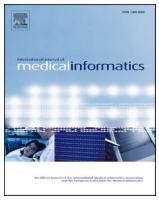

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

International Journal of Medical Informatics 168 (2022) 104897

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



# Assessing data gathering of chatbot based symptom checkers - a clinical vignettes study

Niv Ben-Shabat [a,b,c,*], Gal Sharvit [a], Ben Meimis [d], Daniel Ben Joya [d], Ariel Sloma [a], David Kiderman [e], Aviv Shabat [f], Avishai M Tsur [a,b,c,g], Abdulla Watad [a,b,c,h], Howard Amital [a,b,c]

[a] Sackler Faculty of Medicine, Tel-Aviv University, Israel
[b] Department of Medicine 'B', Sheba Medical Centre, Ramat-Gan, Israel
[c] Zabludowicz Center for Autoimmune Diseases, Sheba Medical Centre, Ramat-Gan, Israel
[d] Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel
[e] Kaplan Medical Center, Rehovot, Israel
[f] Department of Pediatrics A, Edmond and Lily Safra Children's Hospital, Sheba Medical Center, Ramat-Gan, Israel
[g] Israel Defence Forces, Medical Corps, Tel Hashomer, Ramat Gan, Israel
[h] Section of Musculoskeletal Disease, NIHR Leeds Musculoskeletal Biomedical Research Unit, Leeds Institute of Molecular Medicine, University of Leeds, Chapel Allerton Hospital, Leeds, UK

ARTICLE INFO

*Keywords:*
Artificial intelligence
Computer-assisted diagnosis
Diagnosis
Symptom checker
Triage
Chatbots
Medical interview
Data-gathering
Telemedicine

ABSTRACT

*Background:* The burden on healthcare systems is mounting continuously owing to population growth and aging, overuse of medical services, and the recent COVID-19 pandemic. This overload is also causing reduced healthcare quality and outcomes. One solution gaining momentum is the integration of intelligent self-assessment tools, known as symptom-checkers, into healthcare-providers' systems. To the best of our knowledge, no study so far has investigated the data-gathering capabilities of these tools, which represent a crucial resource for simulating doctors' skills in medical-interviews.

*Objectives:* The goal of this study was to evaluate the data-gathering function of currently available chatbot symptom-checkers.

*Methods:* We evaluated 8 symptom-checkers using 28 clinical vignettes from the repository of MSD-Manual case studies. The mean number of predefined pertinent findings for each case was 31.8 ± 6.8. The vignettes were entered into the platforms by 3 medical students who simulated the role of the patient. For each conversation, we obtained the number of pertinent findings retrieved and the number of questions asked. We then calculated the recall-rates (pertinent-findings retrieved out of all predefined pertinent-findings), and efficiency-rates (pertinent-findings retrieved out of the number of questions asked) of data-gathering, and compared them between the platforms.

*Results:* The overall recall rate for all symptom-checkers was 0.32(2,280/7,112;95 %CI 0.31–0.33) for all pertinent findings, 0.37(1,110/2,992;95 %CI 0.35–0.39) for present findings, and 0.28(1140/4120;95 %CI 0.26–0.29) for absent findings. Among the symptom-checkers, Kahun platform had the highest recall rate with 0.51(450/889;95 %CI 0.47–0.54). Out of 4,877 questions asked overall, 2,280 findings were gathered, yielding an efficiency rate of 0.46(95 %CI 0.45–0.48) across all platforms. Kahun was the most efficient tool 0.74 (95 %CI 0.70–0.77) without a statistically significant difference from Your.MD 0.69(95 %CI 0.65–0.73).

*Conclusion:* The data-gathering performance of currently available symptom checkers is questionable. From among the tools available, Kahun demonstrated the best overall performance.

* Corresponding author at: Department of Medicine 'B', Sheba Medical Center, Ramat Gan, 5262100, Israel.
*E-mail addresses:* nivben7@gmail.com, nivbenshabat@mail.tau.ac.il (N. Ben-Shabat).

https://doi.org/10.1016/j.ijmedinf.2022.104897
Received 2 May 2022; Received in revised form 9 October 2022; Accepted 10 October 2022
Available online 22 October 2022

## 1. Introduction

The burden on healthcare systems is mounting worldwide owing to an increase in population growth and aging, the rise in disease incidence, the overuse of medical services, and the recent COVID-19 pandemic [1–3]. These challenges, combined with the severe shortage of healthcare professionals, which is only expected to grow in next years [4,5], are responsible for deteriorating availability of medical services and an increased practitioners' burnout [6,7]. These, in turn, are causing reduced quality of medical care and worsened patients' clinical outcomes [8–12].

Part of the solution is expected to involve the integration of technological advancements into routine healthcare [13], especially from the field of artificial intelligence (AI). One patient-oriented approach gaining popularity is the use of intelligent digital self-assessment tools, known as symptom checkers [14,15]. Most advanced symptom checkers use a conversational "chatbot" format to collect information from the patient and then return an output of likely diagnoses and triage advice [16]. These tools are used in several contexts: independently by patients seeking guidance about health problems, as a part of a service provided by an online or telehealth caregiver, or incorporated into practice in medical centers, primarily in the emergency departments [15,17–19]. These symptom checkers are able to gather and summarize medical information, allocate patients to an appropriate level of care, and suggest potential diagnoses and treatment options. As such, they carry the potential to save trained practitioners time, decrease the overuse of medical services, and minimize unnecessary mistakes [13,19–22]. All of which, theoretically, could reduce the load on healthcare systems and improve healthcare quality.

Although symptom checkers have been investigated in terms of their accuracy for diagnosis and triage [14,23–28], no study had investigated their data-gathering function, which is a cornerstone skill for performing an adequate medical interview [29–31]. The process for gathering clinical evidence is very complex. It's objective is to collect all pertinent information that can corroborate or discredit potential diagnoses (hypotheses); these differential diagnoses are continuously evolving based on new data from the patients' answers [31–34]. The ability of AI-driven tools to simulate these skills has the potential to reduce the burden of healthcare systems by making the medical-interview process more efficient and accurate, while saving precious time for trained personnel. To ensure the useful and safe integration of these tools, a thorough investigation of different aspects of the diagnostic process is needed, as opposed to examining only its endpoints. This may also increase physicians' trust in AI technologies, which has been noted as one of the factors delaying their assimilation [35].

This study aims to evaluate the data-gathering function of currently available symptom checkers, and thereby measure the usefulness and effectivity of these tools.

## 2. Methods

This study was based on data from clinical vignettes. The vignettes were processed and entered into the different symptom-checker platforms by 3 medical students who simulated the role of the patient. For each conversation, we obtained the number of pertinent findings retrieved and the number of questions asked.

### 2.1. Selection of clinical vignettes

The clinical vignettes were selected from the case bank of Merck Sharpe & Dohme Clinical Manuals, known as MSD Manuals [36]. The content of these Manuals is created by independent reviewers who are experts in their fields and goes through a process of peer-review before being published. All 35 vignettes that were available in the MSD repository were screened for eligibility. Of these, 3 vignettes were excluded because they involved patients under the age of 18 (not legally allowed to use symptom-checker applications), 3 were excluded because they involved unconscious patients unable to conduct a conversation with a chatbot, and 1 was excluded because it had no chief complaint (routine checkup scenario). This left 28 clinical vignettes that were included in the analysis.

### 2.2. Creation of case transcripts

We constructed a case transcript for each vignette and divided the data into two sets of information: 1) Voluntary/pre-existing information, 2) Nonvoluntary/extracted information. The first set of information was composed of fixed data which was defined as known about the patient; it included date of birth, gender, comorbidities, medications used, and chief complaint, which is all defined as given voluntarily during the simulated conversation. The second set of information was composed of data defined as needed to be extracted from the patient during the medical interview; this included symptoms (cough, fever, dyspnea, etc.) and the characterization of these symptoms (duration, timing, aggravating and relieving factors, etc.). This second set of information was further divided into present and absent pertinent findings, meaning findings whose inclusion or exclusion are diagnostically valuable for a specific clinical scenario, as determined by the original vignette's creator. The mean number of non-voluntary findings defined per case was 31.8 $\pm$ 6.8, of which 13.4 $\pm$ 6.4 were present findings, significant to be ruled-in, and 18.4 $\pm$ 4.2 were absent (normal) findings, significant to be ruled-out. An example of a case transcript is presented in the supplementary materials (Appendix S1).

### 2.3. Inclusion of symptom checkers

The study aimed to include all advanced popular chatbot symptom-checkers (CSC) publicly available. The search was conducted during January 2022 and was based on the search strategy previously reported by Ceney et al.[14]. We excluded symptom checkers that: used another algorithm provider as their main source, had no ability to convey conversation (chatbot function), focused on a single condition, were not available for Israeli residents, or had no mobile application available. Overall, 8 CSC were eligible for inclusion in the study: Ada [37], Babylon [38], Buoy [39], Kahun [40], K Health [41], Mediktor [42], Symptomate [43], and Your.MD [44].

### 2.4. Procedures and design

During January 2022, 3 fourth-year medical students played the role of the patient in the simulated conversations and entered the case transcripts into the symptom checkers. The students had no prior experience with any of the symptom checkers tested. Each student was assigned to simulate between 9 and 10 cases for each platform. To reduce the potential for bias, the same case was simulated by the same student in all the platforms. For each symptom checker, the latest version of the mobile application (available in Google Play as of January 2022) was tested using a mobile phone carrying the latest version of the Android operating system. The students were instructed to actively enter the simulated patient's voluntary findings, when possible. The non-voluntary findings were delivered only when directly asked for by the chatbot. In addition, the students were instructed to adhere strictly to the case transcripts. If the 'patient' was asked about a symptom that did not appear in the transcript's list of present findings or absent findings, the option "I don't know" was instructed to be chosen. If this option did not exist in the tool tested, they were instructed to answer "no". After each run, the number of questions asked by the chatbot, and the number of pertinent findings gathered during the conversation were recorded.

### 2.5. Measures

Recall – Calculated as the number of pertinent findings retrieved by

the chatbot during the conversation, relative to the total number of pertinent findings predefined in the case. We calculated this measure for all findings, and separately for present findings and absent findings.

Efficiency – Calculated as the number of pertinent findings acquired by the chatbot during the conversation, relative to the number of questions asked by the chatbot. Ininverse relation to the number of redundant questions asked.

### 2.6. Statistical analysis

The 95 % confidence intervals (CI) were calculated assuming a binomial distribution. We compared the results between different symptom checkers using the Pearson Chi-Square test with post-hoc comparison. All p values were two-tailed, and the null hypothesis was considered true if $p \geq 0.05$. All statistical analysis was done using IBM SPSS Statistics version 26 (Armonk, NY: IBM Corp).

## 3. Results

Table 1 presents a comparison of the recall rates measured. The overall data-gathering recall rate for all symptom-checkers was 0.32 (2280/7,112; 95 %CI 0.31 to 0.33) for all pertinent findings, 0.37 (1,110/2,992; 95 %CI 0.35 to 0.39) for present findings, and 0.28 (1,140/4,120; 95 %CI 0.26 to 0.29) for absent findings. A statistically significant difference was observed between the recall rate for present and absent findings ($p < 0.001$); this demonstrates that the symptom checkers generally performed the task of ruling-in present abnormal findings better than excluding relevant absent findings. Among the symptom checkers, Kahun demonstrated the best overall recall rates for findings with 0.51(450/889; 95 %CI 0.47 to 0.54), and the best recall rates for present findings with 0.64 (240/374; 95 %CI 0.59 to 0.69). These results were significantly superior to those of the other tools (Table S1- 2). The highest recall rates for absent findings were observed for Your.MD 0.44 (228/515; 95 %CI 0.40 to 0.49) and Kahun 0.41(210/515; 95 %CI 0.37 to 0.45), with no significant difference between them ($p = 0.471$; Table S3).

Table 2 presents a comparison of efficiency rates. The mean number of questions (interactions) asked during a conversation was 21.8 ± 9.2, showing a large variance between the tools. The highest number of questions was observed for ADA with 29.8 ± 5.8 questions per case, while the lowest was observed for Babylon with only 9.0 ± 6.0 questions per case. The mean number of findings gathered per case was 10.0 ± 5.5. Kahun had the highest number, with 16.1 ± 5.7 findings gathered per case, and Babylon had the lowest, with 5.0 ± 4.2 findings gathered per case. Out of 4,877 questions asked overall, 2,280 findings were gathered, yielding an efficiency rate of 0.46 (95 %CI 0.45 to 0.48) across all platforms. Kahun had the highest rate with 0.74 (95 %CI 0.70 to 0.77) with no statistically significant difference from Your.MD 0.69(95 %CI 0.65 to 0.73) (Table S4).

**Table 2**
Comparison of efficiency rate between symptom checkers.

| | N questions | Mean questions per case | efficiency rate | 95 % CI |
|---|---|---|---|---|
| All | 4976 | 22.2 ± 9.4 | 0.46 | 0.44–0.47 |
| ADA | 833 | 29.8 ± 5.8 | 0.36 | 0.33–0.40 |
| Babylon | 252 | 9.0 ± 6.0 | 0.56 | 0.49–0.62 |
| Buoy | 723 | 25.8 ± 7.9 | 0.37 | 0.34–0.41 |
| Kahun | 709 | 25.3 ± 10.3 | 0.68 | 0.64–0.71 |
| K Health | 739 | 26.4 ± 7.5 | 0.46 | 0.43–0.50 |
| Mediktor | 630 | 22.5 ± 6.8 | 0.25 | 0.22–0.29 |
| Symptomate | 558 | 19.9 ± 5.8 | 0.38 | 0.34–0.42 |
| Your.MD | 532 | 19.0 ± 6.3 | 0.69 | 0.65–0.73 |

Fig. 1 illustrates the relationship between the recall rate, which reflects the comprehensiveness of data-gathering, to the efficiency of data-gathering. A relatively strong linear association was observed between the two ($r = 0.66$). Kahun, K-Health, ADA, and Buoy are all positioned above the trend line. Fig. 2 illustrates the ratio of the number of findings gathered to the number of questions asked per case. A moderate linear correlation was observed between the two variables ($r = 0.43$), suggesting that the quality of the questions asked plays a bigger role in retrieving a patient's finding than the number of questions. Kahun, Your.MD, and K-Health were all positioned above the trend line.

## 4. Discussion

We conducted what we believe to be the first study on the data-gathering capabilities of publicly available chatbot symptom-checkers. Overall, the data-gathering recall (comprehensiveness) and efficiency of the tested symptom checkers were in the low-range with 0.32 and 0.46, respectively; nevertheless, a comparison to these rates for physicians is advisable for a more accurate interpretation. In general, the symptom checkers that were tested favored collecting diagnostically relevant present findings over the exclusion of all relevant absent

**Table 1**
Comparison of pertinent findings recall rates (sensitivity) between symptom checkers.

| | All Findings | | | Present Findings | | | Absent Findings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean findings per case | Recall rate | 95 %CI | Mean findings per case | Recall rate | 95 %CI | Mean findings per case | Recall rate | 95 %CI |
| All | 10.2 ± 5.6 | 0.32 2280/7112 | 0.31–0.33 | 5.0 ± 3.4 | 0.37 1109/2992 | 0.35–0.39 | 5.2 ± 3.6 | 0.28 1171/4120 | 0.27–0.30 |
| ADA | 10.9 ± 3.5 | 0.34 304/889 | 0.31–0.37 | 6.4 ± 3.1 | 0.48 179/374 | 0.43–0.53 | 4.5 ± 2.4 | 0.24 125/515 | 0.21–0.28 |
| Babylon | 5.0 ± 4.2 | 0.16 140/889 | 0.13–0.18 | 1.9 ± 1.8 | 0.14 54/374 | 0.11–0.18 | 3.1 ± 3.0 | 0.17 86/515 | 0.13–0.20 |
| Buoy | 9.7 ± 3.2 | 0.30 271/889 | 0.27–0.34 | 4.7 ± 2.0 | 0.35 132/374 | 0.30–0.40 | 5.0 ± 2.8 | 0.27 139/515 | 0.23–0.31 |
| Kahun | 17.1 ± 5.1 | 0.54 480/889. | 0.51–0.57 | 8.5 ± 4.1 | 0.64 239/374 | 0.59–0.69 | 8.6 ± 3.7 | 0.47 241/515 | 0.42–0.51 |
| K Health | 12.2 ± 5.1 | 0.38 342/899 | 0.35–0.42 | 5.8 ± 3.3 | 0.44 163/374 | 0.39–0.49 | 6.4 ± 3.4 | 0.35 179/515 | 0.31–0.39 |
| Mediktor | 5.7 ± 3.0 | 0.18 160/889 | 0.15–0.21 | 3.3 ± 2.3 | 0.25 93/374 | 0.20–0.29 | 2.4 ± 1.7 | 0.13 67/515 | 0.10–0.16 |
| Symptomate | 7.6 ± 3.5 | 0.24 214/889 | 0.21–0.27 | 3.9 ± 2.6 | 0.29 108/374 | 0.24–0.33 | 3.8 ± 2.0 | 0.21 106/515 | 0.17–0.24 |
| Your.MD | 13.2 ± 5.0 | 0.42 369/889 | 0.38–0.45 | 5.0 ± 3.1 | 0.38 141/374 | 0.33–0.43 | 8.1 ± 3.8 | 0.44 228/515 | 0.40–0.49 |

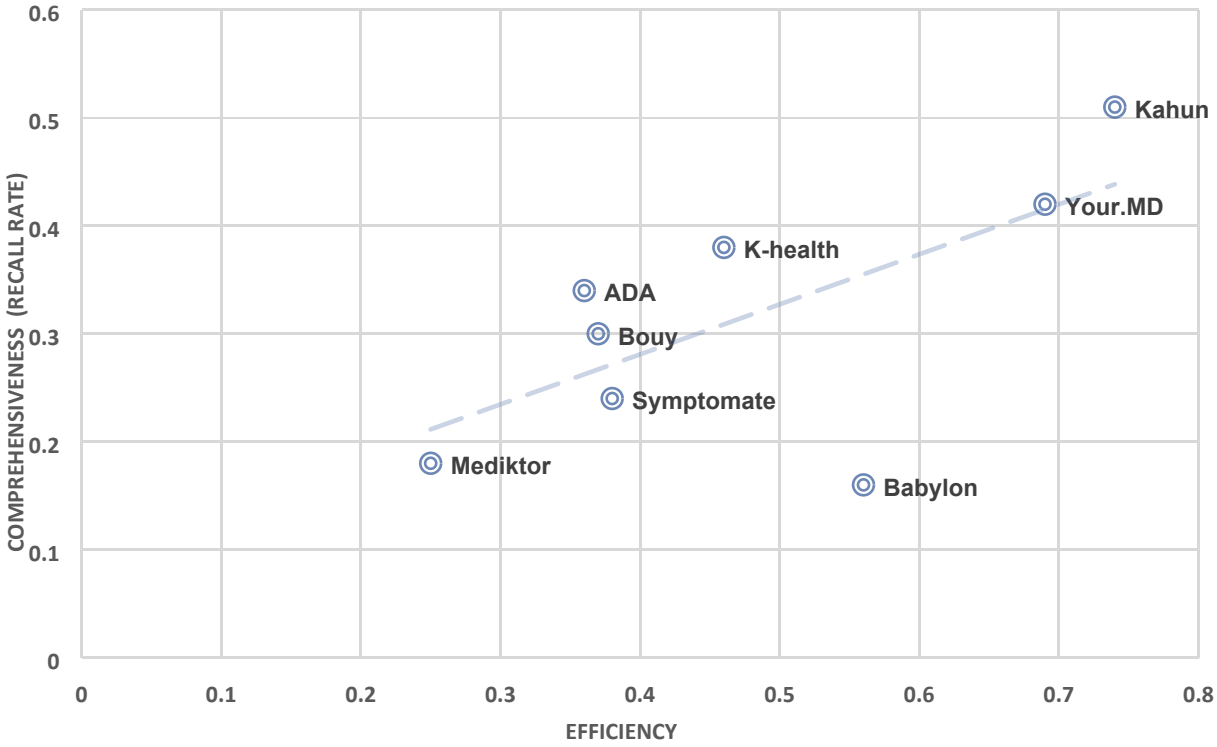



**Fig. 1.** Scatter plot demonstrating the correlation between comprehensiveness of data-gathering (measured as recall rate) to efficiency rate.

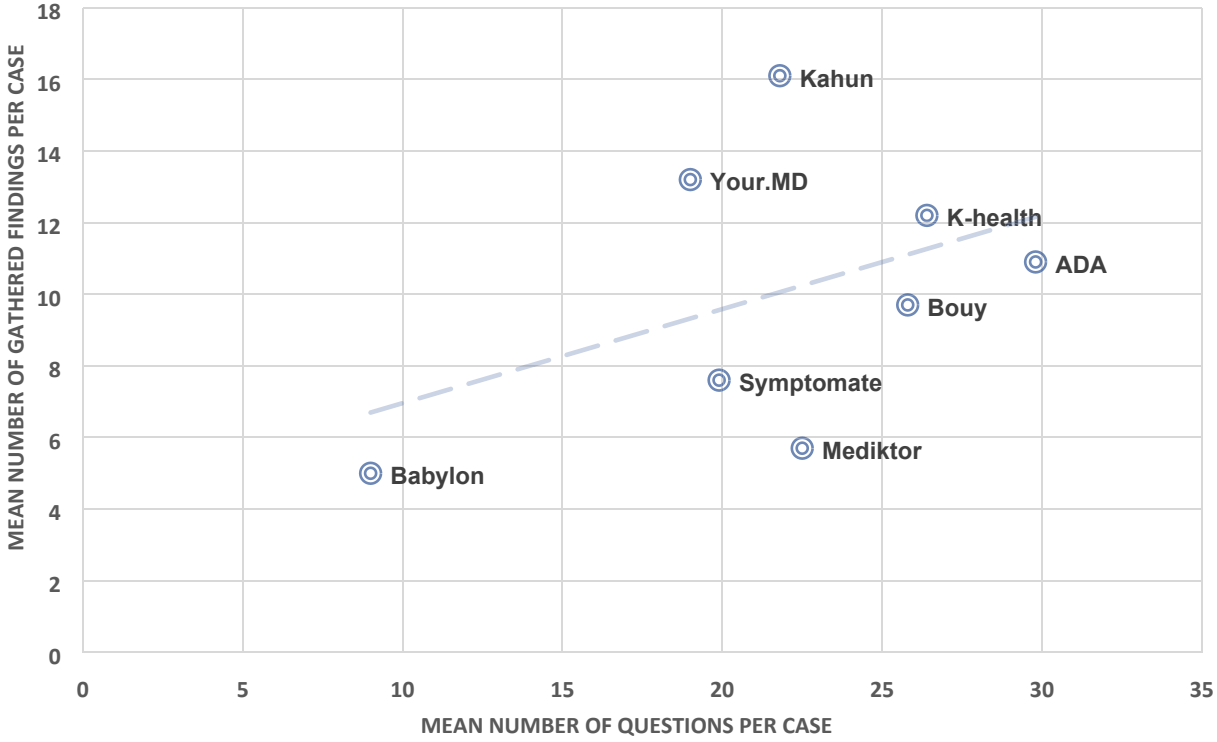



**Fig. 2.** Scatter plot demonstrating the correlation between number of gathered findings and the number of questions asked during a conversation.

findings. This suggests an orientation towards ruling-in differential diagnoses as opposed to ruling them out. Among the symptom-checkers tested, Kahun demonstrated the best data-gathering recall and precision, followed by Your.MD.

To the best of our knowledge, this is the first study to evaluate the data-gathering aspect of advanced intelligent symptom checkers. This area of symptom checkers seems to have been overlooked, as opposed to aspects of diagnostic and triage accuracy, which have been studied extensively. This emphasizes the fact that most available tools perceive their designated utility as keeping patients (users) well-informed about the root causes of their symptoms, suggesting triage advice, and proposing the most likely causes to the provider. We believe these tools

should also be focused on their potential for virtual-intake and aimed at performing the task of a medical interview in a manner similar to a healthcare professional.

The task of performing a relevant and efficient medical interview requires clinical reasoning skills; these skills are very resource demanding and require years of training and experience before they can be mastered by human clinicians [45,46]. Moreover, tremendous variability exists between physicians, depending on their knowledge, beliefs, experience, and training [45,47]; this variability can even exist within the same practitioner on a case to case basis, depending on context [48–50]. A virtual intake-oriented tool that can competently perform a high-quality medical interview in a consistent manner, and provide a relevant summary to the doctor, could reduce the workload of practitioners and allow them to focus on the task of clinical decision-making. Such a tool would also help reduce the variability between doctors and for individual doctors, narrowing the gap between experienced trained physicians and those with less experience, or between the same physician in the first hour of the shift as compared to the last. These implications are especially relevant for telemedicine and home-care practices, which are developing and transforming owing to technological advancements and the effects of the COVID-19 pandemic [17,51].

To achieve the task of high-performance data collection, an AI system must be able to calculate the working differential diagnosis at each point in time, given a set of patient's collected findings, and compute the next best question according to these calculations. This process of hypothesis-driven data collection is imperative for AI systems attempting to conduct high quality medical interviews. Symptom assessment tools using static flow-chart trees or pre-built pathways will only be able to assess the differential at the end of the conversation once the patient completes the pathway of questions. They are thus limited in their ability to adapt the questions according to the changing differential diagnosis. Such static pathways are meticulous and biased in their construction, and restrict the system to a pre-determined set of possible outcomes. Medical records typically include only the reported final diagnosis and not the differential diagnoses hypotheses conducted by the physician during the diagnostic process.

### *4.1. Limitations and future investigations*

Our study has several strengths. Among them are its innovativeness, the use of a validated external source for vignettes, and a design intended to reduce bias. Nevertheless, there are some important limitations to be acknowledged. The first is regarding the design, which is based on structured vignettes in a "sterile" environment rather than actual patients in a true clinical environment. The structured environment is deemed to be inferior to the real-life one as it cannot portray all the aspects of a real-world setting. However, this type of comparison is common and effective in the training and assessment of physicians' medical interview skills. Moreover, the standardized environment assists in reducing noise and confounders. For these reasons most of the studies conducted so far to address diagnostic accuracy have used similar methodology [24–27]. Another limitation is that we did not test the quality of medical history gathered. This was done for several reasons. First, most of the symptom checkers relate to the patient's medical history in a pre-defined, structured manner that is unrelated to the findings. Second, in their designated environment, symptom checkers already have the medical history details of a patient after their first use, or even prior to that through integration with the medical records. Third, for this study, we considered the expert opinion of the healthcare professionals who constructed the vignettes as the gold standard for establishing which findings are pertinent. Similar to any decision based on a "gold standard", this choice is problematic, especially since AI algorithms carry the potential to exceed the standard logic of a physician in determining which findings are most valuable in terms of diagnostics.

Further evaluation is required to overcome the limitations discussed above and to increase the validity and utility of symptom-checkers. Subsequent investigations should compare the data-gathering characteristics between symptom-checkers and physicians with different experience, first in a controlled environment using structured cases and later in a real-life clinical environment. Further work is needed to construct an appropriate investigational framework including standardized definitions, design, outcomes, and measures. This will ensure the studies are informative, comparable, and reproducible. Finally, using machine learning technologies, an attempt should be made to tackle the question "what is an important finding?" One possible method could use a clinical trial comparing relevant clinical outcomes of patients interviewed by an intelligent CSC versus trained physicians. Although not expected in the near future, such investigations could provide new insights to answer the question mentioned above, stretching beyond the scope of clinical reasoning and textbook medicine.

## 5. Conclusions

The comprehensiveness and efficiency of data-gathered by currently available CSCs is questionable. Among available tools, Kahun demonstrated the best overall performance. Investigating and evaluating the data-gathering skills of AI-driven symptom checkers is of great importance because the refinement of these skills to the level of a human professional can impart countless benefits to patients and healthcare systems worldwide.

### *5.1. Summary points:*

- As opposed to their triage and diagnostic accuracy, the data-gathering capabilities of chatbot symptom checkers, which is an important feature, had not yet been evaluated.
- In the present study, we evaluated the efficiency and comprehensiveness of data gathering in all publicly available chatbot symptom checkers and found they are overall unsatisfactory.
- Among the tools tested, Kahun demonstrated the best overall performance.

## Funding

This study was fully funded by Kahun Medical ltd. The funder provided support in the form of salaries for authors [NBS, GS, BM, and AS], but did not play any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

All clinical vignettes were obtained from the MSD Manual Professional Version (known as the Merck Manual in the US and Canada, and the MSD Manual in the rest of the world), edited by Robert Porter. Copyright (2022) by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ. Available at http://www.msdmanuals.com/professional. Accessed (01/01/2022).

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104897.

## References


[1] S. Brownlee, K. Chalkidou, J. Doust, A.G. Elshaug, P. Glasziou, I. Heath, S. Nagpal, V. Saini, D. Srivastava, K. Chalmers, D. Korenstein, Evidence for Overuse of medical services around the world, Lancet (London, England). 390 (10090) (2017) 156–168.

[2] S.L. James, D. Abate, K.H. Abate, S.M. Abay, C. Abbafati, N. Abbasi, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017, Lancet (London, England). 392 (10159) (2018 Nov 10) 1789.

[3] J.L. Dieleman, E. Squires, A.L. Bui, M. Campbell, A. Chapin, H. Hamavid, C. Horst, Z. Li, T. Matyasz, A. Reynolds, N. Sadat, M.T. Schneider, C.J.L. Murray, Factors associated with increases in US health care spending, 1996–2013, JAMA. 318 (17) (2017) 1668.

[4] Closing the gap: Key areas for action on the health and care workforce.

[5] Dall T, West T. 2017 Update The Complexities of Physician Supply and Demand: Projections from 2015 to 2030 Final Report Association of American Medical Colleges. 2017.

[6] N.R. Hoot, D. Aronsky, Systematic review of emergency department crowding: causes, effects, and solutions, Ann. Emerg. Med. 52 (2) (2008) 126–136.e1.

[7] S. Di Somma, L. Paladino, L. Vaughan, I. Lalle, L. Magrini, M. Magnanti, Overcrowding in emergency department: an international issue, Intern. Emerg. Med. 10 (2) (2015 Mar 1) 171–175.

[8] M. Tuczyńska, R. Staszewski, M. Matthews-Kozanecka, A. Żok, E. Baum, Quality of the Healthcare Services During COVID-19 Pandemic in Selected European Countries, Front Public Heal. 12 (10) (2022 May), 870314.

[9] C.S. Dewa, D. Loong, S. Bonato, L. Trojanowski, The relationship between physician burnout and quality of healthcare in terms of safety and acceptability: a systematic review, BMJ Open. 7 (6) (2017) e015141.

[10] S.L. Dickman, D.U. Himmelstein, S. Woolhandler, Inequality and the health-care system in the USA, Lancet (London, England). 389 (10077) (2017 Apr 8) 1431–1441.

[11] P.G. Jones, D. Mountain, R. Forero, Review article: Emergency department crowding measures associations with quality of care: a systematic review, Emerg. Med. Australas. 33 (4) (2021 Aug 1) 592–600.

[12] S.L. Bernstein, D. Aronsky, R. Duseja, S. Epstein, D. Handel, U. Hwang, M. McCarthy, K. John McConnell, J.M. Pines, N. Rathlev, R. Schafermeyer, F. Zwemer, M. Schull, B.R. Asplin, The effect of emergency department crowding on clinically oriented outcomes, Acad. Emerg. Med. 16 (1) (2009) 1–10.

[13] O.H. Salman, Z. Taha, M.Q. Alsabah, Y.S. Hussein, A.S. Mohammed, M. Aal-Nouman, A review on utilizing machine learning technology in the fields of electronic emergency triage and patient priority systems in telemedicine: coherent taxonomy, motivations, open research challenges and recommendations for intelligent future work, Comput. Methods Programs Biomed. 1 (2021 Sep) 209.

[14] Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS One. 2021 Jul 1;16(7).

[15] 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 1;22(11).

[16] Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The Personalization of Conversational Agents in Health Care: Systematic Review. J Med Internet Res. 2019 Nov 1;21(11).

[17] Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of Telehealth During the COVID-19 Pandemic: Scoping Review. J Med Internet Res. 2020 Dec 1;22(12).

[18] Shafaf N, Malek H. Applications of Machine Learning Approaches in Emergency Medicine; a Review Article. Arch Acad Emerg Med. 2019 Jan 1;7(1):e34.

[19] Mueller B, Kinoshita T, Peebles A, Graber MA, Lee S. Artificial intelligence and machine learning in emergency medicine: a narrative review. Acute Med Surg. 2022 Jan;9(1):e740.

[20] Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, 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.

[21] H. Fraser, E. Coiera, D. Wong, Safety of patient-facing digital symptom checkers, Lancet. 392 (10161) (2018 Nov 24) 2263–2264.

[22] Stephanie, Liu RH, Desta BN, Chaurasia A, Ebrahim S. The Use of Artificially Intelligent Self-Diagnosing Digital Platforms by the General Public: Scoping Review. JMIR Med Inf 2019;7(2)e13445 https//medinform.jmir.org/2019/2/e13445. 2019 May 1;7(2):e13445.

[23] H.L. Semigran, J.A. Linder, C. Gidengil, A. Mehrotra, Evaluation of symptom checkers for self diagnosis and triage: Audit study, BMJ. 8 (2015 Jul) 351.

[24] H.L. Semigran, D.M. Levine, S. Nundy, A. Mehrotra, Comparison of physician and computer diagnostic accuracy, JAMA Intern. Med. 176 (12) (2016 Dec 1) 1860–1861.

[25] M.G. Hill, M. Sim, B. Mills, The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia, Med. J. Aust. 212 (11) (2020 Jun 1) 514–519.

[26] A. Baker, Y. Perov, K. Middleton, J. Baxter, D. Mullarkey, D. Sangar, et al., A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis, Front. Artif. Intell. 30 (2020 Nov) 3.

[27] Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, 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 1;10(12): e040269.

[28] A.E. Poote, D.P. French, J. Dale, J. Powell, A study of automated self-assessment in a primary care student health centre setting, J. Telemed. Telecare. 20 (3) (2014 Mar 18) 123–127.

[29] G.L. Engel, Interviewing the patient /, WB Saunders, Philadelphia, 1973.

[30] F.C. Thorne, The evidence gathering process, Clin. Judgm A study Clin. error. 24 (2015 Aug) 101–107.

[31] Cole SA. Function I: gathering data to understand the patient. Third edition. The medical interview : the three function approach /. Philadelphia, PA : Elsevier,; 1991. 23–42 p.

[32] A. Holmes, B. Singh, G. McColl, Revisiting the hypothesis-driven interview in a contemporary context, Australas Psychiatry. 19 (6) (2011 Dec 1) 484–488.

[33] Feinstein AR. An analysis of diagnostic reasoning. II. The strategy of intermediate decisions. Yale J Biol Med. 1973;46(4):264.

[34] J.P. Kassirer, G.A. Gorry, Clinical problem solving: a behavioral analysis, Ann. Intern. Med. 89 (2) (1978) 245–255.

[35] Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J Med Internet Res 2020;22(6)e15154 https//www.jmir.org/2020/6/e15154. 2020 Jun 19;22(6):e15154.

[36] https://www.msdmanuals.com/professional.

[37] https://www.ada.com/.

[38] https://www.babylonhealth.com/.

[39] https://www.buoyhealth.com/.

[40] https://www.kahun.com/.

[41] https://khealth.com/.

[42] https://www.mediktor.com/.

[43] https://symptomate.com/.

[44] https://www.livehealthily.com/.

[45] G.M. Joseph, V.L. Patel, Domain knowledge and hypothesis generation in diagnostic reasoning, Med. Decis Mak. 10 (1) (1990 Jul 2) 31–46.

[46] A.S. Elstein, L.S. Shulman, S.H. Sprafka, S.A. Sprafka, Medical Problem Solving an Analysis of Clinical Reasoning, Harvard University Press, 1978.

[47] S. Fürstenberg, T. Helm, S. Prediger, M. Kadmon, P.O. Berberat, S. Harendza, Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators, BMC Med. Educ. 20 (1) (2020).

[48] S.J. Durning, A.R. Artino, J.R. Boulet, K. Dorrance, C. van der Vleuten, L. Schuwirth, The impact of selected contextual factors on experts' clinical reasoning performance (does context impact clinical reasoning performance in experts?), Adv. Heal. Sci. Educ. 17 (1) (2012 Apr 20) 65–79.

[49] S. Durning, A.R. Artino, L. Pangaro, C.P. van der Vleuten, L. Schuwirth, Context and clinical reasoning: understanding the perspective of the expert's voice, Med. Educ. 45 (9) (2011 Sep 1) 927–938.

[50] K.W. Eva, A.J. Neville, G.R. Norman, Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving, Acad Med. 73 (10 Suppl) (1998).

[51] M.A. Hincapié, J.C. Gallego, A. Gempeler, J.A. Piñeros, D. Nasner, M.F. Escobar, Implementation and Usefulness of Telemedicine During the COVID-19 Pandemic: a Scoping Review, J. Prim. Care Community Health. 11 (2020).