PLOS Digital Health

Evidence and Methods from Studies of Symptom-Assessment Applications Under Real-World Conditions: A Scoping Review

--Manuscript Draft--

Manuscript Number:	PDIG-D-26-00385
Article Type:	Research Article
Full Title:	Evidence and Methods from Studies of Symptom-Assessment Applications Under Real-World Conditions: A Scoping Review
Short Title:	Evidence of Symptom-Assessment Applications Under Real-World Conditions
Corresponding Author:	Marvin Kopka, PhD Imperial College London London, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND
Order of Authors:	Marvin Kopka, PhD
	Sami Altalib
	Benedict Hayhoe
	Austen El-Osta
Keywords:	Symptom assessment; real-world evidence; symptom checkers; methods; review; implementation; accuracy; safety; impact
Abstract:	Background Symptom-assessment applications (SAAs) are widely used as patient-facing digital tools for triage and diagnostic support, yet most evaluations have relied on vignette-based designs that poorly reflect real-world use. Evidence generated under real-world conditions, where users experiencing symptoms enter their own information, remains fragmented and methodologically heterogeneous. Objective We conducted a PRISMA-ScR compliant scoping review to synthesise evidence from studies evaluating SAAs under real-world conditions and to characterise the methods used. Methods Searches of MEDLINE, Embase, Web of Science and medRxiv identified 20 eligible studies published between 2018 and 2025. The aims of the included studies can be clustered into three domains: (i) performance and safety, (ii) user characteristics and usability, and (iii) impact on decision-making and healthcare utilisation. Results Across studies, triage accuracy ranged from 34-85%, with a consistent tendency toward over-triage, that is, assigning a higher urgency than necessary. Reported safety concerns were uncommon but non-negligible, particularly at scale. Reported diagnostic accuracy varied widely depending on operationalisation and reference standards. Users were predominantly younger and female, and acceptability was generally high. Evidence on behavioural and system-level impact showed decreased healthcare utilization and a tendency towards lower perceived urgency, although the evidence remains limited. Methodologically, the studies varied substantially in sampling strategies, reference standards, matching procedures, and blinding, which limits comparability. Conclusions These findings indicate that while SAAs may provide comparatively safe triage advice

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	to users, evidence on their population-level impact is lacking. Future evaluations should prioritise pre-care sampling approaches, standardised methods for assessing accuracy and safety, and longitudinal measurement of behaviour and utilisation to inform safe and effective deployment in health systems.
Suggested Reviewers:	András Meczner meczner@gmail.com Expert on symptom checkers and evaluation methods
	Johannes Knitza johannes.knitza@staff.uni-marburg.de Expert on Symptom Checkers
	Anna-Jasmin Wetzel jasmin.wetzel@uni-konstanz.de Expert on symptom checkers
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Government Employee Are you or any of the contributing authors an employee of the United States government? Manuscripts authored by one or more US Government employees are not copyrighted, but are licensed under a CC0 Public Domain Dedication, which allows unlimited distribution and reuse of the article for any lawful purpose. This is a legal requirement for US Government employees. This will be typeset if the manuscript is accepted for publication.	No - No authors are employees of the U.S. government.
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS Digital Health* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate.	The author(s) received no specific funding for this work.

Funded studies Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? Did you receive funding for this work?	
Competing Interests On behalf of all authors, disclose any competing interests that could be perceived to bias this work. This statement will be typeset if the manuscript is accepted for publication. Review the instructions link below and PLOS Digital Health's competing interests policy to determine what information must be disclosed at submission.	The authors declare no competing interests.
Data Availability Provide a **Data Availability Statement** in the box below. This statement should detail where the data used in this submission can be accessed. This statement will be typeset if the manuscript is accepted for publication. Before publication, authors are required to make all data underlying their findings fully available, without restriction. Review our PLOS Data Policy page for detailed information on this policy. Instructions for writing your Data Availability statement can be accessed via the Instructions link below.	The data extracted from the included articles will be made available in an Open Access Repository (Zenodo) upon acceptance.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Cover Letter

Dear Editors,

We are pleased to submit our manuscript entitled “Evidence and Methods from Studies of Symptom-Assessment Applications Under Real-World Conditions: A Scoping Review” for consideration for publication in PLOS Digital Health.

Symptom-assessment applications (SAAs) are increasingly used as patient-facing digital tools to support self-triage and care-seeking decisions. However, most published evaluations have relied on vignette-based designs that do not reflect how these tools are used in everyday settings. As a result, evidence on their real-world performance, safety, and impact on healthcare utilization remains fragmented and methodologically inconsistent.

In this PRISMA-ScR compliant scoping review, we synthesize evidence from 20 studies evaluating SAAs under real-world conditions, that is, in contexts where users entered their own symptoms they experienced in that situation. Across the included studies, we identified three main areas of investigation: performance and safety, user characteristics and usability, and impact on decision-making and healthcare utilization. Overall, the available evidence suggests that SAAs may provide comparatively safe triage advice and are generally well accepted by users. At the same time, the review shows that evidence on behavioral and system-level impact remains limited, and that substantial methodological heterogeneity persists across studies. Based on the procedures reported in these studies, we suggest best practices for future evaluations and report barriers and facilitators to make future real-world studies more feasible.

We believe this manuscript is a strong fit for PLOS Digital Health because it addresses a clinically and societally relevant class of patient-facing digital health tools and provides a structured synthesis of the current evidence base. Beyond summarizing findings, the review identifies methodological limitations in existing real-world evaluations and outlines concrete steps for future studies. In this way, the manuscript speaks directly to questions of digital health evaluation, implementation, safety, and human interaction with decision-support technologies.

This submission is original, has not been published previously, and is not under consideration elsewhere.

Thank you for considering our manuscript for publication in PLOS Digital Health. Please do not hesitate to contact me if any additional information is required.

Sincerely,



Marvin Kopka for the authors

Manuscript Click here to access/download;Manuscript;Manuscript.docx

Evidence and Methods from Studies of Symptom-Assessment Applications Under Real-World Conditions: A Scoping Review

Marvin Kopka[1,2]*, Sami Altalib[1], Benedict Hayhoe[3], Austen El-Osta[1,4]

[1] Self-Care Academic Research Unit (SCARU), School of Public Health, Imperial College London, London, United Kingdom

[2] Division of Ergonomics, Department of Psychology and Ergonomics (IPA), Technische Universität Berlin, Berlin, Germany

[3] Department of Public Health and Primary Care, Imperial College London, London, United Kingdom

[4] School of Life Course and Population Sciences, King's College London, London, United Kingdom

*Corresponding author

E-mail: m.kopka@imperial.ac.uk

Abstract

Background: Symptom-assessment applications (SAAs) are widely used as patient-facing digital tools for triage and diagnostic support, yet most evaluations have relied on vignette-based designs that poorly reflect real-world use. Evidence generated under real-world conditions, where users experiencing symptoms enter their own information, remains fragmented and methodologically heterogeneous.

Objective: We conducted a PRISMA-ScR compliant scoping review to synthesise evidence from studies evaluating SAAs under real-world conditions and to characterise the methods used.

Methods: Searches of MEDLINE, Embase, Web of Science and medRxiv identified 20 eligible studies published between 2018 and 2025. The aims of the included studies can be clustered into three domains: (i) performance and safety, (ii) user characteristics and usability, and (iii) impact on decision-making and healthcare utilisation.

Results: Across studies, triage accuracy ranged from 34-85%, with a consistent tendency toward over-triage, that is, assigning a higher urgency than necessary. Reported safety concerns were uncommon but non-negligible, particularly at scale. Reported diagnostic accuracy varied widely depending on operationalisation and reference standards. Users were predominantly younger and female, and acceptability was generally high. Evidence on behavioural and system-level impact showed decreased healthcare utilization and a tendency towards lower perceived urgency, although the evidence remains limited. Methodologically, the studies varied substantially in sampling strategies, reference standards, matching procedures, and blinding, which limits comparability.

Conclusions: These findings indicate that while SAAs may provide comparatively safe triage advice to users, evidence on their population-level impact is lacking. Future evaluations should prioritise pre-care sampling approaches, standardised methods for assessing accuracy and safety, and longitudinal measurement of behaviour and utilisation to inform safe and effective deployment in health systems.

Author Summary

Symptom-assessment applications (SAAs) are digital tools that ask users about their symptoms and then suggest possible next steps, such as self-care, seeing a doctor, or seeking urgent or emergency care. They are already widely available, but most research has tested them using written case scenarios rather than with real people who are actually experiencing symptoms. This means that it is still unclear how well these tools perform under everyday conditions and how they may affect healthcare use.

In this study, we reviewed 20 studies that evaluated SAAs in real-world settings, where people entered their own symptoms. We found that these studies mainly examined three areas: how accurate and safe the tools are, who uses them and how acceptable they are to users, and whether they influence care-seeking decisions. Overall, SAAs appeared to provide comparatively safe triage advice and were generally well accepted. They also tended to recommend a higher level of care rather than a lower one. However, there is still little evidence on whether SAAs actually change healthcare use in the long term or reduce pressure on health systems.

Our findings suggest that future research should move beyond accuracy alone and focus more strongly on real-world impact, especially actual care-seeking behavior and healthcare utilization over time.

Introduction

Symptom-assessment applications (SAAs) are widely available patient-facing tools that provide healthcare-seeking advice and, in many instances, differential diagnoses based on the analysis of free-text descriptions of presenting symptoms and responses to individually tailored closed questions(1–3). They can support members of the public in determining likely causes of their symptoms, and making decisions about next steps, and also hold promise for saving healthcare costs through more appropriate use of urgent, emergency, and primary care services (1–3). From an evidence perspective, this combination of wide availability and clinical relevance makes SAAs a recurring target of evaluations (1–3).

Most evidence to date has focused on performance under controlled, pre-clinical conditions, typically using short descriptions of patient cases (3). Across these studies, findings have been mixed and often difficult to compare directly because they vary in the included vignette sets and reference standards (3). Systematic reviews of evaluation studies have concluded that diagnostic accuracy is generally modest and that triage accuracy varies substantially across tools and studies (1,3). For example, Wallace et al. report that diagnostic accuracy of apps ranges from 19% to 38%, while triage accuracy was approximately 49–90% across included evaluations (1). A more recent systematic review of SAAs focused specifically on triage and found the advice to be correct in 12% to 90% of all cases, depending on the exact tool (3). However, well-performing SAAs appeared to make more accurate decisions than users in general (3).

Vignette-based evaluations have been criticised because they often operationalise a real-world, user-facing task (a layperson entering their own symptoms) in an artificial way (4). These imitations arise primarily because the symptom input is often conducted by researchers or clinicians rather than patients experiencing symptoms and the vignettes are constructed by clinicians and do not represent what patients experience (4). This approach introduces several biases including that case vignettes may not include all information that can be entered into SAAs and missing information may be assumed using clinical expertise (4,5). As a consequence, the generalizability of evaluation results is highly questionable. In response, the field has begun to develop new methods, such as standardised input procedures (6), more representative case vignettes (7), procedures for determining reference standards that SAAs are compared against (8) and checklists for standardised reporting (4).

At the same time, few studies are available that specifically assessed SAAs under real-world conditions, while their methods and results seem to vary just as much as in vignette-based studies. For example, some studies assessed the accuracy of SAAs with emergency department patients (9), others assessed the characteristics of SAA users (10), or the impact that SAAs have on patient-physician relationships (11). Although previous reviews already assessed the accuracy of SAAs and the different approaches to evaluation of accuracy (1–3), no such review exists for studies assessing SAAs under real-world conditions, i.e., with patients experiencing acute symptoms and entering the symptoms themselves.

To address this gap in knowledge, the primary aim of this scoping review was to summarise currently available evidence on SAAs under real-world conditions. A

secondary aim was to characterise the specific approaches to benchmarking these tools and to the development of reference standards. Based on this synthesis of existing research, the methods used and identified best practices, this scoping review aims to build a foundation for the design of future real-world studies on SAAs.

Methods

Protocol and Reporting

The scoping review is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) (12). The review was pre-registered on OSF and the protocol is accessible under the registration ID WQ62H (13).

Eligibility Criteria

We aimed to identify all scientific publications in which SAAs were evaluated under real-world conditions. Although the term *real-world evidence* within pharmacology typically refers to routine data collected in conditions outside of clinical trials, such a categorization does not yet exist for digital health applications (14). Digital health tools are typically evaluated regarding different outcomes like privacy, interoperability or accuracy and this is mostly done without actual end user involvement - making the pharmacological 3-phase trial with additional real-world studies not applicable here (3,15). Particularly for SAAs, most studies typically either evaluate them using case vignettes (fictitious and/or from real patient cases) or with involvement from patients or users who enter their symptoms themselves (3). For the purpose of this scoping review, we define real-world evidence of SAAs as (i) actual patients or end-users, who (ii) experienced acute symptoms or used an SAA for another reason at the time of data

collection, and (iii) entered symptoms or used the SAA on their own, not through an unrelated third person entering information on behalf of them.

Peer-reviewed articles and preprints that evaluated SAAs in real-world settings were eligible for inclusion. All potential patient and end-user demographics, age groups and symptom types were eligible. We excluded studies that only evaluated a highly specialised or single-condition SAAs (like COVID-19-only tools) and studies where the SAA was combined with another tool or service (e.g., an SAA that was integrated with a triage hotline and did not function as a stand-alone tool). Studies based solely on vignettes, studies in which participants did not enter symptoms themselves, studies in which participants used the SAA post-hoc, (i.e., not while experiencing their symptoms), as well as studies that do not report empirical outcomes (like opinion articles) were also excluded. Reviews, other non-primary research articles and short articles (such as conference abstracts or posters) without sufficient detail to extract the methods used were also excluded. We included articles that were published in English prior to the date of our search on databases on 6 October 2025.

Information Sources & Search

Following consultation with a research librarian at Imperial College London, we developed and refined a search strategy for Embase on Ovid. We examined relevant articles and screened their reference lists to identify additional relevant terms. The final search string can be found in the Appendix. No time limits or other restrictions were applied. This search string was used to search the data bases MEDLINE, Embase (via Ovid), Web of Science and medRxiv. All searches were conducted on 6 October 2025.

Screening & Data Extraction

Relevant studies were imported into Covidence (16), which automatically removed duplicates. Subsequently, two researchers (MK and SA) screened the titles and abstracts independently from each other to assess their eligibility. Potential disagreements in screening were resolved through arbitration by a third researcher (AEO). Following this, two researchers (MK and SA) independently assessed the full text of each article for eligibility. Disagreements were again resolved through discussion with a third researcher (AEO).

Two authors (MK and SA) independently extracted data from the included studies using a standardised form. This form was developed by MK and pre-tested with three publications to assess whether all relevant variables were included and could be answered. The form included basic information about the paper (the title of the publications, authors, year of publication, DOI), the methods (study aim, study design, study setting, study population, sample size, recruitment strategy, sampling approach, inclusion and exclusion criteria, number and names of SAAs evaluated, how the SAA was accessed, start and end dates, data collection methods, how real-world use was ensured, the gold standard against which the SAA was compared, the procedure for matching SAA recommendation to the gold standard, a potential comparator group for RCTs, blinding of participants and physicians and the corresponding procedures), results (primary outcomes and secondary outcomes), barriers and facilitators for real-world studies reported in the article, limitations and information on potential conflicts of interests (reporting a conflict of interest, the study funder, whether any commercial developer or company was involved in the study, which one and whether a developer

or company author or co-author the article). Both researchers reviewed the final data collection matrix to arrive at the final version.

Synthesis of Results

First, we provide a general overview of the study details according to the Population, Intervention, Comparison, Outcome (PICO) model (17), as well as conflicts of interests and involvement of developers or industry stakeholders/companies. Second, we group the studies according to the outcomes they assessed and summarise the results and the methods used. Third, we report barriers and facilitators for conducting real-world SAA studies that were reported in the included studies. To facilitate this, we grouped all reported barriers and facilitators into several categories according to the Consolidated Framework for Implementation Research (CFIR) (18): Outer setting, inner setting, individuals, the SAAs and the implementation process. The results of this synthesis are presented in a narrative format with support of descriptive summary tables and a figure for the barriers and facilitators.

Results

Study Selection

We identified a total of 1442 articles of which 691 were duplicates. After screening 751 articles, 30 were included for full text screening, of which one was not retrievable due to access limitations. Out of the 29 articles assessed for eligibility, three were excluded because they were abstracts without enough information, four did not examine self-triage use cases (all were regarding professional triage in emergency departments), one was a study based solely on vignettes without a real-world component, and one combined the SAA with another service. This screening and appraisal derived a final list of 20 studies included in the review, see Figure 1.

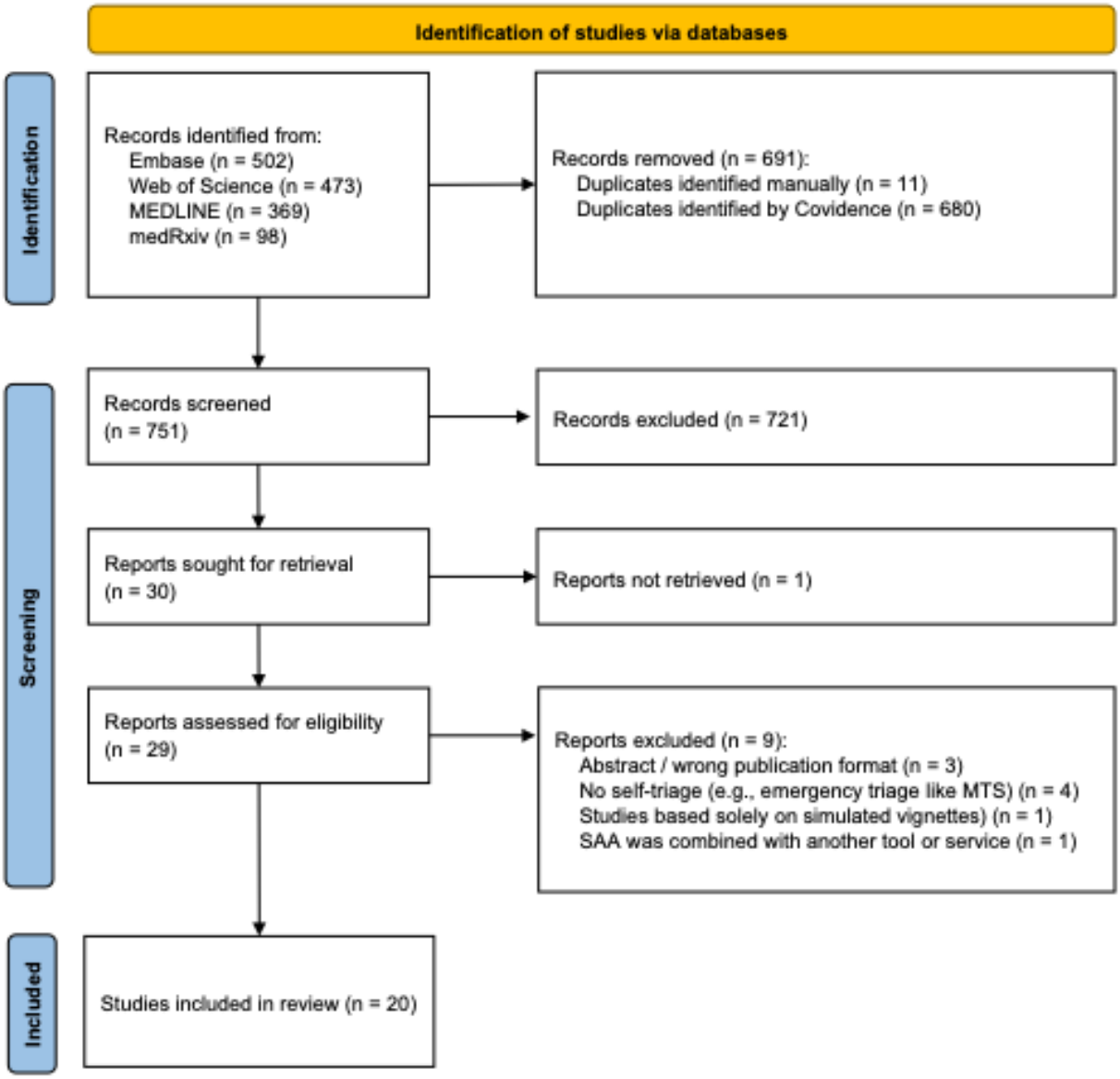


Figure 1: PRISMA flowchart showing identified, screened and included studies.

Study Characteristics

The characteristics of the 20 included studies, all published between 2018 and 2025, are reported in Table 1.

218 Table 1: Characteristics of the included studies

Study	Study Design	Population	Intervention	Comparator	Outcomes	Real-World Element	Developer Involvement
Performance of SAAs							
Verzantvoort et al. (2018) (19)	Prospective observational (cross-sectional)	App users	'Should I see a doctor?' (Dutch)	Triage determined by telephone triage nurse	- Accuracy - Impact on decision-making - User characteristics	Self-entry of symptoms outside of clinical setting (i.e., real-world use)	No (but study was funded by health insurance company)
Chan et al. (2021) (20)	Prospective observational (cross-sectional)	Primary care patients (and emergency care patients)	Self-developed app	Self-triage without symptom checker	- Accuracy - Safety	Self-entry of symptoms in emergency department of general practitioner within clinical trial	No
Cotte et al. (2022) (9)	Prospective observational (cross-sectional)	Emergency care patients	Ada Health	Guideline-based triage gold standard	- Safety - Accuracy - Appropriateness of triage	Self-entry of symptoms in emergency department within clinical trial	Yes, authors work at Ada Health
Faqar-Uz-Zaman et al. (2022) (21)	Prospective observational (cross-sectional)	Emergency care patients with abdominal pain	Ada Health	Hospital discharge diagnosis	- Accuracy	Self-entry of symptoms in emergency department within clinical trial	No
Fraser et al. (2022) (22)	Prospective observational (cross-sectional)	Emergency care patients	Ada Health	Diagnosis/triage determined by external physician panel	- Accuracy - Safety - Appropriateness of triage - Comprehensiveness and Relevance - Usability	Self-entry of symptoms in emergency department within clinical trial	No (although Ada Health adapted the app and created a figure)
Hennemann et al. (2022) (23)	Prospective observational (cross-sectional)	Patients in a psychotherapy outpatient clinic	Ada Health	Diagnosis/triage determined by treating physician	- Accuracy - Usability - Completion time	Self-entry at general practitioner within clinical trial	No
Meer et al. (2024) (24)	Prospective observational (cross-sectional)	Emergency care patients	SMASS (Swiss Medical Assessment System; in4medicine, Inc)	Diagnosis/triage determined by external physician panel	- Safety - Appropriateness of triage	Self-entry of symptoms in emergency department within clinical trial	Yes, authors work at in4medicine
Liu et al. (2024) (25)	Prospective observational (cross-sectional)	Primary care patients	Omaolo	Triage assessment from triage nurse	- Safety - Accuracy - Impact on decision-making - Sensitivity - Specificity	Self-entry of symptoms at general practitioner within clinical trial	Yes, authors work at Duodecim (developer of Omaolo)
Knitza et al. (2024) (26)	Randomised controlled trial	Emergency care patients	Ada Health and Symptoma	Comparison between two SAAs, diagnosis compared to hospital discharge diagnosis	- Accuracy - Usability - Completion Time - Acceptance (Net Promoter Score)	Self-entry of symptoms in emergency department within clinical trial	No
SAA Users							
Morse et al. (2020) (27)	Prospective observational (cross-sectional)	General public app/web users	Ada Health	None	- User demographics - Received recommendation	Self-entry of symptoms outside of clinical setting (i.e., real-world use)	No

Miller et al. (2020) (28)	Prospective observational (cross-sectional)	Primary care patients	Ada Health	None	- Usability - Impact on decision-making - Acceptability	Self-entry of symptoms at general practitioner	Yes, authors work at Ada Health
Katayama et al. (2021) (29)	Retrospective observational	General public app users for pediatric assessment	Name not mentioned, local paediatric app in Osaka Prefecture	None	- User characteristics	Self-entry of symptoms outside of clinical setting (i.e., real-world use)	No
Knitza et al. (2022) (30)	Randomised controlled trial	Outpatient care patients with musculoskeletal symptoms and unknown diagnosis	Ada Health and Rheport	Comparison between two SAAs	- Usability - Impact on decision-making - Completion time - Acceptability	Self-entry of symptoms at general practitioner	Yes, author is member of Rheport
Arellano Carmona et al. (2022) (10)	Prospective observational (cross-sectional)	General public app/web users	Buoy Health	None	- User characteristics - Drivers to use SAAs - Recommendations of the SAA - User perceptions - Intention to follow	Self-entry of symptoms outside of clinical setting (i.e., real-world use)	Yes, Buoy Health funded the study (but were not involved in design, analysis or interpretation)
Gellert et al. (2023) (31)	Prospective observational (cross-sectional)	General public app/web users	Infermedica Triage / Symptomate	None	- Demographic characteristics (and reasons for using it) - Impact on decision-making - Frequency of use - Value derived - Satisfaction with virtual triage	Self-entry of symptoms outside of clinical setting (i.e., real-world use)	Yes, authors work at Infermedica
Mahlknecht et al. (2023) (32)	Prospective observational (cross-sectional)	Primary care patients	No name mentioned	None	- User attitudes and experience with the SAA - Perceived helpfulness - Usability - Future at-home use intentions	Self-entry of symptoms at general practitioner	No
Wetzel et al. (2024) (33)	Prospective observational (longitudinal)	General public app users	Ada Health	None	- Predictors of use - Reasons for SAA usage	Home-use of SAA over 6 weeks, self-entry of symptoms outside of clinical setting	No
Impact of SAAs							
Winn et al. (2019) (34)	Prospective observational (cross-sectional)	General public app users	Buoy Health	Self-triage without symptom checker	- Impact on decision-making - Common symptoms entered	Self-entry of symptoms outside of clinical setting (i.e., real-world use)	Yes, author is advisor for Buoy Health
Schmieding et al. (2025) (11)	Randomised controlled trial	Emergency care patients	Ada Health	Usual care	- Patients' satisfaction with care / patient-physician interaction - Patients' self-rated satisfaction with care - Physicians' satisfaction with care	Self-entry of symptoms in emergency department within clinical trial	No

					- Participants' anxiety levels.		
Cotte et al. (2025) (35)	Prospective observational (longitudinal)	Adult patients in a private healthcare network	Ada Health	Diagnosis/triage determined by external physician panel	- Impact on decision-making	Self-entry of symptoms outside of clinical setting (i.e., real-world use)	Yes, authors work at Ada Health

Fifteen percent of studies (n=7) were conducted in Germany (9,11,21,23,26,30,33), three in the United States (10,22,27), two studies were conducted internationally (31,34) and the remaining studies (n=1 each) were conducted in Portugal (35), the Netherlands (19), Canada (20), the United Kingdom (28), Japan (29), Italy (32), Switzerland (24) and Finland (25).

Most studies (n=18; 90%) were conducted with one SAA, whereas only two studies (10%) compared two SAAs. Ada Health was the most commonly included SAA (n=11; 55%) (9,11,21–23,26–28,30,33,35), followed by Buoy Health (n=2; 10%) (10,34). The remaining SAAs (Symptomate, Symptoma, Omaolo, SMASS, Pathfinder, an unnamed app and a self-developed tool) were each included in one study only (19,20,24–26,29,30,32). In about half of the included studies (n=9; 45%), the developer was involved either through study funding or by conducting or co-authoring the study (9,10,24,25,28,30,31,34,35). Eleven studies (55%) did not report any involvement of the developer (11,19–23,26,27,29,32,33).

The included studies could be naturally grouped as addressing one of three main domains: (i) accuracy and/or safety, (ii) user preferences, usability and acceptability and/or (iii) impact on decision making. Most studies (n=9; 45%) primarily evaluated the accuracy or safety of the SAAs (9,19–26), followed by studies focused on the users of SAAs (demographic characteristics of users, attitudes and reasons for using SAAs

and perceived usability and acceptability, n = 8, 40%) (10,27–33). Only three studies (15%) directly focused on the impact that SAAs have on users' decision-making, their care utilization, or their satisfaction with received care (11,34,35), although an additional five (25%) studies assessed the impact on decision-making as a secondary outcome (19,25,28,30,31).

The study designs differed between these various focuses. Whereas studies evaluating the performance of SAAs were mostly prospective cross-sectional studies (n=8/9; 89%) (9,19–25) including one randomised controlled trial (RCT) (11%) (26), studies focusing on the user included five out of eight (63%) prospective cross-sectional designs (10,27,28,31,32), one prospective longitudinal design (13%) (33), one retrospective observational design (13%) (29), and one RCT (13%) (30).

Studies on the impact of SAAs on users' decision-making included five out of eight (63%) prospective cross-sectional designs (19,25,28,31,34), two (25%) RCTs (11,30) and one (13%) prospective longitudinal study (35). However, these included studies which assessed the impact of SAAs as a secondary outcome only. Among studies that aimed to assess the SAAs' impact on users' decision-making, the study designs were equally distributed: one was designed as a (33%) prospective cross-sectional survey (34), one as a (33%) prospective longitudinal study (35) and one (33%) as an RCT (11).

The most studied populations were primary care patients (7; 35%) (20,23,25,28,30,32,35) and general app users (7, 35%) (10,19,27,29,31,33,34), followed by emergency care patients (7, 35%) (9,11,20–22,24,26).

For studies assessing the performance of SAAs, six out of nine (67%) included emergency care patients (9,20–22,24,26), and three (33%) primary care patients (20,23,25). Notably, only one study evaluated the performance with pre-care patients and general app users (19).

When assessing SAA users, five out of eight (63%) studies included app users (10,27,29,31,33) and three (37%) studies primary care patients (28,30,32). Lastly, five out of eight (63%) studies on the impact of SAAs included primary care patients (19,25,28,30,35), two (25%) app users (31,34), and one study emergency care patients (13%) (11). Among studies with the primary aim of examining the impact, one out of three (33%) included app users (34), one (33%) primary care patients (35), and one (33%) emergency care patients (11).

Accuracy and Performance Evaluations

Findings

The performance of SAAs varied between the exact outcomes that were assessed, see Table 2.

Table 2: Findings regarding the accuracy and performance of symptom-assessment applications.

Study	Triage Accuracy	Safety	Diagnostic Accuracy
Verzantvoort et al. (2018)	Exact match: 81% Under-triage: 11%	Not reported	Not reported

	Over-triage: 8%		
Chan et al. (2021)	Emergency department: Exact match: 61% (172/281) Under-triage: 3% (8/281) Over-triage: 27% (76/281) Primary care: Exact match: 85% (255/300) Under-triage: <1% (1/300) Over-triage: 7% (20/300)	No emergency missed in emergency department 2% (9/581) were potentially harmfully under-triaged in all cases (emergency and primary care combined)	Not reported
Cotte et al. (2022)	Exact match or more conservative: 91% (344/378) Under-triage: 9% (34/378)	Among under-triaged cases: 15% (5/34) judged accurate by 3 physicians 26% (9/34) judged potentially risky based on presented info	Not reported
Faqar-Uz-Zaman et al. (2022)	Not reported	Not reported	Top 5 match: 52% (234/450)
Fraser et al. (2022)	Not reported	Physician-rated triage safety Fully agree: 62% (23/37) Safe but too cautious: 24% (9/37) Unsafe/too risky (by at least 1 physician): 22% (8/37) Unsafe/too risky (by at least 2 physicians): 14% (5/37) Unsafe/too risky (by all 3 physicians): 5% (2/37)	Top 1 match: 30% (9/30) Top 3 match: 63%(19/30) Top 5 match: 70% (21/30)
Hennemann et al. (2022)	Not reported	Not reported	Top 1 match: 51% (25/49) Top 5 match: 69% (34/49)
Meer et al. (2024)	Under-triage: 8% (210/2543) Over-triage: 18% (450/2543)	Potentially harmful under-triage: none	Not reported
Liu et al. (2024)	Exact match: 53.7% (471/877) Sensitivity: 62.6% Specificity: 69.2%	Assessments deemed as safe: 98% (856/877)	Not reported
Knitza et al. (2024)	Ada Health: Exact match: 34% (149/437) Under-triage: 13% (58/437) Over-triage: 53% (230/437)	Ada Health: did not report potentially life-threatening diagnosis in 13% of cases (56/437) Symptoma: did not report potentially life-threatening diagnosis in 14% of cases (61/437)	Ada Health: Top 1 identical match: 14% Top 1 identical or plausible match: 58% Top 5 identical match: 27% Top 5 identical or plausible match: 75% Symptoma: Top 1 identical match: 4%

			Top 1 identical or plausible match: 38% Top 5 identical match: 13% Top 5 identical or plausible match: 64%

Note: The reported metrics differed between the studies. For example, Meer et al. (2024) did not report an exact triage match but only under- and over-triage estimates.

Triage accuracy varied across studies and settings: In emergency care settings, accuracy ranged from 34% to 61% (20,26) whereas in primary care settings, triage accuracy was slightly higher, ranging from 54% to 85% (20,25). Generally, over-triaging was more common (ranging from 8% to 53%) than under-triaging (<1% to 13%) (19,20,26).

Safety was often reported as the proportion of cases not under-triaged or as individual assessments of each case (or under-triaged cases only) by physicians. Overall, SAAs' triage advice appeared to be relatively safe in the included real-world cases: Chan et al. reported that no emergencies were missed among emergency care patients, although a small proportion (2%) were potentially harmfully under-triaged when considering all populations, i.e., emergency, primary care and pre-care patients (20). Similarly, Meer et al. reported that some cases were under-triaged, but none of them were considered harmful after physician review (24). In two other studies, physicians reviewed under-triaged cases and determined 26% (9/34) of them to be potentially risky or harmful based on the presented information, which corresponds to 2% (9/378) of all cases (9,22). One more study assessed triage advice among all cases and found that 98% (856/877) could be considered safe, with safety defined as the same or higher urgency than the gold standard reference (25). Lastly, one study assessed the safety

of the provided diagnoses and found that the two SAAs did not report life-threatening diagnoses of emergency care patients in 13% (56/437) and 14% (61/437) of all cases respectively (26).

Diagnostic accuracy in general ranged from 4% to 75%, depending on the exact operationalization, although most studies that assessed diagnostic performance reported an accuracy of about 50-70% (21–23,26). Interestingly, the diagnostic accuracy appeared to be higher (51% in top 1) for psychiatric diagnoses than for non-psychiatric ones (30%, 4% and 14% as an exact top 1 match across all studies) (22,23,26). Most studies reported whether a diagnosis was within the top one, top three, or top five diagnoses suggested by the SAAs. Accordingly, accuracy in these instances was higher when considering all potential diagnoses. One study examined whether the top 1 and top 5 diagnoses matched the reference standard, and also whether the diagnoses were plausible (26). This procedure improved accuracy substantially since 38% to 75% were considered accurate or plausible, although only 4% to 27% were considered an exact match (26).

Methods

Across the studies, there was substantial heterogeneity in sampling strategies, reference standards against the SAA recommendations were compared, matching procedures and blinding. An overview is provided in Table 3.

332 Table 3: Methods of studies evaluating the accuracy and performance of symptom-assessment applications.

Study	Sampling Frame & Method	Analyzed N	Reference Standard	Matching of SAA Output to Reference Standard	Blinding
Verzantvoort et al. (2018)	All users opening the app, convenience sampling (all app users who used the app for a current symptom)	4456	Telephone triage nurse (triage outcome)	Not reported	Not reported
Chan et al. (2021)	Walk-in patients, sampling method not reported	581 (281 emergency care patients and 300 primary care patients)	Treating physician diagnosis/triage	Physician diagnoses were categorised into emergency / urgency / routine / home-appropriate; SAA triage compared against categorised physician diagnosis	Participants: blinded to SAA recommendation by withholding recommendation Physicians: blinded to patient participation.
Cotte et al. (2022)	Walking-in emergency care patients, consecutive sampling	378	Guideline-based triage gold standard (MTS)	Pre-specified rule: If SAA advice matched MTS or was higher, recommendation was classified as safe. Otherwise, an external physician panel reviewed the case with all collected clinical information (blinded to MTS and SAA advice)	Participants: blinded to SAA recommendation by withholding recommendation Physicians: did not receive app report either Physician panel: blinded to MTS/SAA advice during urgency adjudication.
Faqar-Uz-Zaman et al. (2022)	Walk-in emergency care patients, sampling approach not reported	450	Hospital discharge diagnosis	Not reported	Participants: blinded to SAA recommendation by withholding recommendation Physicians: did not receive app report either
Fraser et al. (2022)	Walk-in emergency care patients, convenience sampling	40	External physician panel assessed each case and determined gold standard triage solution, with majority vote. They also rated reasonableness on a Likert scale.	Not reported	Participants: blinded to SAA recommendation by withholding recommendation (but PDF emailed to study team) Physicians: did not receive app report either

			For diagnostic accuracy, match with at least one discharge diagnosis was used		
Hennemann et al. (2022)	Psychotherapy outpatient clinic, consecutive patients	49	Treating clinician diagnosis, match to ICD-10 4-digit codes (e.g., F40.1)	Text diagnoses were recoded to ICD-10 by a trained clinical psychologist (not otherwise involved in the study) and cross-checked by another psychologist. Disagreements were resolved with a third rater. ICD-10 codes were then compared.	Participants: blinded to SAA recommendation by withholding recommendation (instructed to stop before suggestion is displayed) Therapist: did not receive app report either
Meer et al. (2024)	Walk-in emergency care patients, consecutive patients (between 8am and 5pm)	2543	External physician panel (Panel A) proposed an appropriate triage range	Stepwise multi-panel adjudication: Panel A defined the appropriate triage range, which was then compared to the SAA by a research assistant. If the case was under-triaged, Panel B assessed case as well. If both Panel A and B classified case as under-triaged, Panel C assessed the hazard likelihood on a 4-point Likert scale	Participants: blinded to SAA recommendation by withholding recommendation Physicians: not mentioned Panelists: Unaware of SAA recommendations
Liu et al. (2024)	Walk-in primary care patients, sampling approach not mentioned	877	Triage nurse assessment	Matching performed by external panel (1 nurse employed by developer, 4 physicians, 3 employees working in SAA development, 1 coder responsible for SAA production). Matched nurse triage to SAA recommendation using classification of levels of emergency by the Ministry of Social Affairs and Health and the Finnish Institute for Health and Welfare	Participants: Blinding not mentioned Triage nurse: Blinded to SAA result by not receiving SAA recommendations until after they completed their own assessment
Knitza et al. (2024)	Walk-in emergency care patients, consecutive patients	437	External physician panel determined safety, hospital discharge diagnosis as gold standard diagnosis	6 physicians with emergency department experience blindly reviewed SAA outputs and discharge diagnoses and classified SAA diagnoses as identical, plausible, or diagnostically different, as well as safe or potentially life-threatening	Participants: Blinded, but procedure not detailed Physician: Blinded, but procedure not detailed

					Panelists: Blinded to all other data except SAA diagnoses and final discharge diagnoses

333

Most studies recruited walk-in patients in emergency or primary care settings (9,20–26) and only one study sampled app users during real-world use (19). The sampling procedure was often described as consecutive (or was not described at all), although very few studies described the exact sampling window. Several emergency care studies restricted recruitment to specific hours (24), which limits the representativeness of the full patient population across the study period. Overall, explicit reporting of whether recruitment occurred continuously or was dependent on staff availability was uncommon and many studies did not report a sampling approach at all.

The reference standards that SAA recommendations were compared against also varied widely and included assigned urgency levels by telephone triage nurses (19), treating clinicians (20,23), guideline based triage systems like the Manchester Triage Scale (9), hospital discharge diagnoses (21,22,26) and external physician panels (22,24). Correspondingly, the procedures used to match SAA output to the reference standard also differed. Three out of nine studies (33%) did not report how the output was compared to the reference standard (19,21,22). Among the others, some used pre-defined triage levels (9,20), external physician panels who reviewed each individual case (22,24,25) and one study on psychiatric conditions used ICD-10 codes and matched them using two independent coders (23). Notably, one study compared the SAA recommendation to the hospital discharge diagnosis and used an independent physician panel who blindly reviewed only the SAA outputs and the discharge diagnoses without additional information (26). They subsequently confirmed whether the diagnoses were identical but and also assessed whether the SAA diagnoses were plausible or diagnostically different. Additionally, they classified for each diagnosis whether it was safe or potentially life-threatening.

Most studies (7/9; 78%) mentioned blinding participants to the SAA recommendations (9,20–24,26) and most (7/9; 78%) also blinded treating physicians or therapists (9,20–23,25,26). Although the exact blinding procedure was not always mentioned, participants and clinicians were commonly blinded by not receiving the results from the SAAs (9,20–24). The external physician panels were either blinded to the SAA recommendations (in which case they often received the SAA input or additional case information if they assessed the accuracy for each case individually (22,24)), or they were blinded to the input and all other information and only received the SAA recommendation and discharge diagnoses if their task was to match the diagnoses (26). Pertinently, no single study used two panels where one was used for determining the correct diagnosis (blinded to the SAA output) and the other for matching it (blinded to the SAA input).

User Characteristics, Reasons for Use and Usability

Findings

In most studies, SAA users were described as predominantly female and young (10,27,28). The proportion of female users ranged from 67% in a study using an SAA that was integrated in a large US hospital system (27) to 85% in a study of general users of the SAA Buoy (10). Only in one study of a paediatric SAA, the authors reported no gender difference regarding the children it was used for, with about 53% of patients being male (29). The studies also found that SAA users were more often younger, with one study reporting a mean age of 34 years (27), another study 39 years (10), and a third study reporting that 65% of all users were under 45 (31). One study also reported

additional characteristics, for example that users were more often White and had higher levels of formal education (10).

Symptoms assessed

Most symptoms assessed were mostly non-specific symptoms with pain-related and respiratory complaints being the most common (10,27,29). The most common reported symptoms in an SAA in a US hospital system were abdominal pain (8%), cough (6%), headache (4%) and sore throat (3%) (27), whereas pain (35%), gynaecological concerns (12%), whereas masses or lumps (8%) were the most common concerns among Buoy users (10). In contrast, the most common symptoms that were entered in the paediatric SAA were fever (36%), cough (6%), head and neck injuries (6%) and convulsion (5%) (29).

Advice received

The advice that users received was relatively consistent between studies. In one study, about 29% of respondents were advised to seek emergency or urgent care (27), compared to 34% in a study of Symptomate users (31), and about 10% in a study of paediatric cases (29). Non-emergency care (incl. primary care and pharmacy visits) was the most common advice, with 56% of Buoy users (10), 51% of users of an SAA integrated in a US healthcare system (27), 57% of Symptomate users (31) and 65% of paediatric SAA users (29) receiving this advice, respectively. Self-care advice (i.e., to self-manage symptoms at home, seek advice from a pharmacy or re-assess the situation in a few days) was given in 20% (27), 23% (10), 9% (31) and 25% of all cases (29) in that same order.

Reasons for using SAA

Reasons for using SAAs were most often reported as (a) symptom-triggered information seeking, (b) access convenience and (c) validation or cross-checking (10,31,33). Users mostly started information seeking because of new or potentially embarrassing concerns, they wanted to find out what may cause these symptoms and they wanted help in deciding whether they should see a physician (10,31,33). Obtaining medical advice without visiting a medical professional or curiosity (4%) was only a minor driver of SAA use in one study with Buoy users (10), but more common in another study with Symptomate users (29%) (31). Validating or cross-checking a received diagnosis was common as well, with 9% of Buoy users (10) and 14% of Symptomate users (31) reporting wanting a second opinion as their reason to use the SAA.

Usability and acceptability

The usability and acceptability of SAAs was generally high, with mixed results when SAAs were assessed in routine care (10,28,30–32). Most Buoy users expressed high usefulness and ease of understanding (M=4.18 and 4.64 on a 5-point Likert scale, respectively), as well as feeling less anxious and more empowered (10). Similarly, Symptomate users reported high satisfaction with about 80% likely or highly likely to use the app again (31). In another study, Ada Health and Rheport were tested for usability with the System Usability Scale (36) with which both apps have been rated as acceptable (30). About three quarter of participants indicated they would use the SAA again, whereas most users (about 98%) rated the tools as either very or quite easy to

use (30). In contrast, in an Italian general practice, only 49% of patients and 27% of GPs were satisfied or very satisfied with an unnamed SAA for which the ease and speed of use, the appropriateness of differential diagnoses and the symptom entry quality were cited as the most important concerns (32).

Methods

The included studies showed high heterogeneity in their methodological approach (Table 4).

Table 4: Methods of studies that assessed users, reasons for use and usability of symptom-assessment applications.

Study	Primary Outcome(s)	Sampling	Data Collection	Exposure Group	Analyzed N
Morse et al. (2020)	- User demographics triage advice distrubtion	Full population of completed web-based assessments during study period, recruitment via routine use (all users opening the app/portal)	Logged system data	SAA users only	26,646
Miller et al. (2020)	- Usability - Acceptability	Convenience sample of walk-in primary patients with acute symptoms	Survey after use	Both participants who used and did not use an SAA before	523
Katayama et al. (2021)	- User (patient) demographics	Full population of (paediatric) app users over 4 years, recruitment via routine use	Logged system data	SAA users only	59,375

Knitza et al. (2022)	- Usability - Acceptability - Usefulness	Consecutive newly presenting adult patients, walk-in clinic	Survey after use	Both participants who used and did not use an SAA before	600
Arellano Carmona et al. (2022)	- User demographics - Reasons for use	Recruitment pop-up window shown to all eligible SAA users, self-selection	Survey	SAA users only	2,437
Gellert et al. (2023)	- Use patterns - Value and satisfaction	Survey to users who could self-select, sampling not described	Survey	SAA users only	2,113
Mahlknecht et al. (2023)	- Patients' attitudes - General practitioners' attitudes	Consecutive patients with acute symptoms presenting to a general practitioner, integrated into care	Survey after use	Both participants who used and did not use an SAA before	116
Wetzel et al. (2024)	- Predictors of use	Purposive sampling of SAA users from previous study	Daily diaries	Users only	48

Analytic approach

Two studies analysed full-population routine-use datasets (27,29) and two studies conducted a survey among app users who could self-select to participate (10,31). These studies mostly aimed to describe user demographics, triage advice distribution, reasons for use and satisfaction with the app (10,27,29,31). In contrast, three studies recruited participants in a clinical setting whereby they included current SAA users as well as those who did not use an SAA before (28,30,32). These studies mainly assessed usability and acceptability (28,30,32). Only one study used a unique approach that involved purposively sampling SAA users from a previous study and collected longitudinal data over six weeks to model reasons for SAA use (33).

Sample size

The sample sizes differed between these different approaches: in routine-use log analyses, data from 26,646 and 59,375 users were analysed (27,29). If surveys were distributed among users, this led to lower sample sizes ranging from 2,113 to 2,437 (10,31). In clinical settings that also included non-users, sample sizes were substantially lower, ranging from 116 to 600 (28,30,32). In the longitudinal diary study 48 participants were included.

Impact of Symptom-Assessment Applications

Findings

The impact of SAAs was mainly assessed with respect to three outcomes: the impact on healthcare utilization behaviour, the impact on satisfaction with the patient-physician relationship and the impact on users' decision-making and intentions.

One study assessed users' intentions and healthcare-seeking behaviours (35). Although the SAA changed users' intentions, the behaviours observed after follow-up often diverged from what users initially planned: 41% followed their initial plan, 17% escalated to higher-urgency care and 29% de-escalated, which resulted in a net shift toward lower-acuity care (35). Overall, fewer specialists but more general practitioners were consulted (35). The healthcare-seeking behaviour was rated as appropriate in about 64% of all cases. Among those who initially planned to visit the emergency department but later decided against it, most decisions were appropriate (35).

Another study assessed emergency department patients' satisfaction with the patient-physician interaction and with the care they received (11). Overall, satisfaction was

high among both the group who used the SAA and the group who received care as usual, with no statistically significant difference (11). The same held true for satisfaction with care and anxiety levels(11).

Lastly, an evaluation of Buoy Health users' decision-making showed that users' uncertainty decreased from 34% to 21%, and that about one-third reduced the urgency of their intended care, two-thirds reported no change and 4% increased the acuity of their healthcare-seeking intentions (34). Overall, the intended level of healthcare-seeking shifted toward lower-acuity options after usage and 28% reported intending to self-care, compared to approximately 0% before usage (34). Several other studies assessed the impact of SAAs on participants' intentions as a secondary outcome (19,25,28,30,31). In these studies, most users either reported no change in their intentions (28,30) or – in line with the study of Buoy Health users – a decrease in acuity (19,31).

Methods

The methods used to assess the impact of SAAs mostly focused on surveys, except one study that also extracted data from electronic health records (35) (**Table 5)**.

Table 5: Methods of studies that assessed the impact of SAAs.

Study	Impact Outcome(s)	Sampling	Data Collection	Exposure Group	Analyzed N
Impact as Primary Outcome					
Winn et al. (2019)	- Decision-Making (Pre-post intended level of care)	Sample of completed app assessments during study period,	Survey before and after use	SAA users only	150,083

		sampling approach not specified			
Schmieding et al. (2025)	- Patient-physician interaction - Satisfaction with care - Anxiety (Pre-post)	Self-referred walk-in patients in multiple emergency departments; convenience sample	Survey before and after use	Both participants who used and did not use an SAA before	363
Cotte et al. (2025)	- Decision-Making (Pre-Post intentions) - Actual healthcare-seeking behaviour - Appropriateness of intentions and behaviour	Patients within private healthcare network, recruited via routine app use, sampling approach not specified	Survey before and after use, data extraction from electronic health record	SAA users only	721
Impact as Secondary Outcome					
Verzantvoort et al. (2018)	- Decision-Making (Intention to follow advice)	Convenience sample of app users during study period	Survey after use	SAA users only	4,456
Miller et al. (2020)	- Decision-Making (Change in care intent)	Convenience sample of walk-in primary patients with acute symptoms	Survey after use	Both participants who used and did not use an SAA before	523
Knitza et al. (2022)	- Decision-Making (Self-reported “would have done differently”)	Consecutive newly presenting adult patients, walk-in clinic	Survey after use	Both participants who used and did not use an SAA before	600
Gellert et al. (2023)	- Decision-Making (Intention to follow advice and change in care intent)	Survey to users who could self-select, sampling not described	Survey after use	SAA users only	2,113

Liu et al. (2024)	- Decision-Making (Nurse triage change after SAA recommendation)	Patients at primary care centres, sampling not described	Survey after use	Both participants who used and did not use an SAA before	877

The sampling and data collection strategies varied substantially. One study analysed routine app use by surveying users before and after a symptom assessment to quantify changes in intended level of care (34). Two studies recruited patients in clinical settings and embedded SAA use into care: one study in emergency departments assessed patient–physician interaction, satisfaction with care and anxiety (11), while one longitudinal observational study within a private healthcare network assessed intentions and behaviour (35). The remaining studies assessed the impact of SAAs on decision-making most commonly via post-use surveys among app users (19,31) or among patients recruited in clinical settings (28,30). One study also evaluated the impact an SAA can have in a clinician-facing workflow by examining whether triage nurses revised their initial assessment after seeing the symptom checker's recommendation (25). Similar to the studies assessing SAA users and their use, the samples sizes included in the studies depended on the sample and sampling method and ranged from a few hundred in clinical studies (11,25,28,30,35) to several thousand in SAA user surveys (19,31,34).

Barriers and Facilitators

The included studies reported several barriers & facilitators that mostly relate to the outer setting, the inner setting and to the implementation process and study planning (Figure 2).

On the outer setting, partnerships with external organizations (e.g., developers and companies, clinics or health systems) allowed researchers to sample real-world usage and log data (10,27), as well as to get access to broader user populations than possible in specific healthcare settings (10,27,35). Conversely, a low number of users limited the possibility to conduct impact evaluations and assess system-wide outcomes(29) and safety (20), for example, because rare events (like participants experiencing a true emergency) did not occur often. The regulatory environment also had an impact on data collection since data protection constraints limited the possibility to conduct follow-up investigations and assess the long-term outcomes of patients and users (11,29). Finally, many studies were conducted throughout the COVID-19 pandemic, which slowed down recruitment and may have been a confounder in several studies (11,23–25,32).

Within study sites, the feasibility often depended on how the SAAs were integrated into workflows and whether sufficient resources were available. Embedding study procedures into routine care processes (e.g., enrolling participants at intake) supported the implementation (23). Designated clinic space (25) and on-site technical support (28,32) for using the SAA were also mentioned as facilitators.

The most frequently reported barriers were resource- and staff-related. Fro example, recruitment times were highly dependent on nurse/staff availability, which led to mostly recruiting at daytime (24,26,28) whereas staff shortages lead to not screening all potential participants (11). A high general workload at the recruiting site as well as workflow inconveniences for staff and patients were also reported as barriers (23,24,32).

On an individual level, few barriers and facilitators were reported. Only one study mentioned that patients who felt very unwell were less likely to use the SAA (20), although several studies explicitly excluded patients with more severe symptoms, which may have had an impact on the results (9–11,24,26,32).

The SAAs used themselves were often seen as a facilitator, because they were safely deployable with no changes to usual care (28) and sometimes integrated into existing healthcare networks and health systems, which supported patient recruitment, increased trust and allowed follow-up assessments (27,35).

On the implementation level, multiple studies used public advertising, e.g., on local TV (32), or governmental collaborations to reach potentially eligible participants (29). Incentives were reported as another facilitator, both for physicians whose renumeration was tied to the number of included patients (32), as well as for participants if they received a high compensation (33). Regarding barriers at this level, several articles reported challenges in determining a reference standard or “correct” diagnosis for assessing the accuracy of SAAs (9,22) and some studies only assessed participants’ intentions and could not measure behavioural outcomes (19,30), except for one study that had access to the EHR of patients and could thus access long-term healthcare-seeking (35).

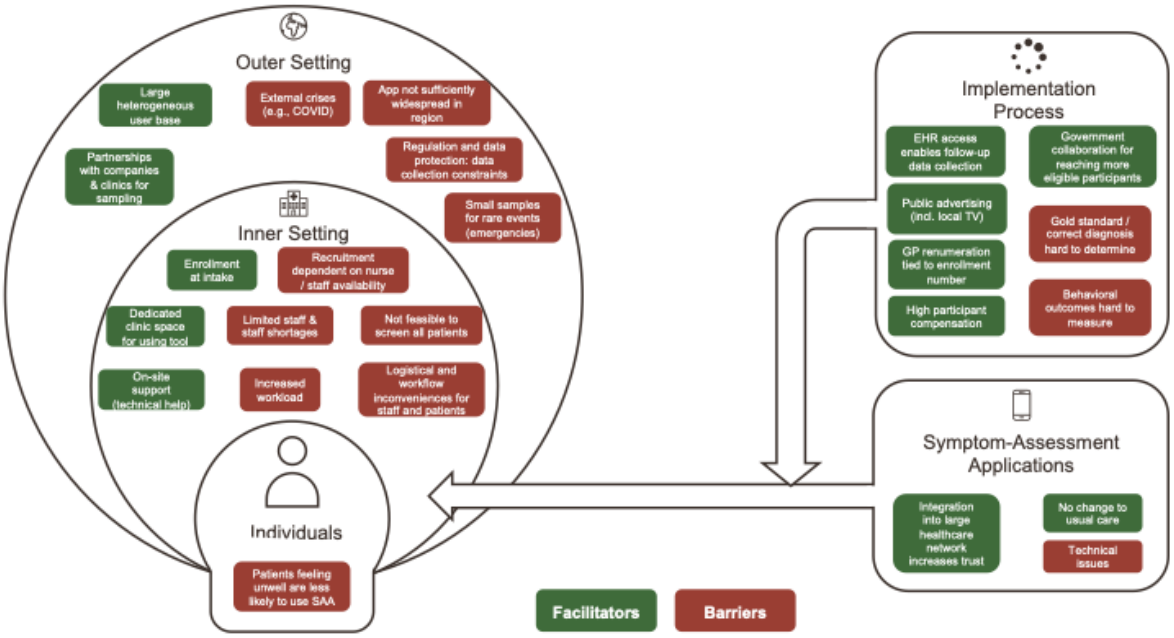


Figure 2: Barriers and facilitators reported in included studies structured according to the Consolidated Framework for Implementation Research (CFIR). This figure was adapted and is based on Damschroder et al. (18).

Discussion

Principal Findings

Across the included studies, real-world evaluations were conducted with three aims: (i) tool-focused evaluations assessing performance and safety, (ii) user-focused evaluations assessing usability and user characteristics, and (iii) system-level evaluations assessing the impact of SAAs. The results suggest that SAAs may deliver comparatively safe triage recommendations and are well accepted by users. However, the evidence base for system-level impact and real-world behaviour remains insufficient and future real-world studies should prioritise impact evaluations. In the following, we first summarise the evidence across these research aims, then discuss

implementation implications and finally derive recommendations for the design of future real-world evaluations.

Overall, real-world performance studies indicate that SAAs provide comparatively accurate triage advice in the mid-range and that they tend to over-triage rather than under-triage, which is in line with results reported in vignette-based evaluations (37–42). Some studies also examined the safety and errors that SAAs made and suggest that clearly unsafe or high-risk advice is relatively uncommon (approximately 2%-5%), although this rate can have a substantial impact when considering population-wide deployment. Diagnostic accuracy was generally lower than triage accuracy, but diagnostic outcomes were also operationalised heterogeneously. Across studies, methods differed substantially and particularly reference standards and matching procedures (to assess whether SAA recommendations were equivalent to the reference standard) were very different between the included studies.

Studies focused on users generally indicate high usability and acceptance and described a user profile skewed toward younger, female and more highly educated individuals (10,27,28,30). However, the interpretability of these studies depends on the exact recruitment strategy: sampling only users is informative for describing real-world users and the advice they receive but may introduce a self-selection bias when testing the usability. Conversely, recruiting in care settings and including both users and non-users can reduce self-selection bias for usability and acceptability, but it is not a representative sampling approach when determining how SAAs impact users' decisions or how the advice users receive is distributed. That is, participants who already decided for a care option and implemented this decision (i.e., visiting the

emergency department or primary care provider) may be subject to a sunk-cost fallacy and thus less likely to change their behaviour than if they would have used the SAA in a pre-care setting.

Evidence on system-level impact remains sparse, but available studies suggest that SAAs can shift intended care-seeking toward lower acuity without adverse effects on anxiety or the patient–physician relationship (11,34,35).

Implications for Implementation

Taken together, the available evidence may support the cautious implementation of SAAs as potentially safe tools that may help guide patients through the healthcare system. However, the limited evidence on the impact of SAAs requires further evaluation before making definitive system-wide recommendations.

First, implementing SAAs requires balancing patient safety and healthcare overutilization. Even if only a small proportion of recommendations are high-risk, a 2%-5% rate of risky advice can be unsafe for a high number of individuals. Conversely, a tendency toward over-triage may reduce individual risk but can increase unnecessary healthcare utilization, increasing costs and contributing to destabilization of already strained healthcare systems (43). However, the outcomes of implementation cannot be inferred from the accuracy or safety of SAA advice alone without humans in the loop, because users appear to integrate SAA recommendations with prior knowledge and experience (44). Prior work suggests that in this information integration process, users may disregard overly cautious advice and override advice that seems implausibly non-urgent (44). The results from the included user and impact studies support this finding:. Although a similar proportion of users was advised to seek

emergency or urgent care (approximately 10%–29%) as to self-care (approximately 9%-25%) (10,27,29,31), impact evaluations indicate that many users decreased their perceived urgency after SAA use, and intentions to self-care increased while intentions to seek emergency care decreased (34,35). In one study assessing the system-level impact, a larger share of participants de-escalated their initial plan than escalated it and de-escalation among those initially planning emergency care was often appropriate (35). Hence, SAA advice distributions are not equivalent to behavioural change; they further support the view that users do not blindly follow recommendations but integrate them with existing beliefs and experiences.

Second, although existing impact evaluations suggest a potentially beneficial effect of SAAs on patients and health systems(34,35), the current impact evidence remains too limited (particularly for behavioural and healthcare utilization outcomes) to confidently infer the impact SAAs may have. The currently available evidence suggests that implementation may reduce overall demand and redirect patients toward lower-acuity options without adverse effects on anxiety or patient–physician interaction, although more evidence is needed to support this conclusion (11,34,35).Impact studies remain scarce as only four studies included real users in pre-care contexts to assess potential or observed healthcare utilization, of which only one study assessed participants' behaviour (19,31,34,35). After having assessed the safety and accuracy extensively, the next relevant research priority is to assess the behavioural impact of SAAs and subsequent healthcare utilization under real-world conditions.

Third, the comparatively lower diagnostic accuracy reported across the included studies suggests that SAAs may have limited utility as tools for obtaining a diagnosis,

particularly if users expect a single explanation for their symptoms. This is perhaps to be expected, given that SAAs do not have access to imaging results, laboratory tests, or results from physical examinations, which will always limit diagnostic accuracy to a certain degree. However, users do not even seem to expect an SAA to give an accurate diagnosis, but rather use them for obtaining general information and for deciding whether and where to seek care (45,46). Thus, future implementations may determine whether providing diagnoses is helpful, or whether SAAs could instead advise users where to seek care and provide more information about their symptoms, as the German *PatientenNavi* by the Association of Statutory Health Insurance Physicians does (47).

When thinking about implementation, researchers should also consider the reality of many users: although some people do use certified SAAs (48,49), many more users now seek health advice from unregulated LLMs (50). As a result, future studies should also evaluate LLM accuracy, safety, advice distributions and effects on care-seeking. Available benchmarking studies suggest that LLM accuracy for healthcare-seeking advice can be comparable to mid-level or well-performing SAAs (3), but that LLMs rarely recommend self-care (51), which may increase healthcare utilization rather than reduce it. At the same time, LLMs provide highly personalised explanations and allow users to ask follow up questions. This may change how users integrate information they receive and makes them more confident in incorrect decisions than when using certified SAAs (44). Given the widespread use of LLMs, real-world impact studies that examine LLMs for symptom-assessment are increasingly needed.

Implications and recommendations for future real-world studies

To conduct future evaluation studies that can inform implementation, the following recommendations are proposed to inform study design, including a congruent sampling approach that fits to the endpoints, and the study aim and to adopting standardised procedures – particularly for determining reference standards, matching SAA advice to reference standards and to blinding.

Generally, the barriers and facilitators identified indicate that real-world evaluations should be planned with staffing, workflow integration, technical support and incentives for participation in mind. In care-based sampling approaches, studies should test potential technical issues and make sure that enough staff is available, both to support participants with tool use and to minimise recruitment dependence on staff availability. Dedicated physical space within the clinic or practice can reduce workflow disruption. Partnerships with developers, clinics, healthcare systems, or even governmental institutions can support recruitment and enable access to relevant data, including usage logs and, where feasible, electronic health record data. Because regulation and data protection laws can make it complicated to access healthcare utilization across different institutions, collaborations with larger local healthcare structures may facilitate longitudinal data collection of routinely collected data. Finally, incentive structures for participants and clinicians can facilitate recruitment, including an appropriately high level of compensation and clinician incentives based on the number of enrolled patients (32,33).

An issue limiting the external validity of many observational studies was the sampling strategy as most analysed certain outcomes but sampling and self-selection bias limits how generalizable the results really are. For example, some studies assessed the impact that SAAs may have on decision-making with a sample of participants visiting the emergency department, which is a very specific sample of patients that have certain symptoms. As these patients decided to visit the emergency department and then acted on their decision (7), they are not representative of SAA users in pre-care settings. In a healthcare facility setting, testing usability and acceptability is appropriate, as both users and non-users are included. Indeed, it would be inappropriate to test usability and acceptability with existing users because those who found the SAA not usable would have discontinued using it. However, when investigating healthcare-seeking decisions and behaviour, using a sample of real users from pre-care settings would be more appropriate. Future studies should therefore more deliberately choose their sampling strategy according to the study aim.

Specifically for accuracy and safety evaluations, we first recommend using pre-care settings more often for determining the accuracy and safety of SAAs to represent the primary use case of pre-care decision support. Second, blinding procedures should be reported in more detail rather than only stating that participants were blinded. Future studies should specify whether participants could view SAA outputs during care, whether treating clinicians had access to SAA recommendations and whether reference standard adjudicators were blinded to SAA inputs and outputs. Third, reference standards and matching procedures should be more standardised. Current approaches range from assigning the discharge diagnosis as the gold standard to independent panels with external clinicians (21,22,25). The same applies to the

matching procedure. An idea to combine the strengths of studies included in this review would be a two-panel procedure: (i) an external clinician panel determines the mostly likely diagnoses and triage recommendations based on all available information, blinded to SAA output, and (ii) a separate panel evaluates whether SAA output is equivalent (and/or plausibly similar) to the diagnosis, blinded to SAA input and any other information on the case (24–26). Finally, whenever possible, sampling procedures should move away from recruitment windows based on staff availability, which can systematically exclude certain patient subgroups (e.g., night-time arrivals) and limit generalizability. A better, although arguably less feasible approach, would be using predefined random sampling intervals or continuous recruitment throughout the recruitment period.

For impact evaluations, we first recommend sampling actual SAA users in pre-care contexts. Second, we recommend assessing behaviour (especially longitudinal healthcare-seeking behaviour) rather than intentions as a proxy, because intentions provide limited inference about system-level impact. At present, there are not enough impact studies to make more recommendations about methodological details.

Given the current field of real-world studies, three future research directions appear most pressing. First, accuracy and safety evaluations in pre-care samples are needed to determine the safety and accuracy of SAAs under typical use conditions. Second, more impact evaluations are needed that assess behavioural outcomes and healthcare utilization with longitudinal follow-up. Third, similar evaluations are needed for unregulated LLM-based health advice, including accuracy, safety, recommendation distributions (particularly self-care versus seeking healthcare) and the impact of LLMs

on decision-making and behaviour. The current evidence on SAAs is not generalizable to LLMs and thus, similar studies should be conducted specifically with LLMs to understand the effects that currently have.

Limitations

This scoping review has several limitations that should be considered when interpreting the findings. The principal limitation of this scoping review is that real-world evidence was defined broadly and for this reason necessarily included studies that tested SAAs under conditions that approximate real-world use (e.g., in emergency departments or primary care). This does not correspond to the definition of real-world evidence in pharmacology (14), but digital health apps typically do not follow a 3-tiered clinical trial scheme with post-market surveillance. Specifically for SAAs, they are typically either evaluated using case vignettes (which we consider pre-clinical), or by patients experiencing the symptoms and entering them themselves (which we considered real-world interaction here). Second, many studies included one single tool (in particular, Ada Health), which means that aggregate conclusions are influenced by the performance of that application and may not generalise to other SAAs. Third, many studies, particularly those examining decision-making, assessed intentions rather than actual healthcare-seeking behaviour. Although we made this explicit at several points in this article, it remains a major limitation of the available impact evidence. Finally, conflicts of interest were common among the included studies, as developers were involved in several studies. This raises the possibility of publication bias by selectively not publishing unfavourable results. Although independent evaluations exist and appear to be increasing, a larger body of independent real-world evidence is needed to strengthen confidence in the overall conclusions.

Conclusions

Real-world evaluations of SAAs have mostly assessed accuracy, user characteristics and acceptance. However, the current state of research does not yet support confident conclusions about system-level effects, because evidence on actual healthcare utilization behaviour and longer-term consequences remains limited. The next phase of research on SAAs should therefore focus on well-designed impact evaluations with pre-care users that measure behaviour and healthcare utilization longitudinally, alongside more standardised approaches to reference standards, matching and blinding for future accuracy and safety studies. In parallel, given the increasing use of unregulated LLMs, similar evaluations are needed for LLMs to quantify their safety and the impact they may have. Until such evidence exists, implementation decisions should focus on certified SAAs. These may be cautiously deployed as decision-support tools with appropriate monitoring, but the system-wide impact still needs to be assessed in more detail before making a definite recommendation.

Conflicts of interest

The authors declare no conflict of interest.

Funding

This project did not receive external funding. Austen El-Osta is grateful for support from the National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) North-West London. Marvin Kopka is grateful for support from the Federal Ministry of Education and Research (BMBF) and the State of Berlin as part of the Excellence Strategy of the German federal and state governments through the Berlin University Alliance. The views expressed in this article are those of the authors and not necessarily those of the NIHR or Department of Health and Social Care.

Author Contributors

All authors provided substantial contributions to the conception (MK, AEO), design (MK, AEO), acquisition (MK, SA) and interpretation of study data and approved the final version of the paper. MK took the lead in planning the study with support from co-authors. MK and SA carried out the data extraction and synthesis. MK is the guarantor.

References

1. 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 Digit Med. 2022 Aug 17;5(1):118. doi:10.1038/s41746-022-00667-w
2. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and Diagnostic Accuracy of Online Symptom Checkers: Systematic Review. J Med Internet Res. 2023 Jun 2;25:e43803. doi:10.2196/43803
3. Kopka M, von Kalckreuth N, Feufel MA. Accuracy of online symptom assessment applications, large language models, and laypeople for self–triage decisions. npj Digit Med. 2025 Mar 25;8(1):1–14. doi:10.1038/s41746-025-01566-6
4. Kopka M, Feufel MA. How to Evaluate the Accuracy of Symptom Checkers and Diagnostic Decision Support Systems: Symptom Checker Accuracy Reporting Framework (SCARF). JMIR Human Factors. 2026 Jan 16;13(1):e76168. doi:10.2196/76168
5. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online Symptom Checkers: Recommendations for a Vignette-Based Clinical Evaluation Standard. J Med Internet Res. 2022 Oct 26;24(10):e37408. doi:10.2196/37408
6. Meczner A, Cohen N, Qureshi A, Reza M, Sutaria S, Blount E, et al. Controlling Inputter Variability in Vignette Studies Assessing Web-Based Symptom Checkers: Evaluation of Current Practice and Recommendations for Isolated Accuracy Metrics. JMIR Form Res. 2024 May 31;8:e49907. doi:10.2196/49907
7. Kopka M, Napierala H, Privoznik M, Sapunova D, Zhang S, Feufel MA. The RepVig framework for designing use-case specific representative vignettes and evaluating triage accuracy of laypeople and symptom assessment applications. Sci Rep. 2024 Dec 23;14(1):30614. doi:10.1038/s41598-024-83844-z
8. El-Osta A, Webber I, Alaa A, Bagkeris E, Mian S, Taghavi Azar Sharabiani M, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. BMJ Open. 2022 Apr;12(4):e053566. doi:10.1136/bmjopen-2021-053566
9. Cotte F, Mueller T, Gilbert S, Blümke B, Multmeier J, Hirsch MC, et al. Safety of Triage Self-assessment Using a Symptom Assessment App for Walk-in Patients in the Emergency Care Setting: Observational Prospective Cross-sectional Study. JMIR mHealth and uHealth. 2022 Mar 28;10(3):e32340. doi:10.2196/32340
10. Arellano Carmona K, Chittamuru D, Kravitz RL, Ramondt S, Ramírez AS. Health Information Seeking From an Intelligent Web-Based Symptom Checker: Cross-sectional Questionnaire Study. J Med Internet Res. 2022 Aug 19;24(8):e36322. doi:10.2196/36322
11. Schmieding ML, Kopka M, Bolanaki M, Napierala H, Altendorf MB, Kuschick D, et al. Impact of a Symptom Checker App on Patient-Physician Interaction Among Self-Referred Walk-In Patients in the Emergency Department: Multicenter, Parallel-Group, Randomized, Controlled Trial. Journal of Medical Internet Research. 2025 Apr 2;27(1):e64028. doi:10.2196/64028
12. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018 Oct 2;169(7):467–73. doi:10.7326/M18-0850

13. Kopka M, El-Osta A. Real-World Evidence on Digital Symptom Checkers: A Scoping Review [Internet]. OSF Registries; 2025 [cited 2026 Jan 15]. Available from: https://osf.io/wq62h/ doi:10.17605/OSF.IO/WQ62H
14. Dang A. Real-World Evidence: A Primer. Pharm Med. 2023 Jan 1;37(1):25–36. doi:10.1007/s40290-022-00456-6
15. Lagan S, Sandler L, Torous J. Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. BMJ Open. 2021 Mar 1;11(3):e047001. doi:10.1136/bmjopen-2020-047001 PubMed PMID: 33741674.
16. Babineau J. Product Review: Covidence (Systematic Review Software). J Can Health Libr Assoc. 2014 Aug 1;35(2):68. doi:10.5596/c14-016
17. Miller SA, Forrest JL. Enhancing your practice through evidence-based decision making: PICO, learning how to ask good questions. Journal of Evidence Based Dental Practice. 2001 Oct 1;1(2):136–41. doi:10.1016/S1532-3382(01)70024-3
18. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implementation Sci. 2022 Oct 29;17(1):75. doi:10.1186/s13012-022-01245-0
19. 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. doi:10.1371/journal.pone.0199284 PubMed PMID: 29944708; PubMed Central PMCID: PMC6019095.
20. 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. Gruneir A, editor. PLoS ONE. 2021 Dec 1;16(12):e0260696. doi:10.1371/journal.pone.0260696
21. Faqar-Uz-Zaman SF, Anantharajah L, Baumartz P, Sobotta P, Filmann N, Zmuc D, et al. The Diagnostic Efficacy of an App-based Diagnostic Health Care Application in the Emergency Room: eRadaR-Trial. A prospective, Double-blinded, Observational Study. Annals of Surgery. 2022 Nov;276(5):935–42. doi:10.1097/SLA.0000000000005614
22. Fraser HSF, Cohan G, Koehler C, Anderson J, Lawrence A, Pateña J, et al. Evaluation of Diagnostic and Triage Accuracy and Usability of a Symptom Checker in an Emergency Department: Observational Study. JMIR Mhealth Uhealth. 2022 Sep 19;10(9):e38364. doi:10.2196/38364
23. Hennemann S, Kuhn S, Witthöft M, Jungmann SM. Diagnostic Performance of an App-Based Symptom Checker in Mental Disorders: Comparative Study in Psychotherapy Outpatients. JMIR Ment Health. 2022 Jan 31;9(1):e32832. doi:10.2196/32832
24. Meer A, Rahm P, Schwendinger M, Vock M, Grunder B, Demurtas J, et al. A Symptom-Checker for Adult Patients Visiting an Interdisciplinary Emergency Care Center and the Safety of Patient Self-Triage: Real-Life Prospective Evaluation. Journal of Medical Internet Research. 2024 Jun 27;26(1):e58157. doi:10.2196/58157
25. Liu V, Kaila M, Koskela T. Triage Accuracy and the Safety of User-Initiated Symptom Assessment With an Electronic Symptom Checker in a Real-Life Setting: Instrument Validation Study. JMIR Human Factors. 2024 Sep 26;11(1):e55099. doi:10.2196/55099
26. Knitza J, Hasanaj R, Beyer J, Ganzer F, Slagman A, Bolanaki M, et al. Comparison of Two Symptom Checkers (Ada and Symptoma) in the Emergency Department: A

Randomized, Crossover, Head-to-Head, Double-Blinded Study (Preprint). Journal of Medical Internet Research. 2024 Jan 19. doi:10.2196/56514
27. 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. Journal of Medical Internet Research. 2020 Nov 30;22(11):e20549. doi:10.2196/20549
28. Miller S, Gilbert S, Virani V, Wicks P. Patients' Utilization and Perception of an Artificial Intelligence-Based Symptom Assessment and Advice Technology in a British Primary Care Waiting Room: Exploratory Pilot Study. JMIR Hum Factors. 2020 Jul 10;7(3):e19713. doi:10.2196/19713 PubMed PMID: 32540836.
29. Katayama Y, Kiyohara K, Hirose T, Matsuyama T, Ishida K, Nakao S, et al. A Mobile App for Self-Triage for Pediatric Emergency Patients in Japan: 4 Year Descriptive Epidemiological Study. JMIR Pediatrics and Parenting. 2021 Jun 30;4(2):e27581. doi:10.2196/27581
30. Knitza J, Muehlensiepen F, Ignatyev Y, Fuchs F, Mohn J, Simon D, et al. Patient's Perception of Digital Symptom Assessment Technologies in Rheumatology: Results From a Multicentre Study. Front Public Health. 2022 Feb 22;10:844669. doi:10.3389/fpubh.2022.844669
31. Gellert GA, Orzechowski PM, Price T, Kabat-Karabon A, Jaszczak J, Marcjasz N, et al. A multinational survey of patient utilization of and value conveyed through virtual symptom triage and healthcare referral. Front Public Health. 2023 Feb 2;10:1047291. doi:10.3389/fpubh.2022.1047291
32. Mahlknecht A, Engl A, Piccoliori G, Wiedermann CJ. Supporting primary care through symptom checking artificial intelligence: a study of patient and physician attitudes in Italian general practice. BMC Prim Care. 2023 Sep 4;24(1):174. doi:10.1186/s12875-023-02143-0
33. Wetzel AJ, Preiser C, Müller R, Joos S, Koch R, Henking T, et al. Unveiling Usage Patterns and Explaining Usage of Symptom Checker Apps: Explorative Longitudinal Mixed Methods Study. Journal of Medical Internet Research. 2024 Dec 9;26(1):e55161. doi:10.2196/55161
34. Winn AN, Somai M, Fergestrom N, Crotty BH. Association of Use of Online Symptom Checkers With Patients' Plans for Seeking Care. JAMA Netw Open. 2019 Dec 27;2(12):1–3. doi:10.1001/jamanetworkopen.2019.18561 PubMed PMID: 31880791; PubMed Central PMCID: PMC6991310.
35. Cotte F, Lourenço FD, Pereira MP, Kini N, Schmude M, Pimenta A, et al. From Advice to Action: Real-World Behaviour of Patients Using an Integrated Clinical AI for Navigating the Healthcare System [Internet]. medRxiv; 2025 [cited 2025 Aug 8]. p. 2025.07.24.25332138. Available from: https://www.medrxiv.org/content/10.1101/2025.07.24.25332138v1 doi:10.1101/2025.07.24.25332138
36. Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction. 2008 Jul 29;24(6):574–94. doi:10.1080/10447310802205776
37. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of Symptom Checkers for Self Diagnosis and Triage: Audit Study. BMJ. 2015 Jul 8;351:1–9. doi:10.1136/bmj.h3480 PubMed PMID: 26157077.
38. Hill MG, Sim M, Mills B. The Quality of Diagnosis and Triage Advice Provided by Free Online Symptom Checkers and Apps in Australia. Med J Aust. 2020;212(11):514–9. doi:10.5694/mja2.50600 PubMed PMID: 32391611.

39. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage Accuracy of Symptom Checker Apps: 5-Year Follow-up Evaluation. J Med Internet Res. 2022 May 10;24(5):e31810. doi:10.2196/31810
40. 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. doi:10.1136/bmjopen-2020-040269 PubMed PMID: 33328258.
41. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. Wilson FA, editor. PLoS ONE. 2021 Jul 15;16(7):e0254088. doi:10.1371/journal.pone.0254088
42. 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 Aug 6. doi:10.7759/cureus.16956
43. Sartini M, Carbone A, Demartini A, Giribone L, Oliva M, Spagnolo AM, et al. Overcrowding in Emergency Department: Causes, Consequences, and Solutions—A Narrative Review. Healthcare. 2022 Aug 25;10(9):1625–38. doi:10.3390/healthcare10091625
44. Kopka M, Wang SM, Kunz S, Schmid C, Feufel MA. Technology-supported self-triage decision making. npj Health Syst. 2025 Jan 25;2(1):1–11. doi:10.1038/s44401-024-00008-x
45. Aboueid S, Meyer S, Wallace JR, Mahajan S, Chaurasia A. Young Adults' Perspectives on the Use of Symptom Checkers for Self-Triage and Self-Diagnosis: Qualitative Study. JMIR Public Health Surveill. 2021 Jan 6;7(1):e22637. doi:10.2196/22637
46. Müller R, Klemmt M, Koch R, Ehni HJ, Henking T, Langmann E, et al. "That's just Future Medicine" - a qualitative study on users' experiences of symptom checker apps. BMC Med Ethics. 2024 Feb 16;25(1):17. doi:10.1186/s12910-024-01011-5
47. Oslislo S, Witt K, Kellner P, Schmidt-Traub H, Wrede C, Somasundaram R. Awareness of the Patient Service Telephone Number 116117 and Willingness to Use a Symptom Checker. Dtsch Arztebl Int. 2025 May;122(11):306–7. doi:10.3238/arztebl.m2025.0065 PubMed PMID: 40633101; PubMed Central PMCID: PMC12539685.
48. Kopka M, Scatturin L, Napierala H, Fürstenau D, Feufel MA, Balzer F, et al. Characteristics of Users and Nonusers of Symptom Checkers in Germany: Cross-Sectional Survey Study. J Med Internet Res. 2023 Jun 20;25:e46231. doi:10.2196/46231
49. El-Osta A, Riboli-Sasco E, Ammouri MA, Altalib S, Neves AL, Majeed A, et al. Factors Influencing the Use of Online Symptom Checkers in the United Kingdom: Cross-Sectional Study. JMIR Formative Research. 2025 Sep 15;9(1):e65314. doi:10.2196/65314
50. Deloitte GmbH Wirtschaftsprüfungsgesellschaft. Digitalisierung im Gesundheitswesen - Aktuelle Erkenntnisse zum Stand der Digitalisierung im Gesundheitswesen [Internet]. 2025 Sep 11 [cited 2025 Oct 18]. Available from: https://www.deloitte.com/de/de/Industries/life-sciences-health-care/research/digitalisierung-im-gesundheitswesen.html
51. Kopka M, He L, Feufel MA. Evaluating the accuracy of ChatGPT model versions for giving care-seeking advice. Commun Med. 2026 Feb 25. doi:10.1038/s43856-026-01466-0

Click here to access/download
Supporting Information
Supporting Information.docx