Longitudinal changes in diagnostic accuracy of a differential diagnosis list developed by an artificial intelligence-based symptom checker: a retrospective observational study

Yukinori Harada, Tetsu Sakamoto, Shu Sugimoto, Taro Shimizu

Submitted to: JMIR Formative Research
on: October 26, 2023

Table of Contents

Longitudinal changes in diagnostic accuracy of a differential diagnosis list developed by an artificial intelligence-based symptom checker: a retrospective observational study

Yukinori Harada[1, 2] MD, PhD; Tetsu Sakamoto[1] MD; Shu Sugimoto[3] MD; Taro Shimizu[1] MBA, MD, MPH, PhD

[1]Department of Diagnostic and Generalist Medicine Dokkyo Medical University Mibu, Simotsuga JP
[2]Department of General Medicine Nagano Chuo Hospital Nagano JP
[3]Department of Internal Medicine Nagano Chuo Hospital Nagano JP

Corresponding Author:
Yukinori Harada MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
880 Kitakobayashi
Mibu, Simotsuga
JP

Abstract

Background: Artificial intelligence (AI) symptom checker models should be trained using real-world patient data to improve their diagnostic accuracy. Given that AI-based symptom checkers are currently employed in clinical practice, their performance should improve over time. However, longitudinal evaluations of the diagnostic accuracy of these symptom checkers are limited.

Objective: This study aimed to assess the longitudinal changes in the accuracy of differential diagnosis lists created by an AI-based symptom checker employed in the real world.

Methods: This was a single-center, retrospective, observational study. Patients who visited an outpatient clinic without an appointment between May 1, 2019 and April 30, 2022 and who were admitted to a community hospital in Japan within 30 days of their index visit were considered eligible. We only included patients who underwent an AI-based symptom checkup at the index visit, and the diagnosis was finally confirmed during follow-up. Final diagnoses were categorized as common or uncommon, and all cases were categorized as typical or atypical. The primary outcome measure was the accuracy of the differential diagnosis list created by the AI-based symptom checker, defined as the final diagnosis in a list of 10 differential diagnoses created by the symptom checker. To assess the change in the symptom checker's diagnostic accuracy over 3 years, we used a chi-squared test to compare the primary outcome over three periods: from May 1, 2019 to April 30, 2020 (first year); from May 1, 2020 to April 30, 2021 (second year); and from May 1, 2021 to April 30, 2022 (third year).

Results: A total of 381 patients were included. Common diseases comprised 257 cases (68%), and typical presentations were observed in 298 cases (78 %). Overall, the accuracy of the differential diagnosis list created by the AI-based symptom checker was 172/381 (45%), which did not differ across the 3 years (first year, 44%; second year, 44%; and third year, 48%; P=.85). The accuracy of the differential diagnosis list created by the symptom checker was low in those with uncommon diseases (24%) and atypical presentations (15%). In the multivariate logistic regression model, disease commonality and presentation typicality were significantly associated with the accuracy of the differential diagnosis list created by the symptom checker.

Conclusions: A 3-year longitudinal survey of the diagnostic accuracy of differential diagnosis lists developed by an AI-based symptom checker, which has been implemented in real-world clinical practice settings, showed no improvement over time. Uncommon diseases and atypical presentations were independently associated with a lower diagnostic accuracy. In the future, symptom checkers should be trained to recognize uncommon conditions. Clinical Trial: Not applicable

(JMIR Preprints 26/10/2023:53985)
DOI: https://doi.org/10.2196/preprints.53985

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Longitudinal changes in diagnostic accuracy of a differential diagnosis list developed by an artificial intelligence-based symptom checker: a retrospective observational study

Abstract

Background: Artificial intelligence (AI) symptom checker models should be trained using real-world patient data to improve their diagnostic accuracy. Given that AI-based symptom checkers are currently employed in clinical practice, their performance should improve over time. However, longitudinal evaluations of the diagnostic accuracy of these symptom checkers are limited.
Objective: This study aimed to assess the longitudinal changes in the accuracy of differential diagnosis lists created by an AI-based symptom checker employed in the real world.
Methods: This was a single-center, retrospective, observational study. Patients who visited an outpatient clinic without an appointment between May 1, 2019 and April 30, 2022 and who were admitted to a community hospital in Japan within 30 days of their index visit were considered eligible. We only included patients who underwent an AI-based symptom checkup at the index visit, and the diagnosis was finally confirmed during follow-up. Final diagnoses were categorized as common or uncommon, and all cases were categorized as typical or atypical. The primary outcome measure was the accuracy of the differential diagnosis list created by the AI-based symptom checker, defined as the final diagnosis in a list of 10 differential diagnoses created by the symptom checker. To assess the change in the symptom checker's diagnostic accuracy over 3 years, we used a chi-squared test to compare the primary outcome over three periods: from May 1, 2019 to April 30, 2020 (first year); from May 1, 2020 to April 30, 2021 (second year); and from May 1, 2021 to April 30, 2022 (third year).
Results: A total of 381 patients were included. Common diseases comprised 257 cases (68%), and typical presentations were observed in 298 cases (78 %). Overall, the accuracy of the differential diagnosis list created by the AI-based symptom checker was 172/381 (45%), which did not differ across the 3 years (first year, 44%; second year, 44%; and third year, 48%; *P*=.85). The accuracy of the differential diagnosis list created by the symptom checker was low in those with uncommon diseases (24%) and atypical presentations (15%). In the multivariate logistic regression model, common disease (odds ratio [OR] 4.13, 95% confidence interval [CI] 2.50-6.98; *P*<.001) and typical presentation (OR 6.92, 95%CI 3.62-14.2; *P*<.001) were significantly associated with the accuracy of the differential diagnosis list created by the symptom checker.
Conclusions: A 3-year longitudinal survey of the diagnostic accuracy of differential diagnosis lists developed by an AI-based symptom checker, which has been implemented in real-world clinical practice settings, showed no improvement over time. Uncommon diseases and atypical presentations were independently associated with a lower diagnostic accuracy. In the future, symptom checkers should be trained to recognize uncommon conditions.
Trial Registration: Not applicable.

Keywords: atypical presentations; diagnostic accuracy; symptom checker; uncommon diseases.

Introduction

Diagnostic errors are a significant global patient safety issue [1]. In outpatient settings, diagnostic errors are evident in 1–5% of cases [2–5]. Notably, the risk of such errors increases

for outpatients unexpectedly admitted shortly after their initial visit [6,7]. The most common factors contributing to diagnostic errors in outpatient settings include problems with data integration, interpretation, and differential diagnosis [3,5,8–10]. To address this, the integration of diagnostic decision support systems, such as differential diagnosis generators, into clinical practice is recommended [11].

Differential diagnosis generators produce possible differential diagnoses by processing clinical information through algorithms, thereby supporting clinicians by reducing the likelihood of overlooking possible diagnoses and countering the cognitive biases inherent to the diagnostic process [12]. Early deployment of differential diagnosis generators can augment an existing list of differential diagnoses, increasing the odds of including the correct diagnosis [13], and can prompt more thorough history-taking [14]. Therefore, current symptom checkers—generators that produce differential diagnoses based on the inputs from patients themselves before they encounter a clinician—are potentially promising tools to reduce diagnostic errors. Indeed, some symptom checkers have already been employed in clinical practice [15–17] and even in national health services, such as the NHS 111 system in the United Kingdom [18].

Given that modern artificial intelligence (AI) is designed to be dynamic and to evolve according to real-world data [19], one might expect the performance of AI-based symptom checkers to improve over time. Importantly, at the same time, a decline in the performance of AI by feedback with data from different populations and settings is also possible. Monitoring such a shift and drift of AI performance is required to use AI-based symptom checkers effectively and safely [19,20]. However, their developers do not usually disclose the data, such as how AI algorithms changed and what types of clinical indicators improved. One set of studies using the same sets of clinical vignettes found that the diagnostic accuracy of symptom checkers improved from 2015 to 2020 [21]. However, because these case vignettes are publicly available, the developers may have trained symptom checker algorithms using these cases. Therefore, it remains unknown whether symptom checkers improve their diagnostic performance over time [12]. Moreover, because clinical vignettes have been found to have considerable inherent limitations when used to assess diagnostic accuracy in comparison with real-world data [22], longitudinal evaluations of the performance of symptom checkers in the real world are needed.

Concerns have arisen regarding the low diagnostic accuracy of current symptom checker output, which often lags behind that of physicians [12,17,23]. Inaccurate initial diagnoses can be detrimental, steering clinicians towards errors [24]. One major hurdle in the accuracy of symptom checker outputs is patient input variability. Differences in symptom interpretation, clinical literacy, input sequencing, and symptom listings can profoundly influence the quality of a symptom checker's output [12,25]. Another challenge is the disparity between simulated and real-world data. Previous research has indicated a diminished accuracy of symptom checker output when applied to real cases instead of fictional vignettes [23]. This could be attributed to the fact that crafted vignettes often provide typical presentations [25], whereas real cases include more atypical presentations and so may contribute to diagnostic errors [26]. Therefore, symptom checkers should be trained using real-world patient data, covering a diverse range of cases and including atypical presentations, to improve their accuracy [12,25]. The real-world application and refinement of these tools post-development are crucial.

Therefore, we conducted this study to assess the changes in the accuracy of differential diagnosis lists created by AI-based symptom checkers in the real world. This article defined AI-based symptom checkers as those using contemporary machine-learning models.

The contributions of our proposed work are summarized as follows:

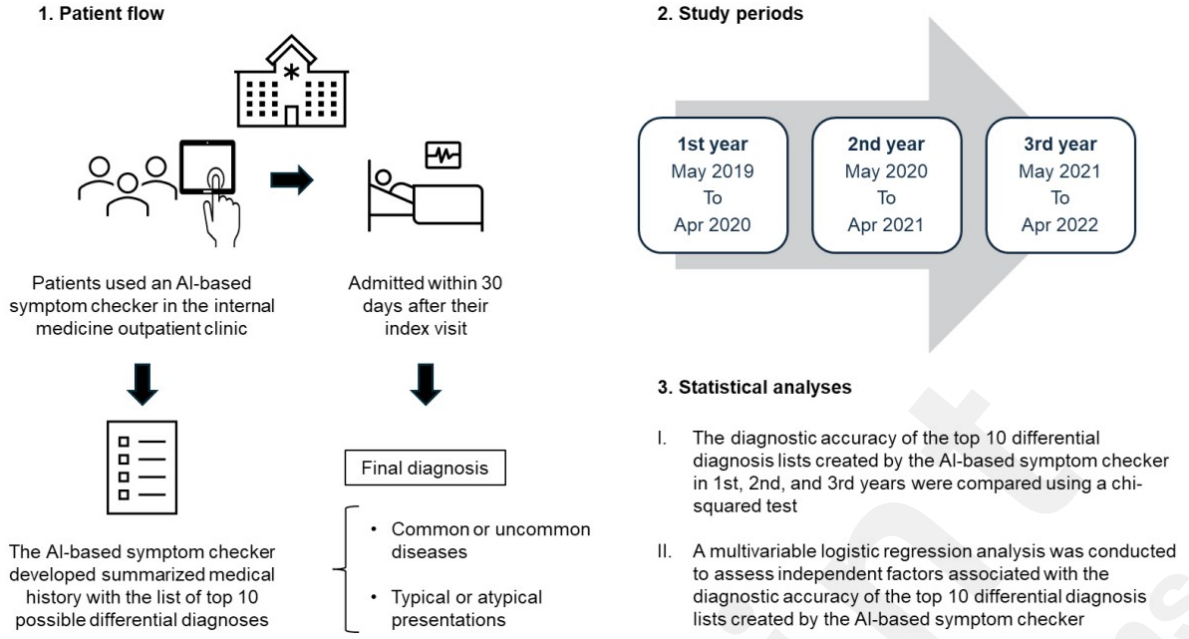
- We provide the data of three-year longitudinal changes in the diagnostic performance of contemporary machine-learning-based symptom checkers.
- We also provide factors related to the diagnostic performance of AI-based symptom checkers.

Methods

Study design and participants

This was a single-center, retrospective, observational study. Patients who visited the internal medicine outpatient clinic at Nagano Chuo Hospital without an appointment between May 1, 2019 and April 30, 2022 and who were then admitted within 30 days after their index visit were considered eligible. We set the inclusion criteria because admission within 30 days after the index visit was considered a useful option to capture the patients with a high risk of diagnostic errors [27–31], in which population diagnostic decision support systems are particularly needed. We included only patients who used an AI-based symptom checker that identified 10 possible differential diagnoses (Ubie, Inc., Japan) at the index visit and excluded patients for whom the AI-based symptom checker produced less than 10 differential diagnoses, whose diagnosis was not confirmed, and those who were admitted for a reason unassociated with their index visit complaint. For patients who used the AI-based symptom checker multiple times at different outpatient visits or who were admitted twice or more, we included only data from the first outpatient visit and admission (others were excluded as duplicates). Overview of this study is shown in Figure 1.

Figure 1. Overview of the study. This study included patients who visited the internal medicine outpatient clinic at a community hospital without an appointment between May 2019 and April 2022 and were admitted within 30 days after their index visit. This study included only patients who used an artificial intelligence (AI)-based symptom checker that identified 10 possible differential diagnoses at the index visit. The final diagnoses were categorized into common or uncommon diseases, and clinical presentations were categorized into typical or atypical. The change in the diagnostic accuracy of the AI-based symptom checker over 3 years was assessed by using a chi-squared test by dividing the study duration into three periods: from May 2019 to April 2020 (first year); from May 2020 to April 2021 (second year); and from May 2021 to April 2022 (third year). A multivariable logistic regression analysis was conducted to assess independent factors with diagnostic accuracy of the top 10 differential diagnosis lists created by the AI-based symptom checker.



Ethical considerations

The study complied with the principles of the Declaration of Helsinki. The Research Ethics Committee of Nagano Chuo Hospital approved this study (NCR202208) and waived the requirement for written informed consent from the participants because of the opt-out method used in this study. We informed the participants by providing detailed information about the study in the outpatient waiting area at Nagano Chuo Hospital and on the hospital's website. The study data are de-identified. There was no compensation for the participants.

AI-based symptom checker

Details of the AI-based symptom checkers assessed in this study have been described previously [7,32]. In brief, the AI-based symptom checker converted the data entered by patients on tablet terminals into medical terms. Patients entered their background information, such as age, sex, and chief complaint, as a free text on a tablet in the waiting room. This AI-based symptom checker asked approximately 20 questions, one by one, tailored to the patient. Based on the previous answers of the same patient, the questions were optimized to generate the most relevant list of potential differential diagnoses. The hospital staff at Nagano Chuo Hospital provided support to the patients when they found it difficult to input information independently. Physicians could view the entered data as a summarized medical history with the top 10 possible differential diagnoses along with their ranks. According to the developer's website, this AI-based symptom checker improved quality through feedback from more than 1,500 medical institutions. However, we could not show the mathematical expression and algorithm of the machine-learning model because the developer did not disclose a detailed machine-learning methodology.

Data collection

We retrospectively collected data from the patients' electronic health records. The following data were collected: date of the index visit, age, sex, medical history recorded by the AI-based symptom checker (including chief complaints, history of present illness, past medical history, family history, and social history), 10 differential diagnoses developed by the AI-based symptom checker, and the final diagnosis. The final diagnosis was judged independently by two researchers (YH and SS) based on the descriptions in the medical records, and disagreements

were resolved through discussion. Final diagnoses were coded by the first author (YH) using the International Classification of Diseases, 11th Revision codes. Final diagnoses were further categorized into common or uncommon diagnoses based on whether the incidence was more than 1 in 2000 (common disease) or not (uncommon disease)[33]; unclear cases were judged by two researchers through discussion (YH, T Sakamoto). According to the final diagnosis and medical history created by the AI-based symptom checker, two researchers (YH, T Sakamoto) independently judged all cases as typical or atypical, and conflicts were resolved by discussion.

Primary outcome

The primary outcome measure was the accuracy of the differential diagnosis list created using the AI-based symptom checker. The accuracy of the differential diagnosis list created by the AI-based symptom checker was defined as the presence of the final diagnosis in the list of 10 differential diagnoses created by the AI-based symptom checker. Two researchers (YH, T Sakamoto) independently judged the accuracy of the differential diagnosis list created by the AI-based symptom checker, and conflicts were resolved through discussion. The accuracy of the AI over 3 years was also assessed in the following the subgroups: age ≥ 65 years and < 65 years, men and women, single and multiple chief complaints, common and uncommon disease, and typical and atypical presentation.

Statistical analysis

Continuous or ordinal data are presented as mean and standard deviation or median and quantiles and compared using a t-test, U-test, or analysis of variance. Categorical or binary data are presented as numbers and percentages and compared using the chi-squared or Fisher's exact test. To assess the change in the diagnostic accuracy of the AI-based symptom checker over 3 years, we compared the accuracy of the differential diagnosis lists created by the AI-based symptom checker using a chi-squared test by dividing the study duration into three periods: from May 1, 2019 to April 30, 2020 (first year); from May 1, 2020 to April 30, 2021 (second year); and from May 1, 2021 to April 30, 2022 (third year). We calculated 108 patients as the minimum required sample size based on an alpha error of .05, power of 0.80, effect size of 0.30 (medium), and degrees of freedom of 2. We also created a multivariable logistic regression model that included the correctness of the differential diagnosis list created by the AI-based symptom checker as an independent variable and the visit year (first, second, and third year), age (as a continuous variable), sex (male or female), typicality of presentation (typical or atypical), and commonality of final diagnosis (common or uncommon) as dependent variables; these variables were selected as confounders because they were considered to be associated with the accuracy of the differential diagnosis list created by the AI-based symptom checker. *P* values below .05 were considered significant. All statistical analyses were performed using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Baseline characteristics

Of the 484 eligible cases, 103 were excluded (duplication, 20; admission unrelated to the index visit, 9; no final diagnosis, 18; AI produced less than 10 differential diagnoses, 56). Therefore, 381 cases were finally included in the analysis. The mean age was 68±18 years, and 205 (54%) were men. One hundred seventy-four patients inputted more than one complaint (46%). Diseases of the digestive system were the most common final diagnosis category (n = 128, 34%), followed by diseases of the circulatory system (n = 55, 14%), respiratory system (n = 44, 12%), neoplasms (n = 42, 11%), and infectious or parasitic diseases (n = 26, 7%). Regarding

commonality and typicality, 257 (68%) were common diseases and 298 (78%) were typical presentations. Typical presentation of common disease was the most common group (205, 54%), followed by typical presentation of uncommon disease (93, 24%), atypical presentation of common disease (52, 14%), and atypical presentation of uncommon disease (31, 9%). The number of patients was higher in the first year than in the second and third years (Table 1) owing to the coronavirus disease (COVID-19) pandemic. Although there was a significant difference in age, no significant differences were observed in other baseline characteristics among the three groups.

Table 1. Baseline characteristics of patients who visited the internal medicine outpatient clinic at Nagano Chuo Hospital without an appointment and then admitted within 30 days for three years from May 2019 to April 2022

	First year N=219	Second year N=72	Third year N=90	*P* value
Age (years±SD)	70±18	63±15	64±17	.002
Men	114 (52%)	40 (56%)	51 (57%)	.72
Multiple chief complaints	104 (48%)	32 (44%)	38 (42%)	.68
Common disease	146 (67%)	45 (63%)	66 (73%)	.32
Typical presentation	164 (75%)	60 (83%)	74 (82%)	.18

[a]SD, standard deviation; first year: May 1, 2019 to April 30, 2020; second year: May 1, 2020 to April 30, 2021; third year: May 1, 2021 to April 30, 2022.

Primary outcome

Overall, the final diagnosis was observed in the top 10 differential diagnosis lists created by the AI-based symptom checker in 172 patients (45%). The accuracy of the differential diagnosis list created by the AI-based symptom checker did not significantly differ among the 3 years (first year, 44%; second year, 44%; and third year, 48%; *P*=.85). There was also no significant difference in the accuracy of AI differential diagnosis among the 3 years in the subgroups (Table 2). In the subgroups with uncommon diseases and atypical presentations, the correct rate of the AI differential diagnosis list was < 30%. Some examples of cases with uncommon diseases and atypical presentations are shown in Table S1 in Multimedia Appendix 1.

Table 2. Proportion of patients with a correct diagnosis included in the top 10 differential diagnosis list generated by artificial intelligence in patients who visited the internal medicine outpatient clinic at Nagano Chuo Hospital without an appointment and then admitted within 30 days for three years from May 2019 to April 2022

	Total N=381	First year N=219	Second year N=72	Third year N=90	*P* value
Overall accuracy	172/381 (45%)	97/219 (44%)	32/72 (44%)	43/90 (48%)	.85
Age ≥ 65 years	110/243 (45%)	69/159 (43%)	15/35 (43%)	26/49 (53%)	.47
Age < 65 years	62/138 (45%)	28/60 (47%)	17/37 (46%)	17/41 (42%)	.87
Men	102/205	53/114	20/40	29/51	.47

	(50%)	(47%)	(50%)	(57%)	
Women	70/176 (40%)	44/105 (42%)	12/32 (38%)	14/39 (36%)	.77
Single chief complaint	103/207 (50%)	53/115 (46%)	23/40 (58%)	27/52 (52%)	.43
Multiple chief complaints	69/174 (40%)	44/104 (42%)	9/32 (28%)	16/38 (42%)	.34
Common disease	142/257 (55%)	79/146 (54%)	27/45 (60%)	36/66 (55%)	.78
Uncommon disease	30/124 (24%)	18/73 (25%)	5/27 (19%)	7/24 (29%)	.67
Typical presentation	160/298 (54%)	88/164 (54%)	29/60 (48%)	43/74 (58%)	.53
Atypical presentation	12/83 (15%)	9/55 (16%)	3/12 (25%)	0/16 (0%)	.10

[a]The first year was from May 1, 2019 to April 30, 2020; the second year was from May 1, 2020 to April 30, 2021; and the third year was from May 1, 2021 to April 30, 2022.

Logistic regression model

In the multivariate logistic regression model, the year of the index visit was not significantly associated with whether the final diagnosis was included in the top 10 differential diagnosis lists created by the AI-based symptom checker (Table 3). By contrast, in the multivariate logistic regression model, the commonality of disease and typicality of presentation were significantly associated with the accuracy of the differential diagnosis list created by the AI-based symptom checker.

Table 3. A logistic regression model for whether the correct diagnosis was included in the differential diagnosis list generated by artificial intelligence in patients who visited the internal medicine outpatient clinic at Nagano Chuo Hospital without an appointment and then admitted within 30 days

Variables	Odds ratio (95% confidence interval)	*P* value
Year of visit		
Second year (reference: first year)	0.84 (0.45-1.54)	.57
Third year (reference: first year)	0.88 (0.51-1.54)	.67
Age (for 1-year increase)	0.99 (0.98-1.00)	.16
Men (reference: women)	1.42 (0.92-2.30)	.11
Multiple complaints (reference: single complaint)	0.70 (0.44-1.11)	.13
Common disease (reference: uncommon disease)	4.13 (2.50-6.98)	<.001
Typical presentation (reference: atypical presentation)	6.92 (3.62-14.2)	<.001

[a]The first year was from May 1, 2019 to April 30, 2020; the second year was from May 1, 2020

to April 30, 2021; and the third year was from May 1, 2021 to April 30, 2022.

Discussion

Principal results

In this study, at a community hospital in Japan, a 3-year longitudinal assessment of the performance of an AI-based symptom checker showed no change in the diagnostic accuracy of its differential diagnosis lists in outpatients admitted within 30 days of their index visit. In the exploratory subgroup and multivariate logistic regression analyses, the commonality of disease and typicality of presentation were significantly associated with the accuracy of the differential diagnosis list created by the AI-based symptom checker.

Implications of the study

This study suggests that current AI-based symptom checkers employed in the real world may not improve their diagnostic performance over time. In this study, no improvement of the diagnostic accuracy of AI was observed, even in the common disease and typical presentation subgroups. Machine learning, using data with reliable teaching labels, is required to improve the accuracy of AI-based symptom checkers. However, patients may not always be able to accurately provide their final diagnosis, which may prevent effective machine learning. In addition, even if symptom checkers are employed in healthcare facilities, reliable feedback may not be guaranteed because of diagnostic uncertainty, low diagnostic quality, and care fragmentation. The results of this study indicate that the developers and users of AI-based symptom checkers should be more responsible for improving the diagnostic quality of AI-based symptom checkers by providing reliable feedback on diagnostic labels.

There can be another perspective for this study’s results. We assumed that the performance of the AI-based symptom checker did not improve over time based on the result that the diagnostic accuracy did not change. However, it is possible that the developer also set indicators other than diagnostic accuracy, such as the impact of service use, clinical and cost-effectiveness, and patient satisfaction, to improve the algorithm of the AI-based symptom checker [17]. Balancing the different outcomes may limit the increase in diagnostic accuracy. Second, since we do not know the ideal and theoretical upper limit of diagnostic accuracy in specific clinical contexts with some restrictions, it is also possible that some AI-based symptom checkers' diagnostic accuracy has already reached the theoretical upper limit of their performance. For example, minimizing questions to save time may reduce the diagnostic performance. Indeed, our previous study showed that physicians' diagnostic accuracy was only 56% when reading the information taken by the same AI-based symptom checker used in this study [34]. Therefore, the judgment that no improvement in diagnostic accuracy was observed in this study may be unfair. We need a standard method with clear indicators for an unbiased and fair evaluation of the improvement of the performance of AI-based symptom checkers.

Comparison with prior work

Longitudinal comparisons of the diagnostic performance of symptom checkers in the real world are scarce; however, several studies have assessed changes in the diagnostic accuracy of symptom checkers using clinical vignettes. According to Schmieding et al., the rate of correct diagnoses listed among the top 10 differential diagnoses of symptom checkers was at least 15% higher in 2020 than in 2015 using the same clinical vignettes [21]. In contrast, other studies suggested that the diagnostic accuracy of symptom checkers did not change from 2015 to 2020 when using some of the new vignettes [21,35]. Considering these and our study results, the

diagnostic accuracy of symptom checkers may be improved for prototypical or standardized patients; however, because there are many variants of demographic patterns and clinical presentations in the real world, slight improvements may not result in the overall improvement of diagnostic accuracy.

In this study, the diagnostic accuracy of the AI-based symptom checker for uncommon diseases was approximately 30% lower than that for common diseases; similarly, approximately 40% lower diagnostic accuracy was observed for atypical presentations than for typical presentations. The diagnostic accuracy of symptom checkers may depend on the urgency of the clinical condition, as well as common and uncommon conditions [17]. Indeed, a previous study also showed that the correct diagnosis was less frequently listed in the top 10 differential diagnoses of symptom checkers for uncommon diseases than for common diseases, with a 60% difference (8% vs. 68%) [35]. Our study provides evidence that atypical presentations, another aspect of uncommon conditions, may also negatively affect the diagnostic accuracy of symptom checkers. Uncommon diseases and atypical presentations are associated with a high risk of diagnostic error [26,36]. Through this perspective, our data indicate that current and future symptom checkers should be further trained with data on uncommon conditions, such as uncommon diseases and atypical presentations, to improve diagnostic quality in clinical practice. According to a previous study, symptom checkers can collect only 30% of all pertinent findings, and are not good at collecting pertinent negative findings [37]. Considering that collecting pertinent findings is vital for diagnosing uncommon conditions, training with data on uncommon conditions and a system of high-quality feedback and reinforcement by expert diagnosticians are warranted for future symptom checkers.

Recent emerging generative AI-related tools such as ChatGPT (OpenAI Inc., San Francisco, CA), a chatbot that uses a large language model, have been studied for their potential as new differential diagnosis generators. Several studies have demonstrated the high diagnostic accuracy of ChatGPT for simple to complex clinical cases using clinical vignettes and published case reports [38–40]. However, these studies input clinical information, including test results. Regarding symptom checking, while one study showed ChatGPT exhibited high accuracy in symptom checking for a broad range of diseases using the Mayo Clinic Symptom Checker as a benchmark [41], another study showed no difference in diagnostic accuracy between current symptom checkers and ChatGPT for patients with urgent or emergent clinical problems [15]. In addition, regarding ChatGPT, there is a concern that the near-infinite range of possible inputs and outputs prevents standardized regulations [15,42]. Furthermore, generative AI did not seem to overcome the problem of current symptom checkers that worsened diagnostic accuracy in cases of uncommon conditions [43]. Therefore, generative AI-related tools cannot be effective symptom checkers right now. However, compared to current symptom checkers, the diagnostic performance of generative AI-related tools can rapidly improve over time. Indeed, some studies showed that ChatGPT-4 outperformed ChatGPT-3.5 in diagnostic performance [38,43,44]. Therefore, generative AI-related tools may be a choice for diagnosis generators before a patient-clinician encounter in the near future.

Limitations

This study has some limitations. First, the modification details of the symptom checker model used in this study, including the type of machine learning methods employed or manual updates used and the frequency at which the model was modified, remained unclear. Secondly, 3 years may not be appropriate for assessing contemporary machine learning model improvement since there is no standard timeframe to assess the improvement of the machine learning model. However, considering that AI-related tools such as ChatGPT show rapid

performance improvement, 3 years can be considered enough. Third, the COVID-19 pandemic may have affected our results due to low participants in the second and third years. Fourth, because this was a single-center retrospective study and we only included patients admitted within 30 days of the index outpatient visit, the results should be interpreted with caution regarding generalizability. Fifth, because there was no validated tool to assess the typicality of the presentation, which was assessed based on the information produced by the AI, the classification of typicality in this study may have been biased. This was also true for disease commonality, which could change if other criteria for uncommon diseases were applied.

Conclusions

A 3-year single-center retrospective observational study of the diagnostic accuracy of differential diagnosis lists developed by an AI-based symptom checker, currently implemented in real-world clinical practice settings, showed no improvement over time. Uncommon diseases and atypical presentations were independently associated with a lower diagnostic accuracy of the differential diagnosis lists generated by the AI-based symptom checker. In the future, symptom checkers should be trained to recognize uncommon conditions.

Acknowledgement

None

Authors' Contributions

YH conceptualized this study, YH, SS, and T Sakamoto collected data, YH wrote the manuscript, YH conducted all statistical analyses. T Shimizu supervised the manuscript creation and revision. All authors reviewed the final manuscript.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Abbreviations

AI: artificial intelligence

References

1. Singh H, Schiff GD, Graber ML, Onakpoya I, Thompson MJ. The global burden of diagnostic errors in primary care. BMJ Qual Saf 2017 Jun;26(6):484-494. PMID:27530239
2. Avery AJ, Sheehan C, Bell B, et al. Incidence, nature and causes of avoidable significant harm in primary care in England: retrospective case note review. BMJ Qual Saf 2021 Dec;30(12):961-976. PMID:33172907
3. Cheraghi-Sohi S, Holland F, Singh H, et al. Incidence, origins and avoidable harm of missed opportunities in diagnosis: longitudinal patient record review in 21 English general practices. BMJ Qual Saf 2021 Dec;30(12):977-985. PMID:34127547
4. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf 2014 Sep;23(9):727-731. PMID:24742777
5. Harada Y, Otaka Y, Katsukura S, Shimizu T. Effect of contextual factors on the prevalence of

diagnostic errors among patients managed by physicians of the same specialty: a single-centre retrospective observational study. BMJ Qual Saf 2023 Jan 23:bmjqs-2022-015436. PMID:36690471
6. Singh H, Giardina TD, Forjuoh SN, et al. Electronic health record-based surveillance of diagnostic errors in primary care. BMJ Qual Saf 2012 Feb;21(2):93-100. PMID:21997348
7. Kawamura R, Harada Y, Sugimoto S, Nagase Y, Katsukura S, Shimizu T. Incidence of diagnostic errors among unexpectedly hospitalized patients using an automated medical history-taking system with a differential diagnosis generator: retrospective observational study. JMIR Med Inform 2022 Jan 27;10(1):e35225. PMID:35084347
8. Singh H, Giardina TD, Meyer AN, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med 2013 Mar 25;173(6):418-425. PMID:23440149
9. Singh H, Thomas EJ, Khan MM, Petersen LA. Identifying diagnostic errors in primary care using an electronic screening algorithm. Arch Intern Med 2007 Feb 12;167(3):302-308. PMID:17296888
10. Gandhi TK, Kachalia A, Thomas EJ, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. Ann Intern Med. 2006 Oct 3;145(7):488-496. PMID:17015866
11. Balogh EP, Miller BT, Ball JR, editors. Improving diagnosis in health care. Washington (DC): National Academies Press (US); 2015.
12. 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 Aug 17;5(1):118. PMID:35977992
13. Sibbald M, Monteiro S, Sherbino J, LoGiudice A, Friedman C, Norman G. Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel. BMJ Qual Saf 2022 Jun;31(6):426-433. PMID:34611040
14. Kämmer JE, Schauber SK, Hautz SC, Stroben F, Hautz WE. Differential diagnosis checklists reduce diagnostic error differentially: a randomised experiment. Med Educ 2021 Oct;55(10):1172-1182. PMID:34291481
15. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR Mhealth Uhealth 2023 Oct 3;11:e49995. PMID:37788063
16. Morse KE, Ostberg NP, Jones VG, Chan AS. Use characteristics and triage acuity of a digital symptom checker in a large integrated health system: population-based descriptive study. J Med Internet Res 2020 Nov 30;22(11):e20549. PMID:33170799
17. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 1;9(8):e027743. PMID:31375610
18. Turner J, Knowles E, Simpson R, et al. Impact of NHS 111 Online on the NHS 111 telephone service and urgent care system: a mixed-methods study. Southampton (UK): NIHR Journals Library 2021 Nov. PMID:34780129
19. Dorr DA, Adams L, Embí P. Harnessing the promise of artificial intelligence responsibly. JAMA 2023 Apr 25;329(16):1347-1348. PMID:36972068
20. 19.20. Embi PJ. Algorithmovigilance-advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. JAMA Netw Open 2021 Apr 1;4(4):e214622. PMID:33856479
21. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res 2022 May 10;24(5):e31810. PMID:35536633

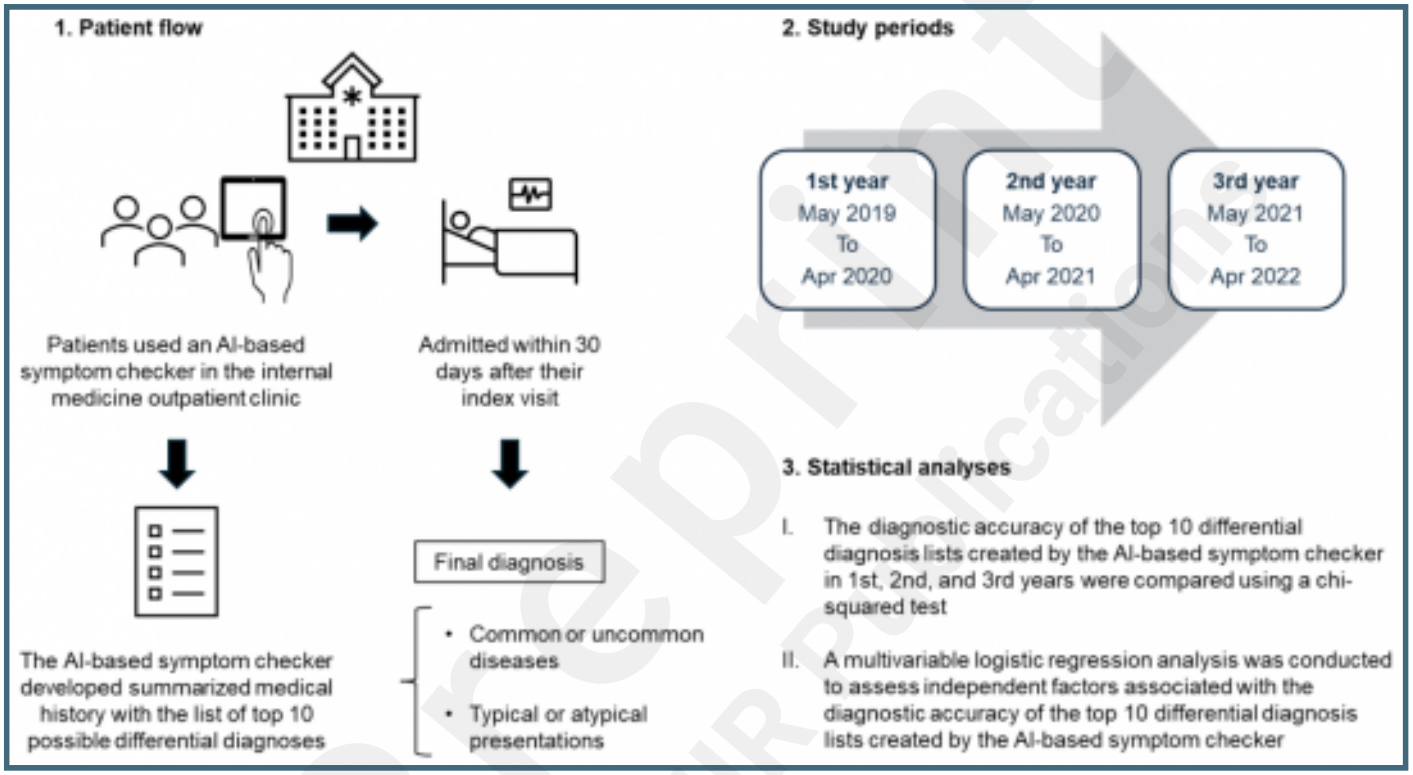
22. El-Osta A, Webber I, Alaa A, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. BMJ Open. 2022 Apr 27;12(4):e053566. PMID:35477872
23. Gilbert S, Mehl A, Baluch A, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open 2020 Dec 16;10(12):e040269. PMID:33328258
24. Meyer FML, Filipovic MG, Balestra GM, Tisljar K, Sellmann T, Marsch S. Diagnostic errors induced by a wrong a priori diagnosis: A prospective randomized simulator-based trial. J Clin Med. 2021 Feb 18;10(4):826. PMID:33670489
25. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. J Med Internet Res 2022 Oct 26;24(10):e37408. PMID:36287594
26. Newman-Toker DE, Peterson SM, Badihian S, et al. Diagnostic errors in the emergency department: A systematic review. Comparative effectiveness review No. 258. (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No. 75Q80120D00003.) AHRQ Publication No. 22(23)-EHC043. Rockville, MD: Agency for Healthcare Research and Quality; December 2022. Errata and Addendum, August 2023. DOI:https://doi.org/10.23970/AHRQEPCCER258
27. Liberman AL, Newman-Toker DE. Symptom-disease pair analysis of diagnostic error (SPADE): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. BMJ Qual Saf 2018 Jul;27(7):557-566. PMID:29358313
28. Sharp AL, Baecker A, Nassery N, et al. Missed acute myocardial infarction in the emergency department-standardizing measurement of misdiagnosis-related harms using the SPADE method. Diagnosis (Berl) 2020 Jul 24;8(2):177-186. PMID:32701479
29. Nassery N, Horberg MA, Rubenstein KB, et al. Antecedent treat-and-release diagnoses prior to sepsis hospitalization among adult emergency department patients: a look-back analysis employing insurance claims data using symptom-disease pair analysis of diagnostic error (SPADE) methodology. Diagnosis (Berl) 2021 Feb 25;8(4):469-478. PMID:33650389
30. Horberg MA, Nassery N, Rubenstein KB, et al. Rate of sepsis hospitalizations after misdiagnosis in adult emergency department patients: a look-forward analysis with administrative claims data using symptom-disease pair analysis of diagnostic error (SPADE) methodology in an integrated health system. Diagnosis (Berl) 2021 Apr 26;8(4):479-488. PMID:33894108
31. Chang TP, Bery AK, Wang Z, et al. Stroke hospitalization after misdiagnosis of "benign dizziness" is lower in specialty care than general practice: a population-based cohort analysis of missed stroke using SPADE methods. Diagnosis (Berl) 2021 Jun 21;9(1):96-106. PMID:34147048
32. Harada Y, Shimizu T. Impact of a commercial artificial intelligence-driven patient self-assessment solution on waiting times at general internal medicine outpatient departments: retrospective study. JMIR Med Inform 2020 Aug 31;8(8):e21056. PMID:32865504
33. Orphanet. Available from: https://www.orpha.net/consor/cgi-bin/index.php
34. Harada Y, Katsukura S, Kawamura R, Shimizu T. Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy of physicians: an open-label randomized controlled study. Int J Environ Res Public Health 2021 Feb 21;18(4):2086. PMID:33669930
35. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Med J Aust 2020 Jun;212(11):514-519. PMID:32391611
36. Blöß S, Klemann C, Rother AK, et al. Diagnostic needs for rare diseases and shared

prediagnostic phenomena: results of a German-wide expert Delphi survey. PLoS One 2017 Feb 24;12(2):e0172532. PMID:28234950
37. Ben-Shabat N, Sharvit G, Meimis B, et al. Assessing data gathering of chatbot based symptom checkers - a clinical vignettes study. Int J Med Inform 2022 Dec;168:104897. PMID:36306653
38. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023 Oct 9;11:e48808. PMID:37812468
39. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by Generative Pretrained Transformer 3 Chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023 Feb 15;20(4):3378. PMID:36834073
40. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 3;330(1):78-80. PMID:37318797
41. Chen A, Chen DO, Tian L. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. J Am Med Inform Assoc 2023 Dec 18:ocad245. PMID:38109889
42. Wachter RM, Brynjolfsson E. Will generative artificial intelligence deliver on its promise in health care? JAMA 2024 Jan 2;331(1):65-69. PMID:38032660
43. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun 2024 Mar 6;15(1):2050. PMID:38448475
44. Luk DWA, Ip WCT, Shea YF. Performance of GPT-4 and GPT-3.5 in generating accurate and comprehensive diagnoses across medical subspecialties. J Chin Med Assoc 2024 Mar 1;87(3):259-260. PMID:38305423

Supplementary Files

Figures

Overview of the study. This study included patients who visited the internal medicine outpatient clinic at a community hospital without an appointment between May 2019 and April 2022 and were admitted within 30 days after their index visit. This study included only patients who used an artificial intelligence (AI)-based symptom checker that identified 10 possible differential diagnoses at the index visit. The final diagnoses were categorized into common or uncommon diseases, and clinical presentations were categorized into typical or atypical. The change in the diagnostic accuracy of the AI-based symptom checker over 3 years was assessed by using a chi-squared test by dividing the study duration into three periods: from May 2019 to April 2020 (first year); from May 2020 to April 2021 (second year); and from May 2021 to April 2022 (third year). A multivariable logistic regression analysis was conducted to assess independent factors with diagnostic accuracy of the top 10 differential diagnosis lists created by the AI-based symptom checker.



Multimedia Appendixes

Table S1. Examples of cases with uncommon diseases or atypical presentations.
URL: http://asset.jmir.pub/assets/52e8d1bddf21698ab5472ea643abe8c8.docx