommendations were less frequently rated as potentially harmful, the lowest score (2.8% [CI, 1.4% to 5.2%]; $\tilde{n} = 12.7$), compared with physician decisions (4.6% [CI, 2.9% to 7.3%]; $\tilde{n} = 21.4$).

Across all symptom types, AI recommendations were rated higher than physician decisions in 14.4% to 40.8% of cases (**Figures 1** to **4** in **Supplement 2**). The highest proportion was observed in urinary symptoms ($\tilde{n} = 89.1$ cases), with 40.8% rating AI better (CI, 33.9% to 48.2%) compared with 9.2% rating physicians better (CI, 4.8% to 16.8%). For respiratory symptoms ($\tilde{n} = 298.8$ cases), AI was rated higher in 15.9% of cases (CI, 12.7% to 19.9%) compared with physicians being rated higher in 11.7% (CI, 8.8% to 15.4%). The lowest proportion was observed for vaginal symptoms ($\tilde{n} = 32.5$ cases), where AI and physician scores were rated higher, respectively, in equal proportions (14.4% AI better [CI, 6.7% to 28.0%] vs. 14.4% physician better [CI, 6.7% to 28.0%]).

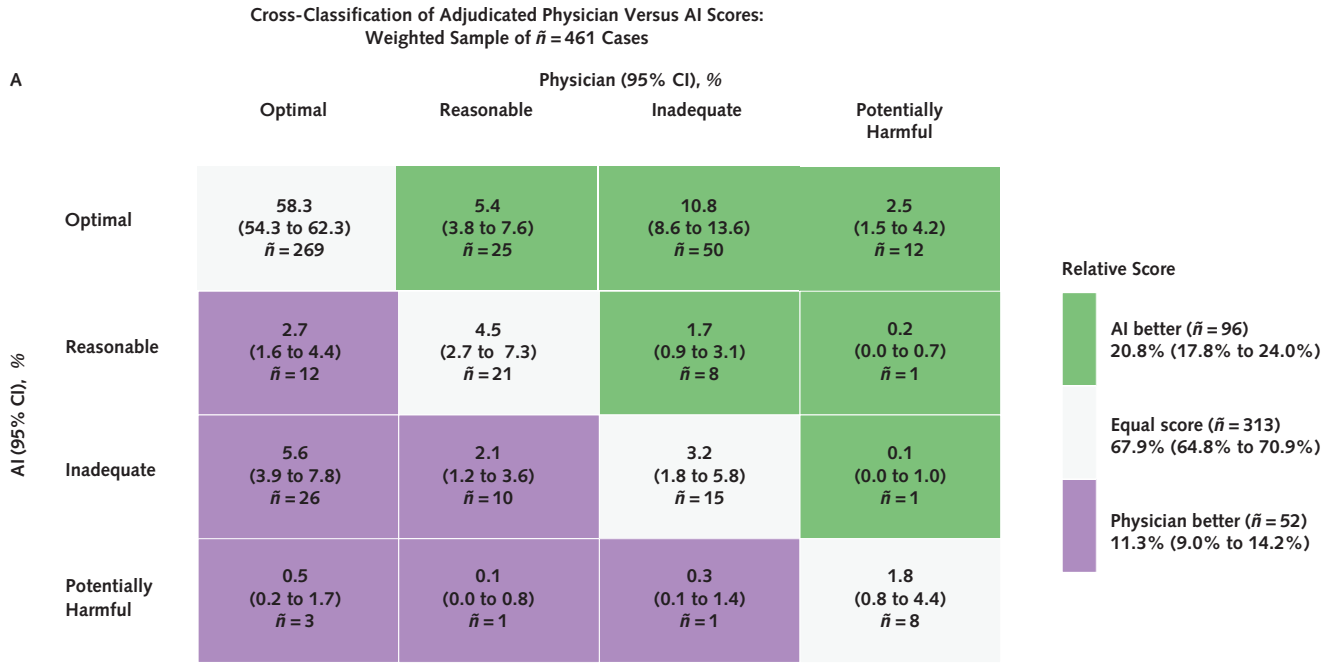
Examples of optimal recommendations include prescribing amoxicillin for a sore throat with a Centor score of 4 or ordering urinalysis and urine culture instead of treating empirically in a patient with recurrent urinary tract infections. Reasonable recommendations include ordering a bacterial culture for a patient with a sore throat and Centor score of 1 (19). Inadequate recommendations involved not ordering urinalysis and urine culture for a patient with recurrent urinary tract infections or prescribing antibiotics for viral upper respiratory infections. Potentially harmful recommendations included failing to refer cases of eye pain with suspected foreign body or COVID-19 with fatigue and shortness of breath to the ED, as well as prescribing medication without accounting for known drug allergies.

Table 2 summarizes adjudicator comments on all 111 nonconcordant cases where adjudicators rated AI and clinician recommendations differently and at least 1 was rated inadequate or potentially harmful. In 64% of these cases ($n = 71$), AI recommendations were rated better than physician decisions. The most common reasons for AI outperformance were physician omission of optimal laboratory and imaging referrals or unjustified empirical treatment (22.8%), deviation from clinical guidelines (16.3%), and omission of necessary in-person referrals to an ED, an urgent care center, or specialists (15.2%). Notably, physicians sometimes overlooked critical risk factors and red flags (4.4%), such as ocular pain with contact lens use. Conversely, physicians were rated better in 36% of cases ($n = 40$). The primary reasons for physician superiority included avoiding inappropriate ED referrals (8.0%), better handling of evolving or inconsistent patient-reported histories (6.2%), making necessary in-person referrals that the AI omitted (5.9%), and correcting AI diagnoses based on virtual physical examination validation (4.4%).

DISCUSSION

We found that initial AI recommendations in virtual urgent care visits for common symptoms were generally concordant or rated by physician adjudicators as better than final physician recommendations. The AI diagnosis and management recommendations were more likely to be rated as optimal (77.1% [CI, 72.7% to 80.9%]) compared with physicians (67.1% [CI, 62.9% to 71.1%]) and less likely to be rated as potentially harmful

Figure 3 . Comparison of adjudicated scores of AI and physician recommendations.



B

Marginal Score Distributions

Score	Physician, ñ	Percentage (95% CI)	AI, ñ	Percentage (95% CI)
Optimal	309	67.1 (62.9 to 71.1)	355	77.1 (72.7 to 80.9)
Reasonable	56	12.1 (9.3 to 15.7)	42	9.1 (6.6 to 12.4)
Inadequate	74	16.1 (13.2 to 19.4)	51	11.1 (8.5 to 14.3)
Potentially harmful	21	4.6 (2.9 to 7.3)	13	2.8 (1.4 to 5.2)

AI = artificial intelligence; ñ = weighted case count. **A.** A 4 × 4 contingency table comparing adjudicator-assigned scores for AI and physician diagnostic and management recommendations across 461 weighted cases. Each cell shows the weighted count (ñ) and proportion (%) of cases assigned the corresponding combination of AI (rows) and physician (columns) scores, with 95% CIs for the proportions in parentheses. Scores were weighted to reflect stratified sampling and the variable number of adjudicators per case, resulting in noninteger counts; details are provided in the text and counts shown are rounded to the nearest integer (see **Figure 5 in Supplement 2** for nonrounded counts). Shading represents relative score categories: AI better, equal score, or physician better. The key includes the proportion of cases in each category, with 95% CIs in brackets. Totals may not sum to 100% due to rounding. **B.** Marginal distributions for AI and physician scores are displayed, with 95% CIs in parentheses. Totals may not sum to 100% due to rounding.

(2.8% [CI, 1.4% to 5.2%] vs. 4.6% [CI, 2.9% to 7.3%]). These findings align with prior studies highlighting AI's potential in medical decision support in radiology, cardiology, and pathology (5, 20, 21) and extends that potential to diagnosis and management of common medical symptoms in a real-world virtual urgent care clinic.

Our observations suggest that AI showed particular strength in adhering to clinical guidelines, recommending appropriate laboratory and imaging tests, and recommending necessary in-person referrals. It outperformed physicians in avoiding unjustified empirical treatments and recognizing key risk factors that may trigger a change in diagnosis or management. Conversely, physicians excelled in adapting to evolving or inconsistent patient narratives, where the information disclosed during the consultations differed from the information provided during the chat intake questionnaire. Physicians also seemed to demonstrate better judgment in avoiding unnecessary ED referrals and in accurately diagnosing conditions requiring visual assessment; AI that is augmented by photographic validation could help address this latter finding.

Our findings suggest that well-designed AI decision support has the potential to improve clinical decision making for common acute symptoms. This may result, for example, from AI's ability to identify relevant clinical EHR information not readily apparent to a physician or by better adherence to evidence-based guidelines. Because the interface in use at the time did not optimize physician viewing of these recommendations and we do not know whether physicians used them, we believe our findings represent a conservative

Table 2. Adjudicator Reasons for Difference in Scores Between AI and Clinician Rating*

Adjudicator Reasons for Difference	Share of Cases, %	Examples
Reason AI was rated better (*n* = 71 [64%])		
Physician omission of optimal laboratory and imaging referrals or unjustified empirical treatment†	22.8	Failure to order urinalysis/urine culture for recurrent urinary tract infections; inadequate evaluation of persistent cough; omission of sexually transmitted infection testing despite reported risk factors
Physician deviation from clinical guidelines	16.3	Inappropriate prescription of antibiotics for viral infections; unwarranted use of oral corticosteroids for viral upper respiratory tract infections; utilization of third-line antibiotics for uncomplicated urinary tract infections
Physician omission of in-person referrals (ED/urgent care/specialist)†	15.2	Lack of referral for in-person evaluation for patients with worsening symptoms or after several telemedicine consultations for the same problem
Physician overlooked risk factors and red flags†	4.4	Oversight of crucial warning signs or risk factors, such as dyspnea in patients with upper respiratory infection symptoms, ocular pain associated with contact lens use, or foreign body presence
Inappropriate physician laboratory test referrals	2.1	Inappropriate recommendation for throat culture when Centor score is 0 to 1
Other reasons	2.1	Oversight of important anamnestic information like very recent COVID-19 infection as a trigger for cough; premature termination of video consultation by patient
Reason physician was rated better (*n* = 40 [36%])		
AI inappropriate ED referrals	8.0	Unnecessary ED referral for young, healthy patients with COVID-19
Evolving or inconsistent patient narratives‡	6.2	Discrepancies in patient-reported information between initial intake and subsequent video consultation, such as recent COVID-19 infection disclosure or contradictory statements about dyspnea
AI omission of in-person referrals (ED/urgent care/specialist)‡	5.9	Failure to refer for otolaryngological assessment after 3 sinus infections within a 12-month period
Incorrect AI diagnosis	4.4	Misalignment between visual findings during video consultation and information recorded during initial intake (for example, subconjunctival hemorrhage or oral candidiasis)
AI omission of required laboratory and imaging referrals or unjustified empirical treatment	3.3	Inadequate evaluation of persistent cough; no throat culture ordered to rule out strep
AI overlooked risk factors and red flags‡	3.1	Oversight of medical background of congestive heart failure in a person presenting with nocturnal cough and orthopnea
Other reasons	2.7	–
Reason not provided	1.8	–

AI = artificial intelligence; ED = emergency department.
* This table summarizes the types of reasons provided by adjudicators for ratings of the weighted sample of 111 cases in which AI recommendations and physician actions were nonconcordant, rated differently by adjudicators, and at least 1 of the 2 was rated either inadequate or potentially harmful.
† Includes cases where 1 or more adjudicators evaluated the diagnosis and management recommendations of the physician as potentially harmful.
‡ Includes cases where 1 or more adjudicators evaluated the diagnosis and management recommendations of AI as potentially harmful.

estimate of the potential for AI to improve care in this setting.

The study's strengths include its real-world setting, the use of several expert adjudicators, and a comprehensive evaluation of both AI and physician decision making across common acute symptoms. However, several limitations warrant consideration. Information about whether physicians used the AI recommendations would result in greater confidence about the potential benefits of AI in this setting. The retrospective design limits insight into how real-time AI recommendations influence physician behavior. Adjudicators had access only to consultation transcripts, not video footage, and were not blinded to the source of the diagnosis and management recommendation, due to their inherently different format (structured AI recommendations and a mix of structured and unstructured physician recommendations), which may have introduced bias. More than half of the visits with nonconcordant recommendations required a third adjudicator, suggesting that determination of the quality of the recommendations had inherent subjectivity. The absence of patient follow-up data also leaves uncertainty about the true impact on care quality or outcomes. Finally, the single-center design, mostly female patients, and limited category of symptoms limits the generalizability of these findings, and the small sample size limits the ability to evaluate for potential algorithmic bias.

Future research should include multicenter prospective studies to validate findings and explore AI integration into clinical workflows, particularly its impact on the appropriateness and safety of clinical decision making, patient outcomes, and health care utilization across diverse settings and patient populations. As AI is increasingly being considered as a tool to support clinical care, careful evaluation is needed to understand how AI capabilities can best complement human decision making and how best to incorporate AI's capabilities into routine clinical workflows while preserving physician oversight.

In conclusion, this study suggests that AI can enhance clinical decision making for common acute symptoms in

a virtual urgent care setting. Further study is needed to understand whether AI can enhance decision making for more complex patient care needs. Thoughtful integration of AI into clinical practice, combining its strengths with those of physicians, could improve the quality of care.

From Tel Aviv University, Tel Aviv, Israel (D.Z.); K Health, New York, New York (Z.K., L.H., T.Brufman, R.I.B., K.L., T.Beer, I.F., R.S.); and Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California (C.G., J.P.).

Acknowledgment: The authors acknowledge Yael Steuerman, PhD; Ishay Bitton; Michal Hershkovitz; Oron Mozes; Yaniv Cohen; and Zachary Siegel for technical support and Kevin Stephens, MD; David Morley, MD; and Neil Brown, MD, for medical support.

Financial Support: By K Health.

Disclosures: Disclosure forms are available with the article online.

Reproducible Research Statement: *Study protocol:* Available upon request from the corresponding author. *Statistical code:* Posted in **Supplement 3** (available at Annals.org). *Data set:* The data used in this study contain identifiable patient information and are not publicly available to protect privacy. However, deidentified data are available for replication purposes upon request, subject to approval by the Cedars-Sinai Institutional Review Board and measures to protect patient confidentiality. Researchers seeking access should contact the corresponding author for details on the data request process.

Corresponding Author: Dan Zeltzer, PhD, Berglas School of Economics, Tel Aviv University, Tel Aviv, Israel 6997801; e-mail, dzeltzer@tauex.tau.ac.il.

Author contributions are available at Annals.org.

References

1. **Esteva A, Kuprel B, Novoa RA, et al.** Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115-118. [PMID: 28117445] doi:10.1038/nature21056
2. **Wenderott K, Krups J, Zaruchas F, et al.** Effects of artificial intelligence implementation on efficiency in medical imaging-a systematic literature review and meta-analysis. NPJ Digit Med. 2024;7:265. [PMID: 39349815] doi:10.1038/s41746-024-01248-9
3. **Siontis K, Noseworthy P, Attia Z, et al.** Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol. 2021;18:465-478. [PMID: 33526938] doi:10.1038/s41569-020-00503-2
4. **Mullainathan S, Obermeyer Z.** Diagnosing physician error: a machine learning approach to low-value health care. The Quarterly Journal of Economics. 2022;137:679-727. doi:10.1093/qje/qjab046
5. **Shafi S, Parwani A.** Artificial intelligence in diagnostic pathology. Diagn Pathol. 2023;18:109. [PMID: 37784122] doi:10.1186/s13000-023-01375-z
6. **Weiss J, Raghu VK, Paruchuri K, et al.** Deep learning to estimate cardiovascular risk from chest radiographs: a risk prediction study. Ann Intern Med. 2024;177:409-417. [PMID: 38527287] doi:10.7326/M23-1898
7. **Dagan N, Magen O, Leshchinsky M, et al.** Prospective evaluation of machine learning for public health screening: identifying unknown hepatitis C carriers. N Engl J Med AI. 2024;1. doi:10.1056/AIoa2300012
8. **Lam T, Cheung M, Munro Y, et al.** Randomized controlled trials of artificial intelligence in clinical practice: systematic review. J Med Internet Res. 2022;24:e37188. [PMID: 35904087] doi:10.2196/37188
9. **Kueper J, Terry A, Zwarenstein M, et al.** Artificial intelligence and primary care research: a scoping review. Ann Fam Med. 2020;18:250-258. [PMID: 32393561] doi:10.1370/afm.2518
10. **Susanto A, Lyell D, Widyantoro B, et al.** Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. J Am Med Inform Assoc. 2023;30:2050-2063. [PMID: 37647865] doi:10.1093/jamia/ocad180
11. **Labkoff S, Oladimeji B, Kannry J, et al.** Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc. 2024;31:2730-2739. [PMID: 39325508] doi:10.1093/jamia/ocae209
12. **Zeltzer D, Herzog L, Pickman Y, et al.** Diagnostic accuracy of artificial intelligence in virtual primary care. Mayo Clinic Proceedings: Digital Health. 2023;1:480-489. doi:10.1016/j.mcpdig.2023.08.002
13. **Watson-Daniels J, Parkes DC, Ustun B.** Predictive multiplicity in probabilistic classification. In: Williams B, Chen Y, Neville J, eds. Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, 7–14 February 2023. AAAI-23 Technical Tracks 9; 2023;37:10306-10314. doi:10.1609/aaai.v37i9.26227
14. **Kompa B, Snoek J, Beam A.** Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4:4. [PMID: 33402680] doi:10.1038/s41746-020-00367-3
15. **Lumley T, Gao P, Schneider B.** survey: Analysis of Complex Survey Samples. 2024. Accessed at https://cran.r-project.org/web/packages/survey/index.html on 13 January 2025.
16. **Korn EL, Graubard BI.** Confidence intervals for proportions with small expected number of positive counts estimated from survey data. Survey Methodology. 1998;24:193-201.
17. **Cohen J.** A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20:37-46. doi:10.1177/001316446002000104
18. **Revelle W.** psych: Procedures for Psychological, Psychometric, and Personality Research [Internet]. 2024. Accessed at https://cran.r-project.org/web/packages/psych/index.html on 13 January 2025.
19. **Centor RM, Witherspoon JM, Dalton HP, et al.** The diagnosis of strep throat in adults in the emergency room. Med Decis Making. 1981;1:239-246. [PMID: 6763125] doi:10.1177/0272989X8100100304
20. **De Fauw J, Ledsam JR, Romera-Paredes B, et al.** Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24:1342-1350. [PMID: 30104768] doi:10.1038/s41591-018-0107-6
21. **Hannun AY, Rajpurkar P, Haghpanahi M, et al.** Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25:65-69. [PMID: 30617320] doi:10.1038/s41591-018-0268-3

Author Contributions: Conception and design: D. Zeltzer, Z. Kugler, L. Hayat, T. Brufman, R. Ilan Ber, K. Leibovich, R. Shaul, C. Goldzweig.
Analysis and interpretation of the data: D. Zeltzer, Z. Kugler, L. Hayat, T. Brufman, R. Ilan Ber, K. Leibovich, T. Beer, I. Frank, C. Goldzweig, J. Pevnick.
Drafting of the article: D. Zeltzer, Z. Kugler, K. Leibovich, C. Goldzweig.
Critical revision of the article for important intellectual content: D. Zeltzer, Z. Kugler, T. Brufman, R. Ilan Ber, K. Leibovich, R. Shaul, C. Goldzweig, J. Pevnick.
Final approval of the article: D. Zeltzer, Z. Kugler, L. Hayat, T. Brufman, R. Ilan Ber, K. Leibovich, T. Beer, I. Frank, R. Shaul, C. Goldzweig, J. Pevnick.
Statistical expertise: D. Zeltzer, R. Ilan Ber, K. Leibovich.
Obtaining of funding: R. Shaul.
Administrative, technical, or logistic support: L. Hayat, T. Beer, I. Frank, J. Pevnick.
Collection and assembly of data: L. Hayat, K. Leibovich, T. Beer.