Triage and diagnostic accuracy of Online Symptom Checkers: a systematic review

Eva Riboli-Sasco, Austen El-Osta, Aos Alaa, Iman Webber, Manisha Karki, Marie Line El Asmar, Katie Purohit, Annabelle Painter, Benedict Hayhoe

Submitted to: Journal of Medical Internet Research
on: October 25, 2022

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Triage and diagnostic accuracy of Online Symptom Checkers: a systematic review

Eva Riboli-Sasco[1] MA; Austen El-Osta[1] MSc, MPA, PhD; Aos Alaa[1] MPH; Iman Webber[1] BSc; Manisha Karki[1] MPH; Marie Line El Asmar[1] MPH, MD; Katie Purohit[1] BCHIR; Annabelle Painter[1] MA, BMed, BCHIR; Benedict Hayhoe[1] MD

[1]Self-Care Academic Research Unit (SCARU) Department of Primary Care and Public Health Imperial College London London GB

Corresponding Author:
Eva Riboli-Sasco MA
Self-Care Academic Research Unit (SCARU)
Department of Primary Care and Public Health
Imperial College London
323 Reynolds Building Charing Cross Hospital
London
GB

Abstract

Background: In the context of a deepening global shortage of health workers, and particularly the COVID-19 pandemic, there is growing international interest in and use of online symptom checkers (OSCs). However, the evidence surrounding the safety and accuracy of OSCs remains inconclusive so far. The triage and diagnostic accuracy of these tools is an essential aspect that needs to be addressed before pushing any further implementation.

Objective: This systematic review aimed to summarize the existing peer-reviewed literature evaluating the triage accuracy (directing users to appropriate services based on their presenting symptoms) and diagnostic accuracy of OSCs aimed at lay users for general health concerns.

Methods: Searches were conducted in Medline, Embase, CINAHL, HMIC and Web of Science. We included peer-reviewed studies published in English between 1 January 2010 and 17 February 2022 with a quantitative assessment of triage and/or diagnostic accuracy of OSCs directed at lay users. We excluded tools supporting health professionals, and disease- or speciality-specific OSCs. Screening and data extraction were carried out independently by two reviewers for each study. We performed a descriptive narrative synthesis.

Results: 21,284 studies were screened and 15 were included. Six studies reported on both triage and diagnostic accuracy, eight focused on triage accuracy, and one on diagnostic accuracy. Diagnostic and triage accuracy varied between studies and OSCs; most studies showed suboptimal diagnostic and triage accuracy. Frequency and urgency of the condition were the main variables that affected the levels of diagnostic and triage accuracy, along with specific features of the OSCs. The impact of each variable differed across tools and studies, making it difficult to draw any solid conclusions. Included studies had either a moderate or high risk of bias according to the revised tool for the Quality Assessment of Diagnostic Accuracy Studies 2.

Conclusions: While OSCs have significant potential to provide accessible and accurate health advice and triage recommendations to users, more research is needed to validate their triage and diagnostic accuracy prior to wide scale adoption in community and healthcare settings. Future studies should aim to use a common methodology and/or agreed standard for evaluation to facilitate objective benchmarking and validation.

(JMIR Preprints 25/10/2022:43803)
DOI: https://doi.org/10.2196/preprints.43803

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

Review

Triage and diagnostic accuracy of Online Symptom Checkers: a systematic review

Eva Riboli-Sasco[1], Austen El-Osta[1], Aos Alaa[1], Iman Webber[1], Manisha Karki[1], Marie Line El Asmar[1], Katie Purohit[1], Annabelle Painter[1], Benedict Hayhoe[1]

[1] Self-Care Academic Research Unit (SCARU), Department of Primary Care and Public Health, School of Public Health, Imperial College London, The Reynolds Building, St Dunstan's Road, London W6 8RP

*Corresponding author:
Eva Riboli-Sasco, MA
Self-Care Academic Research Unit (SCARU)
Department of Primary Care and Public Health
Imperial College London
323 Reynolds Building
Charing Cross Hospital
London W6 8RF
United Kingdom
Phone: + 44 207 594 7604
Email: e.riboli-sasco@imperial.ac.uk

Abstract

Background: In the context of a deepening global shortage of health workers, and particularly the COVID-19 pandemic, there is growing international interest in and use of online symptom checkers (OSCs). However, the evidence surrounding the safety and accuracy of OSCs remains inconclusive so far. The triage and diagnostic accuracy of these tools is an essential aspect that needs to be addressed before pushing any further implementation.

Objective: This systematic review aimed to summarize the existing peer-reviewed literature evaluating the triage accuracy (directing users to appropriate services based on their presenting symptoms) and diagnostic accuracy of OSCs aimed at lay users for general health concerns.

Methods: Searches were conducted in Medline, Embase, CINAHL, HMIC and Web of Science. We included peer-reviewed studies published in English between 1 January 2010 and 17 February 2022 with a quantitative assessment of triage and/or diagnostic accuracy of OSCs directed at lay users. We excluded tools supporting health professionals, and disease- or speciality-specific OSCs. Screening and data extraction were carried out independently by two reviewers for each study. We performed a descriptive narrative synthesis.

Results: 21,284 studies were screened and 15 were included. Six studies reported on both triage and diagnostic accuracy, eight focused on triage accuracy, and one on diagnostic accuracy. Diagnostic and triage accuracy varied between studies and OSCs; most studies showed suboptimal diagnostic and triage accuracy. Frequency and urgency of the condition were the main variables that affected the levels of diagnostic and triage accuracy, along with specific features of the OSCs. The impact of each variable differed across tools and studies, making it difficult to draw any solid conclusions. Included studies had either a moderate or high risk of bias according to the revised tool for the Quality Assessment of Diagnostic Accuracy Studies 2.

Conclusions: While OSCs have significant potential to provide accessible and accurate health advice and triage recommendations to users, more research is needed to validate their triage and diagnostic accuracy prior to wide scale adoption in community and healthcare settings. Future studies should aim to use a common methodology and/or agreed standard for evaluation to facilitate objective benchmarking and validation.

Keywords: Systematic review; digital triage; diagnosis; online symptom checker; safety; accuracy

Introduction

The global shortage of health workers anticipated by the World Health Organization (WHO) is expected to increase from 7.2 million in 2013 to 12.9 million by 2035 [1]. Online symptom checkers (OSCs) have been promoted as a way of saving time and resources for patients while reducing anxiety and allowing users to take more ownership of their health[2] and for health professionals and services by promoting rational use of healthcare services[3], including self-care.

The use of OSCs has exploded in recent years. In the UK, the NHS 111 online service which registered 2 million contacts during 2019, reached 7.5 million visits during the first ten months of 2020, mainly as a consequence of the COVID-19 pandemic [4]. OSCs can be accessed online using a computer, tablet or smartphone, via a website or a smartphone app. Based on responses to a series of questions, OSCs may suggest a possible diagnosis, and/or a triage recommendation to inform the next steps [5]. The triage function guides users on whether they should seek a healthcare assessment, the setting (e.g. emergency department (ED), GP surgery) and the degree of urgency (e.g., immediately, within a few days, or weeks) [6].

The potential benefits of OSCs, whether individual or collective, depend primarily on their safety and accuracy. If inadequately designed, they could misdiagnose and/or misdirect users potentially diverting the user from seeking adequate care or conversely placing additional strain on health systems. Two systematic reviews assessed the literature evaluating OSCs [7, 8] with mostly weak evidence regarding their diagnostic and triage accuracy. One review focused only on urgent health issues, while the other included speciality-specific OSCs, and both were outdated following the recent publication of several eligible studies.

This systematic review aims to update and summarise the peer-reviewed literature evaluating the triage accuracy (defined as directing users to appropriate services based on their presenting symptoms) and diagnostic accuracy of OSCs aimed at lay users for general health concerns.

Methods

This systematic review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [9] (see **Multimedia Appendix 1**).

Eligibility criteria

Any online or digital service designed to provide users with a potential diagnosis and/or promote the rational use of health services (including self-care) among the general population based on self-reported symptoms was accepted as the intervention. We included OSCs readily accessible online and via app providers, as well as prototypes tested in various settings (including general practices and ED). Tools that only provided an asynchronous online consultation (e.g., via email), health advice without diagnosis or triage, and those that were disease- or speciality-specific were excluded. The reference standard (or main comparator) could be telephone or face-to-face consultation with a GP or nurse, or a diagnosis and/or triage attached to a vignette. The primary outcomes of interest were OSC triage and diagnostic accuracy. We reported any additional outcome assessed in included studies, such as coverage or estimated impact on service use.

We included peer-reviewed articles written in English and published between 1 January 2010 and 17 February 2022. Studies could be observational, non-randomised control trials or randomised control trials (RCTs). We excluded dissertations, conference proceedings and non-peer-reviewed manuscripts. Included papers had to provide quantitative data on diagnostic and or triage accuracy of OSCs based on an appropriate reference standard (e.g., telephone triage or face-to-face consultation with a medical professional, or diagnosis attached to the vignette).

Search strategy

A scoping review was conducted after consulting with a research librarian to help establish search terms. An initial list of search terms was compiled and applied to MEDLINE and Embase to confirm the relevance of the results. Reference lists from several relevant studies and similar reviews were manually searched to expand the search terms and refine the search strategies. Subject headings were adapted for each database. Searches were carried out on 17 February 2022 [searching for studies published between 1 January 2010 and 16 February 2022]. We searched the following five databases: Medline, Embase, CINAHL, Health Management Information Consortium and Web of Science. No manual searching was performed, but we screened the references of all studies selected for full-text screening. The final list of search terms for each database is presented in **Multimedia Appendix 2.**

Study Selection

The studies retrieved were first imported into Endnote X7 to help identify and remove duplicates. Included studies were then entered in Covidence, where additional duplicates were removed. Titles and abstracts were screened by two reviewers. The full text of potentially eligible studies was then independently assessed by two reviewers. Studies, where the primary reviewers disagreed, were reviewed independently by a third researcher; any remaining disagreement was resolved through team discussion.

Data extraction

Following full-text screening, data extraction was carried out by two reviewers independently for each study using a comprehensive, standardised extraction form designed for the specific characteristics of this review and refined following the testing of two studies. Key areas of data collection were the study sample size and characteristics; reference standard, measures and levels of triage and diagnostic accuracy, and any additional comparator and reported outcomes. The detailed data extraction tables can be shared upon request.

Risk of bias and applicability

Two reviewers independently assessed the risk of bias and applicability concerns using a revised version of the quality assessment of diagnostic accuracy studies 2 (QUADAS-2) tool [10] for the domains of patient selection, performance of the index test, performance of the reference test, and flow and timing (for risk of bias only). Conflicts were resolved through discussion. No study was excluded based on quality assessment. We also assessed the overall strength of evidence (quality and relevance) for both main outcomes using the method described by Chambers et al. [8] in their review. This involved classifying evidence based on study numbers, design and levels of consistency between findings.

Analysis

We performed a descriptive narrative synthesis, and strength of evidence assessment, structured around the prespecified research questions and outcomes to describe the collective findings of the included studies. Wide variation in design and methodology made meta-analyses impractical.

Results

A total of 21,284 records were identified through initial searches, with an additional 11 studies identified through citation searching. 15 studies were included in the review. The detailed PRISMA flow diagram can be found in **Multimedia Appendix 3**.

Characteristics of included studies

Table 1 shows the main characteristics of included studies published between 2014 and 2022. Six [11-16] were conducted by researchers based in the USA, three in the UK [17-19], two in Australia [20, 21], two in Canada [22, 23], one in the Netherlands [24] and one in Hong Kong [25]. Eight studies [11, 12, 14-16, 19-21] used standardised patient vignettes; several were inspired by or included the 45 vignettes used by Semigran et al. [11, 12] The remaining seven studies used data from real patients either through their medical health records [13, 25] or direct input by users [17, 18, 22-24] in different settings, including primary care and emergent care settings. Population size ranged from 45 to 25,333 patients.

Seven studies evaluated a single OSC [14, 16-18, 22-24], while the other eight tested and compared the performance of two [25] to 36 [20] OSCs. Where provided, the most common justifications for selection were language (English), level of popularity among users, and accessibility (free). The most frequently included OSCs were WebMD, Isabel and Symptomate (included in six studies), followed by Drugs.com, Symcat and FamilyDoctor tested in five of the included studies. The complete list of tested OSCs is presented in **Multimedia Appendix 4**, along with measurements used to assess their diagnostic and triage accuracy.

Table 1: Main characteristics of included studies

Reference	Study design	Nb of OSCs	Population / Sample	Reference standard	Add. Comparator
Poote et al. 2014, UK [17]	Prospective cohort study	1	154 patients from a PC student health centre 17 to 43 y. (mean age: 22) 64.3% F, 35.7% M	7 GPs through F2F consultation	
Semigran et al. 2015, USA [11]	Vignette cohort study	23	45 standardised patient vignettes 4 m to 77 yo 38% F, 62% M	Diagnosis & triage attributed to the vignettes	
Semigran et al. 2016, USA [12]	Vignette cohort study	23	Same as Semigran 2015	Diagnostic attributed to the vignettes	234 GPs through Human Dx platform
Verzantvoort et al. 2018, NL [24]	Prospective, cross-sectional cohort study	1	126 app users 52% F, 48% M	Telephone triage by a nurse	
Berry et al. 2019, USA [13]	Retrospective cohort study	5	168 ED patient records with prior diagnosis of HIV and/ or Hep C 44.9 ± 12.3 yo 36.9% F, 63.1% M 38% Black, 62% White	Diagnosis attributed by ED physician through F2F consultation & triage all deemed emergent as patients presented to ED	
Gilbert et al. 2020, USA [14]	Vignette cohort study	8	200 standardised patient vignettes 1 m to 89 yo M 43% F 57%	Diagnosis & triage attributed to the vignettes	7 GPs through phone consultation; Gold standard set by 2 panels of 3 GPs for diagnosis & triage
Hill et al. 2020, Australia [20]	Vignette cohort study	36	48 standardised patient vignettes (incl. 30 adapted from Semigran 2015) 4 w to 77 yo 43.75% F, 56.25% M	Diagnosis & triage attributed to the vignettes & confirmed by 2 GPs & 1 ED specialist	
Yu et al. 2020, Hong Kong [25]	Retrospective cohort study	2	149 real A&E patient charts; Drugscom (55.6 yo, 58% F, 42%, M) Family Doctor (55.4 yo; 55% F, 45% M)	Triage categories assigned by the triage nurses using A&E Department triage protocols	
Ceney et al. 2021, UK [19]	Vignette cohort study	12	50 standardised patient vignettes, (incl. 44 from Semigran 2015 + additional 6 to account for depression or Covid-19)	Diagnosis attributed to the vignettes & triage recommendation according to NICE guidance	
Chan et al. 2021, Canada [23]	Prospective cohort study	1	581 patients (281 ED patients + 300 primary care patients)	Triage by GP through F2F consultation & reviewed by 2	

			ED patients: 38±16 (16 to 91 yo) PC patients: 48±18 (16 to 91 yo) 63% F, 37% M	physician authors who, by consensus, assigned a corresponding triage recommendation	
Delshad et al. 2021, USA [16]	Vignette cohort study	1	50 standardised patient vignettes 20 to 84 yo 50% F, 50% M	3 consensus triages attributed to the vignettes by a total of 14 GPs	14 individual GPs' triage decision
Gilbert et al. 2021, Australia [21]	Vignette cohort study	1	Same as Hill 2020	Same as Hill 2020 + 1 clinician (with GP & ED experience) decided if it matched	
Schmieding et al. 2021, USA [15]	Vignette cohort study	15	Same as Semigran 2015	Triage attributed to the vignettes	91 USA based adults without professional medical background
Trivedi et al. 2021, Canada [22]	Prospective observational study	1	429 patients mean age 47 yo 50.2% F, 49.8% M	CTAS scores assigned F2F by the dedicated ED triage nurse	
Dickson et al. 2022, UK [18]	Retrospective cohort study	1	25,333 patients 46 yo (30 to 62) 54.2% F 45.8% M	MTS triage categories assigned F2F by a triage nurse	

A&E: Accident & Emergency
ED: Emergency Department
GP: General Practitioner
HIV: Human Immunodeficiency Virus
NICE: National Institute for Health and Care Excellence
m: months / **w:** weeks / **yo:** years old

CTAS: Canadian Triage & Acuity Scale
F2F: face-to-face
Hep C: Hepatitis C
MTS: Manchester Triage System
PC: Primary Care

Diagnostic accuracy

The diagnostic accuracy of the tested OSC was reported in 7 out of 15 of the included studies [11-14, 19-21]. Significant variability in levels of diagnostic accuracy of OSCs was observed between individual OSCs and studies, but it was deemed to be low overall, and on average lower than that of GPs when compared [12, 14]. **Table 2** presents the levels and range of average diagnostic accuracy, defined as listing the correct diagnosis first, as well as the main variables assessed by each study.

There was agreement regarding the general impact of condition frequency with a better average diagnostic accuracy observed for 'common' compared to 'uncommon' conditions in two studies [11, 20] – but findings were conflicting regarding the influence of condition urgency on diagnostic accuracy [11, 19, 20]. Hill et al., [20] also found that the 8 OSCs using artificial intelligence (AI) algorithms had a better diagnostic accuracy overall: they listed the correct diagnosis first for 46% (95% CI, 40-57%) of the vignettes compared to only 32% (95% CI, 26-38%) for the 19 other tested OSCs. However, these authors noted that "information about whether programs employed AI algorithms was drawn solely from that provided in the [O]SC" which is problematic since definitions of AI and algorithms may vary between studies and OSCs, with some authors restricting AI to machine learning methods only while others included Bayesian methods or even simple rules-based algorithms. Finally, the source of the OSC, namely the App Store or Google Play, was found to impact diagnostic accuracy in one instance [20].

Table 2: Levels of average diagnostic accuracy (ADA) & main variables identified

Reference	OSCs	OSCs ADA (listed 1st)	OSCs range of ADA	Main variables identified	Additional comparator's ADA
Semigran et al. 2015, USA [11]	23	34% (95% CI, 31-37%)	5% (MEDoctor) to 50% (DocResponse)	• Urgency ↓ • Frequency ↑ • Demographic data ≈ • Max nb of diagnosis ≈ • Distributor ≈ • Nurse triage protocol ≈	
Semigran et al. 2016, USA [12]	23	34% (95% CI, 31-37%)	5% (MEDoctor) to 50% (DocResponse)		72.1% (95% CI, 69.5-74.8%) (234 GPs on Human Dx platform)
Berry et al. 2019, USA [13]	5	NS	3% (WebMD) to 16.4% (Symcat)		
Gilbert et al. 2020, USA [14]	8	26,1±8.9	18% (Symptomate) to 48% (Ada)	• NHS vignettes (based on transcripts of real calls made to NHS Direct) ↓	71.2±5.6 (7 GPs through phone consultation)
Hill et al. 2020, Australia [20]	36	36% (95% CI, 31–42%)	12% (ePain Assist) to 61% (Symptomate)	• Urgency ⇅ • Frequency ↑ • AI ↑ • Demographic data ↑ • Max number of diagnoses provided ≈ • Apple vs Google ⇅	
Ceney et al. 2021, UK [19]	9 (out of 12)	37.7% (95% CI 33.6–41.7%)	22.2% (CAIDR) to 72.0% (Ada)	• Urgency ↓ • Nb of questions ↑ • Time to complete ↑	
Gilbert et al. 2021, Australia [21]	1 (Ada)	65%		• Australian specific vignettes ↓	

↓ decreases ADA ↑ increases ADA

⇅ mixed impact on ADA ≈ no significant impact on ADA

NS not stated

Triage accuracy

With the exception of one study [12], all others reported on the selected OSCs' triage accuracy which appeared suboptimal overall. Levels of average triage accuracy are presented in **table 2**. A triage was deemed accurate only when it matched the one attributed by one or more clinicians as the "gold standard". In one study however, all cases were "expected to be mostly emergency" since they were records of patients presenting to ED [13]. This was surprising since triage advice, ie whether and where users should seek a healthcare assessment for their presenting symptoms is precisely one of the main functions of OSCs. In addition, as Chan et al. 2021 [23] included in their review and as others have shown [26], patients deciding to present to ED does not automatically qualify them as requiring emergency treatment, thus undermining the pertinence of Berry et al. 2019 [13] findings regarding triage accuracy.

Triage accuracy appeared to be affected by the level of urgency of the condition as shown in six of the included studies [11, 15, 19, 20, 22, 25]. All but two studies [22, 25] found that triage accuracy increased with the urgency of the condition. Results regarding the frequency of the condition were more conflicting depending on the studies and OSCs. According to Hill et at., [20], the accuracy of the five OSCs requiring demographic data (defined as requesting "at least age and sex") was on average greater than for the 14 studies that did not. In Semigran et al. [11], OSCs that used the Schmitt or Thompson Nurse Triage Protocols were more likely to provide appropriate triage decisions. Finally, several studies found that some OSCs (including iTriage, Symcat, Everyday Health, Doctor Diagnose, Symptomate, and Isabel) never recommended 'self-care' and therefore could not match this triage category.

Specific characteristics of the study population may also affect the levels of triage accuracy of the OSCs. Berry et al [13] found that a significantly higher percentage of hepatitis C patients received a "correct diagnosis" than HIV patients, both remaining low, however, thus concluding that current OSCs software algorithms may not account for the complex, immunocompromised HIV and hepatitis C patient populations. Only two studies [17, 22] looked at the impact of users' age and gender on triage accuracy and found diverging results. Finally, methodological choices relating to the type or source of the vignettes also affected diagnostic accuracy: for example vignettes made up by researchers versus vignettes based on transcripts of real calls made to NHS Direct [14] or Australian specific vignettes [21].

Table 3: Levels of average triage accuracy (ATA) and main variables identified

Reference	OSCs	OSCs ATA	OSCs range of ATA	Main variables identified	Add. Comparator's ATA
Poote et al. 2014, UK [17]	1	39%		• Age ≈ • Gender ≈	
Semigran et al. 2015, USA [11]	15 (of 23)	57% (95% CI, 52-61%)	33% (iTriage) to 78% (HMS Family Health Guide)	• Urgency ↑ • Frequency ↓ • Schmitt or Thompson nurse triage protocols ↑	
Verzantvoort et al. 2018, NL [24]	1	81%			
Berry et al. 2019, USA [13]	5	45.8%	NS	• Hep C > HIV	
Gilbert et al. 2020, USA [14]	8	90.1±7.4	80% (Buoy) to 97.8% (Symptomate)	• NHS vignettes (based on transcripts of real calls made to NHS Direct) ↓	97.0%±2.5; (7 GPs through phone consultation)
Hill et al. 2020, Australia [20]	19 (of 36)	49% (95% CI, 44–54%)	17% (Doctor Diagnose) to 61% (Healthdirect)	• Urgency ↑ • Frequency ↑ • Demog data ↑ • AI algorithm ≈ • Max nb of diagnosis provided ≈	
Yu et al. 2020, Hong Kong [25]	2	62%	50% (FamilyDoctor) to 74% (Drugs.com)	• Urgency ↑	
Ceney et al. 2021, UK [19]	10 (of 12)	57.7% (95% CI 53.2–62.2%)	35.6% (CAIDR) to 90.0% (Doctorlink)	• Urgency ↑ • Nb of questions ≈	
Chan et al. 2021, Canada [23]	1	73%			58%; (Patients decision)
Delshad et al. 2021, USA [16]	1	Consensus A: 85% Consensus B: 92% Consensus C: 88%			CA: 82% CB: 69% CC: 80% (14 individual GPs' triage)
Gilbert et al. 2021, Australia [21]	1	63%		• Australian specific vignettes ↓	
Schmieding et al. 2021, USA [15]	15	58.0%; SD 12.8%	32% (iTriage) to 80% (HMS Family Health Guide)	• Urgency ↑	60.9%; SD 6.8% (Lay participants)
Trivedi et al. 2021, Canada [22]	1	27%		• Urgency ↑↓ • Gender: W > M • Age: 20 to 39 y.o.highest • cardiorespiratory problems ↑	
Dickson et al. 2022, UK [18]	1	30.7%			

↓ decreases ADA ↑ increases ADA

↑↓ mixed impact on ADA ≈ no significant impact on ADA **NS** not stated

Additional reported outcomes

Most studies reported on additional outcomes; 10 studies assessed under- and over-triage by OSCs [11, 14, 15, 17, 19, 20, 22-25]. Six studies [11, 15, 17, 22-24] found that OSCs tend to over-triage (i.e. be risk averse), which is defined as encouraging users to seek care in a setting or with a degree of urgency that is not strictly necessary for the presenting symptoms. Over-triage is likely due to concerns about patient safety and product liability. However, most authors observed that under-triage did occur. Yu et al., [25] found that Drugs.com and FamilyDoctor under-triaged 24% (95% CI, 16–34%) and 45% (95% CI, 35–55%) of cases respectively. Chan et al. [23] estimated that compliance with the triage recommendations in their cohort could have reduced hospital visits by 55%, but would also cause potential harm in 2–3% of cases from delayed care. Ceney et al. [19] found that all 12 OSCs tested led to additional resource utilisation, ranging between 12.5% (95% CI 6.1–33.5%) for the lowest impact symptom checker and 87.5% (95% CI 52.8%-100%) for the highest. It is pertinent that such estimates are based on the assumption that users follow the advice provided by the OSC, which none of the included studies assessed. Verzantvoort et al., [24] reported that only 65% of users intended to follow the OSC tool advice. Gilbert et al. [14], reported on each OSC's coverage, comprehensiveness and relevance. Dickson et al., [18] reported that the median time to nurse triage was 17 min (IQR 9–31) compared to 5 min (IQR 4–6) for eTriage.

Risk of bias within studies

The evaluation of the risk of bias and applicability was conducted using the amended QUADAS–2 tool and the results are summarised in **table 4**. This assessment revealed that all studies had at least one area with unclear risk of bias and six had a high risk of bias. For instance, Yu et al. [25] replaced cases with chief complaints not available on the OSCs with more compatible ones, which according to the authors, likely resulted in overestimated OSCs' accuracy levels. Dickson et al [18] acknowledged the possibility of selection bias due to the perceptions of reception staff around the ability of older patients to use the OSC, which resulted in its reduced use by patients above 70 years old. In the study by Poote et al. [17] the GP assessing the patients' conditions had access to the index test results, which means the reference standard was not blinded to index test results. In Hill et al. [20], the lack of data regarding the blinding of the inputters to the diagnostic/triage, as well as their familiarity with the system, introduced a risk of bias regarding the conduct of the index test. The affiliation of authors is another source of bias as several of the included studies were conducted by authors working for OSC developers. For example, all but one of the authors of the 2021 study led by Gilbert worked for the tested app, Ada [21].

Overall strength of evidence assessment

The overall strength of evidence for key outcomes is summarised in **table 5**. While there is strong evidence that the diagnostic accuracy of OSCs tends to be lower than that of health professionals, evidence is more inconsistent regarding triage accuracy.

Table 4: Risk of bias summary using QUADAS-2 risk assessment tool

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	Patient Selection	Index Test	Reference Standard	Flow & Timing	Patient Selection	Index Test	Reference Standard
Poote 2014	🟢	🟢	🔴	🟡	🟡	🟢	🟡
Semigran 2015	🟢	🟢	🟡	🟢	🟡	🟢	🟡
Semigran 2016	🟡	🟢	🟡	🟡	🟡	🟢	🟢
Verzantvoort 2018	🟡	🟢	🟡	🔴	🟢	🟢	🟢
Berry 2019	🟢	🟡	🔴	🟡	🟢	🟡	🟢
Gilbert 2020	🟢	🟡	🟢	🟢	🟢	🟡	🟢
Hill 2020	🟢	🔴	🟢	🟡	🟢	🟡	🟢
Yu 2020	🔴	🟢	🟡	🟡	🟢	🟢	🟢
Ceney 2021	🟢	🟢	🟢	🟡	🟢	🟢	🟢
Chan 2021	🟡	🟢	🟢	🟡	🟢	🟢	🟢
Delshad 2021	🟡	🟡	🟢	🟡	🟢	🟢	🟢
Gilbert 2021	🟢	🟡	🟡	🟢	🟢	🟢	🟢
Schmieding 2021	🟢	🟢	🟡	🟢	🟢	🟢	🟡
Trivedi 2021	🟡	🟡	🟢	🟡	🟡	🟡	🟢
Dickson 2022	🔴	🟢	🟢	🟡	🟡	🟢	🟢

🟢 Low Risk 🔴 High Risk 🟡 Unclear Risk

Table 5: Overall strength of evidence by main outcome

Outcome	Relevant studies	Evidence statement	Strength of evidence
Diagnostic accuracy	-Semigran et al. [11] -Semigran et al. [12] -Berry et al. [13] -Gilbert et al. [14] -Hill et al. [20] -Ceney et al. [19] -Gilbert et al. [21]	Overall diagnostic accuracy was deemed to be low, and always lower than the comparator	Strong
Triage accuracy	-Poote et al. [17] ±Semigran et al. [11] -Verzantvoort et al. [24] -Berry et al. [13] ±Gilbert et al. [14] -Hill et al. [20] -Yu et al. [25] -Ceney et al. [19] +Chan et al. [23] +Delshad et al. [16] -Gilbert et al. [21] ±Schmieding et al. [15]	Inconsistent findings, including within studies. Great variations between OSCs. Usually lower than GPs and even lay persons, but with exceptions.	Inconsistent

	-Trivedi et al. [22] -Dickson et al. [18]		

= no significant difference in outcomes **+** better outcome with OSC **-** worst outcome with OSC
± varying results within study **?** results difficult to interpret in comparative terms.

Discussion

Principal Results

Evidence on the triage and diagnostic accuracy of OSCs suggests they are currently not a viable replacement for other triage and diagnostic options such as telephone triage or in-person consultations. Further, some OSCs performed well on triage but poorly on diagnostic accuracy and vice versa. Studies evaluating various tools also revealed important performance variations between different OSCs. Several studies found that the condition's frequency and urgency could affect diagnostic and/or triage accuracy levels, but with mixed conclusions. In addition, some specific OSC characteristics may also play a role, including the use of AI, self-reported demographic and anthropomorphic data, the maximum number of diagnoses provided or the use of nurse triage protocols. Some characteristics of the "study population" were also shown to impact the level of triage and/or diagnostic accuracy, including the source of the vignettes, but also the health status of patients or the geographical specificity of diseases and symptoms. The safety of the triage recommendation as well as the tendency to over or under triage were important outcomes associated to triage accuracy. These also resulted in some studies estimating the potential impact on service utilisation, which diverged between studies, partly because some tools promoted over utilisation of services whereas others tended to under-triage users.

Strengths and limitations

We conducted a comprehensive search by repeatedly adapting and reviewing our search strategy and search terms, including manually searching reference lists. Highly inclusive searches yielded a significant number of initial results, which we screened in pairs to limit errors. However, we acknowledge that eligible studies might have been excluded or omitted and that relevant papers in grey literature or papers written in languages other than English, or prior to 2010, might also have been excluded due to our selection criteria. Included studies were all conducted in high-income countries, which may limit wider generalizability of findings. Comparison between studies was particularly difficult due to the variety of study designs, outcome measures, populations and tools considered. Additionally, four studies evaluated more than 10 OSCs, adding to the complexity of comparisons. Triage accuracy, which consistently appeared as the main outcome of interest across studies, was measured using varying numbers of categories, different time periods and triage locations, thus limiting further the possibility for objective head-to-head comparisons. The lack of a common methodology for evaluating OSCs strongly limits the possibility of comparison between tools and studies. It is pertinent also that all 15 studies had at least one area with an unclear risk of bias and six studies had a high risk of bias.

Comparison With Prior Work

Two previous systematic reviews assessed the literature on a similar topic. The 2019 systematic review by Chambers et al. [8] included any type of publication, including grey literature, but was limited to studies relating to urgent health issues only. The evidence was assessed as being mostly weak and insufficient to determine the level of safety of digital and online symptom checkers for patients. More recently, Wallace et al. [7] published a systematic review on the diagnostic and triage accuracy of OSCs, including speciality-specific tools but searching only Medline and Web of Science up to 15 February 2021. Both triage and diagnostic accuracy of OSCs were found to be mostly low despite variations. Reliance on these tools was therefore considered as posing a potential clinical risk. The identification of 7 new studies published since mid-February 2021, along with an increasing use of OSCs following the COVID-19 pandemic despite cautionary calls, motivated the conduct of this review.

This review aimed not only to update, but also to strengthen, the quality of the evidence by including only peer-reviewed papers and focusing on OSCs for general health concerns (non-speciality specific). However, the evidence remains inconsistent and calls both for caution in promoting OSCs as well as the need for further studies to improve and inform future development of these tools.

Implications for Research and Practice

Most included studies highlighted that OCS performance tended to remain low and that further improvements, testing and research are needed. While there has been a sense in commentaries and previous studies [8] that OSCs tend to over-triage and thus be considered 'risk averse', our review identified several instances of 'under-triage' amongst OSCs. This finding is concerning because it suggests a risk of delay in accessing care for individuals using these decision support tools. The impact of over-triage on health services must also be considered as this might negatively impact the quality of services provided and thus ultimately represent a risk for service users. Further work is urgently needed to understand the extent and implications of inappropriate triage recommendations of existing publicly available OSCs, which require an assessment of rates of user compliance with the tool's advice.

Four included studies offered suggestions for improvement of OSCs, including incorporating local, regional and/or seasonal epidemiological data and individual clinical data [11, 25], and a more efficient inclusion of demographic data into the algorithm [11]. Authors also suggested alternative uses of current OSCs, such as tracking epidemiological data, self-education of users about their health, improving patient-physician relationships, directing users to appropriate care [20] (especially for tools that are directly linked with health care services), and in supporting the use of AI-based symptom assessment technology in diagnostic decision support for GPs [14].

More studies are needed to clearly assess the triage and diagnostic accuracy of OSCs for all potential users. The lack of consensus on how OSCs should be evaluated by any national/international regulatory body means that developers produce their own evidence to validate products to meet regulatory requirements (UKCA/CE marking).

There is a need for additional research into the methods of evaluating OSCs, including how to establish a gold standard response and determine appropriate accuracy and safety scores in comparison to this gold standard. A consensus agreement on what could be deemed an "acceptable" rate of under or over-triage would also be required. Specific evidence standards should be provided for OSCs to augment existing guidance, such as the NICE evidence standards framework and the evaluation requirements for medical device certification with the MHRA. A set of congruent requirements for standardised vignette-based clinical evaluation process of OSCs has been proposed with this aim [27].

Future studies should ideally be based on the direct input of real-life patients, who would be best placed to enter their own symptoms into the OSCs to allow a better assessment of real-world performances, instead of mostly fictional clinician-authored vignettes or medical records, drafted and entered by researchers, who are likely to be prone to bias. In addition, the study populations should be broad and diverse in terms of race, age, gender, social class, education, and abilities, since these characteristics have been correlated with differential and possibly discriminatory treatment by a healthcare professional (HCP) in real life encounters in multiple countries and settings [28-31]. For several communities and individuals, including ethnic minorities, migrants, women, gender non-conforming and LGBTQ communities [32], the use of an OSC might potentially represent a safer, more accessible and/or more accurate option than a real-life encounter with a HCP. However, if these communities are not included and accounted for in the design and testing of digital technology, including OSCs, such discriminations might be further reinforced [33]. Achieving health equity requires a shift in methodologies and perspectives, including the adoption of a feminist intersectional lens in digital health [34]. Finally, while OSCs may be perceived as useful [35], there may also be issues in understanding and interpreting the recommendations provided [36], making accessibility, usability and interpretability key factors to consider when designing, promoting and evaluating these tools.

In response to the limitations inherent in current evaluations of OSCs, several authors have called for a multistage process evaluation of increasing exposure to real-life clinical environments in proportion to OSC system maturity, taking place both before and after the tools launch, and including the testing of different aspects of the OSC such as usability, effectiveness and safety [37-42].

Conclusion

OSCs have a significant potential to provide accessible and accurate health advice and triage recommendations to patients. If clinical safety is assured through reproducible evidence of diagnostic and triage accuracy, OSCs could have a valuable place in a sustainable health system, with the potential to support individuals to self-care more regularly for self-limiting conditions, whilst also directing them to appropriate healthcare assessment when needed. This arrangement could also help rationalise the use of healthcare products and services and reduce unnecessary pressure on HCPs and health systems in a variety of settings. Our review highlighted inconsistent evidence across the included studies regarding the triage and diagnostic accuracy of OSCs for general health concerns. As the congruent use of these tools continues to increase, especially following the advent of the COVID-19 pandemic, it is essential that researchers, developers and health providers work together to ensure their safety and accuracy prior to their widescale adoption in the home, community and healthcare settings.

PROSPERO registration number: CRD42020215210

Acknowledgements: BH and AEO conceptualised the study. AEO and ERS took the lead in planning the study with support from co-authors. ERS led the data analysis with support from IW, AA, MK, KP, MEA, AP and oversight from BH and AEO. All authors provided substantial contributions to the design, acquisition and interpretation of study data and approved the final version of the paper. BH is the guarantor.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. ERS, AEO, IW, AA & BH are in part supported by the National Institute for Health Research and Care (NIHR) Applied Research Collaboration (ARC) Northwest London. The views expressed are those of the authors and not necessarily those of the NHS or the NIHR or the Department of Health and Social Care.

Conflicts of interests: None declared.

Data sharing statement: template data collection forms; data extracted from included studies and data used for all analyses can be made available by authors upon request

Patient and Public Involvement: No patient was involved in the study

Multimedia Appendix 1
PRISMA checklist

Multimedia Appendix 2
Search strategies

Multimedia Appendix 3
PRISMA Flowchart

Multimedia Appendix 4:
List of OSCs tested & measurements used by included studies

References

1. WHO. Global health workforce shortage to reach 12.9 million in coming decades 2013 [Available from: http://www.who.int/mediacentre/news/releases/2013/health-workforce-shortage/en/
2. Lupton D, Jutel A. 'It's like having a physician in your pocket!' A critical analysis of self-diagnosis smartphone apps. Soc Sci Med. 2015;133:128-35.
3. Alwashmi MF. The Use of Digital Health in the Detection and Management of COVID-19. International Journal of Environmental Research and Public Health. 2020;17(8).
4. Turner J, Knowles E, Simpson R, Sampson F, Dixon S, Long J, et al. Impact of NHS 111 Online on the NHS 111 telephone service and urgent care system: a mixed-methods study. Health Services and Delivery Research. 2021;9(21):1-148.
5. North F, Ward WJ, Varkey P, Tulledge-Scheitel SM. Should you search the Internet for information about your acute symptom? Telemed J E Health. 2012;18(3):213-8.
6. Powley L, McIlroy G, Simons G, Raza K. Are online symptoms checkers useful for patients with inflammatory arthritis? BMC Musculoskeletal Disorders. 2016;17(1):362.
7. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ digital medicine. 2022;5(1):1-9.
8. 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;9(8):e027743.
9. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
10. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Annals of Internal Medicine. 2011;155(8):529-36.
11. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. bmj. 2015;351.
12. Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of physician and computer diagnostic accuracy. JAMA internal medicine. 2016;176(12):1860-1.
13. Berry AC, Cash BD, Wang B, Mulekar MS, Van Haneghan AB, Yuquimpo K, et al. Online symptom checker diagnostic and triage accuracy for HIV and hepatitis C. Epidemiology & Infection. 2019;147.
14. 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;10(12):e040269.
15. Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F. Benchmarking Triage Capability of Symptom Checkers Against That of Medical Laypersons: Survey Study. J Med Internet Res. 2021;23(3):e24475.
16. Delshad S, Dontaraju VS, Chengat V. Artificial intelligence-based application provides accurate medical triage advice when compared to consensus decisions of

healthcare providers. Cureus. 2021;13(8).

17. Poote AE, French DP, Dale J, Powell J. A study of automated self-assessment in a primary care student health centre setting. Journal of telemedicine and telecare. 2014;20(3):123-7.
18. Dickson SJ, Dewar C, Richardson A, Hunter A, Searle S, Hodgson LE. Agreement and validity of electronic patient self-triage (eTriage) with nurse triage in two UK emergency departments: a retrospective study. European Journal of Emergency Medicine. 2022;29(1):49-55.
19. 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;16(7):e0254088.
20. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Medical Journal of Australia. 2020;212(11):514-9.
21. Gilbert S, Fenech M, Upadhyay S, Wicks P, Novorol C. Quality of condition suggestions and urgency advice provided by the Ada symptom assessment app evaluated with vignettes optimised for Australia<a class="reftools" href="#FN1">*</a>. Australian Journal of Primary Health. 2021;27(5):377-81.
22. Trivedi S, Littmann J, Stempien J, Kapur P, Bryce R, Betz M. A Comparison Between Computer-Assisted Self-Triage by Patients and Triage Performed by Nurses in the Emergency Department. Cureus. 2021;13(3):e14002.
23. Chan F, Lai S, Pieterman M, Richardson L, Singh A, Peters J, et al. Performance of a new symptom checker in patient triage: Canadian cohort study. PLOS ONE. 2021;16(12):e0260696.
24. Verzantvoort NCM, Teunis T, Verheij TJM, van der Velden AW. Self-triage for acute primary care via a smartphone application: Practical, safe and efficient? PLoS One. 2018;13(6):e0199284.
25. Yu SWY, Ma A, Tsang VHM, Chung LSW, Leung S-C, Leung L-P. Triage accuracy of online symptom checkers for accident and emergency department patients. Hong Kong Journal of Emergency Medicine. 2020;27(4):217-22.
26. O'Keeffe C, Mason S, Jacques R, Nicholl J. Characterising non-urgent users of the emergency department (ED): A retrospective analysis of routine ED data. PLOS ONE. 2018;13(2):e0192855.
27. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online Symptom Checkers: Recommendations for a vignette-based clinical evaluation standard. Journal of Medical Internet Research. (in-press).
28. Adebowale V, Rao M. It's time to act on racism in the NHS. BMJ. 2020;368:m568.
29. Williams DR, Mohammed SA. Racism and health I: Pathways and scientific evidence. American behavioral scientist. 2013;57(8):1152-73.
30. Bécares L, Kapadia D, Nazroo J. Neglect of older ethnic minority people in UK research and policy. BMJ. 2020;368:m212.
31. Salway S, Holman D, Lee C, McGowan V, Ben-Shlomo Y, Saxena S, et al. Transforming the health system for the UK's multiethnic population. BMJ. 2020;368:m268.
32. McInroy LB, McCloskey RJ, Craig SL, Eaton AD. LGBTQ+ Youths' Community Engagement and Resource Seeking Online versus Offline. Journal of Technology in

Human Services. 2019;37(4):315-33.

33. Noor P. Can we trust AI not to further embed racial bias and prejudice? BMJ. 2020;368:m363.
34. Figueroa CA, Luo T, Aguilera A, Lyles CR. The need for feminist intersectionality in digital health. The Lancet Digital Health. 2021;3(8):e526-e33.
35. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient perspectives on the usefulness of an artificial intelligence–assisted symptom checker: Cross-Sectional survey study. Journal of medical Internet research. 2020;22(1):e14679.
36. Marco-Ruiz L, Bønes E, de la Asunción E, Gabarron E, Aviles-Solis JC, Lee E, et al. Combining multivariate statistics and the think-aloud protocol to assess Human-Computer Interaction barriers in symptom checkers. Journal of Biomedical Informatics. 2017;74:104-22.
37. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. The Lancet. 2018;392(10161):2263-4.
38. Talmon J, Ammenwerth E, Brender J, De Keizer N, Nykänen P, Rigby M. STARE-HI—Statement on reporting of evaluation studies in Health Informatics. International journal of medical informatics. 2009;78(1):1-9.
39. Stead WW, Haynes RB, Fuller S, Friedman CP, Travis LE, Beck JR, et al. Designing medical informatics research and library—resource projects to increase what is learned. Journal of the American Medical Informatics Association. 1994;1(1):28-33.
40. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Elsevier; 2016. p. 843-51.
41. Millenson ML, Baldwin JL, Zipperer L, Singh H. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. Diagnosis. 2018;5(3):95-105.
42. Jutel A, Lupton D. Digitizing diagnosis: a review of mobile applications in the diagnostic process. Diagnosis. 2015;2(2):89-96.

Supplementary Files

Multimedia Appendixes

PRISMA checklist.
URL: http://asset.jmir.pub/assets/222c485f15a568f19c8651981d6a4a55.doc

Search Strategies.
URL: http://asset.jmir.pub/assets/4190b4a90d873c4437e9d1205cbb6e5f.doc

PRISMA Flowchart.
URL: http://asset.jmir.pub/assets/2616dca3a942a56964806e39a8db1064.doc

List of OSCs tested & measurement used by included studies.
URL: http://asset.jmir.pub/assets/fe323b0508d2860b796fc709e4b3b35d.doc