Original Paper

Diagnostic Accuracy of Web-Based COVID-19 Symptom Checkers: Comparison Study

Nicolas Munsch[1], MSc; Alistair Martin[1], DPhil; Stefanie Gruarin[2], MD, PhD; Jama Nateqi[2,3], MD; Isselmou Abdarahmane[1], BSc; Rafael Weingartner-Ortner[1,2], BSc; Bernhard Knapp[1], PhD

[1]Data Science Department, Symptoma, Vienna, Austria

[2]Medical Department, Symptoma, Attersee, Austria

[3]Department of Internal Medicine, Paracelsus Medical University, Salzburg, Austria

Corresponding Author:
Bernhard Knapp, PhD
Data Science Department
Symptoma
Landstraßer Gürtel 3
Vienna, 1030
Austria
Phone: 43 662458206
Email: science@symptoma.com

Abstract

Background: A large number of web-based COVID-19 symptom checkers and chatbots have been developed; however, anecdotal evidence suggests that their conclusions are highly variable. To our knowledge, no study has evaluated the accuracy of COVID-19 symptom checkers in a statistically rigorous manner.

Objective: The aim of this study is to evaluate and compare the diagnostic accuracies of web-based COVID-19 symptom checkers.

Methods: We identified 10 web-based COVID-19 symptom checkers, all of which were included in the study. We evaluated the COVID-19 symptom checkers by assessing 50 COVID-19 case reports alongside 410 non–COVID-19 control cases. A bootstrapping method was used to counter the unbalanced sample sizes and obtain confidence intervals (CIs). Results are reported as sensitivity, specificity, F1 score, and Matthews correlation coefficient (MCC).

Results: The classification task between COVID-19–positive and COVID-19–negative for "high risk" cases among the 460 test cases yielded (sorted by F1 score): Symptoma (F1=0.92, MCC=0.85), Infermedica (F1=0.80, MCC=0.61), US Centers for Disease Control and Prevention (CDC) (F1=0.71, MCC=0.30), Babylon (F1=0.70, MCC=0.29), Cleveland Clinic (F1=0.40, MCC=0.07), Providence (F1=0.40, MCC=0.05), Apple (F1=0.29, MCC=-0.10), Docyet (F1=0.27, MCC=0.29), Ada (F1=0.24, MCC=0.27) and Your.MD (F1=0.24, MCC=0.27). For "high risk" and "medium risk" combined the performance was: Symptoma (F1=0.91, MCC=0.83) Infermedica (F1=0.80, MCC=0.61), Cleveland Clinic (F1=0.76, MCC=0.47), Providence (F1=0.75, MCC=0.45), Your.MD (F1=0.72, MCC=0.33), CDC (F1=0.71, MCC=0.30), Babylon (F1=0.70, MCC=0.29), Apple (F1=0.70, MCC=0.25), Ada (F1=0.42, MCC=0.03), and Docyet (F1=0.27, MCC=0.29).

Conclusions: We found that the number of correctly assessed COVID-19 and control cases varies considerably between symptom checkers, with different symptom checkers showing different strengths with respect to sensitivity and specificity. A good balance between sensitivity and specificity was only achieved by two symptom checkers.

(J Med Internet Res 2020;22(10):e21299) doi: 10.2196/21299

KEYWORDS

COVID-19; symptom checkers; benchmark; digital health; symptom; chatbot; accuracy

Introduction

In the modern world, large numbers of patients initially turn to various web-based sources for self-diagnoses of health concerns before seeking diagnoses from a trained medical professional. However, web-based sources have inherent problems, such as misinformation, misunderstandings, misleading advertisements, and varying quality [1]. Interactive web sources developed to

provide web-based diagnoses are sometimes referred to as symptom checkers or chatbots [2,3]. Based on a list of entered symptoms and other factors, these symptom checkers return a list of potential diseases.

Web-based symptom checkers have become popular in the context of the novel COVID-19 pandemic, as access to physicians is reduced, concern in the population is high, and large amounts of misinformation are circulating the internet [1]. On COVID-19 symptom checker web pages, users are asked a series of COVID-19–specific questions; upon completion, an association between the answers and COVID-19 is given alongside behavioral recommendations, such as self-isolation.

In this context, COVID-19 symptom checkers are valuable tools for preassessment and screening during this pandemic; they can both ease pressure on clinicians and decrease footfall within hospitals. One example is practicing social distancing by not going to a physician's waiting room when feeling sick. The importance of social distancing has been highlighted in the COVID-19 pandemic [4,5], the 2009 H1N1 influenza pandemic [6], and the 1918-1919 influenza pandemic [7] and is reviewed in [8]. Symptom checkers can also ease pressure on medical telephone hotlines [9,10] by reducing the number of human phone operators needed.

A large number of symptom checkers specific to COVID-19 have been developed. Empirical evidence (eg, a newspaper article [11]) suggests that their conclusions differ, with possible implications for the quality of the symptom assessment. To our knowledge, there are no studies comparing and evaluating COVID-19 symptom checkers.

In this paper, we present a study evaluating 10 different web-based COVID-19 symptom checkers using 50 COVID-19 cases extracted from the literature and 410 non–COVID-19 control cases of patients with other diseases. We found that the classifications of many patient cases by the COVID-19 symptom checkers differ. Therefore the accuracies of symptom checkers also differ.

Methods

COVID-19 Symptom Checkers

In April 2020, we conducted a Google search for COVID-19 symptom checkers using the search terms *COVID-19 symptom checker* and *Corona symptom checker*. All ten COVID-19 symptom checkers that we found and that were freely available on the internet between April 3 and 9, 2020, were included in this study (Table 1). Nine checkers were implemented in the English language, while one was in German. These symptom checkers were used in the versions available in this date range, and updates after this date were not considered for analysis.

As a baseline for the performance evaluation of the 10 web-based COVID-19 symptom checkers, we developed two additional simplistic symptom checkers. These two checkers evaluate and weigh the presence of COVID-19 symptom frequencies provided by the World Health Organization (WHO) [12] (see Multimedia Appendix 1) based on vector distance (SF-DIST) and cosine similarity (SF-COS). These approaches can be implemented in a few lines of code (see Multimedia Appendix 2).

Table 1. List of web-based COVID-19 symptom checkers included in this study.

Name	Reference
Ada	[13]
Apple	[14]
Babylon	[15]
CDC[a]	[16]
Cleveland Clinic	[17]
Docyet	[18]
Infermedica	[19]
Providence	[20]
Symptoma	[21]
Your.MD	[22]

[a]CDC: US Centers for Disease Control and Prevention.

Clinical Cases

We used a total of 460 clinical cases to evaluate the performance of the COVID-19 symptom checkers. Each case lists both symptoms and the correct diagnosis alongside the age and sex of the patient when available. Details of the two case sets used are given below and in Table 2.

Table 2. Number of symptoms and age and sex distributions in each case set (N=460).

Characteristic	Case set	
	COVID-19, n=50	Control, n=410
Number of symptoms		
Mean (SD)	8.4 (4.1)	9.8 (4.4)
Median	7	9
Age (years)		
Mean (SD)	45.6 (16.9)	38.6 (22.4)
Median	45	38
Sex, n (%)		
Male	25 (50)	238 (58)
Female	21 (42)	160 (39)
Unknown	4 (8)	12 (2.9)

COVID-19 Cases

A total of 50 COVID-19 cases were extracted by three trained physicians from the literature in March and April 2020 and are listed in Multimedia Appendix 3. Each case describes one patient's medical situation (ie, symptoms experienced or COVID-19 contacts). The symptoms of each case were extracted separately from the COVID-19 engine construction and evaluation. The physicians entering the symptoms did not know how the engine would react to their symptom lists. To the best of our knowledge, we included all cases available at the time except for severe edge cases (eg, several severe comorbidities causing unrelated symptoms). Changes to the initial symptom lists were not allowed later.

Control Cases

The COVID-19 case data enabled us to evaluate the sensitivity of the symptom checkers. To evaluate the specificity, 410 control cases from the *British Medical Journal* (BMJ) were also sourced [23,24]. To allow a fair assessment, we only used cases containing at least one of the COVID-19 symptoms reported by the WHO [12] (see Multimedia Appendix 4). Classifying nonrelevant cases (eg, a fracture) would overestimate the symptom checkers' specificity. Furthermore, these patients would not consult a web-based COVID-19 symptom checker. None of the 410 BMJ cases lists COVID-19 as the diagnosis, as the cases where collected before the COVID-19 outbreak.

Mapping of Symptoms and Addressing Missing Inputs and Questions

Each of the symptom checkers has a different interface and different question sequences to reach the diagnosis. Therefore, we mapped the exhibited symptoms across our cases to the constrained input allowed by each checker via a synonym table and hierarchy created by a trained physician. For example, if a checker asked for "shortness of breath" but the case description listed "respiratory distress" or "(acute) dyspnea", the symptom was still correctly used for this case and symptom checker.

Not all cases contained answers to all the questions of a checker. In such cases, the answer "I don't know" was chosen; if the "I don't know" answer option did not exist in a checker, "No" was used. In contrast, if a case contained information that did not fit any of the questions of the checker, this information was not used for this checker.

Accuracy Evaluation

For statistical analysis, we used the following classification:

- True-positive : COVID-19 case classified as COVID-19
- False-positive: non–COVID-19 case classified as COVID-19
- True-negative: non–COVID-19 case classified as non–COVID-19
- False-negative: COVID-19 case classified as non–COVID-19

For each symptom checker, we calculated the following metrics:

Sensitivity (true-positive rate):

$$\frac{TP}{TP+FN}$$

Specificity (true-negative rate):

$$\frac{TN}{TN+FP}$$

F1 score (harmonic mean of the precision and recall):

$$\frac{2.TP}{2.TP+FP+FN}$$

Matthews correlation coefficient (MCC):

$$\frac{TP.TN-FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}$$

Classification of the Outputs of the Symptom Checkers

Most COVID-19 symptom checkers return human-readable text that contains an association between the entered symptoms and COVID-19. We classified these associations into three different categories: high risk, medium risk, and low risk. Respective examples of a high, medium, and low risk classification are "There is a high risk that COVID-19 is causing your symptoms," "Your symptoms are worrisome and may be related to

COVID-19," and "There's nothing at present to suggest that you have coronavirus (COVID-19). Please practice physical/social distancing." Our full mapping of text outputs to risk for all symptom checkers and all text outputs is given in Multimedia Appendix 5.

Some symptom checkers only have two possible outputs: COVID-19 risk or no COVID-19 risk. To compare symptom checkers with three and two risk levels, we performed two different analyses: (1) medium risk and high risk were treated as COVID-19–positive (and low risk was treated as COVID-19 as negative), and (2) high risk was treated as COVID-19–positive (and low risk and medium risk were treated as COVID-19–negative).

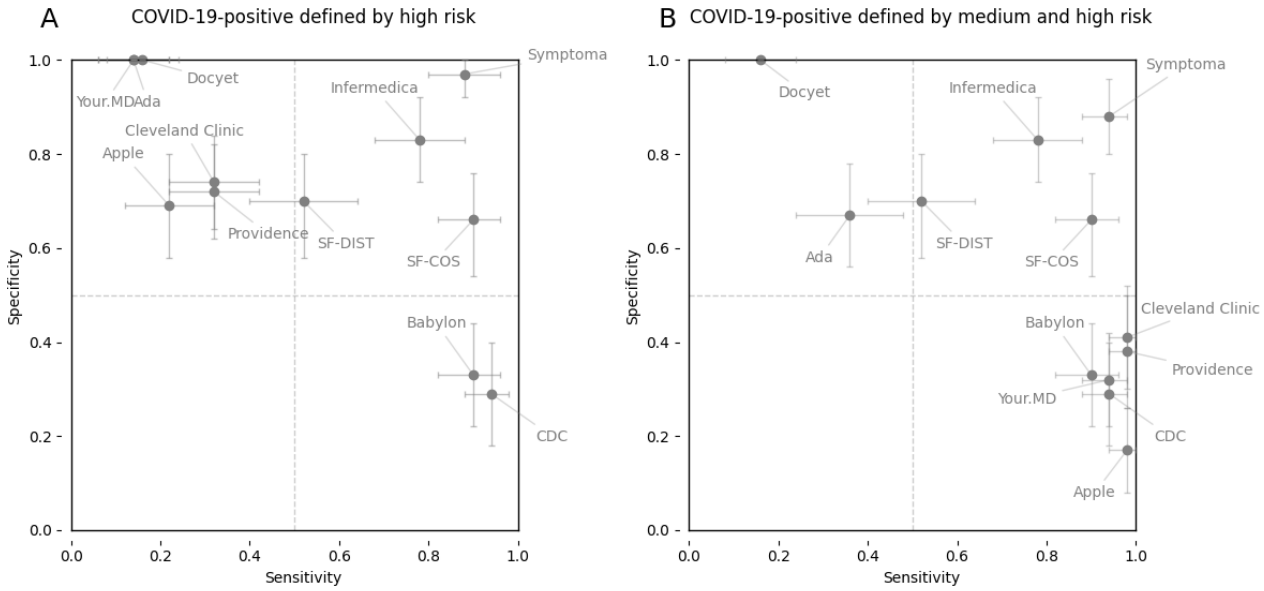
Bootstrapping

To evaluate the robustness of our statistical measures and account for the unbalanced dataset, we performed bootstrapping across our cases. A total of 3000 random samples consisting of 50 COVID-19 cases and 50 control cases was created by sampling with replacement from the original set of 50 COVID-19 cases and the 410 control cases.

Results

To analyze the performance of the 10 web-based symptom checkers, we calculated the sensitivity and the specificity of each symptom checker based on the cases described in the method section. Scatterplots of the sensitivity and specificity to COVID-19 of the different symptom checkers are given in Figure 1, and detailed numerical values are provided in Multimedia Appendix 6 and Multimedia Appendix 7. These symptom checkers fall approximately into four groups: upper left corner, lower right corner, central region, and upper right corner.

Figure 1. Sensitivities and specificities of web-based COVID-19 symptom checkers to COVID-19 cases and controls. The means of the 3000 random samples and 90% bootstrap CIs are reported as dots and crosses, respectively. (A) High risk: A COVID-19–positive prediction is defined only by a high risk result returned by a symptom checker. (B) Medium-high risk: A COVID-19–positive prediction is defined by either a medium risk or high risk result returned by a symptom checker. CDC: US Centers for Disease Control and Prevention; SF-COS: symptom frequency based on cosine similarity; SF-DIST: symptom frequency based on vector distance.



Further analysis of the true and false case classifications of these groups shows that the group in the upper left corner is composed of symptom checkers that require the presence of one (or few) highly specific symptoms to classify a case as COVID-19–positive (eg, "intensive contact with a COVID-19–positive person"). In this way, these symptom checkers miss many patients who are positive for COVID-19 who did not exactly report this highly specific symptom. In contrast, these highly specific symptoms are rarely present in non–COVID-19 cases. This results in low sensitivity and high specificity.

The group in the lower right corner is composed of symptom checkers that predict a case as COVID-19–positive based on the presence of one or few COVID-19 associated symptoms (eg, the presence of fever or cough is sufficient to predict a patient to be COVID-19–positive). These checkers classify almost every patient that has a respiratory disorder or viral infection as COVID-19–positive. As such, they do not miss many patients with COVID-19 but wrongly predict many patients who do not have COVID-19 to be COVID-19–positive. This results in low specificity and high sensitivity.

The group in the more central region is composed of symptom checkers that use a more balanced prediction but exhibit limited success in correctly classifying patients with and without COVID-19.

The group in the upper right corner is composed of symptom checkers that also use a more balanced model to associate symptoms to COVID-19; however, in this case, the classification of patients with and without COVID-19 is more successful.

Discussion

Principal Findings

We classified 50 COVID-19 case descriptions from the recent literature as well as 410 non–COVID-19 control cases using 10 different web-based COVID-19 symptom checkers. Only 2/10 symptom checkers showed a reasonably good balance between sensitivity and specificity (Figure 1). Most other checkers were either too sensitive, classifying almost all patients as COVID-19–positive, or too specific, classifying many patients with COVID-19 as COVID-19–negative (Figure 1). For example, our BMJ control cases included a patient suffering from a pulmonary disease who presented with various symptoms, including fever, cough, and shortness of breath, which are the three most frequent symptoms associated with COVID-19. Additional symptoms and risk factors were not considered by most checkers. Namely, loss of appetite, green sputum, and a history of smoking can be used to discern a correct diagnosis of COVID-19–negative.

Furthermore, in terms of F1 score, most of the symptom checkers were outperformed by a simplistic symptom frequency vector approach; the F1 scores were 0.57 and 0.79 for SF-DIST and SF-COS, respectively. Notably, the cosine version showed surprisingly good results, outperforming 8/10 symptom checkers based on the F1 score.

In contrast, it could also be argued that sensitivity is more important for a COVID-19 symptom checker than specificity (ie, numerous false-positive COVID-19 diagnoses are not of concern as long as no COVID-19 infections are missed). However, it is not difficult to create a symptom checker that is 100% sensitive by simply returning every test as COVID-19–positive. While no checker does this 100% of the time, some checkers tend to declare every person who reports any flu-like symptom to be COVID-19–positive. This assesses every patient with allergic asthma (“shortness of breath”), heatstroke (“fever”), or heavy smoker (“cough”) to be COVID-19–positive. Therefore, we believe that a healthy balance between sensitivity and specificity is necessary for a useful checker. However, from the figure in this paper, readers can decide for themselves which balance between sensitivity and specificity is most useful and select the corresponding checker.

An additional aspect is that the developers of the 10 checkers may have had different purposes in mind during development. For example, they may have envisioned the checker to be a self-triage and recommendation tool or a likelihood predictor (as classified in [2]). In our study, we found that most checkers provide a certain likelihood as well as recommendations; therefore, classification is difficult. Therefore, we did not further subgroup the checkers in our analysis.

To our knowledge, this is the first scientific evaluation of web-based COVID-19 symptom checkers; however, there are a number of related studies evaluating symptom checkers. These include a study that evaluated 23 general-purpose symptom checkers based on 45 clinical case descriptions across a wide range of medical conditions and found that the correct diagnosis was listed among the top 20 results of the checkers in 58% of all cases on average [2]. The aforementioned study design was extended to five additional symptom checkers using ear, nose, and throat (ENT) cases, showing similar results [25]. Other evaluations include a study of symptom checkers used for knee pain cases; based on 527 patients and 26 knee problems, it was found that the physician’s diagnosis was present within the prediction list in 89% of the cases, while the specificity was only 27% [26]. In another study, an analysis of an automated self-assessment triage system for university students prior to an in-person consultation with a physician found that the system’s urgency rating agreed perfectly in only 39% of cases; meanwhile, for the remaining cases, the system tended to be more risk-averse than the physician [27]. Also, the applicability of web-based symptom checkers for 79 persons aged ≥50 years based on “think-aloud” protocols [28], deep learning algorithms for medical imaging [29], and services for urgent care [3] were evaluated.

The acceptability of the performance of a web-based symptom checker depends on the perspective and use of the results. In the case of COVID-19, a web-based assessment cannot fully replace a polymerase chain reaction (PCR) test, as some people are asymptomatic while others presenting with very specific COVID-19 symptoms may in fact have a very similar but different disease. Regardless, web-based COVID-19 symptom checkers can act as a first triage shield to avoid in-person physician visits or ease pressure on hospitals. Symptom checkers could even replace telephone triage lines in which non–medically trained personnel read a predefined sequence of questions. Although this was not part of this study, the authors believe that COVID-19 symptom checkers (if appropriately maintained and tested) may also be more reliable than the direct use of search engines such as Google or information via social media.

Strengths and Limitations

The strength of this study lies in the fact that it is based on a large number of real patients’ case descriptions from the literature (n=460) and a detailed evaluation of the best performing symptom checker in terms of F1 score (Multimedia Appendix 8). In contrast, a potential weakness of this study lies in its use of real literature-based cases, which may have biased the test set to rather severe cases of COVID-19 because mild and uninteresting cases are usually not found in the literature. We countered this bias by not including extreme edge cases from the literature in our 50 COVID-19 cases. A limitation of our study is that the benchmarking represents a specific point in time (April 2020; see Methods) and underlying algorithms may change. However, this temporal limitation is present in all benchmarking studies as knowledge increases and software is updated. Another bias may be that our control case descriptions do not report a COVID-19 contact, even though, for example, a person with a cold may have had a COVID-19 contact (and did not become infected). Another limitation of this study is the nonstraightforward mapping of the symptom checker outputs to risk levels (Multimedia Appendix 5). The interpretation of the textual output is debatable in some cases. We countered this by allowing three different risk levels and merging them in two different ways (see Figure 1A and Figure 1B). Also, every

symptom checker output was classified by multiple persons until consensus was reached.

Conclusion

Symptom checkers are being widely used in response to the global COVID-19 pandemic. As such, quality assessment of these tools is critical. We show that various web-based COVID-19 symptom checkers vary widely in their predictive capabilities, with some performing equivalently to random guessing while others show strength in sensitivity, specificity, or both.

Acknowledgments

This study received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 830017.

Conflicts of Interest

All authors are employees of Symptoma GmbH. JN holds shares in Symptoma.

Multimedia Appendix 1

Symptom frequencies used in Multimedia Appendix 2.
[PDF File (Adobe PDF File), 28 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Pseudocode of symptom frequencies based on vector distance (SF-DIST) and cosine similarity (SF-COS).
[PDF File (Adobe PDF File), 49 KB-Multimedia Appendix 2]

Multimedia Appendix 3

List of the COVID-19 cases.
[PDF File (Adobe PDF File), 50 KB-Multimedia Appendix 3]

Multimedia Appendix 4

List of COVID-19 symptoms according to the World Health Organization.
[PDF File (Adobe PDF File), 13 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Mappings between output texts and risk levels of the symptom checkers. All mappings were independently performed by two different persons, and conflicts were resolved by a third person's opinion.
[PDF File (Adobe PDF File), 33 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Full table of sensitivities, specificities, accuracies, F1 scores, and Matthews correlation coefficients for all symptom checkers (COVID-19–positive was defined by "high risk" for nonbinary symptom checkers).
[PDF File (Adobe PDF File), 24 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Full table of sensitivities, specificities, accuracies, F1 scores, and Matthews correlation coefficients for all symptom checkers (COVID-19–positive was defined by "medium risk or "high risk" for nonbinary symptom checkers).
[PDF File (Adobe PDF File), 24 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Constraining symptoms for Symptoma.
[PDF File (Adobe PDF File), 101 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Sensitivity vs specificity for all symptom checkers and Symptoma input constraint respectively by each symptom checker.
[PDF File (Adobe PDF File), 253 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Full table of sensitivities, specificities, accuracies, F1 scores, and Matthews correlation coefficients for Symptoma constrained by each symptom checker (COVID-19–positive was defined by “high risk” for nonbinary symptom checkers).
[PDF File (Adobe PDF File), 33 KB-Multimedia Appendix 10]

Multimedia Appendix 11

Full table of sensitivities, specificities, accuracies, F1 scores, and Matthews correlation coefficients for Symptoma constrained by each symptom checker (COVID-19–positive was defined by “medium risk” or “high risk” for nonbinary symptom checkers).
[PDF File (Adobe PDF File), 33 KB-Multimedia Appendix 11]

Multimedia Appendix 12

Pairwise comparisons between all symptom checkers and Symptoma based on the MCC only if the subset of symptoms used by one checker is also used for Symptoma.
[PDF File (Adobe PDF File), 79 KB-Multimedia Appendix 12]

References

1. Tasnim S, Hossain MM, Mazumder H. Impact of Rumors and Misinformation on COVID-19 in Social Media. J Prev Med Public Health 2020 May;53(3):171-174 [FREE Full text] [doi: 10.3961/jpmph.20.094] [Medline: 32498140]
2. Semigran H, Linder J, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 08;351:h3480. [doi: 10.1136/bmj.h3480] [Medline: 26157077]
3. Chambers D, Cantrell A, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review. Health Services and Delivery Research 2019;7(29):online. [doi: 10.3310/hsdr07290] [Medline: 31433612]
4. Chu DK, Akl EA, Duda S, Solo K, Yaacoub S, Schünemann HJ, COVID-19 Systematic Urgent Review Group Effort (SURGE) study authors. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. Lancet 2020 Jun 01:1973-1987 [FREE Full text] [doi: 10.1016/S0140-6736(20)31142-9] [Medline: 32497510]
5. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science 2020 May 22;368(6493):860-868 [FREE Full text] [doi: 10.1126/science.abb5793] [Medline: 32291278]
6. Lee B, Haidari L, Lee M. Modelling during an emergency: the 2009 H1N1 influenza pandemic. Clin Microbiol Infect 2013 Nov;19(11):1014-1022 [FREE Full text] [doi: 10.1111/1469-0691.12284] [Medline: 23800220]
7. Caley P, Philp DJ, McCracken K. Quantifying social distancing arising from pandemic influenza. J R Soc Interface 2008 Jun 06;5(23):631-639 [FREE Full text] [doi: 10.1098/rsif.2007.1197] [Medline: 17916550]
8. Ahmed F, Zviedrite N, Uzicanin A. Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. BMC Public Health 2018 Apr 18;18(1):518 [FREE Full text] [doi: 10.1186/s12889-018-5446-1] [Medline: 29669545]
9. Kristal R, Rowell M, Kress M, Keeley C, Jackson H, Piwnica-Worms K, et al. A Phone Call Away: New York's Hotline And Public Health In The Rapidly Changing COVID-19 Pandemic. Health Aff (Millwood) 2020 Aug 01;39(8):1431-1436. [doi: 10.1377/hlthaff.2020.00902] [Medline: 32525707]
10. Judson T, Odisho A, Neinstein A, Chao J, Williams A, Miller C, et al. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. J Am Med Inform Assoc 2020 Jun 01;27(6):860-866 [FREE Full text] [doi: 10.1093/jamia/ocaa051] [Medline: 32267928]
11. Ross C. I asked eight chatbots whether I had Covid-19. The answers ranged from ‘low’ risk to ‘start home isolation’. STAT. 2020 Mar 23. URL: https://www.statnews.com/2020/03/23/coronavirus-i-asked-eight-chatbots-whether-i-had-covid-19/ [accessed 2020-09-30]
12. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). World Health Organization. 2020 Feb. URL: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf [accessed 2020-09-30]
13. COVID-19 screener. ada. URL: https://ada.com/covid-19-screener/ [accessed 2020-04-09]
14. COVID-19 Screening Tool. Apple. URL: https://www.apple.com/covid19 [accessed 2020-04-09]
15. COVID-19 Symptom Checker. Babylon. URL: https://www.babylonhealth.com/ask-babylon-chat [accessed 2020-04-09]
16. Symptoms of Coronavirus. US Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html [accessed 2020-04-09]
17. Coronavirus Self-Checker. Cleveland Clinic. URL: http://covid19chat.clevelandclinic.org/ [accessed 2020-04-09]
18. Corona information. Webpage in German. Docyet. URL: https://corona.docyet.com/client/index.html [accessed 2020-04-09]
19. Symptomate. Infermedica. URL: https://symptomate.com/covid19/checkup/en/ [accessed 2020-04-09]

20. Coronavirus assessment tool. Providence. URL: https://coronavirus.providence.org/ [accessed 2020-04-09]
21. COVID-19 Chatbot Test. COVID-19. URL: https://www.symptoma.com/covid-19 [accessed 2020-04-09]
22. Your.MD. URL: https://webapp.your.md/login [accessed 2020-04-09]
23. BMJ Best Practice. URL: https://bestpractice.bmj.com/info/ [accessed 2020-04-09]
24. BMJ Case Reports. BMJ Journals. URL: https://casereports.bmj.com [accessed 2020-09-30]
25. Nateqi J, Lin S, Krobath H, Gruarin S, Lutz T, Dvorak T, et al. From symptom to diagnosis-symptom checkers re-evaluated: Are symptom checkers finally sufficient and accurate to use? An update from the ENT perspective.. HNO 2019 May 16;67(5):334-342. [doi: 10.1007/s00106-019-0666-y] [Medline: 30993374]
26. Bisson LJ, Komm JT, Bernas GA, Fineberg MS, Marzo JM, Rauh MA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. Am J Sports Med 2014 Oct;42(10):2371-2376. [doi: 10.1177/0363546514541654] [Medline: 25073597]
27. Poote AE, French DP, Dale J, Powell J. A study of automated self-assessment in a primary care student health centre setting. J Telemed Telecare 2014 Mar 18;20(3):123-127. [doi: 10.1177/1357633x14529246]
28. Luger TM, Houston TK, Suls J. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. J Med Internet Res 2014 Jan 16;16(1):e16 [FREE Full text] [doi: 10.2196/jmir.2924] [Medline: 24434479]
29. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689. [doi: 10.1136/bmj.m689] [Medline: 32213531]

Abbreviations

BMJ: British Medical Journal
ENT: ear, nose, and throat
MCC: Matthews correlation coefficient
SF-COS: symptom frequency based on cosine similarity
SF-DIST: symptom frequency based on vector distance
WHO: World Health Organization

Edited by T Rashid Soron, G Eysenbach; submitted 15.06.20; peer-reviewed by T Judson, E Bellei; comments to author 11.07.20; revised version received 27.07.20; accepted 14.09.20; published 06.10.20

Please cite as:
Munsch N, Martin A, Gruarin S, Nateqi J, Abdarahmane I, Weingartner-Ortner R, Knapp B
Diagnostic Accuracy of Web-Based COVID-19 Symptom Checkers: Comparison Study
J Med Internet Res 2020;22(10):e21299
URL: http://www.jmir.org/2020/10/e21299/
doi: 10.2196/21299
PMID: 33001828