Accuracy of Online Symptom-Assessment Applications, Large Language Models, and Laypeople for Self-Triage Decisions: A Systematic Review

Marvin Kopka[1*], Niklas von Kalckreuth[1], and Markus A. Feufel[1]

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

[*] Correspondence: marvin.kopka@tu-berlin.de

Abstract

Symptom-Assessment Application (SAAs, e.g., NHS 111 online) that assist medical laypeople in deciding if and where to seek care (*self-triage*) are gaining popularity and their accuracy has been examined in numerous studies. With the public release of Large Language Models (LLMs, e.g., ChatGPT), their use in such decision-making processes is growing as well. However, there is currently no comprehensive evidence synthesis for LLMs, and no review has contextualized the accuracy of SAAs and LLMs relative to the accuracy of their users. Thus, this systematic review evaluates the self-triage accuracy of both SAAs and LLMs and compares them to the accuracy of medical laypeople. A total of 1549 studies were screened, with 19 included in the final analysis. The self-triage accuracy of SAAs was found to be moderate but highly variable (11.5 – 90.0%), while the accuracy of LLMs (57.8 – 76.0%) and laypeople (47.3 – 62.4%) was moderate with low variability. Despite some published recommendations to standardize evaluation methodologies, there remains considerable heterogeneity among studies. The use of SAAs should not be universally recommended or discouraged; rather, their utility should be assessed based on the specific use case and tool under consideration.

Introduction

Symptom-assessment Applications (SAAs, also known as online symptom checkers or digital triage tools) are digital platforms accessible via smartphones or websites that analyze symptoms using various methods[1,2]. They provide a diagnosis and recommendation whether and where medical care should be sought, a process referred to as *self-triage*[2]. SAAs are potentially useful for various stakeholders: health protection agencies may use the symptom input for syndromic surveillance[3], general practitioners and clinics can implement SAAs for patient (re-)direction[4,5] and medical laypeople can use them for assistance in health-related decisions[6]. Hence, they could make healthcare resource distribution more efficient and ultimately increase healthcare access and health equity by providing health advice and recommendations regardless of a person's socioeconomic status, education, or other determinants of health.

SAAs are increasingly used worldwide. For instance, the United Kingdom's National Health System (NHS) launched *NHS 111 online* in 2017[7] and Germany's Association of Statutory Health Insurance Physicians supplemented their triage hotline with the digital *PatientenNavi* in 2021[8,9]. Consequently, these tools perform millions of assessments annually, with about 7% of the German population using SAAs[7,10]. However, some studies raised concerns about their real-world utility and cost-effectiveness, as they did not seem to reduce healthcare utilization in an NHS evaluation study[11]. This is no surprise, as SAAs tend to be risk-averse and frequently provide users with a recommendation of higher urgency than necessary, making them seek care more often[2,12]. The opposite of this *over-triage* is *under-triage* – where users receive a recommendation of lower urgency than warranted – and poses potential safety risks to users[12,13]. Hence, both the safety and accuracy of SAAs have been subjects of several studies. Three systematic reviews have been published to synthesize the available evidence on SAAs so far and show that SAA accuracy is generally far from perfect, but they demonstrate a high variability between different apps[14–16].

As an alternative to SAAs, Large Language Models (LLMs) have been proposed in some studies[13,17]. After becoming available in 2022, they quickly garnered interest in the medical community for passing state licensing exams and, as a result, are now suggested as potential clinical decision support system [18–20]. Some studies have also tested LLMs with cases developed for SAAs and suggest them as decision support tools for medical laypeople as well[21,22]. Nevertheless, an evidence synthesis that reports the accuracy of LLMs for self-triage decisions is still missing.

All these studies on SAAs and LLMs have in common that they view these tools as sole decision-makers, and most researchers recommend or discourage their use without considering the accuracy of actual users. This perspective might overlook scenarios where – if users alone perform poorly – even suboptimal SAAs could be beneficial. Conversely, if users generally make very good decisions, SAAs might not offer any effective assistance. Although one study compared the accuracy of SAAs directly with that of laypeople[23], an evidence synthesis contextualizing the accuracy of SAAs and LLMs with the accuracy of laypeople is missing.

Therefore, this systematic review aims to extend previous reviews on SAA accuracy[14–16] by including studies on the accuracy of LLMs as an alternative to SAAs and medical laypeople as users. This comparison shifts the focus from SAAs and LLMs as the sole decision-making entity to considering their user group of medical laypeople as a benchmark against which their accuracy should be interpreted. Since specific diagnoses are of no use for medical laypeople – and are ultimately made and treated by medical professionals anyway – this review focuses on self-triage decisions only and deliberately excludes diagnostic accuracy to focus on user utility.

Methods

Eligibility Criteria

This study was preregistered on PROSPERO (ID: CRD42024563111) and adheres to the PRISMA reporting guideline[24]. Following a previous systematic review on SAAs[16], we included studies published from 2010 onward. We included all primary research articles (including preprints) that were published in English language. Our inclusion criteria comprised all patient demographics (including both vignette-based and real-world evidence studies) and various symptoms, but we excluded studies that focused solely on highly specialized tools or cases, such as only COVID-19 SAAs or COVID-19 cases only[25]. Our inclusion criteria required interventions to examine the self-triage advice of SAAs, LLMs, or laypeople. We excluded any studies that evaluated multiple tools being used simultaneously (e.g., SAAs combined with a telephone triage hotline) or tools that did not offer self-triage advice. Each study needed a gold standard solution for each case as a comparator. Studies that only rated the appropriateness of the received self-triage advice (e.g., on a 5-point Likert scale) without providing a direct solution to a case were excluded. Lastly, studies were required to quantitatively report (self-)triage accuracy by advising the most appropriate care facility, as this recommendation is the purpose that SAAs are developed for[2]. We excluded any studies that exclusively reported triage accuracy for emergency departments (e.g., using the Manchester Triage Scale or Emergency Severity Index) without considering other care facilities. Studies that reported only diagnostic accuracy without corresponding self-triage accuracy were excluded as well.

For the synthesis, we grouped studies according to the agent for which they provided self-triage accuracy estimates, i.e., for SAAs, LLMs and/or laypeople.

Search Strategy & Information Sources

We conducted our search on July 09, 2024, using the databases Web of Science, MEDLINE / Pubmed, and Scopus to identify relevant articles. The search was limited to studies published from 2010 onward and included English articles only. We developed an initial search string based on previously published systematic reviews of SAAs[14–16] and adapted it to focus on self-triage accuracy and to include LLMs and laypeople. This search string was refined until it identified all studies reporting self-triage that previous systematic reviews reported. The same refined search string was applied across all databases. The search string for Web of Science read:

> *AB=(app OR apps OR application OR artificial intelligence OR AI OR online OR web-based OR chatbot OR mobile OR computer-assisted OR internet OR smartphone OR phone OR web) OR AB=(symptom checker OR symptom check* OR symptom assessment app* OR symptom-assessment app* OR webmd OR symptomate OR ada OR yourmd OR mediktor OR buoy OR self-refer*) OR AB =(human OR layperson OR laypeople OR lay OR user OR non-professional OR non-clinician) OR AB =(GPT-3 OR ChatGPT OR GPT-4 OR GPT-4o OR Large Language Model OR LLM OR Claude OR Google Bard OR Mistral OR GPT)) AND AB =(self-triage OR triage OR symptom urgency OR dispositional advice OR self-assess*) AND AB=(accuracy OR correct)*

After identifying relevant articles, we conducted both forward and backward citation searches to identify additional studies, particularly preprints, that were not initially retrieved from the databases.

Data Extraction & Data Analysis

The studies were retrieved and imported into PicoPortal, where they were deduplicated. The titles and abstracts were screened by two researchers (MK & NvK) independently on PicoPortal. In cases of disagreement, both researchers re-examined the title and abstract and resolved conflicts through discussion. Afterwards, the full texts of each eligible study were independently screened by both researchers according to the pre-specified inclusion and exclusion criteria. Cases of disagreement were examined again, and conflicts were resolved through further discussion.

The data were extracted by both researchers independently using a standardized Excel template. The primary outcome focused on assessing the self-triage accuracy of SAAs, LLMs, and laypeople. For the secondary outcomes, the researchers extracted reported accuracy across the urgency levels and the specific self-triage accuracy of each individual SAA and LLM. To gain insights into differences in methodology, the data extraction form included the number of SAAs, LLMs, and laypeople, a brief description of the methods used, the number of triage levels in the study, the number of cases examined, the gold standard assignment process, the number of data inputters, other reported outcomes with respect to self-triage, as well as any conflicts of interest and funding sources. Any instances of missing data were coded as '*not available'*. Due to the varying methodologies among the included studies, the data were analyzed using narrative synthesis, as the estimates of the studies are not directly comparable. Nonetheless, a quantitative summary of the accuracy across all included studies is provided to show overall trends.

Risk of Bias

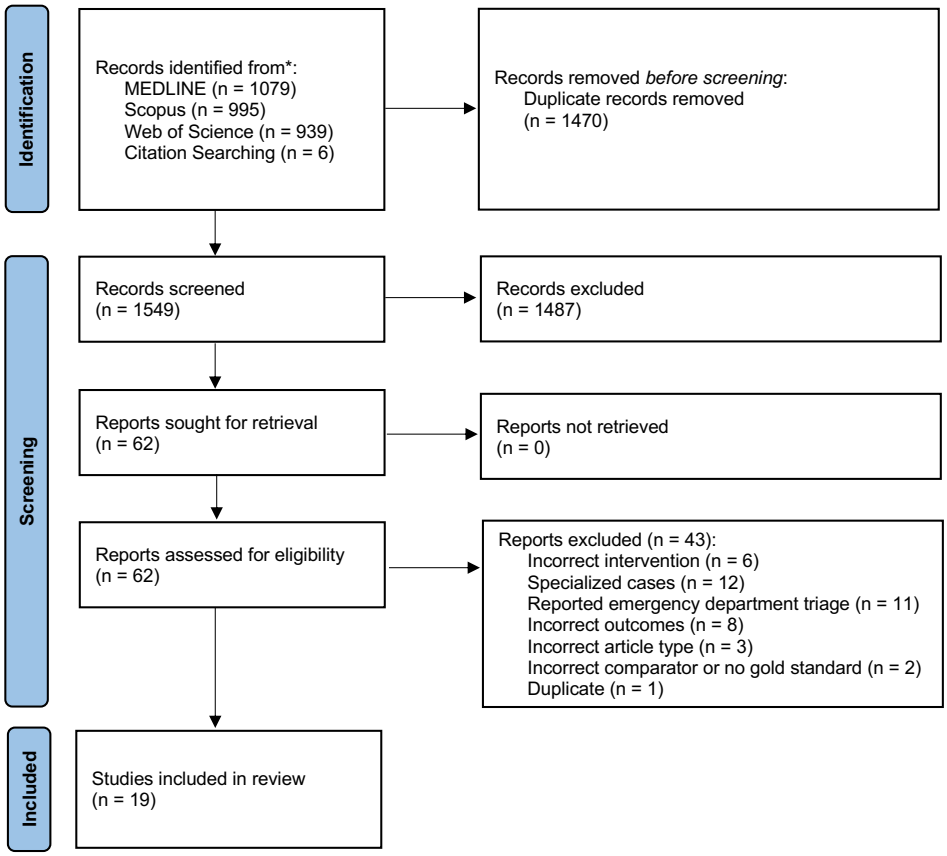
The risk of bias was assessed by two authors (MK & NvK) independently using the Quality Assessment of Diagnostic Accuracy Studies-2 tool (QUADAS-2)[26]. Any discrepancies were resolved through discussion again. QUADAS-2 uses four dimensions to rate the risk of bias and three dimensions to rate the applicability of a study to the research question. The risk of bias and applicability concerns were categorized into 'low', 'some concerns' and 'high'.

Results

Included Studies

In total, 3019 potentially eligible studies were identified (3013 using the database search and 6 using citation search). After excluding ineligible studies, for example because they referred to emergency department triage only[27], 19 studies were included in the review, see Figure 1.

Figure 1. PRISMA flow diagram detailing the study search and selection process.

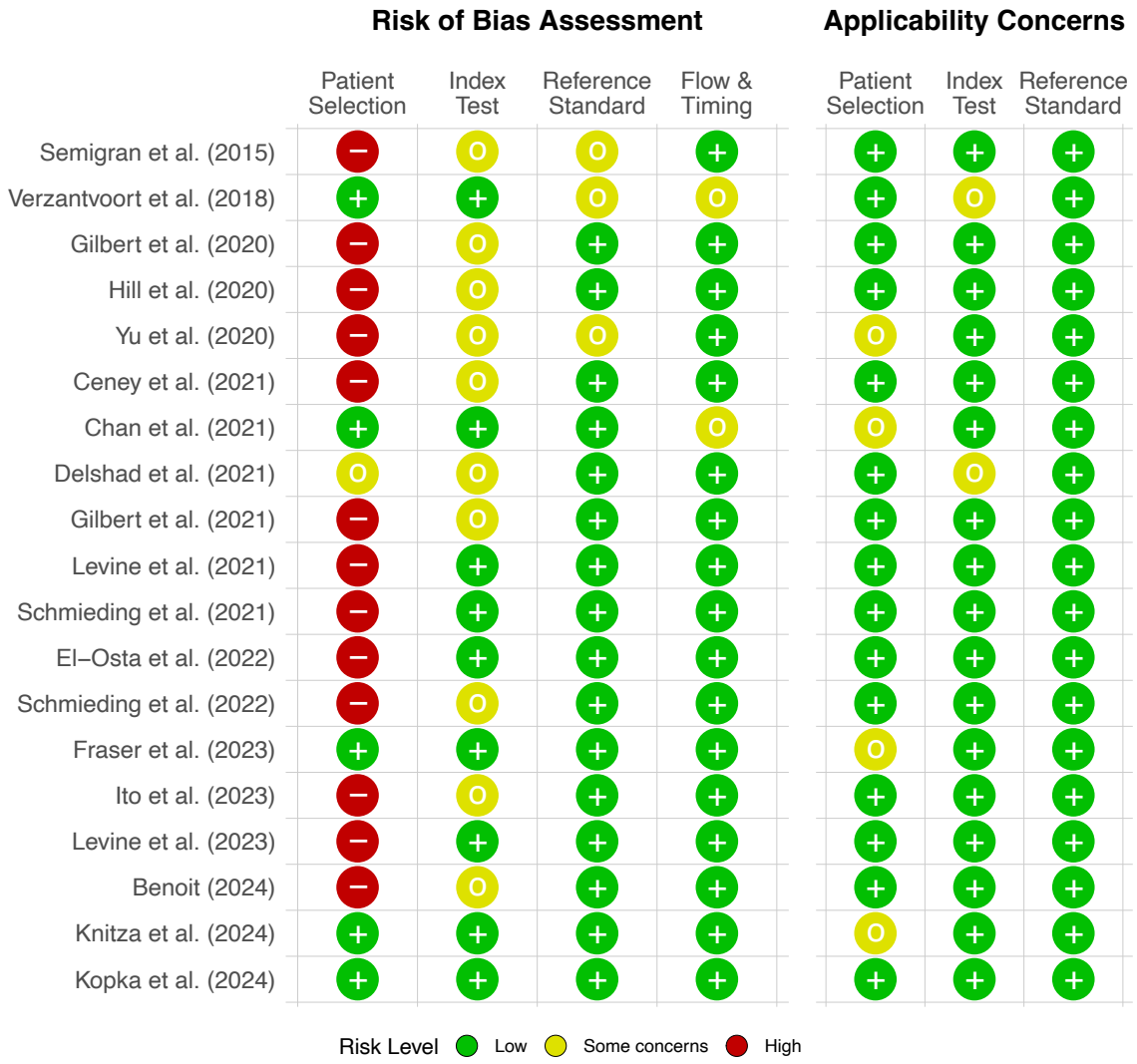


Most included studies (89%, 17/19) had at least one domain with a high risk of bias or some concerns, see Figure 2. The domain with the highest risk of bias was patient selection, as most studies used fictitious vignettes that were not based on real patient cases. For example, Semigran et al. used cases from text books and other medical resources that have a clear diagnosis assigned[2] and other studies used cases that were completely made up by clinicians based on their experience[28–30]. Both methods do not represent real cases that SAAs are normally approached with[17,31,32]. Only five studies had a low risk of bias because they included real patient cases: three studies used patients from emergency departments and primary care settings[13,33,34], one study directly surveyed SAA users[4], and one study used real patient cases from medical laypeople that were actively making self-triage decisions and sought technical assistance for this decision[17].

Index test was another domain in which many studies (53%, 10/19) have some risk of bias. Most of them did not report blinding of the inputter[2,12,21,22,29,35–39]. One study did not report how results from SAAs were obtained at all[36]. In the reference standard domain, only few studies have a moderate or high risk of bias (16%, 3/19). Those with concerns did not report how their gold standard was determined or used the judgement of one person only, e.g., the triage nurse in the emergency department[2,4,37]. Studies with some concerns regarding flow and timing had follow-up contact after several hours with a patient after using an app[4] or did not mention when cases were reviewed[34].

Applicability concerns were generally low. Most concerns comprised patient selection in studies that only used cases from the emergency department or a general practitioner setting, without including self-care cases[13,33,34,37]. Two studies had some concerns regarding the applicability of the index test, as one study used binary decisions only (visit a medical professional or not)[4] and another study did not provide information how SAA results were determined[36].

Figure 2. Risk of bias assessment and applicability concerns using QUADAS-2.



Study Characteristics

In total, 14 (74%) studies analyzed the self-triage accuracy of SAAs[2,4,12,13,17,29,33–40], four (21%) studies the accuracy of laypeople[17,23,30], and four (21%) studies the accuracy of LLMs[17,21,22,28]. For SAAs, three (21%) studies let patients enter their symptoms directly[4,33,34], three (21%) used real patient cases that were entered retrospectively[13,17,37], and the remaining 8 (57%) studies used fictitious case vignettes developed by medical professionals[2,12,29,35,36,38–40]. For studies on laypeople, one study (25%) asked participants how

they would rate the urgency of their own symptoms[34], one (25%) used real patient cases that were presented to laypeople[17] and two (50%) used fictitious vignettes phrased by medical professionals[23,30]. For LLMs, no study let patients enter symptoms themselves, one (25%) used real patient cases retrospectively[17] and three (75%) used fictitious vignettes[21,22,28].

Six (43%) studies examined only one SAA[4,33,34,36,39,40], two (14%) studies examined two SAAs[13,37] and six (43%) studies examined multiple SAAs[2,12,17,29,35,38], ranging from seven[29] to 23 different SAAs[2]. Studies on laypeople used sample sizes between 91 participants[23] and 5000 participants[30]. For LLMs, three (60%) studies examined only one LLM[21,22,28], whereas two (40%) studies examined multiple LLMs, ranging from two[13] to five models[17]. The most frequently included SAA was Ada Health and the most frequently included LLM was GPT-4. All study characteristics are summarized in Table 1.

Table 1. Characteristics of the included studies.

Authors (year)	Study Design	Description of cases	Number of SAAs, laypeople, and LLMs	Number of cases or vignettes
Semigran et al. (2015)	Cross-sectional vignette study	Fictitious cases derived from various medical resources	SAAs: 23 Laypeople: None LLMs: None	45
Verzantvoort et al. (2018)	Prospective cross-sectional cohort study	Patients used an app and entered their own symptoms	SAAs: 1 Laypeople: None LLMs: None	126
Gilbert et al. (2020)	Cross-sectional vignette study	Vignettes were created based on NHS triage calls (32%) and supplemented with fictitious vignettes developed from medical professionals (68%)	SAAs: 7 Laypeople: None LLMs: None	200
Hill et al. (2020)	Cross-sectional vignette study	Fictitious cases from Semigran et al. extended to include Australia-specific vignettes	SAAs: 19 Laypeople: None LLMs: None	48
Yu et al. (2020)	Retrospective cohort study	Real cases from the emergency department were transcribed to case vignettes	SAAs: 2 Laypeople: None LLMs: None	100
Ceney et al. (2021)	Cross-sectional vignette study	Fictitious case vignettes from Semigran et al.	SAAs: 10 Laypeople: None LLMs: None	50
Chan et al. (2021)	Prospective cohort study	Patients in the emergency department and family practices entered their symptoms into an app	SAAs: 23 Laypeople: None LLMs: None	581
Delshad et al. (2021)	Cross-sectional vignette study	Fictitious case vignettes were developed	SAAs: 1 Laypeople: None LLMs: None	50
Gilbert et al. (2021)	Cross-sectional vignette study	Fictitious case vignettes from Hill et al.	SAAs: 1 Laypeople: None	48

			LLMs: None	
Levine et al. (2021)	Cohort study	Fictitious case vignettes developed based on Semigran et al. and Hill et al.	SAAs: None Laypeople: 5000 LLMs: None	48
Schmieding et al. (2021)	Longitudinal vignette study	Fictitious case vignettes from Semigran et al.	SAAs: None Laypeople: 91 LLMs: None	45
El-Osta et al. (2022)	Cross-sectional vignette study	Fictitious vignettes created by medical professionals	SAAs: 1 Laypeople: None LLMs: None	139
Schmieding et al. (2022)	Cross-sectional vignette study	Fictitious case vignettes from Semigran et al.	SAAs: 17 Laypeople: None LLMs: None	45
Fraser et al. (2023)	Clinical data analysis	Patients in an emergency department entered their symptoms. Reports from the app were used to evaluate the tools	SAAs: 2 Laypeople: None LLMs: 2	37
Ito et al. (2023)	Cross-sectional vignette study	Fictitious case vignettes from Semigran et al.	SAAs: None Laypeople: None LLMs: 1	45
Levine et al. (2023)	Cross-sectional vignette study	Fictitious case vignettes developed based on Semigran et al. and Hill et al.	SAAs: None Laypeople: None LLMs: 1	48
Benoit (2024)	Cross-sectional vignette study	Fictitious case vignettes from Semigran et al.	SAAs: None Laypeople: None LLMs: 1	45
Knitza et al. (2024)	Cross-over randomized trial	Patients in the emergency department entered their symptoms	SAAs: 1 Laypeople: None LLMs: None	437
Kopka et al. (2024)	Retrospective cohort study	Real patient cases from an ‘ask the doctor’ platform where laypeople asked for help in their self-triage decision	SAAs: 12 Laypeople: 198 LLMs: 5	45

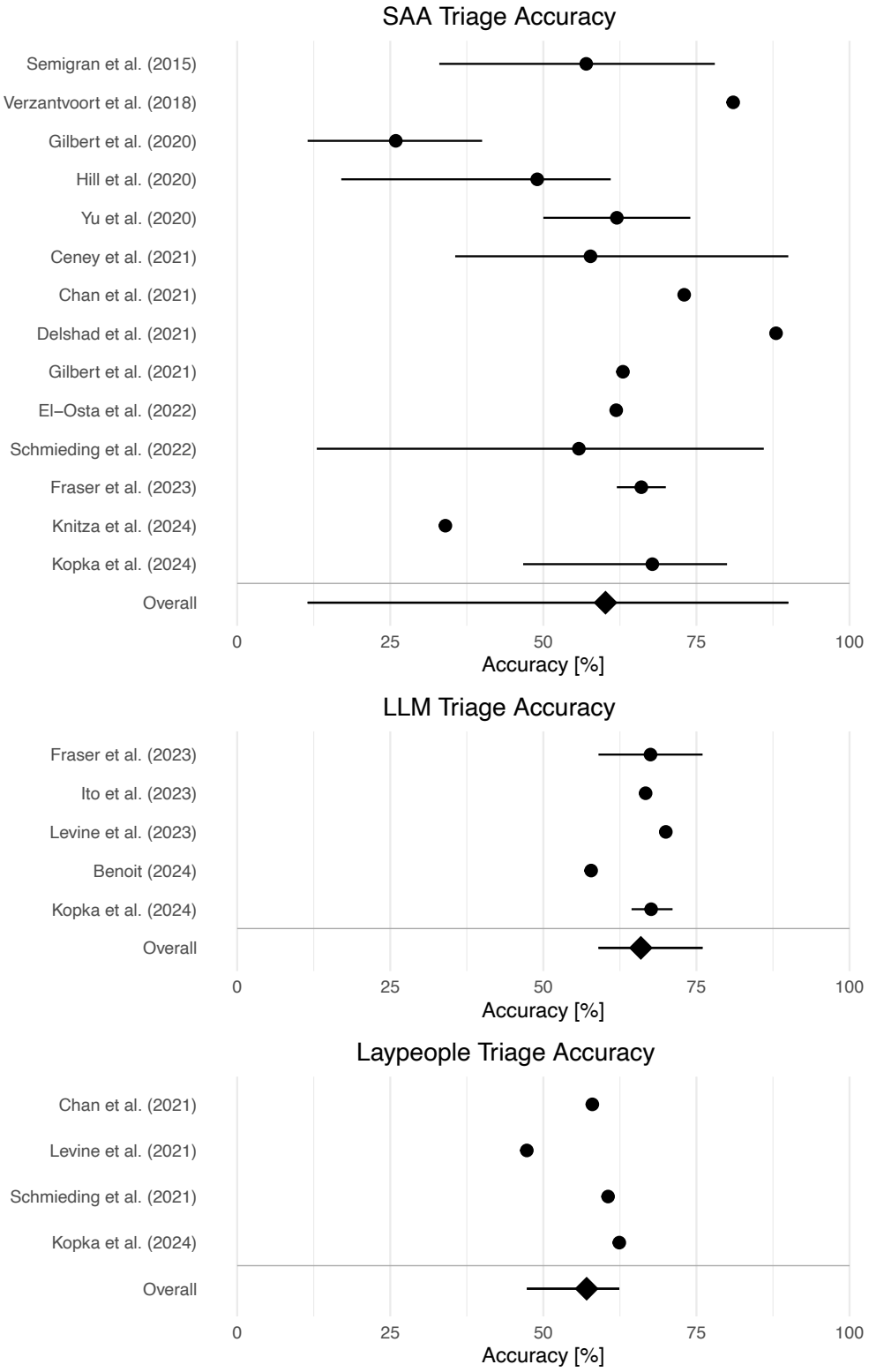
Self-Triage Accuracy

The reported average accuracy of SAAs ranged from 25.9% in a study by Gilbert et al.[29] to 88.0% in a study by Delshad et al.[36], see Figure 3. However, the self-triage accuracy varies widely between different systems: The lowest individual SAA accuracy of 11.5% was reported in the study by Gilbert et al.[29], while the highest accuracy of 90.0% was reported in a study by Ceney et al.[38].

The average accuracy of LLMs ranged from 57.8% in a study by Benoit[21] to 70.0% in a study by Levine et al.[28]. Individual accuracy estimates for LLMs had a relatively low variation compared to SAAs and ranged from 57.8% in the study by Benoit[21] to 76.0% in a study by Fraser et al.[13].

The reported average accuracy of laypeople had a lower variation and ranged from 47.3% to 62.4%, see Figure 3. No study reported the individual accuracy, making a comparison of worst- and best-performing individuals with SAAs and LLMs impossible.

Figure 3. Overview of reported self-triage accuracy estimates for Symptom-Assessment Applications (SAAs), laypeople and Large Language Models (LLMs)



Note. Points indicate the reported mean and lines indicate reported minimum and maximum accuracy values within a study. Studies on laypeople reported means only without information on minimum and maximum values.

Most studies reported not only average accuracy but also average accuracy across different self-triage levels. For all three agents, accuracy differed between different urgency levels. SAAs generally had a high accuracy for emergency cases (74.5%, with a range from 57% to 100%) and a lower accuracy for urgent (53.3% range from 23.0% to 92.2%) and non-emergent cases (69.7%, range from 55.0 to 82.5%)[2,33,34,37]. Their accuracy was the lowest for self-care cases (42.1%, range from 0.0% to 74.0%)[4,33].

LLMs had a moderate to high accuracy in emergency cases (66.7%, range from 50% to 86.7%) and reliably identified non-emergency cases (94.1%, range from 87% to 100%)[17,21,22,28]. However, they had a very low accuracy for self-care cases (10.8%, range from 6.15% to 16.7%)[17,28].

Laypeople had a relatively high accuracy in identifying emergency cases (67.9%, range from 57.5% to 78.6%) and non-emergency cases (70.8%, range from 68.4% to 73.2%)[17,23,30]. For self-care cases, they had a low accuracy (35.6%, range from 25.4% to 46.7%)[23,30], see Table 2.

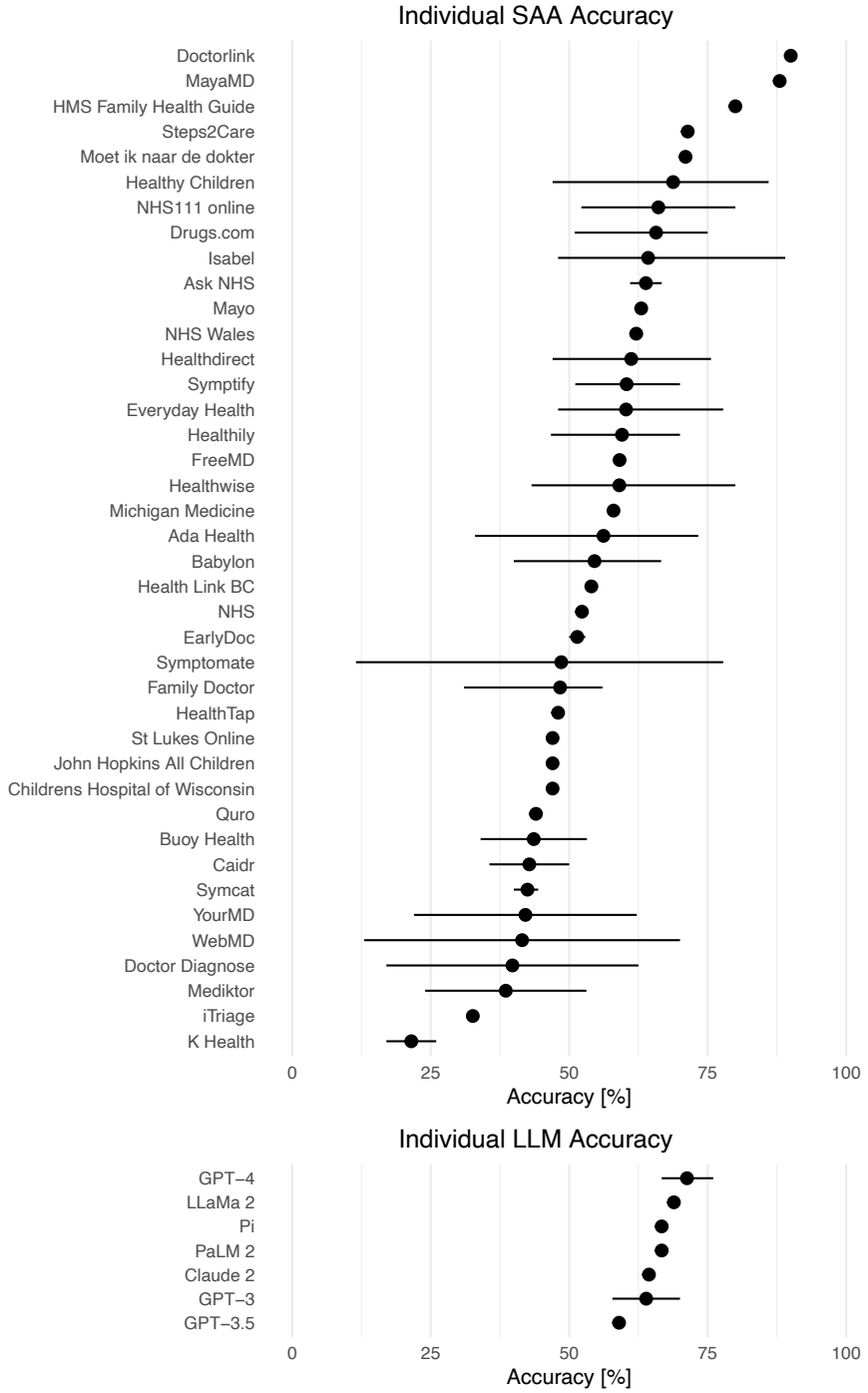
Table 2. Reported self-triage accuracy of Symptom-Assessment Applications, Laypeople, and Large Language Models across different self-triage levels.

Self-Triage Level	Symptom-Assessment Applications, % (Range)	Large Language Models, % (Range)	Laypeople, % (Range)
Emergency	74.5% (57%-100%)	66.7% (50.0% - 86.7%)	67.9% (57.5% - 78.6%)
Urgent Care	53.3% (23.0% - 92.2%	16.7% (n.a.)	50% (n.a.)
Non-Emergency/Non-Urgent	69.7% (55.0% - 82.5%)	94.1% (87% - 100%)	70.8% (68.4% - 73.2%)
Self-Care	42.1% (0.0 – 74.0%)	10.8% (6.15% - 16.7%	35.6% (25.4% - 46.7%)

Individual SAAs demonstrated a high variability: Doctorlink – which was examined in one study only – had the highest accuracy with 90.0%, whereas K Health had the lowest accuracy with 21.5%. When only examining SAAs that were tested across multiple studies, Healthy Children (68.8%) and NHS111 online (66.1%) had the highest accuracy among all SAAs. The spread between the accuracy reported in different studies for the same SAA was high as well. For example, accuracy values for Symptomate ranged from 11.5% to 77.8% (with a mean of 48.6%), see Figure 4.

For LLMs, the spread was relatively low. Although GPT-4 had the highest accuracy (71.3%), all of them scored between 59.0% and 71.3%. The accuracy between different studies only ranged from 66.7% to 76.0% for GPT-4, and 57.8% to 70.0% for GPT-3.

Figure 4. Overview of accuracy values reported for individual Symptom-Assessment Applications (SAAs) and large language models (LLMs) across multiple studies.



Note. Points indicate the mean of reported accuracy values and lines indicate minimum and maximum reported accuracy values. SAAs and LLMs without a line were examined in one study only. Since the methodology between studies differ and some are sponsored by the developer, the accuracy of these SAAs/LLMs should be interpreted with caution.

Methodology

The methodology varied between studies. Although most studies assigned the gold standard for each case using a physician panel of two or more physicians that independently rated the cases and resolved disagreements through discussion[17,33,35], some studies omitted independent ratings and directly used a physician discussion panel without letting them rate cases independently beforehand[29,36]. In other studies, the authors (who are physicians) assigned the gold standard themselves[28,30] or used the decision of a single triage nurse[4,37]. Most studies used only one person to input data into SAAs and LLMs[2,12,13,21,22,28,35,38,39]. Two studies employed two people[17,37], one study six people[40] and one study eight people[29]. Some studies used medical professionals as inputters[2,29], while others specifically let laypeople enter the symptoms[12,17]. Notably, only two studies mentioned blinding inputters to the gold standard solution[13,17]. Although most studies used three self-triage levels in their assignment[2,12,13,17,21–23], some used only two[4,37] (e.g., emergency or no emergency), and one study even used six levels[29]. Additionally reported self-triage outcomes varied between studies: One study used metrics from signal detection theory[4], three studies reported the comprehensiveness of an SAA[17,29,38], seven studies reported the inclination to over-/ and undertriage[12,13,17,23,33,34,37], and five studies reported the safety of advice[13,17,29,38,40]. One study additionally reported the Capability Comparison Score, which was developed specifically to compare SAAs[17,41].

Table 1. Methodological details of the included studies.

Authors (year)	Gold Standard	Number of inputters	Number of self-triage levels	Other outcomes reported
Semigran et al. (2015)	Correct diagnosis was part of medical resource; no information on self-triage level	1	3	None
Verzantvoort et al. (2018)	Triage nurse determined self-triage level after telephone interview	n.a.	2	Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value
Gilbert et al. (2020)	Assigned by physician panel	8	6	Comprehensiveness, Safety
Hill et al. (2020)	Two physicians and one emergency specialist rated cases, disagreement resolved through discussion	1	4	None

Yu et al. (2020)	Assigned triage level by triage nurse upon visiting emergency department	2	2	Over-/Undertriage
Ceney et al. (2021)	Taken from Semigran et al. and assessed against National Institute for Health and Care Excellence summaries	1	4	Comprehensiveness, Safety
Chan et al. (2021)	Decision from treating physician was reviewed by two physicians	n.a.	4	Over-/Undertriage
Delshad et al. (2021)	Several physicians from different institutions were asked about the most appropriate self-triage level. They were asked to develop a consensus	n.a.	4	None
Gilbert et al. (2021)	Taken from Hill et al.	1	4	None
Levine et al. (2021)	Assigned by two physicians	n.a.	4	None
Schmieding et al. (2021)	Taken from Semigran et al.	n.a.	3	Over-/Undertriage
El-Osta et al. (2022)	Multiple gold standards tested: general practitioners (GPs) that developed vignettes also assigned solution. 3 independent GPs were asked about correct self-triage level. Both solutions were pooled	6	3	Safety
Schmieding et al. (2022)	Taken from Semigran et al.	1	3	Over-/Undertriage, Binary Self-Triage decision
Fraser et al. (2023)	Three emergency department physicians rated each case	1	3	Safety, Overcaution
Ito et al. (2023)	Taken from Semigran et al.	1	3	None
Levine et al. (2023)	Assigned by two physicians	1	4	None
Benoit (2024)	Taken from Semigran et al.	1	3	None
Knitza et al. (2024)	Two physicians rated each case	n.a.	4	Over-/Undertriage
Kopka et al. (2024)	Panel of 2 physicians rated independently, disagreement resolved through discussion	2	3	Safety, Over-/Undertriage, Comprehensiveness, Capability Comparison Score

Discussion

This systematic review aimed to synthesize available evidence on self-triage accuracy, focusing not only on SAAs but also on LLMs as an alternative, and on laypeople as the user group. Our findings indicate that SAAs have a relatively low accuracy on average, but they also show that accuracy is highly dependent on the specific tool used. Most studies report a high spread between different SAAs and there is also high heterogeneity between the studies. However, when assessing individual SAAs across different studies, some tools seem to consistently perform well. For example, NHS 111 online was included in multiple studies and consistently showed moderate to high accuracy. Conversely, Mediktor showed a consistently low performance across multiple studies. Surprisingly, LLM accuracy does not have a high spread in comparison – all studies report values between 58% and 76% and the individual spread for LLMs across studies is minimal as well. The same holds true for laypeople: All the included studies report an accuracy between 47% and 62%, indicating that laypeople make decisions better than chance level, but far from perfect.

Our review, while including more recent studies, aligns with the findings from previous systematic reviews. These reviews consistently report that SAA accuracy is relatively low, but note that the variability between the tools is very high[14–16]. This variation is understandable, considering that SAAs are developed by different institutions, each using different methods and working with varying levels of funding. For example, some developers use simple rule-based algorithms, while others use Bayesian networks[42]. Based on varying accuracy levels, all reviews conclude that SAAs pose a safety risk and suggest that their use might not be encouraged. While the safety concerns are valid and important, it is noteworthy that laypeople also tend to make decisions with only moderate accuracy. Hence, leaving them unassisted in their self-triage decisions might not be a viable solution either and the interaction between laypeople and digital tools for self-triage warrants further investigation. Because the interaction is not fully understood, there's a risk that their combined errors could lead to even worse decisions than if each made a decision separately. Alternatively, their correct decisions might complement each other and increase the overall self-triage accuracy beyond the accuracy of each agent alone. Since humans make the final decision in the end, it is also important to understand how they include and compensate incorrect advice. One previous study on human-SAA interaction suggests that laypeople can increase their accuracy with well-performing SAAs, but not to the level of the SAA's isolated accuracy[43]. However, users were able to compensate incorrect recommendations and were not entirely dependent on the system. Thus, the study overall suggests that errors do not add up, but rather that laypeople can successfully use SAAs – even if the system's accuracy is not perfect – and compensate incorrect recommendations.

When comparing SAAs, LLMs and laypeople, it is also important to examine the specific decisions that are made. The accuracy of all three agents differed drastically between the urgency levels of the presented cases. Whereas all performed relatively well in identify emergencies (with laypeople and SAAs showing very similar accuracy), their accuracy in self-care cases varied drastically. SAAs had a variation between 0 and 74%, while laypeople solved between 25% and 47% of these cases correctly. LLMs rarely advised self-care at all and thus had an accuracy below 20%. These findings indicate that laypeople may not require assistance in identifying emergencies but could profit from support in identifying self-care cases. However, LLMs are not well-suited for this task and only certain SAAs can be helpful in this regard. Some previous studies suggest dividing the urgency levels into two steps to better reflect how laypeople make self-triage decisions: First, they determine whether their symptoms require medical attention at all, and if so, they then decide where to seek care[12,44]. Considering our findings, laypeople may need more assistance in determining whether their symptoms require medical attention rather than deciding where to seek care; this could be the decision in which SAAs and other tools could be more beneficial. Thus, it is

not universally advisable to recommend or dismiss using SAAs. Rather, recommendations should depend on the specific implementation use case. When deciding between emergency and non-emergency care, LLMs might be helpful due to their high accuracy in this regard. However, when deciding if care is needed at all, LLMs generally do not offer any assistance and only some SAAs are useful.

For users, a general recommendation for using any SAA or LLM is not advisable. However, some tools might be helpful depending on the specific decision they need to make. For instance, when deciding between emergency and non-emergency care, LLMs and specifically GPT-4 might be beneficial, as it has been found to be relatively safe and accurate in this decision[13,17]. On the other hand, if users want to determine whether their symptoms warrant any medical attention at all, using a tool like NHS 111 online could be helpful due to its high accuracy in this decision. Nevertheless, users should always use these tools with caution and cross-verify the recommendations with additional information sources and critical thinking.

For evaluators (such as other researchers, implementers or policymakers), a standardized evaluation process is essential. The primary quality risk in current evaluations is the use of fictitious case vignettes that do not represent real patients[15–17,45]. Although using these vignettes is convenient and resource-efficient, they often yield results that are not generalizable to real-world settings[17,45]. Since using real patients who enter their own symptoms might not be feasible and cannot be applied to evaluate multiple SAAs, a cost-effective alternative could involve using real patient descriptions that are entered into SAAs. A procedure for that is available with the RepVig framework[17]. Afterwards, specific SAAs can be tested with actual patients in a clinical trial to validate positive findings. Alongside the type of cases used in testing SAAs, other methodological variations influence the outcomes, such as the number of inputters, the gold standard assignment, the metrics, and the number of self-triage levels that are reported. Although recent studies provide specific recommendations for these issues[17,40,41,46], they are rarely being applied yet. For example, Meczner et al. examined inputter variability and suggest using standardized instructions, multiple inputters, and a pooled accuracy metric to reflect the recommendations that multiple inputters receive[46]. El-Osta et al. investigated the gold standard assignment process and concluded that pooling decisions of two independent physician panels gets closer to the best solution than using one physician panel or a single person only[40]. Kopka et al. reviewed the metrics reported in other studies and proposed a set of standardized metrics to better understand the strengths and weaknesses of an SAA[41,45]. Lastly, standardizing the number of self-triage levels could improve comparability both within and between studies. Most studies use three or four levels, yet not all SAAs provide an ‘urgent care’ recommendation[12]. Thus, we suggest that using three triage levels – as originally proposed by Semigran et al.[2] – might increase comparability.

This review has several limitations. First, unlike previous systematic reviews, we focused solely on self-triage accuracy rather than diagnostic accuracy. This choice was motivated by the relevance to laypeople: While a preliminary diagnosis might lead to further information-seeking, a correct diagnosis often requires medical tests or more details that are not accessible to laypeople[47]. Ultimately, diagnoses are made by medical professionals anyway. As noted in several studies already, aiding laypeople in finding the most suitable care pathway is a more effective use case for these tools[2,15]. This perspective is also reflected in the included studies, as only one study involving laypeople assessed their diagnostic accuracy[30] – unliked numerous SAA studies that typically evaluate both diagnostic and self-triage accuracy[15,16].

Another limitation concerns the number of included studies: Although many studies test the accuracy of SAAs, only few studies examine the accuracy of laypeople. A potential reason might be the novelty of the field, and that researchers thus initially focus on evaluating the technological aspects before progressing

to more realistic scenarios that include human participants – akin to lab studies that first assess effects under controlled conditions and then move on to observational studies to confirm these effects in the real world. Similarly, the number of studies evaluating LLMs was also low. Because LLMs were first released to the public with ChatGPT in 2022, the technology can be considered relatively new and there has been limited time to conduct and publish studies on their accuracy. Although these is a vast body of medical research on LLMs already, most of it has focused on their ability to pass pre-specified exams like board tests or other diagnostic tasks[20,48]. As more time passes, we can expect to see additional evidence on the self-triage accuracy of LLMs and conducting an updated systematic review on their accuracy might be insightful. This is particularly relevant, because LLMs seem to quickly improve their accuracy across various tasks with new iterations[49].

Lastly, the methodologies varied among the included studies, which makes the direct comparison of accuracy estimates complicated. While differences in methods are more pronounced for diagnostic accuracy – e.g., some studies evaluate only the first diagnosis while others consider the top 3, 5, or 10[15,16] – the methods for self-triage accuracy also vary. A major issue concerns using fictitious vignettes in most studies that were phrased by clinicians and developed based on clear case descriptions from medical education resources or from physicians' experience. Although these vignettes represent clear cases with a definitive solution, they do not accurately reflect real cases that SAAs are approached with[17,32,50,51]. As a result, generalizability of most included studies is questionable.

Conclusions

In conclusion, the performance of SAAs compared to laypeople varies; some SAAs outperform laypeople, while others do not. Therefore, universally recommending SAAs to the public may not be advisable, but well-performing SAAs might warrant a recommendation if their safety is assured. LLMs showed less variability and higher accuracy than many SAAs in handling both emergency and non-emergency cases, which suggests a potential usefulness in these scenarios. Nonetheless, they rarely recommend self-care and can thus not be universally endorsed either.

Deciding which tools to use should be based on the specific use case. For users confident that their symptoms require medical attention, a high-performing SAA or LLM could be beneficial. However, for those uncertain whether their symptoms warrant medical attention at all, SAAs that effectively differentiate between self-care and medical care could be useful, while LLMs in their current form do not provide any assistance in this decision-making process. Although general endorsement of SAAs or LLMs is not recommended, their use should not be outright discouraged either. The appropriateness of these tools depends on the specific use case and the particular tool that is considered.

References

1. Napierala, H. *et al.* Examining the impact of a symptom assessment application on patient-physician interaction among self-referred walk-in patients in the emergency department (AKUSYM): study protocol for a multi-center, randomized controlled, parallel-group superiority trial. *Trials* **23**, 791 (2022).
2. Semigran, H. L., Linder, J. A., Gidengil, C. & Mehrotra, A. Evaluation of Symptom Checkers for Self Diagnosis and Triage: Audit Study. *BMJ* **351**, 1–9 (2015).
3. Elliot, A. J. *et al.* Internet-based remote health self-checker symptom data as an adjuvant to a national syndromic surveillance system. *Epidemiol. Infect.* **143**, 3416–3422 (2015).
4. Verzantvoort, N. C. M., Teunis, T., Verheij, T. J. M. & van der Velden, A. W. Self-Triage for Acute Primary Care via a Smartphone Application: Practical, Safe and Efficient? *PLoS One* **13**, e0199284 (2018).
5. Poote, A. E., French, D. P., Dale, J. & Powell, J. A study of automated self-assessment in a primary care student health centre setting. *J Telemed Telecare* **20**, 123–127 (2014).
6. Pairon, A., Philips, H. & Verhoeven, V. A scoping review on the use and usefulness of online symptom checkers and triage systems: How to proceed? *Front. Med.* **9**, 1040926 (2023).
7. Turnbull, J., Prichard, J., MacLellan, J. & Pope, C. eHealth Literacy and the Use of NHS 111 Online Urgent Care Service in England: Cross-Sectional Survey. *Journal of Medical Internet Research* **26**, e50376 (2024).
8. Wetzel, A.-J. *et al.* 'Better see a doctor?' Status quo of symptom checker apps in Germany: A cross-sectional survey with a mixed-methods design (CHECK.APP). *DIGITAL HEALTH* **10**, 20552076241231555 (2024).
9. Zentralinstitut Kassenärztliche Versorgung. „Patienten-Navi online" der 116117 bietet Hilfesuchenden digitale Selbsteinschätzung medizinischer Beschwerden. https://www.zi.de/das-zi/medien/medieninformationen-und-statements/detailansicht/7-dezember-2021 (2021).
10. Kopka, M. *et al.* Characteristics of Users and Nonusers of Symptom Checkers in Germany: Cross-Sectional Survey Study. *J Med Internet Res* **25**, e46231 (2023).
11. Simpson, R. M., Jacques, R. M., Nicholl, J., Stone, T. & Turner, J. Measuring the impact introducing NHS 111 online had on the NHS 111 telephone service and the wider NHS urgent care system: an observational study. *BMJ Open* **12**, e058964 (2022).
12. Schmieding, M. L. *et al.* Triage Accuracy of Symptom Checker Apps: 5-Year Follow-up Evaluation. *J Med Internet Res* **24**, e31810 (2022).
13. Fraser, H. *et al.* Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study. *JMIR Mhealth Uhealth* **11**, e49995 (2023).
14. Chambers, D. *et al.* Digital and Online Symptom Checkers and Health Assessment/Triage Services for Urgent Health Problems: Systematic Review. *BMJ Open* **9**, e027743 (2019).
15. Wallace, W. *et al.* The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *npj Digit. Med.* **5**, 118 (2022).
16. Riboli-Sasco, E. *et al.* Triage and Diagnostic Accuracy of Online Symptom Checkers: Systematic Review. *J Med Internet Res* **25**, e43803 (2023).
17. Kopka, M. *et al.* Evaluating self-triage accuracy of laypeople, symptom-assessment apps, and large language models: A framework for case vignette development using a representative design approach (RepVig). 2024.04.02.24305193 Preprint at https://doi.org/10.1101/2024.04.02.24305193 (2024).
18. Jung, L. B. *et al.* ChatGPT passes German state examination in medicine with picture questions omitted. *Deutsches Ärzteblatt international* (2023) doi:10.3238/arztebl.m2023.0113.
19. Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* **2**, e0000198 (2023).
20. Liu, M. *et al.* Performance of ChatGPT Across Different Versions in Medical Licensing

Examinations Worldwide: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research* **26**, e60807 (2024).
21. Benoit, J. R. A. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. Preprint at https://doi.org/10.1101/2023.02.04.23285478 (2023).
22. Ito, N. *et al.* The Accuracy and Potential Racial and Ethnic Biases of GPT-4 in the Diagnosis and Triage of Health Conditions: Evaluation Study. *JMIR Med Educ* **9**, e47532 (2023).
23. Schmieding, M. L., Mörgeli, R., Schmieding, M. A. L., Feufel, M. A. & Balzer, F. Benchmarking Triage Capability of Symptom Checkers Against That of Medical Laypersons: Survey Study. *J Med Internet Res* **23**, e24475 (2021).
24. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev* **10**, 89 (2021).
25. Munsch, N. *et al.* Diagnostic Accuracy of Web-Based COVID-19 Symptom Checkers: Comparison Study. *Journal of Medical Internet Research* **22**, e21299 (2020).
26. Whiting, P. F. *et al.* QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* **155**, 529–536 (2011).
27. Karlafti, E. *et al.* Support Systems of Clinical Decisions in the Triage of the Emergency Department Using Artificial Intelligence: The Efficiency to Support Triage. *AML* **30**, 2 (2023).
28. Levine, D. M. *et al. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model*. http://medrxiv.org/lookup/doi/10.1101/2023.01.30.23285067 (2023) doi:10.1101/2023.01.30.23285067.
29. Gilbert, S. *et al.* How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. *BMJ Open* **10**, e040269 (2020).
30. Levine, D. M. & Mehrotra, A. Assessment of Diagnosis and Triage in Validated Case Vignettes Among Nonphysicians Before and After Internet Search. *JAMA Netw Open* **4**, e213287 (2021).
31. 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* **24**, e37408 (2022).
32. Arellano Carmona, K., Chittamuru, D., Kravitz, R. L., Ramondt, S. & Ramírez, A. S. Health Information Seeking From an Intelligent Web-Based Symptom Checker: Cross-sectional Questionnaire Study. *J Med Internet Res* **24**, e36322 (2022).
33. Knitza, J. *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) doi:10.2196/56514.
34. Chan, F. *et al.* Performance of a new symptom checker in patient triage: Canadian cohort study. *PLoS ONE* **16**, e0260696 (2021).
35. Hill, M. G., Sim, M. & Mills, B. The Quality of Diagnosis and Triage Advice Provided by Free Online Symptom Checkers and Apps in Australia. *Med J Aust* **212**, 514–519 (2020).
36. Delshad, S., Dontaraju, V. S. & Chengat, V. Artificial Intelligence-Based Application Provides Accurate Medical Triage Advice When Compared to Consensus Decisions of Healthcare Providers. *Cureus* (2021) doi:10.7759/cureus.16956.
37. Yu, S. W. Y. *et al.* Triage Accuracy of Online Symptom Checkers for Accident and Emergency Department Patients. *Hong Kong Journal of Emergency Medicine* **27**, 217–222 (2020).
38. Ceney, A. *et al.* Accuracy of online symptom checkers and the potential impact on service utilisation. *PLoS ONE* **16**, e0254088 (2021).
39. Gilbert, S., Fenech, M., Upadhyay, S., Wicks, P. & Novorol, C. Quality of condition suggestions and urgency advice provided by the Ada symptom assessment app evaluated with vignettes optimised for Australia. *Aust. J. Primary Health* **27**, 377–381 (2021).
40. El-Osta, A. *et al.* What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. *BMJ Open* **12**, e053566 (2022).
41. Kopka, M. & Feufel, M. A. Software symptomcheckR: an R package for analyzing and visualizing symptom checker triage performance. *BMC Digit Health* **2**, 43 (2024).

42. Ćirković, A. Evaluation of Four Artificial Intelligence–Assisted Self-Diagnosis Apps on Three Diagnoses: Two-Year Follow-Up Study. *J Med Internet Res* **22**, e18097 (2020).
43. Kopka, M., Wang, S. M., Kunz, S., Schmid, C. & Feufel, M. A. Technology-Supported Self-Triage Decision Making: A Mixed-Methods Study. 2024.09.12.24313558 Preprint at https://doi.org/10.1101/2024.09.12.24313558 (2024).
44. Kopka, M., Feufel, M. A., Balzer, F. & Schmieding, M. L. The Triage Capability of Laypersons: Retrospective Exploratory Analysis. *JMIR Form Res* **6**, e38977 (2022).
45. Kopka, M., Feufel, M. A., Berner, E. S. & Schmieding, M. L. How suitable are clinical vignettes for the evaluation of symptom checker apps? A test theoretical perspective. *DIGITAL HEALTH* **9**, 20552076231194929 (2023).
46. Meczner, A. *et al. Accuracy as a Composite Measure for the Assessment of Online Symptom Checkers in Vignette Studies: Evaluation of Current Practice and Recommendations (Preprint)*. http://preprints.jmir.org/preprint/49907 (2023) doi:10.2196/preprints.49907.
47. Von Lengerke, T. Distinctiveness of disease prototypes in lay illness diagnosis: An exploratory observational study. *Psychology, Health & Medicine* **10**, 108–121 (2005).
48. Garg, R. K. *et al.* Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. *Health Promot Perspect* **13**, 183–191 (2023).
49. Lin, J. C., Younessi, D. N., Kurapati, S. S., Tang, O. Y. & Scott, I. U. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. *Eye* **37**, 3694–3695 (2023).
50. Mosier, K. L. & Kirlik, A. Brunswik'S Lens Model in Human Factors Research: Modern Applications of a Classic Theory. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **48**, 350–354 (2004).
51. Nadler, I. & Sanderson, P. M. Using Brunswik's Probabilistic Functionalism to Test How Clinicians Make Judgments in Simulated Neonatal Resuscitation Scenarios. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **55**, 743–747 (2011).

Competing interests

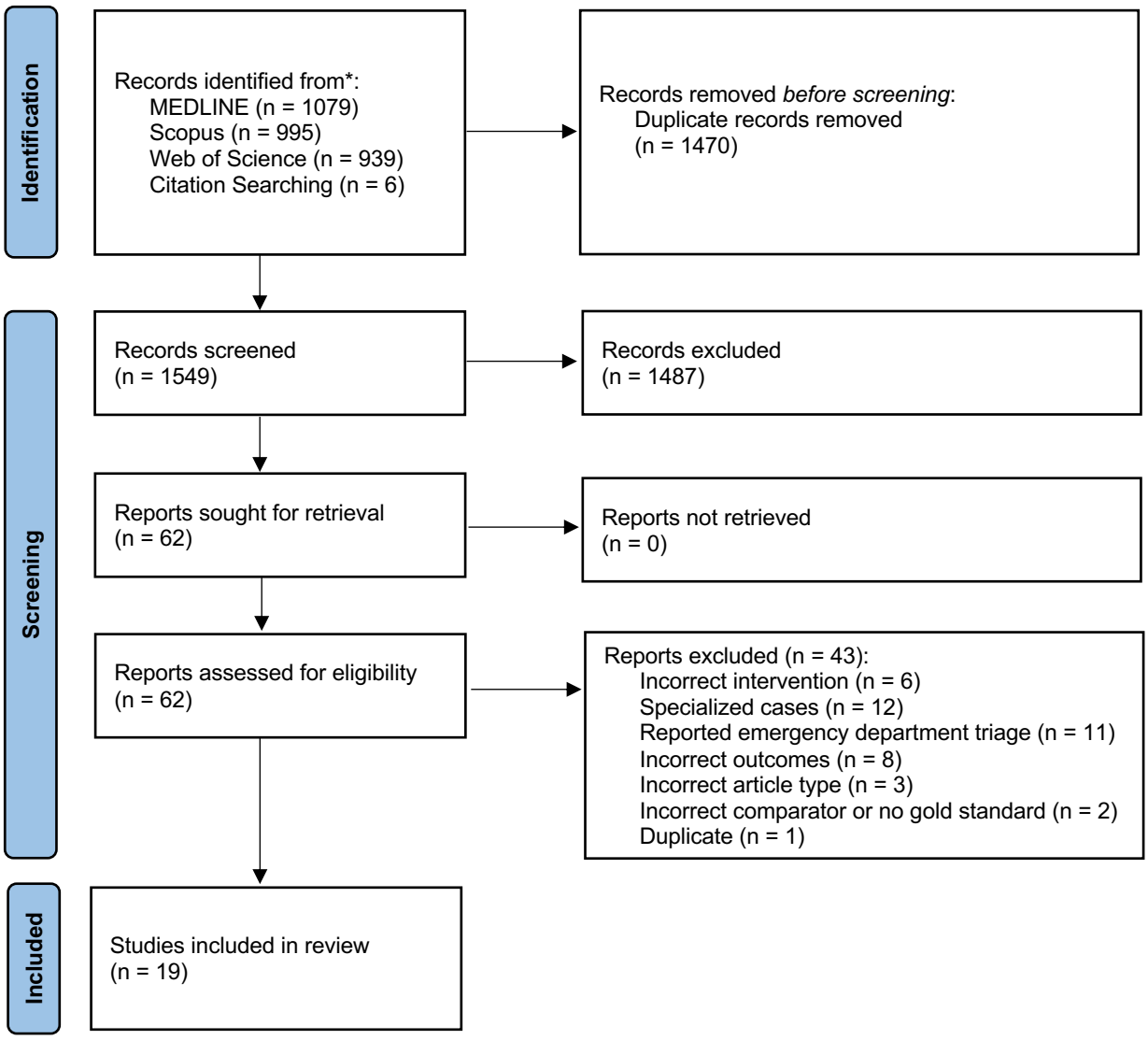
The authors declare no competing interests.

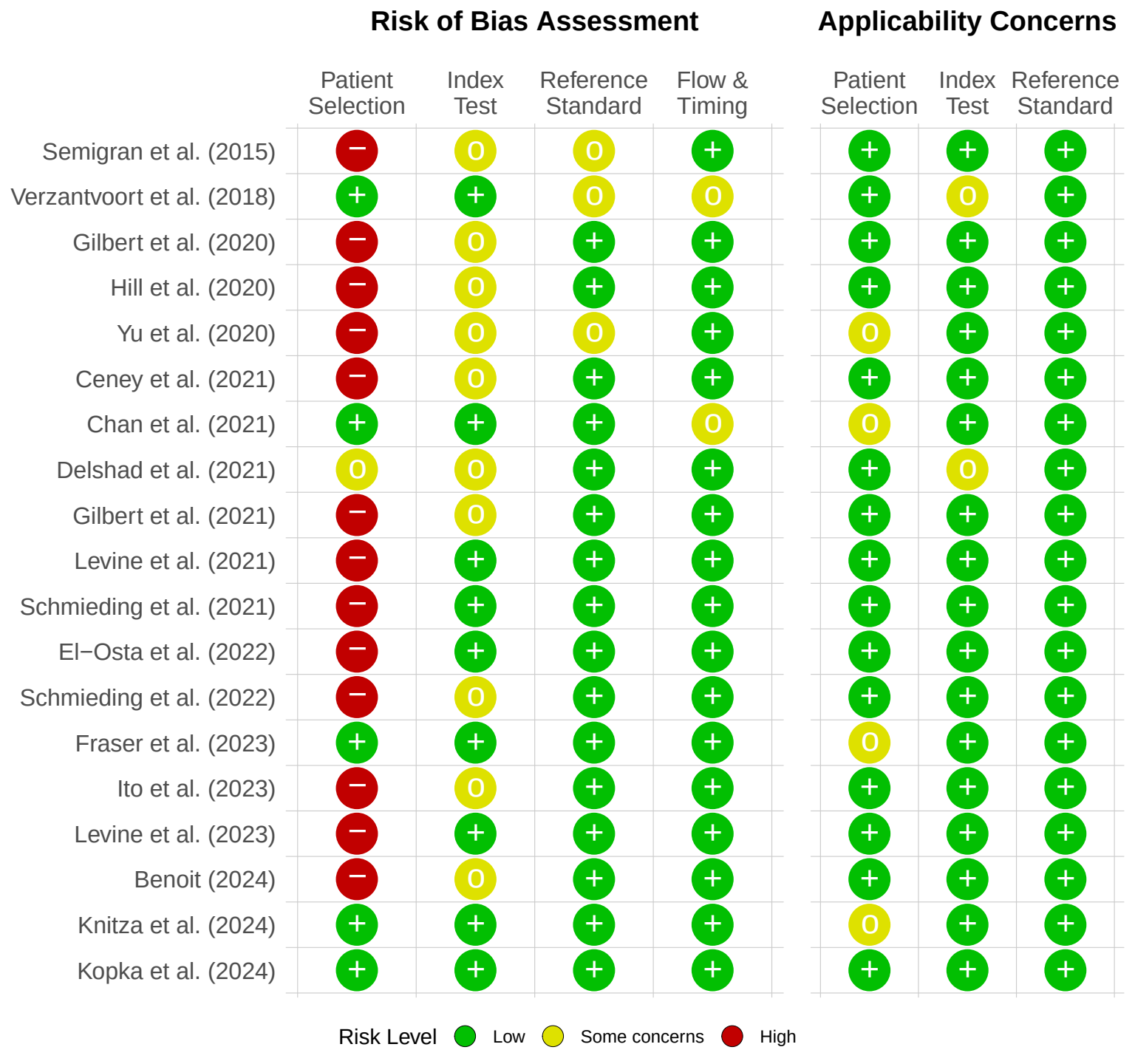
Author contributions

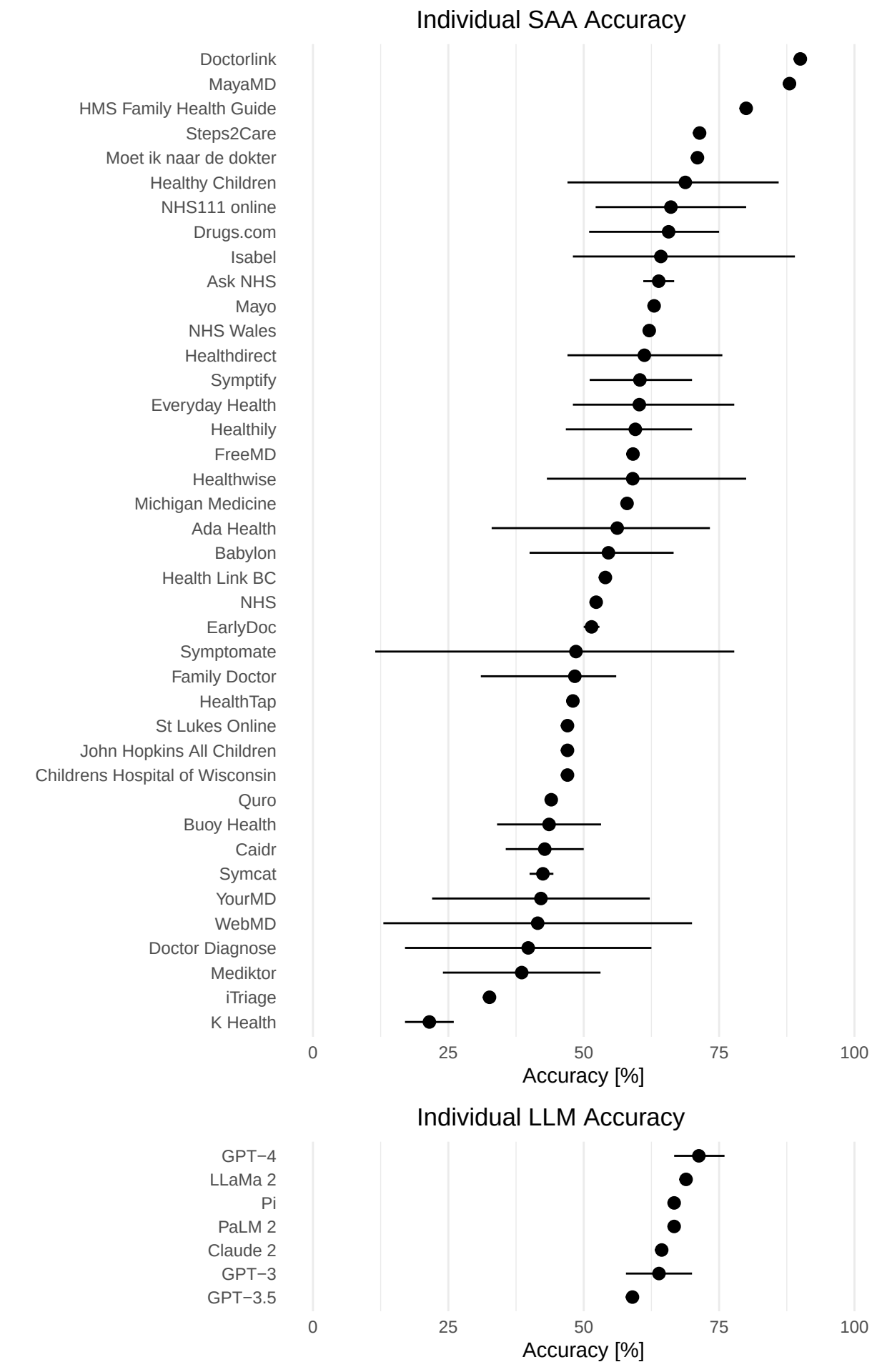
MK conceived of the study. MK and NvK conducted the screening and data extraction. MK conducted the data analysis and wrote the first draft of the manuscript. All authors provided critical input and worked on manuscript development.

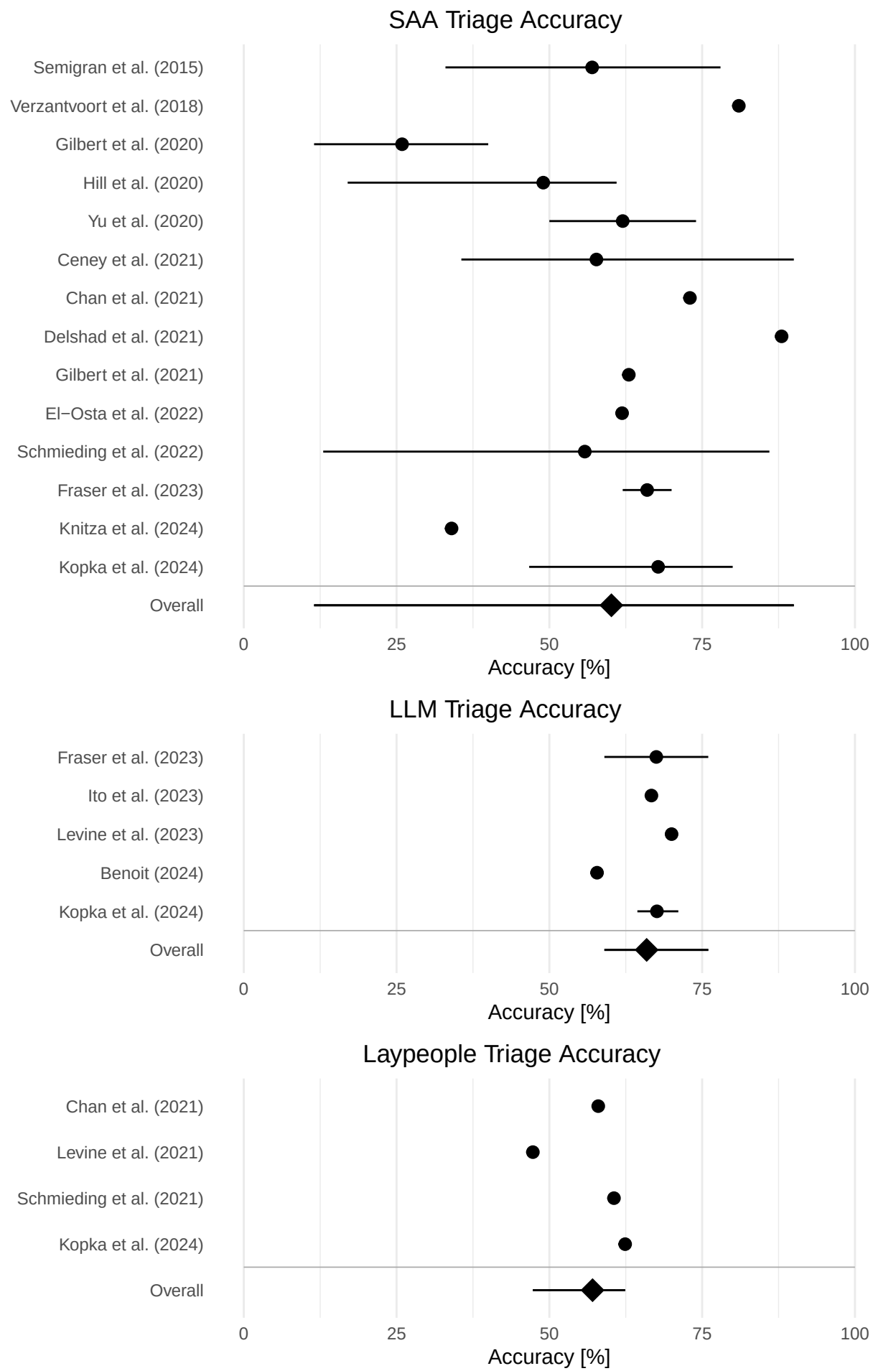
Data availability

The search strategy can be found in the Methods section and all information on studies are cited. Any additional data is available upon request.




Risk of Bias Assessment
Applicability Concerns
Patient Selection
Index Test
Reference Standard
Flow & Timing
Patient Selection
Index Test
Reference Standard
Semigran et al. (2015)
Verzantvoort et al. (2018)
Gilbert et al. (2020)
Hill et al. (2020)
Yu et al. (2020)
Ceney et al. (2021)
Chan et al. (2021)
Delshad et al. (2021)
Gilbert et al. (2021)
Levine et al. (2021)
Schmieding et al. (2021)
El−Osta et al. (2022)
Schmieding et al. (2022)
Fraser et al. (2023)
Ito et al. (2023)
Levine et al. (2023)
Benoit (2024)
Knitza et al. (2024)
Kopka et al. (2024)
Risk Level
Low
Some concerns
High


Individual SAA Accuracy
Doctorlink
MayaMD
HMS Family Health Guide
Steps2Care
Moet ik naar de dokter
Healthy Children
NHS111 online
Drugs.com
Isabel
Ask NHS
Mayo
NHS Wales
Healthdirect
Symptify
Everyday Health
Healthily
FreeMD
Healthwise
Michigan Medicine
Ada Health
Babylon
Health Link BC
NHS
EarlyDoc
Symptomate
Family Doctor
HealthTap
St Lukes Online
John Hopkins All Children
Childrens Hospital of Wisconsin
Quro
Buoy Health
Caidr
Symcat
YourMD
WebMD
Doctor Diagnose
Mediktor
iTriage
K Health
0
25
50
75
100
Accuracy [%]
Individual LLM Accuracy
GPT−4
LLaMa 2
Pi
PaLM 2
Claude 2
GPT−3
GPT−3.5
0
25
50
75
100
Accuracy [%]


SAA Triage Accuracy
Semigran et al. (2015)
Verzantvoort et al. (2018)
Gilbert et al. (2020)
Hill et al. (2020)
Yu et al. (2020)
Ceney et al. (2021)
Chan et al. (2021)
Delshad et al. (2021)
Gilbert et al. (2021)
El−Osta et al. (2022)
Schmieding et al. (2022)
Fraser et al. (2023)
Knitza et al. (2024)
Kopka et al. (2024)
Overall
0
25
50
75
100
Accuracy [%]
LLM Triage Accuracy
Fraser et al. (2023)
Ito et al. (2023)
Levine et al. (2023)
Benoit (2024)
Kopka et al. (2024)
Overall
0
25
50
75
100
Accuracy [%]
Laypeople Triage Accuracy
Chan et al. (2021)
Levine et al. (2021)
Schmieding et al. (2021)
Kopka et al. (2024)
Overall
0
25
50
75
100
Accuracy [%]