ORIGINAL ARTICLE

MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

MAYO CLINIC

Diagnostic Accuracy of Artificial Intelligence in Virtual Primary Care

Dan Zeltzer, PhD; Lee Herzog, MD; Yishai Pickman, PhD; Yael Steuerman, PhD; Ran Ilan Ber, PhD; Zehavi Kugler, MD; Ran Shaul, BASc; and Jon O. Ebbert, MD MSc

Abstract

Objective: To evaluate the diagnostic accuracy of artificial intelligence (AI)-generated clinical diagnoses.

Patients and Methods: A retrospective chart review of 102,059 virtual primary care clinical encounters from October 1, 2022, to January 31, 2023 was conducted. Patients underwent an AI medical interview, after which virtual care providers reviewed the interview summary and AI-provided differential diagnoses, communicated with patients, and finalized diagnoses and treatment plans. Our accuracy measures were agreement between AI diagnoses, virtual care providers, and blind adjudicators. We analyzed AI diagnostic agreement across different diagnoses, presenting symptoms, patient demographic characteristics such as race, and provider levels of experience. We also evaluated model performance improvement with retraining.

Results: Providers selected an AI diagnosis in 84.2% (n = 85,976) of cases and the top-ranked AI diagnosis in 60.9% (n = 62,130) of cases. Agreement rates varied by diagnosis, with greater than or equal to 95% provider agreement with an AI diagnosis for 35 diagnoses (47% of cases, n = 47,679) and greater than or equal to 90% agreement for 57 diagnoses (69% of cases, n = 70,697). The average agreement rate for half of all presenting symptoms was greater than or equal to 90%. Adjusting for case mix, diagnostic accuracy exhibited minimal variation across demographic characteristics. The adjudicators' consensus diagnosis, reached in 58.2% (n = 128) of adjudicated cases was always included in the AI differential diagnosis. Provider experience did not affect agreement, and model retraining increased diagnostic accuracy for retrained conditions from 96.6% to 98.0%.

Conclusion: Our findings show that agreement between AI and provider diagnoses is high in most cases in the setting of this study. The results highlight the potential for AI to enhance primary care disease diagnosis and patient triage, with the capacity to improve over time.

 ■ Mayo Clin Proc Digital Health 2023;1(4):480-489

From the Berglas School of Economics, Tel Aviv University, Tel Aviv, Israel (D.Z.); K Health Inc, New York, NY (L.H., Y.P., Y.S., R.I.B., Z.K., R.S.); and Department of Medicine, Mayo Clinic, Rochester, MN (J.O.E.).

Artificial intelligence (AI) has emerged as a powerful tool with the potential to transform health care domains, including diagnosis and clinical decision-making. The presence of AI in community-based primary health care has focused on diagnosis, detection, and surveillance.[1,2] AI in clinical primary care has focused on the development of AI methods to support provider diagnosis and treatment recommendations, but few studies have evaluated AI applications in the real-world setting, particularly in virtual primary care.[3,4] Previous assessments of digital symptom checkers and medical chatbots have often relied on small sets of clinical vignettes or patient encounters.[5–7] Small sample sizes limit generalizability and the ability to evaluate differences in diagnostic accuracy across diagnoses and patient groups.

To address this gap, we evaluated the diagnostic accuracy of AI-based recommendations by analyzing over 102,059 AI-augmented clinical encounters from a large virtual primary care practice. We assessed the accuracy of AI recommendations across diagnoses, presenting symptoms, and patient demographic characteristics. We also evaluated AI performance improvement after retraining of the prediction model with new clinical data from the

virtual primary care service. By analyzing concurrence between AI-suggested and virtual care provider diagnoses, we sought to identify opportunities for AI to support clinical case resolution.

METHODS

Clinical Setting

We used data from K Health Inc, a technology company that operates an affiliated virtual primary care practice across 48 continental US states.[8] The practice uses AI for patient intake and diagnostics. Patients access the service through the web or a mobile application and initiate the visit by typing their medical concern and entering structured information about their demographic characteristics. An AI medical chat asks patients about medical risk factors and presenting symptoms, conducting a structured dynamic interview to gather symptom-related information and medical history, in which, on average, 25 questions are asked over 5 minutes. Patients receive a list of possible conditions related to their symptoms and can then choose to proceed to a virtual visit with a provider. At the start of the visit, the provider reviews, through the electronic medical record (EMR) system, a detailed intake summary and the AI-generated differential diagnosis, predicted on the basis of patient-provided information. The differential diagnosis consists of a maximum of 5 (median 2, mean 2.5, SD 1.0) most likely diagnoses, ordered by likelihood. Offering providers a list of potential diagnoses broadens their scope of consideration, potentially aiding the recognition of less common diseases. On the platform, virtual care providers are instructed to use their independent clinical judgment when rendering care. They conduct patient interviews by text or video before making their final diagnosis and treatment decisions. Case management may include test orders, medication prescriptions, follow-ups, and referrals to various care settings.

The baseline version of the medical chat and diagnostic algorithms was initially trained on an EMR dataset of 2 million patients.[7] The medical chat uses simplified diagnostic predictions to ask structured questions most related to reported symptoms. The current diagnostic model consists of an ensemble of expert diagnostic classifiers, each for a group of related conditions (eg, genitourinary conditions). As the platform continues to handle new cases, the classifiers are routinely updated through retraining and replacement processes that leverage the collective experience of providers on the platform. The diagnostic model now covers 156 common adult primary care conditions, including the top reasons for primary care visits globally.[9] Additional information on model training is included in Supplementary Appendix A (available online at https://www.mcpdigitalhealth.org/).

Sample and Main Variables

The study population included all virtual primary care short-term care patient visits between October 1, 2022, and January 31, 2023. This period was chosen to minimize the potential effect on accuracy measurements because of changes in the provider's EMR platform user experience and condition sets that the AI model can diagnose. We focused on the most recent period during which the platform was stable, with no such changes. The study period also includes one instance where the diagnostic algorithm for 6 common genitourinary conditions among women, originally trained on data sourced from external datasets, was retrained on a new batch of clinical data sourced from the same virtual primary care setting where the algorithm has been used. Model accuracy for these conditions was assessed before and after the release of the new model on November 8, 2022.

For each case, we observed the patient's self-reported age, sex, race or ethnicity, existing long-term conditions, presenting symptoms, and other related symptoms. Patient demographic characteristics, clinical histories, and self-reported symptoms were documented and de-identified to ensure patient confidentiality. We also observed the full set of algorithmic diagnosis suggestions made to the provider and the International Classification of Diseases (ICD)-10 diagnosis recorded by the provider at the end of the visit. During our study period, 586,819 users started a conversation with the medical chat, of which 83% (n=490,705) completed the conversation and 20% (n=117,795) proceeded to complete a clinical encounter

with a provider. Of these encounters, we excluded 8.4% (n=9,946) of cases in which providers recorded ICD-10 codes for nonspecific conditions (eg, A49.9 bacterial infection, unspecified), symptoms (eg, R30.0 dysuria), or administrative tasks (eg, Z76.0 encounter for issue of repeat prescription), as in such cases, no diagnosis was made by the provider. We also excluded from the main study sample 5.4% (n=5,790) of the remaining cases in which providers recorded more than 1 diagnosis, as in such cases defining diagnostic agreement is complicated by the multiplicity of provider diagnoses. The resulting sample includes 102,059 visits (cases) with 133 primary care providers. Of these providers, 103 (77%) were physicians and 30 (23%) were advanced practice providers. Among physicians, the most common specialties were family medicine (51.5%, n=53), internal medicine (26.2%, n=27), and emergency medicine (22.3%, n=23). Among physicians, 10.7% (n=11) were double board-certified. Physicians had a minimum of 2 years of post-residency clinical experience. Age was available for 75.9% of providers, and average age was 46.0 years (SD 9.8, n=101). Of the providers 88.3% were women (physicians: 61.2%, n=63; advanced practice providers: 93.3%, n=28). The average time spent practicing on the virtual care platform was 20.6 months (SD 4.9, n=133).

Table 1 summarizes the main study sample. The patient population exhibits diversity in age, sex, race or ethnicity, and presenting symptoms, mirroring patient panels typically observed in primary care populations. The cases included a wide variety of clinical conditions commonly encountered in primary care settings. The cases include 215 presenting symptoms and 992 distinct ICD-10 diagnoses. Diagnoses and presenting symptoms are listed in Supplementary Tables 1 and 2 (available online at https://www.mcpdigitalhealth.org/).

To enhance our evaluation of the correct diagnosis in each case and to assess the potential over-reliance of providers on the AI model, 3 clinicians adjudicated a random subset of 220 cases from the main study sample. These cases were stratified by patient demographic characteristics to represent the same mixture of patients and cases as the original study population (Supplementary Table 3, available online at https://www.mcpdigitalhealth.org/). For each case, adjudicators were provided with a detailed AI medical interview and instructed to select the single most probable diagnosis. To mitigate potential bias, adjudicators were blinded to the AI diagnostic model suggestions, the provider-selected diagnosis, the conversation between the provider and the patient, and the identity of the provider in each case.

TABLE 1. Patient and Visit Characteristics of 102,059 Virtual Primary Care Cases[a,b]

Patient Characteristics	Number of Cases	Share of Cases (%)
Sex		
Woman	77,129	75.6
Man	24,930	24.4
Age group (y)		
18-29	32,974	32.3
30-39	35,437	34.7
40+	33,648	33.0
Race or ethnicity		
Caucasian	62,004	60.8
Black or African American	8956	8.8
Hispanic or Latino	8206	8.0
All other[c]	8811	8.6
Declined	14,082	13.8

Visit Characteristics	Mean	SD	Median	IQR
Patient age (y)	36.2	11.1	34	28-43
Number of patient-reported symptoms	6.6	6.6	6	4-9
Number of long-term conditions	0.5	0.7	0	0-1
Number of AI-suggested diagnoses	2.5	1.0	2	2-3

[a]ICD, International Classification of Diseases; IQR, interquartile range
[b]Cases represent 88,535 patients and 133 providers, covering 215 presenting symptoms and 992 unique ICD-10 diagnoses.
[c]Includes patients identifying as Asian (3.1%), American Indian or Alaskan Native (0.4%), Middle Eastern (0.4%), Other (1.7%), and patients reporting multiple races (3%).

Ethics and Privacy Protection

All patient data were de-identified in accordance with the Health Insurance Portability and Accountability Act requirements. The study was approved by the Institutional Review Board of Tel Aviv University (#2023-0006499-1). Informed consent was waived on the basis of the determination that this retrospective study of de-identified patient data meets the minimal risk criteria.

Diagnostic Agreement Measures

We assessed AI diagnostic accuracy by evaluating the agreement between the AI-proposed diagnoses and the diagnosis chosen by the provider for each case. We used 3 measures: (1) the percentage of cases in which the provider-selected diagnosis matched any of the diagnoses proposed by the AI model; (2) the percentage of cases in which the provider-selected diagnosis was top-ranked by the AI model; and (3) the percentage of cases in which the top-ranked diagnosis was selected by the provider (measures [2] and [3] use different denominators. That is, for a given diagnosis *d*, [2] is the ratio of cases where both the provider's diagnosis and the AI's top-ranked diagnosis were *d* to the total number of cases where the provider's diagnosis was *d*, whereas [3] is the same numerator, but divided by the total number of cases where the AI's top-ranked diagnosis was *d*) (Supplementary Appendix B, available online at https://www.mcpdigitalhealth.org/ provides an example of these calculations). The first measure emphasizes AI's potential to guide providers in diagnostic decision-making, whereas the second and third measures highlight AI's role in identifying the most probable diagnosis. When averaging these measures over multiple diagnoses, we calculated micro-average recall and precision. In these cases, measures (2) and (3) were identical.[10]

Adjudication Analyses

Although our primary analysis shed light on the AI model's diagnostic accuracy within a real-world clinical setting, we acknowledge that not every provider's diagnostic decision may be accurate. Errors in provider diagnosis can stem from 2 factors: variance and bias. First, a considerable body of research documents the high variance in expert decisions, including in the medical field.[11] Consequently, to the degree that providers make diagnostic errors, measures of disagreement with the AI may not accurately reflect the AI's true diagnostic accuracy. Second, given that virtual care providers were privy to the AI recommendations, these suggestions may have influenced their decisions, potentially introducing a bias if providers overly rely on AI. These considerations complicate our understanding of provider-AI agreement as a gauge of AI accuracy.

To address these concerns, we conducted a secondary analysis that compares the diagnoses made by the AI and the original case providers with those made by independent blind adjudicators, using the adjudication sample previously described. Two measures were calculated: (1) Majority Agreement—the average agreement of the AI's top diagnosis suggestion and the provider's diagnosis with the majority-selected adjudicator diagnosis; (2) Average Agreement—the average agreement among the adjudicator's diagnosis, AI top-ranked diagnosis, and provider diagnosis. This approach sheds light on the variance in provider diagnoses and potential biases introduced by sharing the AI diagnostic model's suggestions with the virtual care provider.

Subgroup Analyses

First, we calculated agreement measures separately for specific conditions defined on the basis of the provider-selected diagnosis. Granular analysis by diagnosis facilitates a more comprehensive understanding of the AI's diagnostic capabilities and potential role in varied clinical contexts. Second, we calculated average diagnostic accuracy separately for different cases sharing the same presenting symptom. This analysis assesses the AI model's potential for accurately recommending diagnoses based solely on information collected from patients before any input from a provider at the earliest stage of patient care. Recent work has documented the large potential benefits of such preliminary triage.[12]

To investigate whether the model exhibited consistent performance across different patient

TABLE 2. Virtual Care Provider and AI Diagnostic Agreement Rates in 102,059 Virtual Encounters

	Provider Diagnosis in AI Differential Diagnosis List	Provider Diagnosis Top-Ranked By the AI[a]	Top-Ranked AI Diagnosis Selected By the Provider[a]	N Cases	Share of Cases (%)
Diagnostic agreement					
All cases	84.2%	60.9%	60.9%	102,059	100.0%
Most common diagnoses[b]					
Bladder infection	98.9%	97.0%	95.0%	20,469	20.1%
Upper respiratory infection	97.3%	62.4%	55.4%	17,614	17.3%
Acute sinusitis	91.5%	57.0%	57.1%	11,057	10.8%
Dental infection	89.8%	82.5%	98.9%	7,399	7.2%
Vaginal yeast infection	96.4%	76.7%	70.2%	3,799	3.7%
Most common presenting symptoms[c]					
Burning with urination	95.9%	91.4%	91.4%	11,065	10.8%
Cough	82.1%	51.0%	51.0%	8,574	8.4%
Sore throat	88.1%	39.1%	39.1%	7,377	7.2%
Dental pain	90.7%	85.4%	85.4%	6,703	6.6%
Nasal congestion	92.1%	54.1%	54.1%	6,346	6.2%

Agreement Before and After Model Retraining[d]	Before Retraining	After Retraining	N Cases (Before)	N Cases (After)
Vaginal yeast infection	94.8%	97.5%	1,194	2,596
Pyelonephritis	90.0%	97.1%	219	349
Bladder infection	99.0%	99.4%	8,081	12,209
Bacterial vaginosis	78.6%	95.2%	859	2,733
Genital herpes	97.3%	97.7%	112	310
Vulvovaginitis	96.3%	79.8%	667	440
Average (weighted by number of cases)	96.6%	98.0%	11,132	18,637

[a]When averaging across multiple diagnoses, micro-average recall and precision (columns 2 and 3) are identical by definition. See methods section for details.
[b]Grouping by diagnosis is on the basis of the provider-selected diagnosis.
[c]Grouping by presenting symptoms refers to the average diagnostic accuracy calculated for groups of cases sharing the same patient-reported presenting symptoms.
[d]Accuracy measures before and after a model retraining occurring on November 8, 2022, for only female patients.

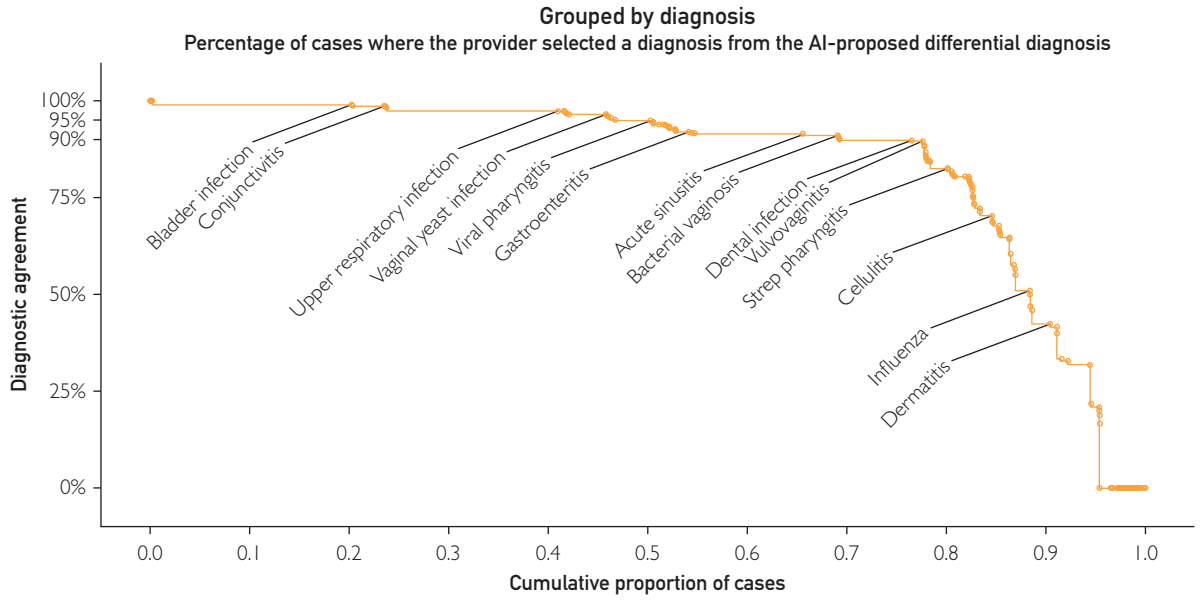
groups, we evaluated AI accuracy by patient age, sex, and race or ethnicity. We estimated a fixed-effects multivariate regression with controls for age group, race, and sex controls and fixed effects for the provider-selected diagnosis, adjusting for potential differences in case mix across different subgroups (R package fixest: Fast Fixed-Effects Estimations, version 0.11.1). To investigate whether overall agreement varies on the basis of physician tenure, we estimated the same fixed-effects multivariate regression with provider experience on the platform in months as the exposure variable. To investigate performance improvement related to model retraining, we compared AI accuracy before and after retraining, focusing on 6 diagnoses for which a new version of the model was released.

RESULTS

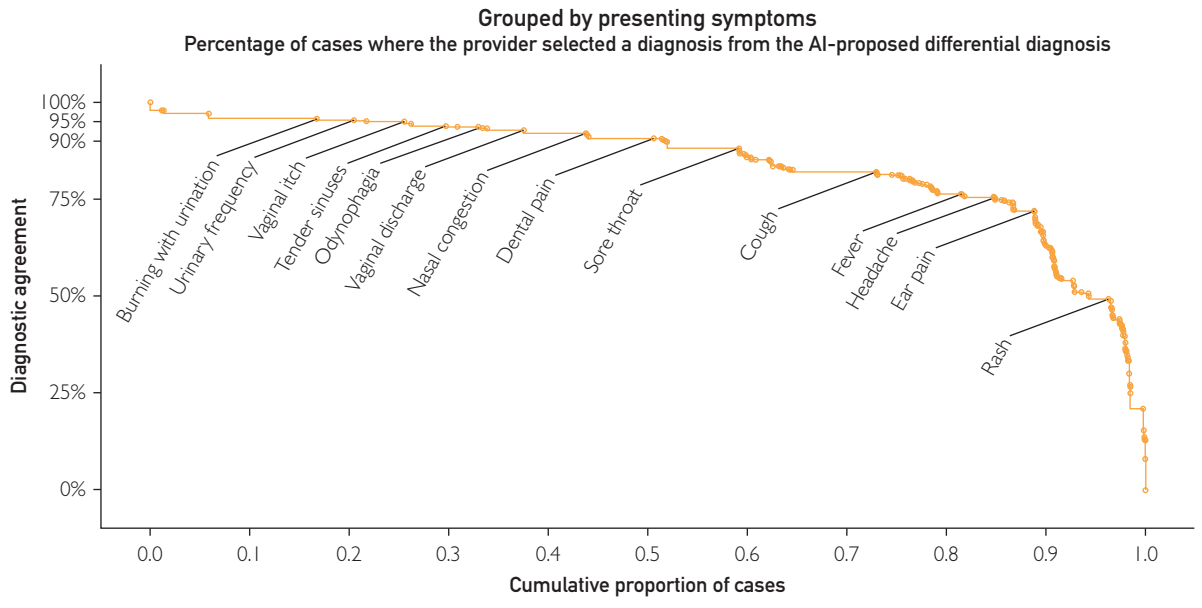
Overall Diagnostic Agreement

In 84.2% (n = 85,976) of cases, providers selected a diagnosis presented by the AI based on the automated interview (Table 2). In 60.9% (n = 62,130) of cases, the provider-selected diagnosis was top-ranked by the AI model. However, the provider agreement with the AI differential diagnoses varied across different provider-selected diagnoses, with high agreement rates for bladder infection (98.9% agreement) (Supplementary Table 1),

Mayo Clin Proc Digital Health ■ December 2023;1(4):480-489 ■ https://doi.org/10.1016/j.mcpdig.2023.08.002
www.mcpdigitalhealth.org



A



B

FIGURE 1. Accuracy distribution: AI and provider diagnostic agreement for 102,059 cases. The figure summarizes agreement between providers and AI-recommended differential diagnosis. The y-axis shows the rate of agreement, defined as the percentage of cases for which the provider selected a diagnosis from the AI-proposed differential diagnosis list. The x-axis shows the cumulative proportion of cases with diagnoses exceeding each level of agreement. (A) panel shows results grouped by each of the provider-selected diagnoses in the sample (the most common are labeled). (B) panel shows results grouped by each of 215 presenting symptoms in the sample as reported by the patient during automated intake (the most common are labeled). See text for detailed sample and measure definitions. AI, artificial intelligence.

TABLE 3. Multivariate Regression Analyses of Differences in AI and Provider Agreement Across Different Patient Subgroups

	Provider Agreement With Differential Diagnosis			
	Without Adjusting for Differences Between Groups in Demographic Characteristics and Case Mix[a]			Adjusting for Differences Between Groups in Demographic Characteristics and Case Mix[b]
A. Sex				
Woman	comparator			comparator
Man	−.092*** (.003)			.003 (.01)
B. Race/ethnicity				
Caucasian		comparator		comparator
Black or African American		−.03*** (.004)		−.01* (.005)
Hispanic or Latino		−.03*** (.004)		−.01* (.003)
Other ethnicity		−.02*** (.003)		−.002 (.003)
C. Age group				
18-29			comparator	comparator
30-39			−.01* (.003)	.0001 (.004)
40+			−.003 (.003)	−.003 (0.01)
Fixed-effects for diagnosis	no	no	no	yes
Observations	102,059	102,059	102,059	102,059
R^2	.012	.001	<.0001	.48

[a]Columns 1-3 show univariate regression estimates for the unadjusted difference in agreement with differential diagnosis relative to the comparator category, with standard errors in parentheses.
[b]Column 4 shows multivariate fixed-effect regression estimates for the same comparison, with adjustments for diagnosis and demographic characteristic differences among groups, with standard errors (in parentheses) clustered by diagnosis.
*P<.05, **P<.01, ***P<.001

conjunctivitis (98.7% agreement), and upper respiratory infection (97.3%), and lower agreement rates for conditions such as dermatitis (42.4% agreement), acute bronchitis (41.7% agreement), asthma (32.9% agreement), and unspecified abdominal pain (20.8% agreement).

Diagnostic Agreement by Presenting Symptom

Variations in accuracy were also seen across patient-reported symptoms. The AI model performs well in cases presenting with eye symptoms (eg, eye discharge or redness, 98.0% and 95.3% agreement, respectively) and urinary symptoms (eg, urinary urgency, burning with urination, and urinary frequency, 97.2%, 95.9%, and 95.5% agreement, respectively) (Supplementary Table 2). However, accuracy rates are lower for skin symptoms (eg, rash and itchy skin, 49.3% and 44.0% agreement, respectively) and mouth sores (21.0% agreement).

Distribution of Diagnostic Agreement Rates

Diagnostic agreement between the AI model and virtual care providers varies by diagnosis. For diagnoses at the 75th and 25th percentiles of the accuracy distribution, providers selected an AI diagnostic recommendation in 97.3% and 89.8% of cases, respectively, and the top-ranked diagnosis in 84.6% and 51.5% of cases, respectively (Figure 1A). For the most accurate 47% of cases covering 35 diagnoses, including common diagnoses such as bladder infection, conjunctivitis, and upper respiratory

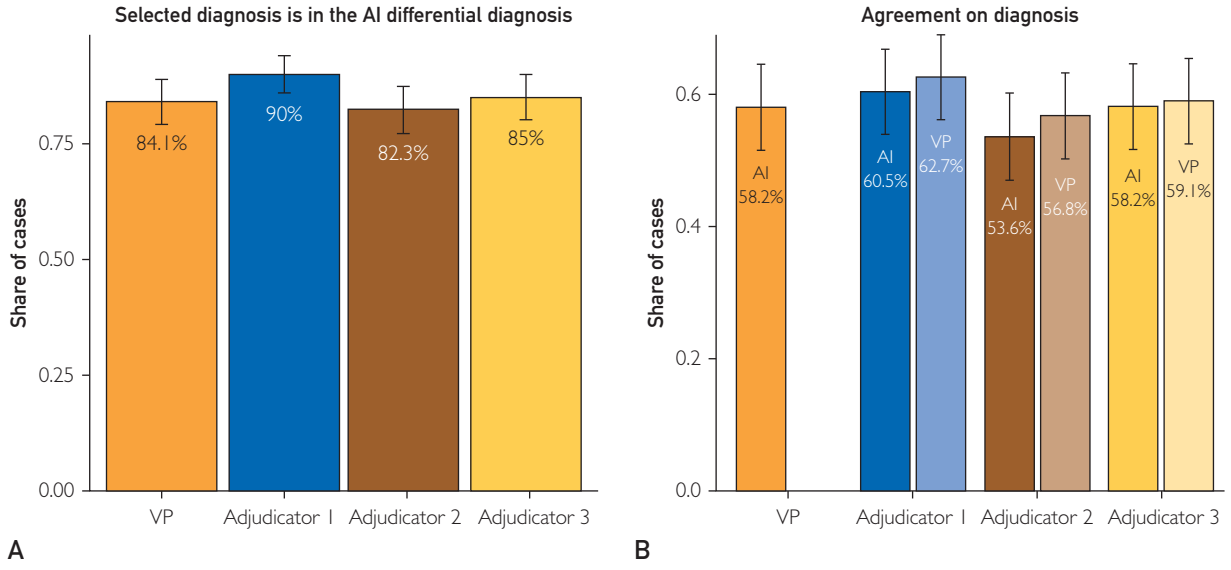


FIGURE 2. Accuracy distribution: AI and provider diagnostic agreement for 102,059 cases. (A) panel shows the share of cases in which the virtual care provider and each of the adjudicators selected a diagnosis that is within the AI list of differential diagnoses. (B) panel shows the share of cases in which the virtual care provider and adjudicators selected the same diagnosis as the virtual care provider or the AI top-ranked diagnosis. Error bars show 95% CI. Y-axis scales differ between panels. The sample is described in detail in Supplementary Table 3 (available online at https://www.mcpdigitalhealth.org/). AI, artificial intelligence; VP, virtual care provider.

infection, the rate at which providers selected 1 of the AI-recommended diagnoses was ≥95%. For 69% of cases covering 57 diagnoses, the rates of agreement were ≥90%. Accuracy varied across presenting symptoms, with providers selecting an AI recommendation in 95.2% and 81.3% of cases at the 75th and 25th percentiles, respectively (Figure 1B). Detailed results by diagnosis and presenting symptoms are summarized in Supplementary Tables 1 and 2. Diagnostic agreement did not vary significantly by provider experience on the platform (Supplementary Table 4, available online at https://www.mcpdigitalhealth.org/).

Diagnostic Agreement Across Demographic Variables

Without adjusting for differences between groups in demographic characteristics and case mix, provider agreement with AI is 9.2% ($P<.001$) higher in cases involving women compared with men and 3.2% and 3% points lower, respectively, in cases involving Black or African American patients and Hispanic or Latino patients compared with Caucasian ($P<.001$ in both cases) (Table 3). Differences were smaller across age groups. Including fixed effects for provider-selected diagnosis that controlled for differences in other demographic characteristics, the sex difference in performance was no longer significant, and differences across racial or ethnic groups decreased in size and significance.

Case Adjudication

Figure 2 summarizes the agreement rates between the virtual care provider, AI, and each of the 3 adjudicators for a subset of 220 cases. All adjudicators selected the same diagnosis as the virtual care provider and the AI at a similar rate (85.7% average adjudicator-AI agreement). Rates of agreement on the top diagnosis between the AI or virtual care provider and adjudicators (58.2% average adjudicator-AI agreement) were similar to the top diagnosis agreement rate between the virtual care provider and the AI. Similar rates of agreement with AI among adjudicators and the original providers mitigate concerns that virtual care providers were systematically biased when selecting case diagnoses.

Adjudicators achieved consensus in 128 cases (58.2% of all adjudicated cases) (Supplementary Table 5, available online at https://www.mcpdigitalhealth.org/). The consensus diagnosis was included in the AI differential in all these 128 cases (100%), top-ranked by the AI in 97 (75.8%) of these cases, and matched the provider diagnosis in 95 (74.2%) of these cases. Adjudicators exhibited a 2-to-1 majority agreement in 73 cases (33.2% of all adjudicated cases). The majority diagnosis was included in the AI differential diagnosis list in all these cases (n = 73, 100%), was top-ranked by the AI in 32 (43.8%) of these cases, and matched the provider diagnosis in 42 (57.5%) of these cases. Overall, the majority (consensus or 2-to-1) adjudicated diagnosis was always (100%, n = 201) in the AI differential diagnosis, top-ranked by the AI in 129 (64.2%) of these cases, and selected by the provider in 137 (68.2%) of these cases.

Model Retraining

For all but 1 condition, model retraining was associated with increased provider agreement. On average, the share of cases among female patients in which the providers selected an AI-recommended diagnosis in the set of retrained conditions increased from 96.6% to 98.0% (Table 2). The drop in agreement for vulvovaginitis is possibly because of its redefinition from a broad category covering all vaginal conditions to a specific identification, excluding vaginal yeast infections and bacterial vaginosis.

DISCUSSION

In this investigation, we observed a high agreement rate between AI and virtual care providers for most clinical cases presenting to virtual primary care with short-term symptoms. In all cases with a consensus among independent blinded adjudicators, the AI differential diagnoses successfully included the consensus diagnosis. Adjusting for case mix differences, agreement rates did not differ substantially across age, sex, and racial or ethnic groups. We also observed that model retraining can improve AI performance.

The use of AI in clinical practice has predominantly focused on disease detection.[6] We evaluated the use of AI for medical interviews and assistance in differential diagnoses for patients presenting to a virtual care provider with undifferentiated short-term complaints. Primary care is challenged by significant time demands to deliver preventive, long-term disease, and short-term care.[13] Innovative solutions to increase clinical efficiencies in short-term care may alleviate the burden on primary care practices. Our investigation envisions possible futures of how AI could work with clinicians to optimize primary care clinical workflows through automated patient interviews and triage, with AI diagnosing and presenting management recommendations to providers for presenting symptoms for which it has proven and reliable accuracy. For clinical conditions with lower reported accuracy, AI retraining can improve clinical diagnostic performance with more clinical case experience.

Our study's strengths include the analysis of a large, diverse sample of clinical cases and the use of blinded adjudication. These ensured a comprehensive evaluation of AI model diagnostic accuracy across different real-world clinical scenarios representative of the national population engaging with digital clinical care platforms.[14]

Our study also has several limitations. First, the lack of a non-AI-exposed comparison group, possibly leading to bias if virtual care providers were unduly influenced by AI suggestions. However, adjudication results and consistent agreement across varying provider experiences suggest providers did not significantly over rely on the AI model. Second, patient self-selection to progress from the AI medical interview to a virtual encounter could have influenced the type of diagnoses and diagnostic accuracy tested. Third, the small size of the randomly selected adjudicator sample may restrict our findings' generalizability. Larger future samples could further substantiate our results.

CONCLUSION

Our research highlights AI's potential to enhance clinical care efficiency through automated medical interviews and differential diagnoses. Further studies should continue to improve AI diagnostic accuracy for common clinical conditions and expand its ability to diagnose complex diseases and multiple comorbidities. Further research is

needed to advance our understanding of the effect of AI on physician decisions and patient outcomes and its broader implications for health care.

POTENTIAL COMPETING INTERESTS

During the development and conduct of this study, Zeltzer, Herzog, Pickman, Steuerman, Ilan Ber, Kugler, and Shaul received consulting fees or were paid employees and held stocks or stock options from K Health Inc. Dr Ebbert's institution received consulting fees from K Health Inc and EXACT Sciences. Dr Ebbert also received payments, royalties, and travel support from Applied Aerosol Technologies, MedInCell, and EXACT Sciences.

ACKNOWLEDGMENTS

For in-depth perspective about the AI models, we acknowledge Yaara Arkin, Tom Beer, Tamar Brufman, Ilan Frank, and Itay Manes, who took part in the development and training of the clinical AI models. We acknowledge funding from K Health Inc.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **AI**, artificial intelligence; **EMR**, electronic medical record; **ICD-10**, International Classification of Diseases, 10th Revision

Correspondence: Address to Dan Zeltzer, PhD, 214 Berglas Hall, School of Economics, Tel Aviv University, Tel Aviv 6997801, Israel (dzeltzer@tauex.tau.ac.il).

ORCID

Dan Zeltzer: https://orcid.org/0000-0002-4140-7531

REFERENCES

1. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. *J Med Internet Res*. 2021;23(9):e29839.
2. Huang S, Ribers MA, Ullrich H. Assessing the value of data for prediction policies: the case of antibiotic prescribing. *Econ Lett*. 2022;213:110360.
3. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. *Ann Fam Med*. 2020;18(3):250-258.
4. Liaw WR, Westfall JM, Williamson TS, Jabbarpour Y, Bazemore A. Primary care: the actual intelligence required for artificial intelligence to advance health care and improve health. *JMIR Med Inform*. 2022;10(3):e27691.
5. Wallace W, Chan C, Chidambaram S, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *NPJ Digit Med*. 2022;5(1):118.
6. Mirbabaie M, Stieglitz S, Frick NRJ. Artificial intelligence in disease diagnostics: a critical review and classification on the current state of research guiding future direction. *Health Technol*. 2021;11(4):693-731.
7. Koren G, Souroujon D, Shaul R, et al. A patient like me—an algorithm-based program to inform patients on the likely conditions people with symptoms like theirs have. *Med (Baltim)*. 2019;98(42):e17596.
8. 24/7 Access to high-quality medicine. KHealth. Accessed July 22, 2023. https://www.khealth.com.
9. Finley CR, Chan DS, Garrison S, et al. What are the most common conditions in primary care? systematic review. *Can Fam Physician*. 2018;64(11):832-840.
10. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. Preprint. Posted online August 13, 2020. arXiv 2008.05756. https://doi.org/10.48550/arXiv.2008.05756.
11. Kahneman D, Sibony O, Sunstein CR. *Noise: a flaw in human judgment*. 1st ed. Hachette Book Group; 2021.
12. Raghu M, Blumer K, Corrado G, et al. The algorithmic automation problem: prediction, triage, and human effort. Preprint. Posted online March 28, 2019. arXiv 1903.12220. https://doi.org/10.48550/arXiv.1903.12220.
13. Porter J, Boyd C, Skandari MR, Laiteerapong N. Revisiting the time needed to provide adult primary care. *J Gen Intern Med*. 2023;38(1):147-155.
14. Reed ME, Huang J, Graetz I, et al. Patient characteristics associated with choosing a telemedicine visit vs office visit with the same primary care clinicians. *JAMA Net Open*. 2020;3(6):e205873.