Aus dem Institut für Medizinische Informatik
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

A benchmarking comparison of triage capability between 15 symptom checker apps and medical laypersons

Vergleichende Beurteilung der Fähigkeit die Dringlichkeit medizinischer Beschwerdebilder einzuschätzen zwischen 15 S*ymptom Checker Apps* und medizinischen Laien

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Malte Schmieding

Datum der Promotion: 26. Juni 2022

Inhaltsverzeichnis

1 Abstract

1. Deutschsprachige Synopse

Hintergrund: Symptom Checker Apps sind digitale Anwendungen (Smartphone Apps oder Webseiten) die Laiennutzer bei klinischen Entscheidungen unterstützen. Neben der Einschätzung, welche Diagnosen für ein beschriebenes Beschwerdebild wahrscheinlich seien, geben sie auch oft Empfehlungen, ob und wo ein Nutzer medizinische Hilfe aufsuchen sollte ("Triage Empfehlung"). Obwohl die Genauigkeit von Symptom Checker Apps in unabhängigen Studien bisher eher als unzureichend bewertet wurde, erfreuen sich solche Apps wachsender Beliebtheit. Bisher wurde noch nicht verglichen, ob solche Apps medizinischen Laien bei der Triage-Einschätzung überlegen sind.

Methoden: Auf Amazon MTurk haben wir 91 US-amerikanische Probanden rekrutiert. In einer Online-Umfrage schätzten die Probanden die Dringlichkeit von 45 fiktiven, kurzen Fallvignetten ein. Daten zu 15 Symptom Checker Apps, die anhand der gleichen 45 Fallvignetten getestet wurden, wurden aus einer vorigen Studie übernommen. Wir verglichen die Genauigkeit der Triage-Einschätzung zwischen Symptom Checker Apps und den Laien, bezogen auf alle 45 Fallvignetten und pro Dringlichkeitsstufe. Zudem wurde bestimmt, ob die Apps und Laien eher über- oder untertriagieren. Explorativ haben wir analysiert, ob Alter, Geschlecht und Bildungshintergrund einen Einfluss auf die Triage-Genauigkeit und die Neigung zur Übertriage haben bei den Laien haben.

Ergebnisse: Im Gesamtdurchschnitt waren die Triage-Genauigkeiten der Probanden (60.9%; 95% KI 59.5%-62.3%) und Symptom Checker Apps (58%) sehr ähnlich. Der Mehrheit der Probanden gelang es, besser als zehn von 15 Symptom Checker Apps zu triagieren. Sowohl die Symptom Checker als auch die Laien machten mehr Übertriage-Fehler als Untertriage-Fehler. Einen Einfluss soziodemographischer Merkmale auf die Triage-Genauigkeit bei den Laien zeigte sich nicht. Das Verhältnis von Übertriage- zu Untertriage-Fehlern war bei Frauen (2:1) höher als bei Männern (1.2:1).

Diskussion: Während die meisten Symptom Checker Apps keine höhere Triage-Genauigkeit hatten als der durchschnittliche Proband, gab es fünf Apps, die der deutlichen Mehrheit der Probanden überlegen war. Ob die Verwendung von Symptom Checker Apps nützlich ist, hängt nicht nur ab von der Fähigkeit solcher Apps, sondern

auch von denen ihrer Nutzer sowie den spezifischen Anwendungsfall. Weitere Studien sollten untersuchen, wie Symptom Checker Apps die Defizite ihrer Nutzer ausgleichen können, ohne sie fehlzuleiten, wenn die Nutzer in ihrer Einschätzung richtig liegen. Erkenntnisse dazu, in welchen Fällen und warum Nutzer den Einschätzungen von Symptom Checker Apps trauen, werden hierbei sehr wertvoll sein.

2. Englischsprachiges Abstract

Background: Symptom checkers are digital health applications (smartphone applications or website-based applications) to support laypersons in clinical decision making. Besides providing suggestions on probable diagnoses, symptom checkers appraise the urgency of patient reported medical complaints (triage recommendation). Despite past studies rating the accuracy of symptom checkers as deficient, these apps are becoming increasingly popular among the general public. Until now, no study has evaluated whether symptom checker triage accuracy is superior to that of their intended user group, that is laypersons.

Methods: In an online survey, participants had to assess the treatment urgency of 45 fictitious, short patient descriptions (case vignettes). We recruited 91 US participants via the platform Amazon Mechanical Turk. Data on triage accuracy for 15 symptom checkers on the same case vignettes was provided by a previous study. We compared the triage accuracy between symptom checkers and laypersons, for all 45 vignettes and for each of three urgency levels. We further investigated whether laypersons and symptom checkers are inclined towards over-triage or under-triage. In exploratory analyses we searched for effects of age, gender and level education on participants' triage accuracy and inclination towards over-triaging.

Results: On average, participants' triage accuracy (60.9%; 95% CI 59.5%-62.3%) was similar to that of symptom checkers (58%). The majority of participants outperformed ten out of 15 symptom checkers in terms of overall triage accuracy. Both participants and symptom checkers were inclined towards over-triage rather than under-triage. We detected no influence of socio-demographic variables on participants' triage accuracy. Female participants had a higher ratio of over-triage to under-triage (2:1) errors than male participants (1.2:1).

Discussion: While on average symptom checkers have no superior triage accuracy than laypersons, five symptom checkers outperformed the majority of participants. Whether symptom checker usage is beneficial, depends not only on the symptom checker, but also on the user and the specific use case. Future studies should investigate how symptom checkers can balance out laypersons' deficits and blind spots while not misleading them when their own intuition proves correct. Future research on when and why laypersons trust symptom checker appraisals will prove valuable.

2 Manteltext

1. Introduction

1.1. General Introduction

Grounding clinical decisions in solid science is at the core of evidence-based medicine. But making those decisions requires not only data but also the means for drawing coherent and reproducible conclusions from that data. Even before the digitization of medical research and clinical documentation made health-related data available in abundance, the magnitude and complexity of available evidence of in-print material was already a challenge for the clinical decision maker. As a result, algorithms and applications supporting clinical decision-making were envisioned as a necessity very early in the course of evidence-based medicine in the dawning digital era [1,2].

The idea of what a decision aid should encompass has closely followed the developing capabilities of computational methods: while early computer-based systems were constructed on rule-based algorithms, the past decade has seen a surge in research on more complex algorithms of machine learning (e.g., recurrent neural networks, random forests, Bayesian networks) for clinical use cases. Most of these (notional) use cases have in mind a user who is a healthcare professional as the clinical decision maker, for example a radiologist who is being supported by pattern-detection algorithms.

Apart from the scientific advances in medicine that have led to improved options for therapy, diagnosis, and prevention, another shift in medicine has been to incorporate the patient into clinical decision-making processes [3], an approach called shared decision allocation or shared decision-making, an explicit renunciation of what is now described as medical paternalism, where clinical decisions lay with healthcare professionals alone [4]. There are other reasons why patients and/or medical laypersons are now confronted with clinical decisions. For example, healthcare systems have grown ever more complex, including providing a patient with more options for where to seek care — but also burdening them with the responsibility of determining which option is best.

Clinical decision-making is complex for healthcare professionals and laypersons alike, and a general shift towards encouraging patients and/or the general public to take the responsibility for their own health has consequently led to the marked rise in the availability of clinical decision-support systems geared toward laypersons. One

example of such clinical decision-support systems is that of symptom checker applications, which are the focus of the research presented here.

1.2. Symptom checker apps

There is no standard definition of what constitutes a symptom checker app. They have been defined as tools "used by patients seeking guidance about an urgent health problem. These services [i.e., symptom checkers] generally provide people with possible diagnoses and/or suggest a course of action based on their reported symptoms" [5], and as an alternative to generic health-related online keyword searches attempt to provide patients at home with differential diagnoses and triage advice based on self-reported symptoms [6]. Thus there exists a common understanding of the purpose as well as the target group of symptom checkers, that is, supporting laypersons in their self-assessment by naming probable diagnoses and/or providing estimates of the urgency of the care needed. In their functionality they can, however, be very different one from another. While some symptom checker apps are web applications, others are smartphone applications, and some are both. Among the 23 symptom checkers assessed in a 2015 audit study [7], some require the user to input his or her complaints as free text, while others are set-up as chatbots asking only closed questions. The underlying algorithms are presumably quite diverse, too. Of the 36 symptom checkers assessed by Hill et al. in 2020, probably only a minority based their reasoning on machine learning algorithms, while the remainder did not, with some working from simple ruled-based systems [8].

Lastly, symptom checkers vary widely in their scope: next to general-purpose symptom checkers, capable of consulting on a broad range of chief complaints, some cater only to specific chief complaints or disorders (e.g., knee pain [9]) or patient groups (e.g., adult patients).

Symptom checker apps' foremost functionality is to aid patients in self-assessing their complaints, rating the urgency of these complaints and thereby helping them to identify where best to seek care, and also educating them about (possible) diagnoses. Some have described public health surveillance as another possible use case of symptom checker apps: the aggregated data on symptoms entered by users could help identify regional patterns such as outbreaks of infectious diseases [10,11]. Online self-assessment apps have also been suggested as a supplement to or replacement for telephone triage services, or as an assistive tool for the community health workers who are triaging patients, especially in resource-limited countries [12]. Thus the notional

utility of symptom checkers lies with both the micro level of the individual patient and the macro level of the healthcare system.

1.3. Patients' and healthcare professionals' perspectives on symptom checkers

The current literature concerning healthcare professionals' opinions on symptom checker use by patients is ambiguous. A Finnish study reports a majority of surveyed healthcare professionals (HCPs) agreeing that symptom checker apps can increase availability of services to patients by guiding them through the available healthcare services and be beneficial in the individual care of patients, while about half of HCPs doubted that patients are capable of using symptom checkers or even willing to use them [13]. A German qualitative interview study also documents clinicians' disapproval of health-related online searches by patients, with a majority in that study presuming patients' general inability to adequately judge the retrieved information [14]. Similarly, a US study reports that HCPs commonly do not recommend websites or apps to their patients, and that they dislike patients bringing results from online searches to a consultation [15]. Ironically, the same study, a randomized controlled trial on patients in an emergency department's waiting room using the search engine *Google*, found that in the specific instances of patients bringing up information from a previous online search during a consultation, the clinicians considered it helpful, and that neither the clinicians' satisfaction with the care provided, nor the patients' satisfaction with the care received, nor the patient-clinician relationship was compromised by patients' prior online health searches. Two observational studies reported that online health information searching prior to visiting a doctor improves the patient-clinician interaction and increases patient confidence in the physician [16,17]. In other words, the reported subjective concerns of HCPs have not been shown to be justified by objective findings thus far.

A large proportion of patients and the public in general consult online sources for health information, a fact that has made *Doctor Google* an accepted term in everyday language as well as in scientific literature. Laypersons consider symptom checker apps a less well-known alternative to *Doctor Google* [18]. Handling the volume of unfiltered and mostly inapplicable healthcare information that can be obtained using *Google* searches is a challenge for anyone, whereas symptom checker apps promise to provide advice tailored more specifically to the patient and his or her complaints [18].

Symptom checker apps are not commonly regarded as an alternative to seeking professional medical care in the first place [18]. A systematic review summarises that patients commonly consult symptom checker apps for what they consider trivial (non-serious) complaints for which a visit to the doctor would be inappropriate, persistent complaints that remain undiagnosed after they have sought professional medical care, and complaints considered potentially embarrassing [19]. The complaints for which symptom checkers are most usually consulted resemble those for which patients commonly seek primary healthcare office visits [20-22], being foremost respiratory or ear, nose, and throat (ENT) complaints such as the common cold, a runny nose, cough, and sore throat, or abdominal pain and nausea, etc.

Users report a desire to receive advice on whether and where to seek a healthcare professional and self-education on what can potentially cause complaints as the main motivation for using a particular online symptom checker (*Isabel*) [23]. Similarly, a 2009 study names gaining information on self-management and the reduction of uncertainty as reasons motivating patients to use symptom checker apps [20]. This study argues that "taboo complaints," defined by the authors as complaints relating to the genitourinary tract, might be an exception, where patients prefer consulting a computer system rather than a healthcare professional.

1.4. Prevalence of symptom checker app use

Despite limited evidence for the benefits of symptom checker app use, they are nonetheless widely used as one of several options for seeking health information online. A 2013 US study showed that in the general population, one in three people regularly seeks healthcare information online, and attempting to diagnose oneself is one of the most common use cases [24]. A 2020 German study reports that 20% of the population uses the internet as the primary source of healthcare information [25]. Among patients, the proportion of those consulting online health information sources prior to seeking professional medical care is reported to be between one-third and two-thirds according to several studies [15-17]. Symptom checkers are already used by many and the number is growing quickly: the German EPatient Survey 2020 estimates the proportion of Germans using diagnostic apps at 13%, it having doubled in the last 5 years, and presumably this trend has been accelerated by the surge of digital health applications during the COVID-19 pandemic [26,27]. Studies reporting on the characteristics of symptom checker app users suggest that a typical user is female and

well-educated, and that older people use or would use apps less frequently than younger people [22-24,28,29].

1.5. Evaluation of symptom checker apps

Several studies have been published on assessing online-based decision-support tools similar to symptom checker apps, for example differential diagnostic generators aimed at healthcare professionals as the user group [9,30,31]. However, unlike tools intended to aid in pharmaceutical or diagnostic laboratory testing, no framework has yet evolved on how to assess the performance and safety of symptom checker apps [32]. What has been published thus far in the literature in terms of approaches for evaluating symptom checkers varies. One common approach is to utilize patient descriptions (case vignettes), either fictitious [7,8] or based on clinical documentation of real patients [6], and test the symptom checker's response to these. An alternative approach lets patients or healthcare professionals enter signs and symptoms for themselves or their patients, respectively [33]. With either of these approaches, some studies focus on a limited range of symptoms (for example symptom checker performance concerning only abdominal [33], knee [34], or ophthalmic complaints [35]), while others aim at capturing the entire breadth of complaints for which symptom checkers might be consulted. Independent of the chosen approach, these studies commonly face study-specific limitations, for example how the gold standard for the correct diagnoses and adequate triage level was set regarding the composition of the pool of case vignettes, or how ambiguities during the entering of information from the vignettes were handled, which hinders comparability of results between different studies. Furthermore, many studies evaluating symptom checkers are conducted by the developers of such systems themselves [34–37] and are therefore at risk of reporting overly promising results. Two literature reviews on symptom checker apps for that reason have concluded that there is only limited evidence that the apps are accurate and aid in providing better care [5,38]. Thus it is argued that establishing a framework for evaluating the performance of symptom checkers and the effects of their utilization is necessary [32,39].

1.6. Contribution of this thesis

In 2015, Semigran et al. reported that, on average, symptom checker apps return correct triage advice in 57% of their triage evaluations [7]. Though Semigran et al. do not provide a benchmark for sufficient accuracy, they judge this performance as deficient [7]. In a subsequent study, Semigran et al. found that physicians

outperformed symptom checker apps in terms of diagnostic accuracy [40]. Other studies have also compared symptom checker apps' or differential diagnosis generators' (geared towards healthcare professionals as users) diagnostic or triage accuracy against that of healthcare professionals [33,36,41]. Such study designs imply that symptom checker apps would have to perform equally as well as healthcare professionals in order to be deemed useful. This would, however, only be reasonable if symptom checker apps were used or promoted as alternatives to seeking advice from a healthcare professional, which is not actually what the user is looking for [18,23], and only one app developer has publicly hinted that this is its own aim [41].

Symptom checker apps can be considered useful when they enable their target user groups to make better clinical decisions than users could make on their own. As the target user groups of symptom checker apps generally consist of laypersons, contrasting the performance of symptom checker apps and laypersons is a more adequate comparison for establishing a benchmark criterion. Although some studies have focused on laypersons' ability to self-diagnose with the help of online tools [15,34,42,43], triage advice is the more important feature of symptom checkers [8], both in terms of patient safety and the potential to make the providing of healthcare more efficient, and thus the comparison of the triage abilities of laypersons and symptom checker apps is the more relevant.

Research performed in a different field provided further inspiration for the main question that our own study, Schmieding et al. [44], explored. Dressel and Farid (2018) evaluated the performance of a commercial algorithm in predicting recidivism rates for criminal defendants in the US [45]. Though the commercial algorithm is highly complex and bases its recommendations on more than one hundred features, that is, variables of information inputted into the algorithm, laypersons were able to achieve a similar accuracy when provided with only seven features on which to base their decisions. Thus the question arises whether symptom checker apps, too — despite their complexity — achieve no higher level of accuracy than laypersons do.

A previous study assessing laypersons' ability to triage using case vignettes shows that laypersons do struggle with triaging, but that their performance is better than that of someone guessing at random [46]. Following this approach, our study reports on the triage capability of laypersons using the same case vignettes with which the symptom checker apps were tested by Semigran et al. [7] and thereby provides a benchmark that apps need to surpass as a first criterion for being considered useful.

Apart from suggesting a criterion for assessing symptom checker apps' usefulness, our approach also yields detailed insights into where the strengths and weaknesses of laypersons' triage abilities lie, which can contribute to identifying the use cases where symptom checker apps are most and least necessary.

2. Methods

The following sections on Methods, Results and Discussion will summarise and discuss the main findings from Schmieding et al. [44], but include two more in-depth analyses which are not part of the published paper.

2.1. Summary of methods of Schmieding et al.

Our research builds upon a study from 2015 that examined the accuracy of the triage recommendations of 15 publicly available symptom checker apps: Semigran et al. based their evaluation on 45 fictitious descriptions of patients and their complaints (case vignettes), 15 for each of three triage levels, that is, the gold-standard rating of a respective case's urgency of need for treatment (emergency care, non-emergency care, and self-care) [7].

We modified these 45 case vignettes to make the information they provide comprehensible to a lay audience, for example by rephrasing *rhinorrhea* as *runny nose*. Three experts modified the vignettes, two of whom were physicians and two of whom were native English speakers. In that way, we were able to ensure that the modified vignettes remained medically correct while at the same time were comprehensible to an English-speaking layperson. We embedded the modified vignettes in an online survey, asking the participants to rate the urgency of the need for treatment of the fictional patients using the three-tiered scale (emergency care, non-emergency care, and self-care). Prior to the case vignettes being presented, these urgency levels were explained to the participants by providing the definition for each level as phrased by Semigran et al. [7]. Additionally, we surveyed three demographic variables of the participants: age, sex, and level of education.

In addition to the case vignettes and the triage-level definitions, we also used Semigran et al.'s data on the symptom checker apps' triage capability [7]. In other words, for the comparison of the triage accuracy of the participants to that of the symptom checkers, we did not collect our own data for the symptom checkers. Some symptom checkers had limitations as to which cases they were capable of evaluating. For example, the app *Healthy Children* [47] only considers paediatric patients and thus could only provide a triage appraisal for 15 of the 45 cases. Similarly, other apps do not address

a vignette character's chief complaint and so could not be used with that particular case vignette. As a consequence, since not every one of the 15 apps was able to evaluate all of the 45 case vignettes, the total number of vignette evaluations by the apps provided in the Semigran et al. study amounts to 532 [7]. This stands in contrast to the data we collected on laypersons' triage accuracy, where each of the 91 participants assessed the urgency of all 45 vignettes, yielding 4,095 case-vignette evaluations in total. As four of the 15 apps sampled by Semigran et al. [7] never recommended the least urgent triage level (self-care), probably because they were not designed to do so, we conducted our main analyses twice, including and excluding these apps, to ensure that our results were not skewed by this subset of symptom checkers erring on one-third of the vignettes presumably because of their design.

In March 2020, we recruited participants via the online platform Amazon Mechanical Turk (MTurk) [48]. As the case vignettes' gold-standard solutions were designed to be appropriate in the US healthcare system, we chose only to recruit participants with permanent residence in the United States. Participants were remunerated $4.00 US for their participation, and a bonus payment of $3.00 US was awarded to those participants who accurately assessed more than 26 of the 45 case vignettes. This threshold was chosen as it roughly corresponds to outperforming the average app as assessed by Semigran et al. [7].

Our primary outcome measure was the difference between the mean triage accuracy of participants and apps for all case vignettes (overall triage accuracy). In sub-analyses to the main outcome measure, we evaluated whether the mean accuracies differed between the three triage levels for participants and apps. To determine 95% confidence intervals for the mean triage accuracy of participants, we chose a bootstrapping approach, which returns stable estimates of confidence intervals even when the underlying data is non-normally distributed [49].

As secondary outcome measures, we determined: (a) the proportion of participants outperforming each app (discussed and presented in the published paper, but not in this synopsis); (b) the degree of difficulty of each case vignette (discussed and presented in the published paper, but not in this synopsis); (c) and the types of errors for apps and participants. Finally, we juxtaposed the triage estimates of apps and laypersons on an app-by-app basis (not included in the published paper, but described in this synopsis).

For the analyses of types of errors, we evaluated the proportion of evaluations where the participants and apps over-triaged and under-triaged. Over-triage was defined as rating a case vignette's triage level as higher (more urgent) than appropriate, for example assigning emergency care to a case vignette where non-emergency care was set as the gold-standard solution, while under-triage was defined as the opposite.

The required sample size for our participant sample could only be roughly estimated, as no sufficient data basis on distribution, variance, and mean triage accuracy of laypersons existed at that time. The minimum sample size was estimated by calculating the sample size required if a two-sided t-test to detect the difference between a participants' mean accuracy and a constant (the apps' mean accuracy) of an effect size of Cohen's d = 0.4 had been planned to be utilized with an alpha level = 0.05 and a power = 0.8, and assuming equal variance for symptom checkers and laypersons. This yielded a required sample size of 52 participants. We used this sample size calculation as an estimate of the minimum number of participants required to acquire meaningful results but decided to continue sampling more participants within the earmarked budget in order to have a larger sample for the exploratory secondary analyses. We considered the benefits anticipated from the secondary analyses higher than the risks associated with oversampling (i.e., burden to study participants, risk of detecting significant yet irrelevant differences) in our case.

2.2. Juxtaposition of symptom checker apps and participant triage evaluations

We observed that some case vignettes were challenging to symptom checkers but correctly evaluated by most participants, and vice versa [44]. Thus symptom checker apps might prove beneficial when they are able to correctly evaluate scenarios with which laypersons tend to struggle. Conversely, an app's advice is less helpful to users when it provides correct advice only for scenarios that the users are commonly able to rate correctly by themselves.

To address this issue, we juxtaposed the triage evaluations of each symptom checker app with those of every participant for each case vignette: for every symptom checker app, we selected the subset of vignettes it evaluated (as most apps evaluated only a portion of vignettes). For every case vignette in this subset, we juxtaposed the app's triage recommendation against the appraisals of each of the 91 participants. We compared the app's triage appraisal with each participant's triage appraisal by determining whether the app and the participant agreed in their ratings, and whether both, neither, or only one of them was correct in their rating. We summarised these

comparisons by determining for each app (a) the proportion of evaluations where it concurred with participants, that is, both app and participant assigned a case vignette the same urgency level, and (b) the proportion of evaluations where the app disagreed with participants. Based upon a) and b) we calculated the percentage of evaluations where the app was correct when it (c) concurred or (d) disagreed with participants. These two latter measures are of interest for different use cases of symptom checker apps: an app that proves to be commonly correct when disagreeing with its users is valuable when the user is seeking advice in order to question his or her own judgement, while an app that more commonly affirms a user's own correct judgement rather than falsely affirming a user's incorrect judgement might prove beneficial to users seeking reassurance. Although an ideal app that never misjudges a patient's symptomatic urgency can serve both purposes, imperfect apps might prove better in one function than the other, depending on whether they are mistaken in the same cases as their users or in different ones.

2.3. Analyses of demographic factors influencing the triage capability of laypersons

Our sample size did not allow us to obtain a participant sample representative of the US population. To rule out that the demographic composition of our participant sample influenced the external validity of our results on laypersons' triage accuracy, we performed a logistic regression with correct triage advice as a dependent variable and the three surveyed demographic variables (age, gender, and level of education) as independent variables. For this synopsis we further conducted post-hoc analyses on these three demographic variables' univariate influence on risk-aversion by means of descriptive statistics. If the descriptive statistics indicated a relevant influence, we conducted a post-hoc Pearson Chi^2-test using R 4.0.0 [50].

3. Results

The following summarises the main results from the publication Schmieding et al. [44], and the subsequent sections present two analyses not included in the publication.

3.1. Summary of results of the publication

In total, 91 participants are included in our sample. Our sample includes relatively more male participants and more persons with higher levels of education than the general US population. With each of the 91 participants assessing all 45 case vignettes, 4,095 case evaluations were produced by the participants.

With all three triage levels taken together, the average participant's triage accuracy (60.9%, SD 6.8%) was comparable to the average symptom checker's (58%, SD

12.8%). While a symptom checker's triage ability decreased with lower urgency of triage level, that is, that they did best with emergency vignettes and worst with self-care vignettes, our participants showed a different pattern: their triage accuracy was about equal for emergency and non-emergency vignettes, and considerably lower for self-care vignettes. In comparison, the apps outperformed the participants in detecting emergencies but were less reliable than the participants in the remaining two categories. Upon exclusion of the four symptom checker apps that never suggested self-care, the symptom checkers' average accuracy did improve, but the overall pattern remained unchanged.

Both the participants and the apps were inclined towards over-triage, that is, they more commonly erred by assigning a higher triage level than appropriate than by assigning a lower than appropriate triage level. This inclination is more pronounced in symptom checker apps than in human decision makers: while our participants over-triaged in 23.3% of case-vignette evaluations, symptom checkers did so in 34.8%. Additionally, the symptom checker apps rated approximately every fourth self-care vignette as an emergency (i.e., in 24.7% of case evaluations), while the participants only misjudged vignettes in that way in 4.1% of case evaluations.

While the participants had in common with the symptom checkers a general inclination towards over-triage, their patterns of over-triage differed on closer inspection of the data; see Table 1. Differentiating between self-care and non-emergency, participants made more over-triage errors (53.3% of self-care vignette evaluations) than under-triage errors (14.9% of non-emergency vignette evaluations). In contrast, when differentiating between emergencies and non-emergencies, participants erred more towards under-triaging (32.6% of emergency vignette evaluations) than over-triaging (16.7% of non-emergency vignette evaluations). Thus calling the participants risk-averse does not grasp the full picture: when deciding whether care should be sought at all (self-care versus non-emergency care) the participants are risk-averse, but when deciding where to seek care (emergency versus non-emergency care), participants show no such risk-aversion.

The pattern is simpler for symptom checker apps: both when deciding between whether care is necessary at all and where to seek care, they more often err towards the more urgent triage level than towards the less urgent one. This holds true upon exclusion of the four symptom checker apps that do not include all three triage levels.

	Triage Appraisal								
	All SCs (n=15)			Select 11 SCs (n=11)			Participants (n=91)		
Solution	Em	NE	S-c	Em	NE	S-c	Em	NE	SC
Em	80.3% (147/183)	16.9% (31/183)	2.2% (5/183)	79.2% (103/130)	16.9% (22/130)	3.8% (5/130)	67.5% (921/1365)	28.4% (387/1365)	4.2% (57/1365)
NE	37.7% (66/175)	54.9% (96/175)	7.4% (13/175)	32.0% (41/128)	57.8% (74/128)	10.2% (13/128)	16.7% (228/1365)	68.4% (934/1365)	14.9% (203/1365)
S-c	24.7% (43/174)	42.0% (73/174)	33.3% (58/174)	18.1% (23/127)	36.2% (46/127)	45.7% (58/127)	4.1% (56/1365)	49.2% (672/1365)	46.7% (637/1365)

Table 1. Confusion matrices of triage appraisals for symptom checkers (SCs) and participants. "Select 11 SCs" refers to the subset of symptom checkers providing self-care advice at least once. Em*,* NE*, and* S-c *refer to the three triage levels* emergency*,* non-emergency *and* self-care*. Table is based on data from Schmieding et al. (2021) [51] and Semigran et al. (2015) [7].*

3.2. Further results

3.2.1. Influence of demographic variables on risk aversion

A logistic regression on the influence of the three demographic variables (age, gender, level of education) on triage accuracy yielded no significant results [44]. A post-hoc analysis, however, hints at differences concerning risk-averseness, that is, the proportion of over-triage errors among erroneous triage appraisals. On average, female participants made twice as many over-triage errors as under-triage errors (407:195 case evaluations); see Figure 1. The ratio of over-triage to under-triage errors was less marked for males, at 1.2:1 (549:452 case evaluations). This difference in risk-aversion was also significant in a post-hoc Chi²-test ($\chi^2_1 = 24.9$, $p<0.001$). The level of education and the age of participants, on the other hand, showed no such effect on risk-aversion; see Figures 2 and 3.

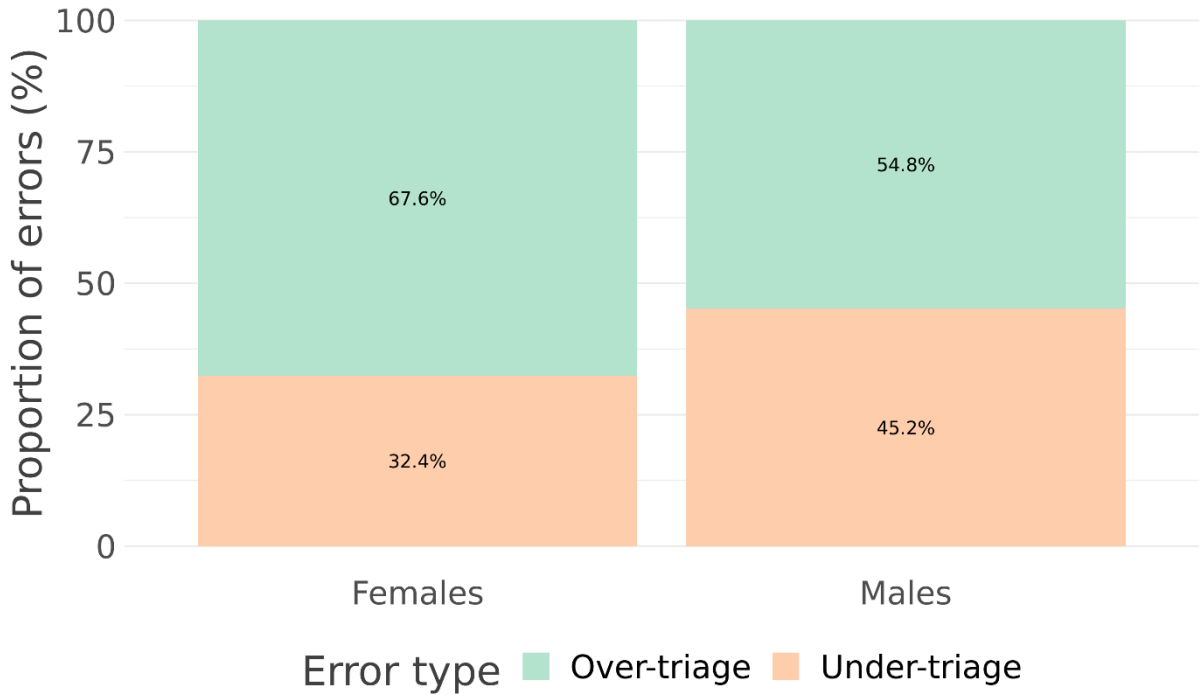


Figure 1. Inclination towards under- and over-triage (risk-aversion) *among female and male participants. Data from Schmieding et al. [51]. Illustration by Malte Schmieding.*

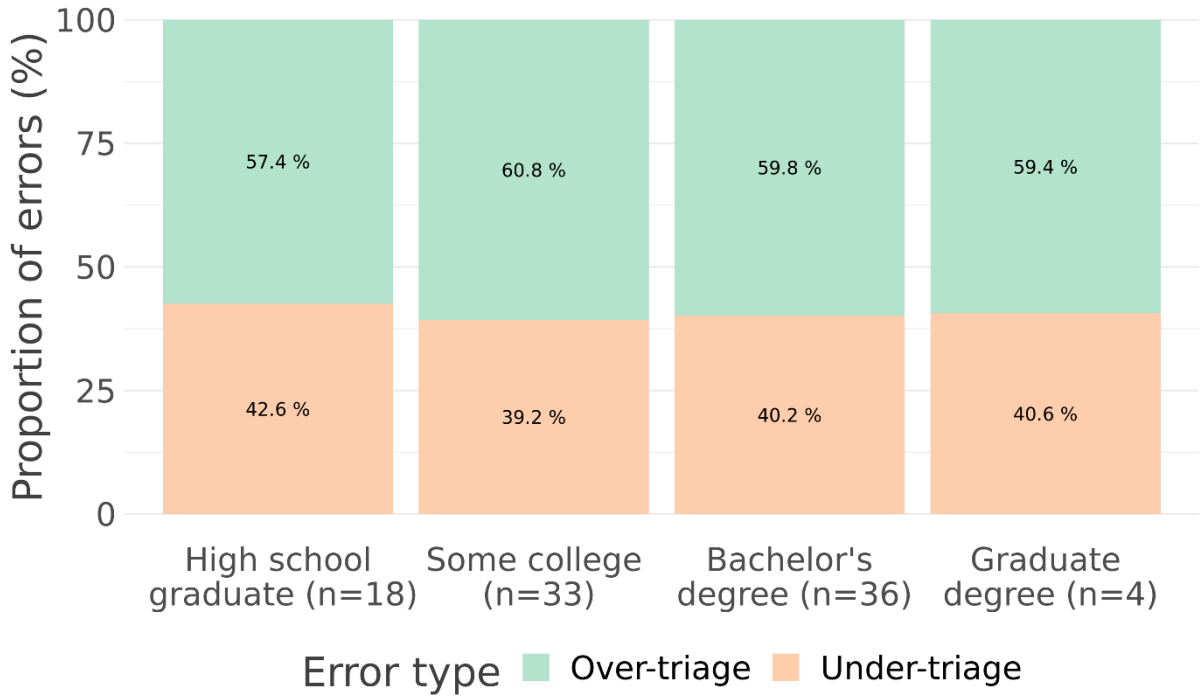


Figure 2. Inclination towards under- and over-triage (risk-aversion) *by a participant's highest level of education attained. Data from Schmieding et al. [51]. Illustration by Malte Schmieding.*

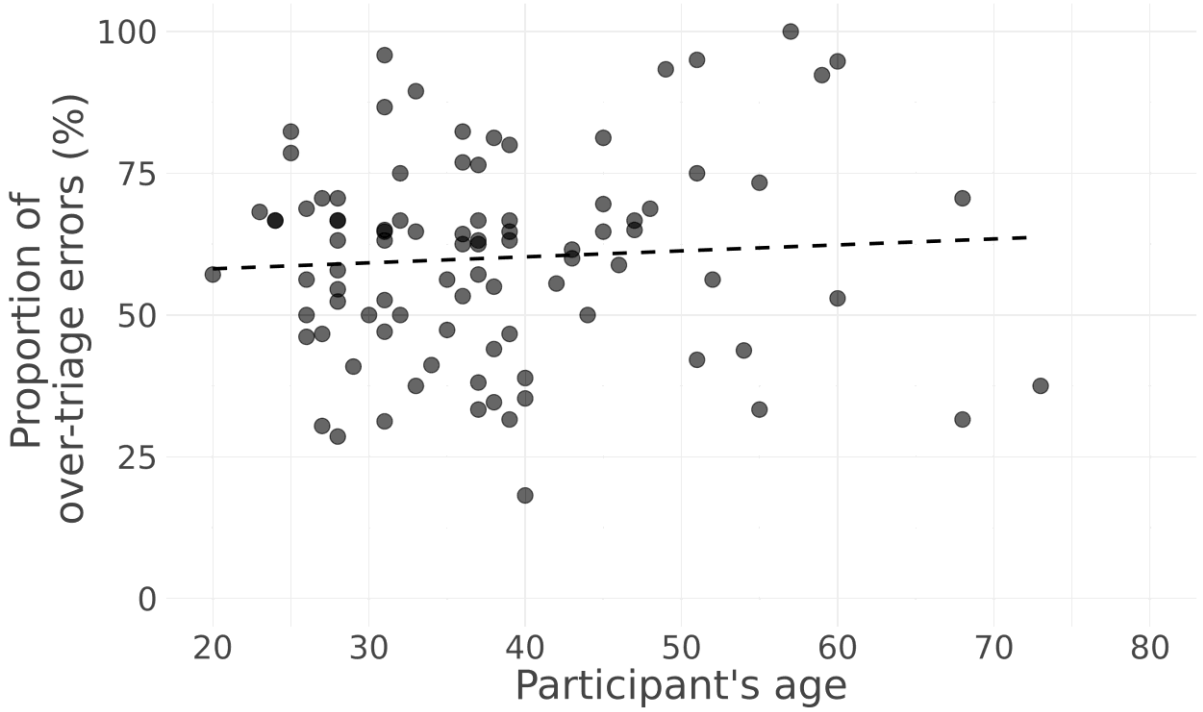


Figure 3. Inclination towards over-triage (risk-aversion) *by age of participants. Proportion of over-triage errors refers to the proportion of over-triage errors in relation to all errors made by the respective participant. The dashed line represents a linear model. Data from Schmieding et al. [51]. Illustration by Malte Schmieding.*

3.2.2. Juxtaposing participants' and symptom checkers' triage appraisals

When participants and symptom checkers agreed in their triage appraisals, those appraisals were likely to be correct, especially when the symptom checker had a high accuracy rate on its own, but also at times when the symptom checker had a low accuracy rate: the minimum combined accuracy rate for case vignettes on which participants and symptom checkers agreed in their assessments was 62.6%, and the maximum was 87.7%. The accuracy when participants agreed with a symptom checker was always superior to both the participants' stand-alone accuracies and the participants' accuracies when disagreeing with the SC; see Table 2.

When a symptom checker's triage appraisal did not match that of the participant's (disagreement), the best five symptom checkers ranked by accuracy were more often correct than the average participant. Concerning the less accurate apps, the participants were more reliable when disagreeing with them; see Table 2.

SC name	n assessments[a]	Stand-alone accuracies		Rate of agreement[c]	SC and P disagree			SC and P agree	
		SC's accuracy	Participants' accuracy[b]		SC correct	P correct	Both incorrect	Correct	Incorrect
HMS Family Health Guide	3640	80% (32/40)	59.5% (SD: 7.1%)	57.3%	69.7%	21.6%	8.7%	87.7%	12.3%
Healthy Children	1365	73.3% (11/15)	49.9% (SD: 10.7%)	64.2%	78.5%	12.9%	8.6%	70.5%	29.5%
Steps2Care	3822	71.4% (30/42)	59.7% (SD: 7.2%)	55.5%	59.5%	33.1%	7.4%	81%	19%
Symptify	3640	70% (28/40)	60.2% (SD: 7.2%)	62.3%	56.1%	30.2%	13.7%	78.4%	21.6%
Symptomate[d]	1274	64.3% (9/14)	60.9% (SD: 11.6%)	55.2%	50.6%	43.1%	6.3%	75.4%	24.6%
Drugs.com	3822	59.5% (25/42)	60.6% (SD: 6.5%)	58%	43%	45.6%	11.3%	71.5%	28.5%
FreeMD	4004	59.1% (26/44)	60.2% (SD: 6.7%)	58.9%	41.8%	44.7%	13.5%	71.1%	28.9%
Doctor Diagnose	1456	62.5% (10/16)	69.5% (SD: 10.9%)	60.1%	38.7%	56.4%	4.8%	78.2%	21.8%
Family Doctor	3731	53.7% (22/41)	58.1% (SD: 7%)	39.1%	40.9%	48.2%	11%	73.6%	26.4%
Early Doc	1547	52.9% (9/17)	63.4% (SD: 11.4%)	43.5%	37.4%	56%	6.6%	73.1%	26.9%
Isabel[d]	4095	51.1% (23/45)	60.9% (SD: 6.8%)	53.6%	32.9%	54%	13.2%	66.9%	33.1%
NHS	4004	52.3% (23/44)	62% (SD: 6.9%)	44.4%	32.9%	50.4%	16.7%	76.6%	23.4%
Symcat[c]	4095	44.4% (20/45)	60.9% (SD: 6.8%)	49.6%	26.5%	59.2%	14.3%	62.6%	37.4%
Healthwise	4004	43.2% (19/44)	61.2% (SD: 7%)	40.4%	22.1%	52.4%	25.5%	74.2%	25.8%
iTriage[d,e]	3913	32.6% (14/43)	60.5% (SD: 6.9%)	28.5%	14.9%	54%	31%	76.7%	23.3%

Table 2. Summary of the juxtaposition of triage appraisals by symptom checkers (SC) and participants (P). [a] *An assessment is the juxtaposition of each participant's appraisal with that of the symptom checker. Hence the number of assessments is the product of the number of cases an SC considered and the participant sample size (n=91).* [b] *Accounting only for cases that the respective SC considered.* [c] *Mean proportion of assessments where the respective SC and the participants provide the same triage appraisal.* [d] *SCs that never suggested self-care.* [e] *SC always advises to go to the emergency department. Data from Semigran et al.* [7] *and Schmieding et al.* [51].

4. Discussion

4.1. Principal results

Our main outcome suggests that the triage capability of laypersons is comparable to that of the 15 symptom checker apps as evaluated by Semigran et al. in 2015 [7]. This result is also in line with Dressel and Farid's finding that laypersons' judgement can keep up with that of complex computational algorithms in some use cases [45]. However, looking beyond the comparison of both samples' mean accuracies, our data reveal that five of the 15 symptom checkers outperform a large majority of the participants. Furthermore, apps' and laypersons' strengths and weaknesses do not fully overlap, which might enable users to make better decisions when supported by an app than they would on their own. Thus concluding that symptom checkers are not of benefit is short-sighted.

Female participants demonstrated a greater risk-aversion than males in our sample of laypersons. This concurs with previous studies describing gender differences in risk perception and risk-taking in general and in a health-related circumstance in particular [52,53]. With symptom checkers being risk-averse on average, the on average risk-taking male user group might benefit from their use to a greater extent than female users.

Our analyses further identified that for some case vignettes, symptom checkers reliably provided accurate advice, but that laypersons failed to appraise these vignettes correctly. This is a valuable insight, as it indicates the possibility that symptom checkers and their users balance out each other's shortcomings and blind spots, that is, that symptom checkers and users enable better decision-making when working together rather than by themselves. To further investigate this idea, we juxtaposed symptom checkers' advice and laypersons' urgency appraisal. This comparison revealed that when a user's triage appraisal is in agreement with a symptom checker app, be it a high-performing one or not, his or her chance of making the correct decision always increases. Thus this juxtaposition suggests that symptom checkers can be beneficial when used as confirmation of a user's own evaluation (reassurance). However, upon disagreement between app and layperson, only the top five apps managed to be correct more often than the average layperson. Consequently, users should generally trust their own judgement more than that of an app unless the app has a proven high triage capability.

These results emphasize that the usefulness of a symptom checker must be assessed with an eye to its specific users. Different user groups might benefit to a greater or

lesser extent from a particular symptom checker app, or might benefit for different reasons. Furthermore, the results demonstrate that although symptom checkers may provide useful decision support and thereby potentially simplify clinical decision-making for laypersons, the user is burdened with two new decisions: which app to trust and when to trust that app.

4.2. Discussion of Methods

4.2.1. Aggregate statistics

Semigran et al. reported an overall triage accuracy of 57% for the symptom checker apps [7]. This percentage is the fraction of all correct triage evaluations by symptom checker apps (n=301) over all of the vignettes for which the apps provided a triage recommendation (n=532). Of the 15 apps, only two evaluated all 45 vignettes, hence the remaining 13 apps evaluated only a subset of vignettes, each subset of different number and composition. For example, the app *Doctor Diagnose* evaluated only 16 vignettes, 10 of them correctly, whereas the app *Isabel* evaluated all 45 vignettes, 23 correctly. Consequently, those two apps contribute differently to the denominator of Semigran et al.'s fraction: *Isabel*'s accuracy has an impact on the overall accuracy reported by Semigran et al. nearly three times higher than that of *Doctor Diagnose* [7]. For the analyses in our study, the overall purpose of which is to compare the capabilities of symptom checkers and laypersons a different approach is more suitable. We first calculated the triage accuracy of each app, that is, the fraction of correct evaluations that the app made over all of the vignettes evaluated, and then determined the mean of these 15 fractions. Thus each app has an equal weight in our calculation of the apps' aggregated triage accuracy. This explains why we report an average triage accuracy of 58% for the apps [44] while Semigran et al. report 57% [7].

The analysis approach of Semigran et al. [7] seeks to answer the following question: if a user sought medical advice from all available apps, how often would he be advised correctly? Our approach to reporting the triage accuracy of the symptom checker sample seeks to answer a slightly different question: if a user sought medical advice from any one given app, how often on average would the returned advice be correct, omitting instances where the app did not provide a recommendation at all?

The results from comparing the apps' and participants' mean averages using our approach are still somewhat skewed, as the case vignettes contribute unequally to the apps' aggregated average: some vignettes were evaluated by more apps than others (and no vignette was evaluated by all), allowing each vignette's influence on the overall

triage accuracy reported to vary. Therefore we considered any aggregation of the apps' triage accuracy as a simplification best avoided by reporting results on an app-to-app basis. Consequently, our paper's analyses focus on comparing the participants with each app rather than with the apps' sample average. This is a more nuanced approach that does not try to answer whether apps or laypersons are more capable of triaging in general, but which apps are better or worse than the average layperson, which we consider a suitable benchmarking criterion.

4.2.2. Inferential statistics for the sample of symptom checker apps

Studies evaluating the capabilities of symptom checker apps commonly provide inferential statistics, such as confidence intervals and p-values, in their analyses [7,8,36]. For example, Semigran et al. calculated the 95%-confidence intervals for the symptom checker apps' aggregated triage accuracy based on a binomial distribution [7]. However, the assumptions underlying a binomial distribution are not met when considering a sample of symptom checkers and their triage evaluations: inferential statistics are used to infer insights on a population of which only a fraction was studied (sample). The sample must represent a random sampling of the population since otherwise the inferences would be biased.

Symptom checker apps share the same purpose and target user group, but in all other respects are highly heterogenous, for example, in terms of their underlying algorithms or user interface, and arguably have fewer commonalities than differences. Thus there is doubt as to whether each app can be regarded as members of the same class of app (population). Regarding all symptom checker apps as belonging to one class (in statistcal terms: "stemming from one population") is analogous to assessing the efficacy of a limited sample of painkillers and then extrapolating the conclusions of that study to all substance classes of painkillers, including those with a different pharmaceutical agent.

Even if one assumes that each symptom checker app can be regarded as an individual from a larger population of symptom checker apps, only a randomly drawn sample provides a valid basis for inferences on the larger population. Semigran et al., however, state that they sampled the apps purposefully and not at random and thus compiled a sample that is not appropriate for inference [7]. Secondly, the binomial distribution Semigran et al. assumed [7] requires that the events, in this case the evaluations (n=532), are not clustered, but that they all share the same probability, that is, the probability of being evaluated correctly. The case evaluations, however, are grouped

both by the symptom checker (n=15) that provides the recommendation and the case vignette to which the recommendation refers (n=45). Thus a binomial model oversimplifies the complexity of the data. Models capable of accounting for these dependencies between the case evaluations, such as generalized linear mixed-effect models with crossed random effects, potentially prove to be a better fit to describe the data on symptom checkers' case evaluations. For the analyses in this study such intricate models are not suitable for data with a single data point for each cluster combination (combination of app or participant and case vignette) and thus mostly descriptive statistics for the symptom checker apps are provided.

4.3. Limitations

To our knowledge, our study represents the first attempt to directly compare the triage accuracy of symptom checkers with their potential users, that is laypersons. Our study does come with limitations. Some are due to the study's design and thus difficult if not impossible to mitigate without using an entirely different one, while other of its shortcomings could be avoided in future studies.

First, our study used the data on symptom checkers' accuracy from Semigran et al., which was published in 2015 [7]. In that study, they entered the vignettes into the symptom checkers in 2014, which means that their data was six-and-a-half years old when our paper came to press. The triage capability of symptom checkers may have changed significantly during that time. Four of the 15 symptom checker apps assessed by Semigran et al. [7] have been discontinued since, as Yu et al. point out [6]. On the other hand, many of the currently prominent symptom checker apps were not included in Semigran et al.'s analyses, as a comparison to much more recent study on symptom checkers suggests [8].

The compilation of case vignettes came from Semigran et al. [7], too. This was necessary in order to allow a fair comparison between symptom checkers and laypersons' triage capability. However, it restricted us to recruiting US participants, as the case vignettes' gold-standard triage levels might not be appropriate in another healthcare system. Some of their case vignettes also lacked proper diagnoses, naming only chief complaints instead (e.g., back pain, vomiting), which makes it difficult to independently re-evaluate the appropriateness of their gold-standard triage levels. Additionally, the Semigran et al. vignettes [7] cannot be regarded as a representative sample of acute-care cases: while a disproportionate number (13/45) of case vignettes' chief complaint is a respiratory one (e.g., shortness of breath, cough), none of the

fictitious patients suffers from a mental-health condition. With 15 case vignettes for each of the three triage levels, the levels of urgency are equally distributed, while in acute care the less urgent cases are the more common ones. Calculating metrics like overall triage accuracy, but also sensitivity, specificity, and negative and positive predictive values (which have not been calculated by us but are reported in other studies [6,36]) are therefore heavily influenced by the composition of the pool of case vignettes. These metrics are very useful and common in test theory, but without consideration of the pool of case vignettes with which the apps were tested, a comparison of these metrics between studies is unreliable.

A more fundamental limitation is that laypersons potentially base their decisions on different information than symptom checkers. Both the symptom checkers and our participants have been provided with the same information—however, in real life, a human decision maker might notice more information than what is provided in a case vignette or overlook other information if not explicitly prompted by a symptom checker to provide such information. Furthermore the laypersons were provided the information as matter-of-fact text. They might have decided differently if the information had been presented differently. For example, one case vignette soberly states that the patient is in severe pain (eight on a scale of one to ten). In a real-life setting, this patient would be in great despair, and such strong emotions would probably influence a human decision maker’s decisional context, but not a symptom checker app's, since it is not influenced by social and emotional contextual factors.

Semigran et al. had the case vignettes that they used entered into the app by a medical layperson [7], while other vignette-based studies have had vignettes entered by medical professionals [6,36,54], raising the question of which approach is the more suitable to appropriately assess a symptom checkers capability. Laypersons are the target user group of symptom checkers, not medical professionals, which speaks for the latter providing greater validity. However, just as healthcare professionals are not the intended target group, laypersons entering fictitious clinical vignettes is not the use case of symptom checkers, either. Arguably, patients might enter their own symptoms into an app differently and potentially with greater ambiguity than when they simply transfer information from a given case vignette into a symptom checker app. The sources and effects of such ambiguity are out of the scope of the assessment capabilities of vignette-based studies. Rather, the strength of vignette-based studies is that they allow very controllable laboratory conditions: the information provided can be

controlled and a gold-standard solution can be set for the case vignette. Other types of study designs commonly used to assess the accuracy of symptom checker apps, such as using non-fictitious cases from medical records [1] or having real patients enter their complaints into an app [55], face greater difficulties in reaching these two laboratory conditions.

Consequently, it is worth noting that vignette-based studies commonly ignore contextual factors in decision-making and in fact name it as one of the study's limitations [36], but really, it should be regarded as a strength of vignette-based studies rather than a weakness: ultimately, vignette-based studies can be used to estimate a decision maker's highest attainable capability (here the triage accuracy of a symptom checker) when factors potentially compromising adequate decision-making are disregarded: For the human decision maker, contextual factors such as emotional or social context can be minimized. For the symptom checker app, information should be entered in a manner that is unambiguous and most suited to the respective app, as pioneered by Berner et al. in a study on medical expert systems [31]. Thus vignette-based studies are a fruitful first step in assessing a symptom checker's capability despite their limited ecological validity, that is the limited generalisability from an app's performance in a case vignette-based assessment to its usefulness for its intended use case.

4.4. Open questions for further research

The evaluation of symptom checker apps is a relatively new scientific endeavour, and no framework has yet been established on how best to assess symptom checker stand-alone capabilities or how to estimate the risks and opportunities of integrating such apps into routine clinical care [32]. Building on our own contribution to the field, the following outlines two directions that future research can take: a) addressing the question of how the methodology of vignette-based studies can advance to become more reproducible and more meaningful, and b) addressing the interaction of users and apps and the factors influencing this interaction.

4.4.1. Advancing the methodology of vignette-based studies on triage accuracy

Numerous studies have tested symptom checker apps or differential diagnostic generators with case vignettes. Unfortunately, their results are often not directly comparable to each other for many reasons, three of which are addressed here and remedies suggested for improving generalisability.

4.4.1.1. Dichotomising triage appraisals

Our study follows Semigran et al. in categorising the case vignettes into three levels of urgency, or triage levels, namely emergency-care, non-emergency-care and self-care [7]. Other studies assessing the triage ability of symptom checker apps have used different classifications of urgency, for example only two levels (emergency and non-emergency) [6], four levels [8], or even six levels [36]. Naturally, the number of triage levels, that is, the number of possible answers, influences the accuracy found in the tested symptom checker app and thus our judgement of it: with a binary triage decision, even random guessing would result in an accuracy rate of around 50%, while with six triage levels, an attained accuracy rate of 50% would be far above what can be expected from a poorly performing system providing nothing more than randomly chosen estimates.

Coming to a consensus on how to subdivide and define triage levels will prove difficult, as healthcare systems differ widely regarding what levels and providers of care the patient can choose from and how accessible these are. One approach to this issue of differing triage-level definitions is to evaluate symptom checkers on a set of binary triage questions, for example, (a) can the symptom checker apps reliably discern between complaints requiring emergency care versus non-emergency care and no professional care, and (b) can symptom checker apps reliably discern between whether complaints need professional care at all or whether self-care is appropriate instead. These and similar binary questions would allow utilising analysis methods of signal detection theory, such as determining sensitivity, specificity, and the area under the curve of the receiver operating characteristic, or more modern approaches such as the decision curve analysis [56]. These analyses of binary—and more actionable recommendations — are of great value in identifying the potential use cases of symptom checkers. For example, their advice on where to seek care might be unreliable, but the provided advice on whether care should be sought at all might be reliable and safe, and thus beneficial to users seeking advice on that specific question.

4.4.1.2. Differentiating vignettes

The number of case vignettes included in case vignette-based evaluation studies varies greatly, from as little as four [54] to as many as 200 [36]. The more vignettes used, the more meaningful the results ostensibly are, as a high number of diverse vignettes can better resemble the distribution and wide variety of complaints for which patients seek advice in a primary-care setting. However, as case vignettes are primarily a means for discerning symptom checker apps with high performance rates from those

with low performance rates, they themselves should be assessed on their suitability for this task. Our analyses on the data provided by Semigran et al. show that some case vignettes were assessed correctly by all the apps and others by none [7,44]. Such vignettes cannot contribute to differentiating between high- and low-performing apps. Similarly, case vignettes on which high-performing apps fail but low-performing apps do not might potentially have been assigned an erroneous gold-standard solution or be highly ambiguous, thus not providing a sufficient basis for a sound decision. Irrespective of the reasons why this occurs, such vignettes are not helpful in differentiating between high- and low-performing apps, either.

Accordingly, future studies should focus on developing metrics to assess the suitability of case vignettes to attain a higher informative value, instead of (solely) relying on increasing the quantity of them used in studies.

4.4.1.3. Transparent, pre-defined instructions on entering symptoms into apps

The user interface and reasoning engines underlying symptom checker apps differ substantially: some symptom checker apps prompt the user to provide specific information, while others expect the user to provide all input with no prompts at all. For example, the symptom checker *Isabel* lets the user enter input as free text and only asks a limited number of fixed, that is, non-adaptive questions, for example the respective patient's age and sex [57]. In contrast, symptom checker apps like *Symptomate* [58] and *Ada – check your health* [59] let the user choose from a finite set of symptoms as chief complaints and, based on these, prompt the user in a chatbot format to provide more information, for example on the onset of these symptoms or whether certain additional symptoms are present. These differences in the interaction between user and symptom checker apps provides an obstacle to meaningfully comparing the performance of symptom checker apps against each other since they require the test user to make many decisions: How should a symptom checker app's question be answered on which the case vignettes provide no information? Should all of a vignette's information be presented unhesitatingly by the test user, or should some information be provided only if specifically asked for by the app, as a real user might engage? No matter how elaborately the case vignettes are designed, they will most likely still contain some ambiguities or omit information that one or many symptom checker apps specifically address in their assessments. Accordingly, specifying instructions on how test users should handle such instances when interacting with the

assessed symptom checkers and publishing these instructions as given should become an essential part of the methods published in articles assessing symptom checker apps. At the moment, this is not standard practice, which hinders researchers' ability to reproduce, but also comparability between study results.

4.4.2. Interaction between layperson users and apps

Apart from assessing the performance of symptom checker apps and establishing benchmarking criteria to help in judging their capabilities, the potential benefit of these apps can only be estimated when one looks at whether and when their users would follow the presented advice and which groups of users are most and least in need of this advice. Hence, after having elaborated three suggestions above on how the methods of symptom checker app performance evaluation can improve, the following describes two paths that future research on the interaction between symptom checker apps and their users might follow.

4.4.2.1. Subpopulations of laypersons benefitting differently

Our results indicate that neither age nor level of education showed any influence on risk-aversion but that gender did. Gender differences probably are not the only factor influencing laypersons' triage accuracy and triage behaviour; health literacy and risk-aversion in general might also be influencing factors. Thus when approaching the question of whether the use of symptom checkers is beneficial to a layperson, a more nuanced answer indicates that their use is potentially beneficial to some but not all laypersons. For example, the given data on symptom checkers characterises them as rather risk-averse and thus very good at detecting emergencies. Such risk-averse symptom checkers might be beneficial to laypersons currently inclined to underestimate risks, while less useful to laypersons sensitive towards emergencies themselves. For that reason, further research could follow the path of analysing who would benefit from such apps, and whether users can reliably judge what kind of decision support they require and when. For example, users might benefit from a tool reassuring them that self-care is appropriate, while a tool that is more risk-averse than the users themselves could cause more anxiety than benefit, an effect termed cyberchondria in the literature [60,61]. Also, if users cannot adequately judge when they would benefit from consulting a self-assessment app and when they would not, then even high-performing symptom checker apps would be of only limited value, and research focus should turn towards educating the laypersons rather than improving the apps.

4.4.2.2. Influences on trust

A symptom checker app's reliability in providing safe and accurate advice does not on its own provide sufficient proof that it might be beneficial to a user in supporting his or her decision-making. The user must in fact trust the symptom checker app in order for it to benefit the user.

Current symptom checker apps use different methods to attain their users' trust. Some apps accompany their advice with explanations, illustrating how the information provided by the user contributes to the app's suggestions, for example by using Sankey diagrams. Others include anthropomorphic elements (e.g., icons resembling human physicians) in the app's user interface to emulate a patient-physician interaction. Studies on user-machine interaction in other use cases provide evidence of such measures affecting trust in automation [62,63]. It cannot be assumed, however, that the findings from these studies apply to the interaction between users and symptom checkers, as trust in automation is dependent on the specific use case and stereotypical assumptions about the advising agent's expertise [64]. Research on trust in automation-related aspects of symptom checker apps will prove of value for two major reasons: A landmark study on trust in automation has shown that users tend to "under-trust" advice from software when it is not perfectly accurate, even when it is much more accurate than the user is [65]. Thus, even highly accurate and safe symptom checker apps worthy of the user's trust still need to actively employ other means of acquiring the user's trust in order to fulfil their potential for guiding the user through the healthcare system and disburdening over-utilized healthcare services such as emergency departments. On the other hand, efficacious measures for acquiring the user's trust pose a threat when they are utilized by symptom checker apps not worthy of attaining that level of trust, that is, such measures might impede the user from critically appraising the app's advice. Measures proven to influence the user's trust in the app poorly calibrated to the app's ability could be used in a manipulative manner. Accordingly, independent research should investigate how efficacious different trust-building measures actually are in convincing a user to follow an app's correct advice, while at the same time considering whether such measures aggravate the risk of users being misled by an app's wrong advice.

3 Literaturangaben

1. Ledley RS, Lusted LB. Reasoning Foundations of Medical Diagnosis. Science 1959 Jul 3;130(3366):9–21. [doi: 10.1126/science.130.3366.9]

2. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. Computers and Biomedical Research 1975 Aug;8(4):303–320. [doi: 10.1016/0010-4809(75)90009-9]

3. Chewning B, Bylund CL, Shah B, Arora NK, Gueguen JA, Makoul G. Patient preferences for shared decisions: A systematic review. Patient Education and Counseling 2012 Jan;86(1):9–18. [doi: 10.1016/j.pec.2011.02.004]

4. Veatch RM. Models for Ethical Medicine in a Revolutionary Age. The Hastings Center Report 1972 Jun;2(3):5. [doi: 10.2307/3560825]

5. Chambers D, Cantrell A, Johnson M, Preston L, Baxter SK, Booth A, Turner J. Digital and online symptom checkers and assessment services for urgent care to inform a new digital platform: a systematic review. Health Serv Deliv Res 2019 Aug;7(29):1–88. [doi: 10.3310/hsdr07290]

6. Yu SWY, Ma A, Tsang VHM, Chung LSW, Leung S-C, Leung L-P. Triage accuracy of online symptom checkers for Accident and Emergency Department patients. Hong Kong Journal of Emergency Medicine 2020 Jul;27(4):217–222. [doi: 10.1177/1024907919842486]

7. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 8;h3480. [doi: 10.1136/bmj.h3480]

8. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Medical Journal of Australia 2020 May 11;mja2.50600. [doi: 10.5694/mja2.50600]

9. Farmer N, Schilstra MJ. A Knowledge-based Diagnostic Clinical Decision Support System for Musculoskeletal Disorders of the Shoulder for Use in a Primary Care Setting. Shoulder & Elbow 2012 Apr;4(2):141–151. [doi: 10.1111/j.1758-5740.2011.00165.x]

10. Elliot AJ, Kara EO, Loveridge P, Bawa Z, Morbey RA, Moth M, Large S, Smith GE. Internet-based remote health self-checker symptom data as an adjuvant to a national syndromic surveillance system. Epidemiol Infect 2015 Dec;143(16):3416–3422. [doi: 10.1017/S0950268815000503]

11. Mehl A, Bergey F, Cawley C, Gilsdorf A. Syndromic Surveillance Insights from a Symptom Assessment App Before and During COVID-19 Measures in Germany and the United Kingdom: Results From Repeated Cross-Sectional Analyses. JMIR Mhealth Uhealth 2020 Oct 9;8(10):e21364. [doi: 10.2196/21364]

12. Morita T, Rahman A, Hasegawa T, Ozaki A, Tanimoto T. The Potential Possibility of Symptom Checker. Int J Health Policy Manag 2017 Apr 5;6(10):615–616. [doi: 10.15171/ijhpm.2017.41]

13. Kujala S, Hörhammer I, Hänninen-Ervasti R, Heponiemi T. Health Professionals' Experiences of the Benefits and Challenges of Online Symptom Checkers. Stud Health Technol Inform 2020 Jun 16;270:966–970. PMID:32570525

14. Oslislo S, Heintze C, Schmiedhofer M, Möckel M, Schenk L, Holzinger F. How to decide adequately? Qualitative study of GPs' view on decision-making in self-referred and physician-referred emergency department consultations in Berlin, Germany. BMJ Open 2019 Apr;9(4):e026786. [doi: 10.1136/bmjopen-2018-026786]

15. Martin SS, Quaye E, Schultz S, Fashanu OE, Wang J, Saheed MO, Prem Ramaswami, de Freitas H, Ribeiro-Neto B, Parakh K. A randomized controlled trial of online symptom searching to inform patient generated differential diagnoses. npj Digit Med 2019 Dec;2(1):110. [doi: 10.1038/s41746-019-0183-0]

16. Cocco AM, Zordan R, Taylor DM, Weiland TJ, Dilley SJ, Kant J, Dombagolla M, Hendarto A, Lai F, Hutton J. Dr Google in the ED: searching for online health information by adult emergency department patients. Medical Journal of Australia 2018 Oct;209(8):342–347. [doi: 10.5694/mja17.00889]

17. Van Riel N, Auwerx K, Debbaut P, Van Hees S, Schoenmakers B. The effect of Dr Google on doctor–patient encounters in primary care: a quantitative, observational, cross-sectional study. Br J Gen Pract Open 2017 Jul 10;1(2):BJGP-2017-0833. [doi: 10.3399/bjgpopen17X100833]

18. 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]

19. Mueller J, Jay C, Harper S, Davies A, Vega J, Todd C. Web Use for Symptom Appraisal of Physical Health Conditions: A Systematic Review. J Med Internet Res 2017 Jun 13;19(6):e202. [doi: 10.2196/jmir.6755]

20. Nijland N, Cranen K, Verlinden SFF, Kelders SM, Boer H, Gemert-Pijnen JEWC van. Computer Generated Self-Care Advice via Web-Based Triage of Complaints in Primary Care. 2009 International Conference on eHealth, Telemedicine, and Social Medicine [Internet] Cancun, Mexico: IEEE; 2009 [cited 2020 Mar 13]. p. 129–135. [doi: 10.1109/eTELEMED.2009.17]

21. North F, Varkey P, Laing B, Cha SS, Tulledge-Scheitel S. Are e-Health Web Users Looking for Different Symptom Information Than Callers to Triage Centers? Telemedicine and e-Health 2011 Jan;17(1):19–24. [doi: 10.1089/tmj.2010.0120]

22. Morse KE, Ostberg NP, Jones VG, Chan AS. Digital Symptom Checker Usage and Triage: Population-Based Descriptive Study in a Large North American Integrated Health System. 2020 Nov 7 [cited 2020 Nov 20]; [doi: 10.2196/20549]

23. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient Perspectives on the Usefulness of an Artificial Intelligence–Assisted Symptom Checker: Cross-Sectional Survey Study. J Med Internet Res 2020 Jan 30;22(1):e14679. [doi: 10.2196/14679]

24. Fox S, Duggan M. Health Online 2013 [Internet]. Washington, D.C.: Pew Research Center; 2013 [Accessed 2021 Mar 6]. Available from: https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/PIP_HealthOnline.pdf

25. Baumann E, Czerwinski F, Rosset M, Seelig M, Suhr R. Wie informieren sich die Menschen in Deutschland zum Thema Gesundheit? Erkenntnisse aus der ersten Welle von HINTS Germany. Bundesgesundheitsbl 2020 Sep;63(9):1151–1160. [doi: 10.1007/s00103-020-03192-x]

26. EPatient Analytics GmbH. EPatient Survey 2020 [Internet]. 2020 [Accessed 2021 Mar 6]. Available from: https://www.hcm-magazin.de/epatient-survey-2020-digital-health-studie/150/10992/407743

27. Doctolib GmbH. Mehr Informationen zu digitalen Services im Gesundheitswesen notwendig [Internet]. 2020 [Accessed 2021 Mar 6]. Available from: https://info.doctolib.de/blog/mehr-informationen-zu-digitalen-services-im-gesundheitswesen-notwendig/

28. Nijland N, Cranen K, Boer H, van Gemert-Pijnen JEWC, Seydel ER. Patient use and compliance with medical advice delivered by a web-based triage system in primary care. J Telemed Telecare 2010 Jan;16(1):8–11. [doi: 10.1258/jtt.2009.001004]

29. Using technology to ease the burden on primary care [Internet]. Healthwatch Enfield; 2019. Report No.: Rep-4398. [Accessed 2021 Mar 6]. Available from: https://www.healthwatch.co.uk/sites/healthwatch.co.uk/files/reports-library/20190122_Enfield_%20Using%20technology%20to%20ease%20the%20burden%20on%20primary%20care.pdf

30. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential Diagnosis Generators: an Evaluation of Currently Available Computer Programs. Journal of General Internal Medicine 2012 Feb;27(2):213–219. [doi: 10.1007/s11606-011-1804-8]

31. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, Ball EV, Cobbs CG, Dennis VW, Frenkel EP, Hudson LD, Mancall EL, Rackley CE, Taunton OD. Performance of Four Computer-Based Diagnostic Systems. N Engl J Med 1994 Jun 23;330(25):1792–1796. [doi: 10.1056/NEJM199406233302506]

32. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. The Lancet 2018 Nov;392(10161):2263–2264. [doi: 10.1016/S0140-6736(18)32819-8]

33. Berry AC, Cash BD, Mulekar MS, Wang B, Melvin A, Berry BB. Symptom Checkers vs. Doctors, the Ultimate Test: A Prospective Study of Patients Presenting with Abdominal Pain. Gastroenterology 2017 Apr;152(5):S852–S853. [doi: 10.1016/S0016-5085(17)32937-2]

34. Bisson LJ, Komm JT, Bernas GA, Fineberg MS, Marzo JM, Rauh MA, Smolinski RJ, Wind WM. How Accurate Are Patients at Diagnosing the Cause of Their Knee Pain With the Help of a Web-based Symptom Checker? Orthopaedic Journal of Sports Medicine 2016 Feb;4(2):232596711663028. [doi: 10.1177/2325967116630286]

35. Shen C, Nguyen M, Gregor A, Isaza G, Beattie A. Accuracy of a Popular Online Symptom Checker for Ophthalmic Diagnoses. JAMA Ophthalmol 2019 Jun 1;137(6):690. [doi: 10.1001/jamaophthalmol.2019.0571]

36. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, Millen E, Multmeier J, Pick F, Richter C, Tuerk E, Upadhyay S, Virani V, Vona N, Wicks P, Novorol C. Original research: How accurate are digital symptom assessment apps for suggesting conditions and urgency advice?: a clinical vignettes comparison to GPs. medRxiv [Internet] Cold Spring Harbor Laboratory Press; 2020; [doi: 10.1101/2020.05.07.20093872]

37. Moreno Barriga E, Pueyo Ferrer I, Sánchez Sánchez M, Martín Baranera M, Masip Utset J. [A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application]. Emergencias 2017 Dic;29(6):391–396. PMID:29188913

38. Gottliebsen K, Petersson G. Limited evidence of benefits of patient operated intelligent primary care triage tools: findings of a literature review. BMJ Health Care Inform 2020 May;27(1):e100114. [doi: 10.1136/bmjhci-2019-100114]

39. Lewis TL, Wyatt JC. mHealth and Mobile Medical Apps: A Framework to Assess Risk and Promote Safer Use. J Med Internet Res 2014 Sep 15;16(9):e210. [doi: 10.2196/jmir.3133]

40. Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of Physician and Computer Diagnostic Accuracy. JAMA Internal Medicine 2016 Dec 1;176(12):1860. [doi: 10.1001/jamainternmed.2016.6001]

41. Middleton K, Butt M, Hammerla N, Hamblin S, Mehta K, Parsa A. Sorting out symptoms: design and evaluation of the 'babylon check' automated triage system. 2016 Jun 7;6. [https://arxiv.org/abs/1606.02041]

42. Schembri G, Schober P. The Internet as a diagnostic aid: the patients' perspective. International Journal of STD & AIDS 2009 Apr;20(4):231–233. [doi: 10.1258/ijsa.2008.008339]

43. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a Chatbot (Ada) in the Diagnosis of Mental Disorders: Comparative Case Study With Lay and Expert Users. JMIR Form Res JMIR Publications; 2019 Oct 29;3(4):e13863–e13863. [doi: 10.2196/13863]

44. Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F. Benchmarking Triage Capability of Symptom Checkers Against That of Medical Laypersons: Survey Study. J Med Internet Res 2021 Mar 10;23(3):e24475. [doi: 10.2196/24475]

45. Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Science Advances 2018 Jan;4(1):eaao5580. [doi: 10.1126/sciadv.aao5580]

46. Mills B, Hill M, Buck J, Walter E, Howard K, Raisinger A, Smith E. What constitutes an emergency ambulance call? Australasian Journal of Paramedicine [Internet] 2019 Mar 22 [cited 2020 Jul 13];16. [doi: 10.33151/ajp.16.626]

47. Healthy Children [Internet]. American Academy of Pediatrics; [cited 2021 Mar 21]. Available from: https://www.healthychildren.org/english/tips-tools/symptom-checker/Pages/default.aspx

48. Amazon Mechanical Turk [Internet]. [Accessed 2021 March 6]. Available from: https://www.mturk.com/

49. Yu CH. Resampling methods: Concepts, Applications, and Justification. University of Massachusetts Amherst; [cited 2020 Dec 22]; [doi: 10.7275/9CMS-MY97]

50. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/

51. Schmieding, Malte L, Mörgeli, Rudolf, Schmieding, Maike A L, Feufel, Markus A, Balzer, Felix. Data set supplementing "Benchmarking triage capability of symptom checkers against that of medical laypersons: Survey study" [Internet]. Zenodo; 2021 [cited 2021 Jan 21]. [doi: 10.5281/ZENODO.4454538]

52. YoungHo K, Park I, Kang S. Age and gender differences in health risk perception. Cent Eur J Public Health 2018 Mar 30;26(1):54–59. [doi: 10.21101/cejph.a4920]

53. Harris CR, Jenkins M. Gender Differences in Risk Assessment: Why do Women Take Fewer Risks than Men? Judgment and Decision Making 2006;1(1):16.

54. Ćirković A. Evaluation of Four Artificial Intelligence–Assisted Self-Diagnosis Apps on Three Diagnoses: Two-Year Follow-Up Study. J Med Internet Res 2020 Dec 4;22(12):e18097. [doi: 10.2196/18097]

55. 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 [Internet] 2020 Jul 10; [doi: 10.2196/19713]

56. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Making 2006 Nov;26(6):565–574. [doi: 10.1177/0272989X06295361]

57. Isabel Symptom Checker [Internet]. Isabel Healthcare; [cited 2021 Mar 21]. Available from: https://symptomchecker.isabelhealthcare.com/suggest_diagnoses_advanced/landing_page

58. Symptomate Symptom Checker [Internet]. Infermedica; [cited 2021 Mar 21]. Available from: https://symptomate.com/diagnosis/

59. Ada - check your health [Internet]. Ada Health GmbH; [cited 2021 Mar 21]. Available from: https://play.google.com/store/apps/details?id=com.ada.app&hl=en_US&gl=US

60. Doherty-Torstrick ER, Walton KE, Fallon BA. Cyberchondria: Parsing Health Anxiety From Online Behavior. Psychosomatics 2016 Jul;57(4):390–400. [doi: 10.1016/j.psym.2016.02.002]

61. Harding KJ, Skritskaya N, Doherty E, Fallon BA. Advances in understanding illness anxiety. Curr Psychiatry Rep 2008 Aug;10(4):311–317. [doi: 10.1007/s11920-008-0050-1]

62. de Visser EJ, Monfort SS, McKendrick R, Smith MAB, McKnight PE, Krueger F, Parasuraman R. Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied 2016 Sep;22(3):331–349. [doi: 10.1037/xap0000092]

63. Pak R, Fink N, Price M, Bass B, Sturre L. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. Ergonomics 2012;55(9):1059–1072. PMID:22799560

64. Hertz N, Wiese E. Good advice is beyond all price, but what if it comes from a machine? Journal of Experimental Psychology: Applied 2019 Sep;25(3):386–395. [doi: 10.1037/xap0000205]

65. Van Dongen K, Van Maanen P-P. Under-reliance on the decision aid: A difference in calibration and attribution between self and aid. 50th Annual Meeting of the Human Factors and Ergonomics Society, San Francisco, CA; 2006.

4 Eidesstattliche Versicherung

„Ich, Malte Schmieding, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema *A benchmarking comparison of triage capability between 15 symptom checker apps and medical laypersons - Vergleichende Beurteilung der Fähigkeit die Dringlichkeit medizinischer Beschwerdebilder einzuschätzen zwischen 15* symptom checker apps *und medizinischen Laien* selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Erstbetreuer/in, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

5 Ausführliche Anteilserklärung an der erfolgten Publikation

Publikation:

Malte Schmieding, Rudolf Mörgeli, Maike Anna Lena Schmieding, Markus A. Feufel and Felix Balzer. *Benchmarking Triage Capability of Symptom Checkers Against That of Medical Laypersons: Survey Study*. Journal of Medical Internet Research 2021;23(3). 10. März 2021.

Beiträge im Einzelnen:

- Thema und Idee (Malte Schmieding)
- Auswertungsmethodik (Malte Schmieding, Feufel, Balzer)
- Entwicklung des Fragebogens (Malte Schmieding)
- Modifikation der Fallvignetten (Malte Schmieding, Mörgeli, Maike Schmieding)
- Rekrutierung der Studienteilnehmer (Malte Schmieding)
- Datenaufbereitung (Malte Schmieding)
- Statistische Auswertung (Malte Schmieding)
- Erstellung aller Graphiken und Tabellen (Malte Schmieding)
- Diskussion der Methode und Ergebnisse (Malte Schmieding, Feufel)
- Erster Manuskriptentwurf (Malte Schmieding)
- Inhaltliches Korrekturlesen (Feufel, Mörgeli, Balzer, Maike Schmieding)
- Stilistisches Korrekturlesen (Frances Lorié)
- Präsentation auf Konferenzen: 2. Versorgungsforschungskongress der Charité (Malte Schmieding)

_____
Unterschrift, Datum und Stempel des erstbetreuenden Hochschullehrers

_____
Unterschrift des Doktoranden

6 Auszug aus der Journal Summary List

Journal Data Filtered By: **Selected JCR Year: 2019** Selected Editions: SCIE,SSCI
Selected Categories: **"MEDICAL INFORMATICS"**
Selected Category Scheme: WoS
Gesamtanzahl: 27 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	IEEE Journal of Biomedical and Health Informatics	5,472	5.223	0.012910
2	JOURNAL OF MEDICAL INTERNET RESEARCH	16,349	5.034	0.029410
3	ARTIFICIAL INTELLIGENCE IN MEDICINE	2,953	4.383	0.003370
4	JMIR mHealth and uHealth	4,226	4.313	0.010020
5	JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	9,959	4.112	0.017380
6	COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE	8,014	3.632	0.011370
7	JMIR Serious Games	350	3.526	0.000660
7	JOURNAL OF BIOMEDICAL INFORMATICS	8,253	3.526	0.011190
9	Internet Interventions-The Application of Information Technology in Mental and Behavioural Health	996	3.513	0.002720
10	JOURNAL OF MEDICAL SYSTEMS	5,695	3.058	0.007050

7 Originalpublikation

- Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F. *Benchmarking Triage Capability of Symptom Checkers Against That of Medical Laypersons: Survey Study*.J Med Internet Res 2021;23(3):e24475.

  https://doi.org/10.2196/24475

Original Paper

Benchmarking Triage Capability of Symptom Checkers Against That of Medical Laypersons: Survey Study

Malte L Schmieding[1,2], MD; Rudolf Mörgeli[1], MD; Maike A L Schmieding[3]; Markus A Feufel[4*], Dipl-Ing (FH), MSc, PhD; Felix Balzer[1,2*], MSc, PhD, MD

[1]Department of Anesthesiology and Operative Intensive Care, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[2]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[3]Department of Biology, Chemistry, and Pharmacy, Institute of Pharmacy, Freie Universität Berlin, Berlin, Germany

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

[*]these authors contributed equally

Corresponding Author:
Felix Balzer, MSc, PhD, MD
Institute of Medical Informatics
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin
Charitéplatz 1
Berlin, 10117
Germany
Phone: 49 30 450 5704
Email: felix.balzer@charite.de

Abstract

Background: Symptom checkers (SCs) are tools developed to provide clinical decision support to laypersons. Apart from suggesting probable diagnoses, they commonly advise when users should seek care (*triage advice*). SCs have become increasingly popular despite prior studies rating their performance as mediocre. To date, it is unclear whether SCs can triage better than those who might choose to use them.

Objective: This study aims to compare triage accuracy between SCs and their potential users (ie, laypersons).

Methods: On Amazon Mechanical Turk, we recruited 91 adults from the United States who had no professional medical background. In a web-based survey, the participants evaluated 45 fictitious clinical case vignettes. Data for 15 SCs that had processed the same vignettes were obtained from a previous study. As main outcome measures, we assessed the accuracy of the triage assessments made by participants and SCs for each of the three triage levels (ie, *emergency care*, *nonemergency care*, *self-care)* and overall, the proportion of participants outperforming each SC in terms of accuracy, and the risk aversion of participants and SCs by comparing the proportion of cases that were overtriaged.

Results: The mean overall triage accuracy was similar for participants (60.9%, SD 6.8%; 95% CI 59.5%-62.3%) and SCs (58%, SD 12.8%). Most participants outperformed all but 5 SCs. On average, SCs more reliably detected emergencies (80.6%, SD 17.9%) than laypersons did (67.5%, SD 16.4%; 95% CI 64.1%-70.8%). Although both SCs and participants struggled with cases requiring self-care (the least urgent triage category), SCs more often wrongly classified these cases as emergencies (43/174, 24.7%) compared with laypersons (56/1365, 4.10%).

Conclusions: Most SCs had no greater triage capability than an average layperson, although the triage accuracy of the five best SCs was superior to the accuracy of most participants. SCs might improve early detection of emergencies but might also needlessly increase resource utilization in health care. Laypersons sometimes require support in deciding when to rely on self-care but it is in that very situation where SCs perform the worst. Further research is needed to determine how to best combine the strengths of humans and SCs.

(J Med Internet Res 2021;23(3):e24475) doi: 10.2196/24475

KEYWORDS

digital health; triage; symptom checker; patient-centered care; eHealth apps; mobile phone; decision support systems; clinical; consumer health information; health literacy

Introduction

Use of Symptom Checkers

Patients obtain health-related information from health care professionals, but more frequently, information for patients is provided in print; on the web; and, most recently, via smartphone apps. Patients not only use these resources to supplement information received from health care professionals but also as a decision-support tool to advise them on whether and where to seek adequate health care, especially as health care pathways grow more complex. Symptom checkers (SCs) are tools developed to provide support to laypersons. Users can enter their complaints and, with some SCs, demographic or health-related information (eg, age, sex, and past medical history) to obtain advice on the urgency of their complaints (*triage advice*) and the most likely diagnosis. The demand for this type of support is evident; in the United States, 1 in 3 people reported resorting to the internet for self-diagnosis [1], and a study from 2019 found that half of the patients involved in that study had investigated their symptoms with an online search engine before going to an emergency department [2].

Evidence on SCs

Despite their popularity, there is no established framework to evaluate the performance of SCs [3,4]. The use of case vignettes, based on real or fictitious patients, has been a common approach for rating SCs [5-9]. The 2 most recent non–industry-funded audit studies using this methodology rated SC triage capability as unreliable, with an average of only 49% and 58% of appraisals deemed correct [10,11]. In line with these findings, a 2020 literature review concluded that most investigated SCs offered limited benefits [12].

A study showing that laypersons are just as capable of predicting criminal recidivism as a complex commercial algorithm [13] inspired us to compare the triage capability of SCs with that of participants with little or no medical training: are SCs merely a more complicated means of pointing out what an untrained individual could just as easily deduce? Is there an advantage to consulting SCs instead of relying on one's own judgment?

In addition to advising the individual user, SCs are also said to have the potential to reduce the burden on health care services. Unfortunately, not only has this potential benefit not materialized yet [3] but also there is evidence of the opposite effect, as overly risk-averse SCs promote more visits to emergency care services [14]. To address this issue, we also analyzed whether SCs were more risk averse than our participants. Although SCs can also provide diagnostic suggestions, we considered triage advice to be more relevant for assessing the impact of SC on use of health care resources and patient safety.

The purpose of this study is to benchmark the triage capability of SCs against that of their potential users, that is, laypersons.

Methods

Ethics Approval and Consent to Participate

This study was approved by the Ethics Committee of the Department of Psychology and Ergonomics (Institut für Psychologie und Arbeitswissenschaft) at Technische Universität Berlin (tracking number: FEU_03_20180615). Participants volunteered to participate in the survey, and informed consent was required.

Data Collection

This investigation builds on a prior study by Semigran et al [11], who evaluated SC triage performance based on case vignettes. We used their results on the performance of SCs as well as their case vignettes. Data were collected to determine the triage ability of medical laypersons, which was then used as a benchmark for comparing laypersons' performance with that of SCs.

Participants

All participants were US residents, at least 18 years of age, and had no professional medical background. Our investigation was limited to US residents, as the triage level definitions and the gold standard solutions assigned to the case vignettes by Semigran et al [11] might only be applicable to the US health care environment and might not apply to other health care systems with different service provider options.

Survey

We created an online survey with UNIPARK (QuestBack GmbH) [15] containing questions on demographics (age, sex, US residency, and highest level of completed formal education), past online searching behavior for medical information, 45 randomly ordered clinical case vignettes, and 5 attention checks (see *Procedure* for further details). We used the 45 case vignettes compiled and adjusted by Semigran et al [11], which are between 1 and 3 sentences long and describe a patient's signs and symptoms and occasionally mention elements of the patient's past medical history.

Participants were asked to classify each vignette into 1 of 3 triage categories, as defined by Semigran et al [11]: *emergency care*, involving "the advice to call an ambulance, go to an emergency department, or see a general practitioner immediately"; *nonemergency care*, which encompasses "advice to call a general practitioner or primary care provider, see a general practitioner or primary care provider, go to an urgent care facility, go to a specialist, go to a retail clinic, or have an e-visit"; and *self-care*, which is "advice to stay at home or go to a pharmacy." The definition of each triage level was explained at the beginning of the survey. The understanding of these definitions by participants was ascertained by 3 control questions given before the case vignettes were presented. The questionnaire was piloted with 12 participants and refined

according to their feedback to ensure readability and understandability.

Preparing the Case Vignettes

The 45 standardized case vignettes included 15 cases for each triage level. The vignettes, as chosen by Semigran et al [11], included both common and uncommon conditions with a wide range of chief complaints. The vignettes stemmed from various clinical sources, including material used to educate health care professionals.

For the purpose of our study, the vignettes were adapted to increase the comprehensibility of lay individuals. First, we transformed the bullet points into complete sentences. Second, we paraphrased technical terms. For example, we replaced "rhinorrhea" with "runny nose" and "tender" with "painful to the touch." In very few cases, explanations required elaboration. Our overall aim was to provide participants with the same information used by Semigran et al [11] to assess SCs. We deemed 1 case vignette vague regarding a crucial piece of information and had to supplement it with a detail left out in the Semigran et al [11] version of the vignette (see Multimedia Appendix 1 [11] for details). We retained the classification of the 45 case vignettes into 3 triage levels.

Understandability and paraphrasing were cross-validated by two native English speakers: one was a medical professional (RM) and the other was without a professional medical background (MALS). The adapted vignettes are shown in Multimedia Appendix 1.

Procedure

We recruited the participants through Amazon Web Service *Amazon Mechanical Turk* (MTurk), as it provides an established means to recruit US-based participants for sociopsychological surveys and is easy to access for researchers working outside of the United States [16]. Each participant received US $4.00 for completing the survey and a US $3.00 bonus if their overall accuracy in assigning the correct triage level was greater than or equal to 58%. The bonus was intended to provide an incentive for participants to pay close attention to the case vignettes and to assess a case's urgency as accurately as possible. The chosen threshold of 58% corresponds to outperforming the SC average reported by Semigran et al [11].

Two methods were employed to ensure that the participants paid close attention to the survey questions. First, we added 5 attention checks to the set of 45 case vignettes. These attention checks were formatted similarly to the case vignettes but included prompts to choose specific answer options. Participants were excluded from the analysis if they answered any of the 5 attention checks incorrectly. Second, upon completion of the survey, participants were asked to affirm that they were attentive and honest to improve the reliability of our data, as suggested in a reliability analysis on MTurk data [17]. We assured participants that they would be compensated for completing the survey even if they stated that they had responded inattentively or dishonestly. We analyzed data only from participants who affirmed their honesty and attentiveness.

The survey on MTurk was published on 3 different days (March 21, 2020, at 2 PM Pacific Daylight Time [PDT]; March 22, 2020, at 1:45 PM PDT; and March 29, 2020, at 1 PM PDT). By selecting the weekend day and early afternoon PDTs, we attempted to reach an MTurk population as diverse as possible, following a 2017 study on the intertemporal variation of the MTurk population [18]. On each day, participants were recruited within a few hours of publishing the survey.

Due to limited funding, the sample size was ultimately determined by the availability of funds and the number of participants who performed well enough to earn a bonus.

Data Analysis

Data were cleaned and explored using *R* 4.0.0 [19] and *tidyverse* packages [20]. Inferential analysis was conducted using the packages *lme4* [21] and *infer* [22]. Figures were created using the package *ggplot2* [23]. The data set containing participants' triage assessments and their demographic variables was made publicly available [24].

Following Semigran et al [11], we refer to each instance of an SC or a participant assessing a case vignette as a "case evaluation." For example, 2 participants each assessing all 45 case vignettes yielded 90 case evaluations.

Participant Characteristics

To assess the effects of demographic variables (age, sex, and educational level), a logistic regression was performed with the correct triage of a case vignette as a dependent variable. We calculated 95% CIs for the marginal probabilities of the fixed effects using the Wald method to assess whether demographic variables had a significant effect on participants' accuracy. The α level was set at .05.

Comparing SCs and Participants

For the comparison of SCs and participants, we performed (1) a comparison between participants and all rated SCs aggregated and (2) between participants and individual SCs.

Aggregate Comparison of SCs and Participants

The performance of the SCs was obtained from the appendix of the audit study by Semigran et al [11]. Comparisons were made between SCs and participants in terms of (1) triage accuracy, (2) tendency to overtriage (*risk aversion*), and (3) how difficult each case vignette was for the respective group (SCs and participants). Of the 15 SCs, 4 (*iTriage*, *Isabel*, *Symcat*, and *Symptomate*) were designed to never suggest self-care, with 1 SC (*iTriage*) always advising users to seek emergency care. To ensure that our results were not skewed by these special SCs, we conducted the main aggregate analyses twice, including and excluding those 4 SCs, and reporting results for both.

Triage Accuracy

Following Semigran et al [11], we compared the performance of SCs and participants at an aggregate level and for each triage level separately and overall. This was performed by calculating the sample's mean accuracy for SCs and participants, with accuracy defined as the proportion of vignettes solved correctly. For the participants, the standard error of the sampling mean with 95% CIs was estimated by bootstrapping the participant

data with 15,000 replications. The limits of the CI were calculated using the quantile method (2.5th and 97.5th quantile of the bootstrap sample means). The CIs for the SC sample were not calculated, as Semigran et al [11] sampled the SCs purposefully, that is, they selected which SCs to evaluate with care and not randomly.

Risk Aversion

The risk aversion of the SCs and the participants was determined using the ratio of overtriaged vignettes to undertriaged vignettes. We deemed a ratio greater than 1:1, which is more case vignettes overtriaged than undertriaged, as risk averse. To determine what type of triage mistakes were most likely to occur, we calculated the proportion of triage recommendations given in each triage category by SCs and by participants (eg, the proportion of evaluations in which participants recommended emergency care when self-care was appropriate or the proportion of evaluations in which SCs recommended nonemergency care when emergency care would have been the correct solution) and compared these proportions using the Pearson χ^2 test.

Difficulty of Case Vignettes

To analyze whether SCs and participants were challenged by the same case vignettes, the degree of difficulty of a case was calculated using the proportion of SCs and participants correctly triaging it. For example, if a case vignette was solved correctly by every SC, the vignette's degree of difficulty for SCs was 100%. SCs that did not evaluate the respective case vignette for technical reasons were not included in the denominator. A linear correlation analysis was then conducted to determine the relationship between case difficulty for SCs and case difficulty for participants.

Comparing Individual SCs With Participants

As users are likely to use only one or very few SCs, there is no basis for recommendations about using or not using SCs on an aggregated analysis alone. Therefore, additional analyses compared the performance of the participant group with each SC. Considering that most SCs did not evaluate every case vignette (due to technical reasons, see the study by Semigran et al [11]), the triage accuracy of the participants was calculated using only the cases evaluated by a specific SC, enabling a direct comparison. The CIs for participants' mean accuracy were calculated as described above. We also determined the proportion of participants that managed to achieve higher accuracy across cases than the respective SC. Furthermore, risk aversion was also evaluated, given the specific set of case vignettes for any given SC by plotting the proportion of vignettes that were overtriaged against the proportion of those undertriaged for participants versus SC.

Results

Participant Characteristics

Our survey was accessed 142 times in 3 days during which it was available in total, 51 participants were excluded, either for failing attention checks (n=41) or for not fulfilling the eligibility criteria (n=10). All the remaining participants affirmed that they had paid close attention during the survey and answered honestly. This yielded a total of 91 participants, each having assessed all 45 case vignettes, which totaled 4095 case evaluations by participants, 1365 for each triage level (Table 1).

The median time for completion of the survey (excluding the time for obtaining informed consent) was 20 minutes and 12 seconds (1st quartile=15 minutes:43 seconds; 3rd quartile=27 minutes:23 seconds). There was no significant difference in the participants' mean accuracy between the 3 sampling days. We detected no statistically significant influence of demographic variables on participants' triage accuracy.

Table 1. Participant characteristics (N=91).

Characteristics	Values
Age (years), median (range)	37 (20-73)
Gender, n (%)	
Female	36 (40)
Male	55 (60)
Education, n (%)	
Non–high school graduate	0 (0)
High school graduate	18 (20)
Some college	33 (36)
Bachelor's degree	36 (40)
Graduate degree	4 (4)
Recent[a] triage experience, n (%)	
Recently consulted an SC	20 (22)
Recently faced triage decision	23 (25)
Neither faced triage decision nor consulted an SC recently	62 (69)
Medical training, n (%)	
No training	80 (88)
Basic first aid training	11 (12)

[a]Recent was defined as "in the last 6 months."

Comparing SCs' and Participants' Triage Performance

Participant Performance

Overall, the participants triaged 3 out of 5 case vignettes correctly (2462/4065, 60.57%), and most participants qualified for the bonus payment (56/91, 62%). Their mean accuracy varied with triage level, roughly balanced for emergency and nonemergency situations (67.5% and 68.4%, respectively) but dropped below 50% for self-care vignettes. Of the 39.43% (1603/4065) of incorrect assessments, the majority (956/4065, 23.52%) were *overtriaged*, that is, participants assigned a more urgent triage level than necessary. Only about every sixth case vignette was *undertriaged* (647/4065, 15.92%), that is, participants assigned a less urgent triage level than necessary.

Aggregated Comparison Analyses

As most SCs were unable to evaluate at least one of the case vignettes, the 15 SCs assessing the 45 case vignettes yielded only 532 case evaluations (see the study by Semigran et al [11] for details): 183 for emergency vignettes, 175 for nonemergency vignettes, and 174 for self-care vignettes.

Triage Accuracy

At the aggregate level, SCs (58.0%; SD 12.8%) and participants (60.9%; SD 6.8%) showed very similar mean accuracies (Table 2). This remains to be the case when excluding the 4 SCs that did not suggest self-care (adjusted mean for the 11 SCs; 61.6%; SD 11.0%). Table 2 shows that differences become apparent when evaluating the triage levels separately: for emergency case vignettes, SCs outperformed the participants, whereas the participants outperformed the average SC in the nonemergency and self-care cases. For the least urgent triage level, this difference decreases when excluding those SCs that never recommend self-care.

Table 2. Mean triage accuracy of symptom checkers and participants.

Triage level	Percent triage accuracy, mean (SD)			95% CI
	All 15 SCs[a]	Subset of 11 SCs[b]	Participants[c]	
Emergency cases	80.6 (17.9)	79.8 (17.2)	67.5 (16.4)	64.1-70.8
Nonemergency cases	58.5 (29.1)	61.6 (27.8)	68.4 (13.8)	65.6-71.2
Self-care cases	30.6 (25.7)	41.8 (20.3)	46.7 (15.9)	43.4-49.8
Overall	58.0 (12.8)	61.6 (11.0)	60.9 (6.8)	59.5-62.3

[a]SC: symptom checker.

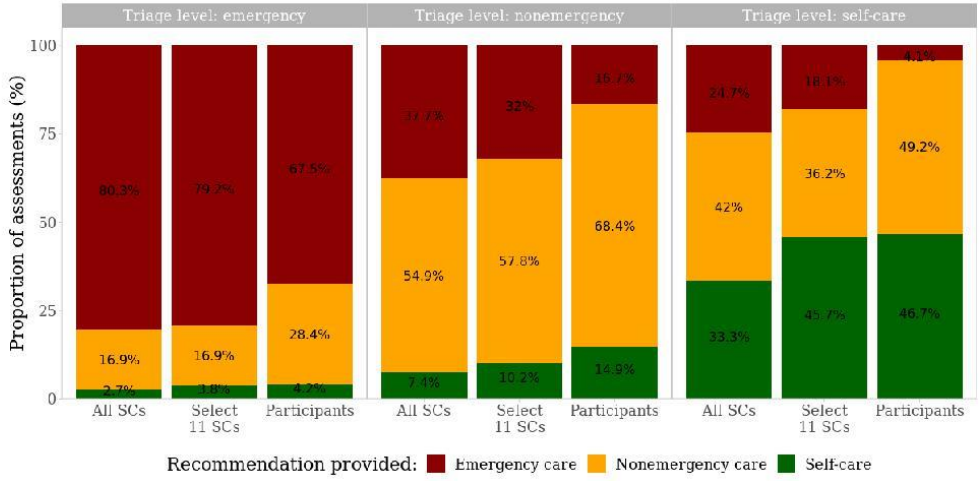
[b]For the subset of 11 SCs, SCs never recommending self-care or always recommending emergency care by design were excluded.

[c]For the participant sample, 95% CIs were calculated using bootstrapping.

Risk Aversion

The SCs were risk averse and overtriaged in more than a third of the evaluations (182/532, 34.2%), whereas undertriaging occurred in only 9.2% (49/532). Although participants also tended to be risk averse, this tendency was less pronounced (Figure 1). The ratio of overtriage to undertriage errors was 1.5:1 for participants whereas it was 3.5:1 for SCs. The SCs misclassified self-care cases as emergencies 6 times more often than participants did (43/174, 24.7% vs 56/1365, 4.10%) and 4.5 times more often (23/127, 18.1% vs 56/1365, 4.1%) when considering the subset of 11 SCs. The pair-wise differences in recommendations per triage level were statistically significant between participants and SCs (P=.002 for triage-level emergency [χ^2_2=12.5]; P<.001 for nonemergency [χ^2_2=46.3] and self-care [χ^2_2=109.6]). This holds true when comparing the participants' performance with the subset of 11 SCs (P=.02 for an emergency [χ^2_2=8.1] and P<.001 for a nonemergency [χ^2_2=19.0] and for self-care [χ^2_2=47.1]).

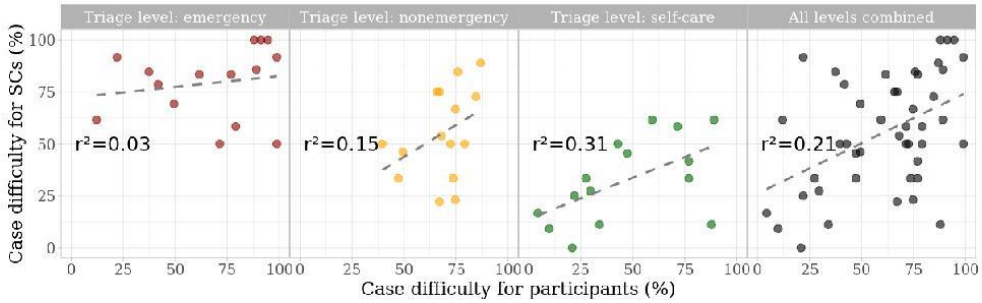
Figure 1. Triage evaluations by participants and SCs and triage level. "11 SCs" refers to the SC sample after exclusion of SCs that never recommend self-care (the least urgent triage level). SC: symptom checker.



Comparing Case Vignette Difficulty for SCs and for Participants

How challenging a case vignette was for SCs and participants varied widely: 3 vignettes were solved correctly by every SC and 1 vignette by none. Similarly, 4 vignettes were solved correctly by more than 90% of the participants and 2 by less than 10%. At every triage level, a broad variation in the degree of difficulty among case vignettes was observed. A very weak or no relationship could be detected for SCs and participants regarding case difficulty within each triage level (Figure 2).

Figure 2. Distribution of case difficulty for participants and SCs. Case difficulty is defined as the proportion of the group (SC or participants) evaluating the respective case correctly. The dashed line models a linear relationship. SC: symptom checker.



Comparing Individual SCs With Participants

As previously mentioned, an aggregated analysis of SCs is less meaningful than a direct comparison between the participant population and each SC, as users are likely to consult only one or very few SCs. The overall trend shows that the accuracy of both participants and SCs decreases for self-care vignettes (Figure 3).

A total of 5 SCs (*HMS* [Harvard Medical School] *Family Health Guide*, *Healthy Children*, *Steps2Care*, *Symptify*, and *Symptomate*) managed to outperform the participant sample, achieving an overall accuracy greater than the mean of the participants and its CI's upper limit (Table 3; see yellow dots in Figure 3). Five SCs had a triage capability lower than 80% (73/91) of the participants. This finding is partially explained by 3 of them apparently designed to never recommend self-care, hence failing in one-third of the cases owing to their design. One of these 3 SCs (*Isabel*) was outperformed only by a minority of participants (17/91, 18%), when self-care case vignettes were excluded from the analysis. The remaining 2 SCs (*Symcat* and *iTriage*) were still outperformed by most participants when self-care case vignettes were excluded. The participants' mean accuracy was stable at approximately 60%, independent of the slightly different samples of vignettes assessed by the SCs, with 2 exceptions: the participants were challenged by the sample of vignettes evaluated by *Healthy Children*, reaching a mean accuracy that was approximately 10% lower than in the other samples; conversely, the participants fared much better in assessing the vignette sample considered by *DoctorDiagnose*.

All but 2 SCs (*Family Doctor* and *Drugs.com*) were risk averse, making more overtriage errors than undertriage errors. Although the best 5 SCs were inclined toward overtriage, only one of them overtriaged more vignettes than the average participant (*Symptomate*; Figure 4).

Figure 3. Accuracy of SCs and participants by triage level (Em), nonemergency, and S-c. The accuracy of individual participants is indicated with blue dots. The aggregate accuracies of participants are shown as box plots. Em: emergency; SC: symptom checker; S-c: self-care.

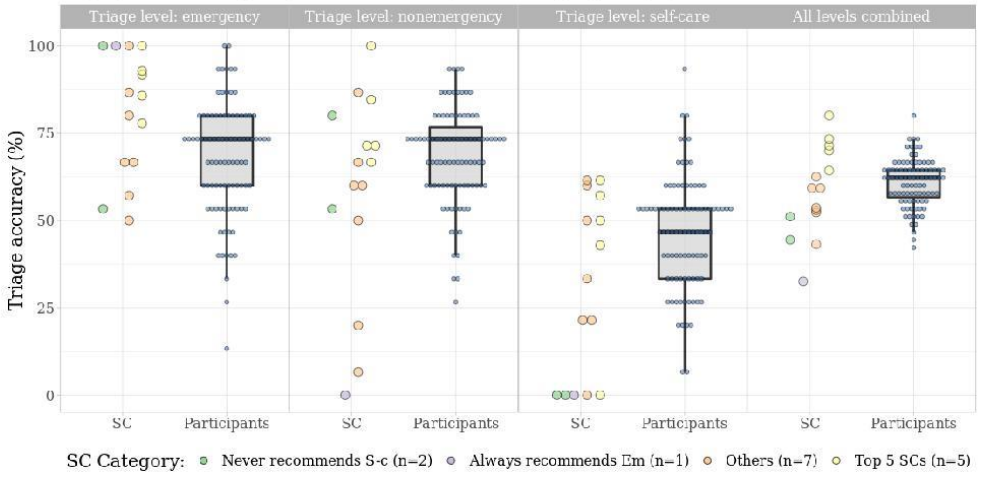


Table 3. Comparison of accuracy between symptom checkers and participants.

SC[a,b] name	Accuracy[c], n (%)	Participants		Comparison
		Percent accuracy[d,e], mean (SD)	95% CI	Percentage of participants outperforming the SC (95% CI)[d,e]
HMS[f] Family Health Guide, n=40	32 (80)	59.5 (7.1)	58.0-60.9	0 (0-0)
Healthy Children, n=15	11 (73)	49.9 (10.1)	47.7-52.1	1.1 (0-3.3)
Steps2Care, n=42	30 (71)	59.7 (7.2)	58.2-61.1	1.1 (0-3.3)
Symptify, n=40	28 (70)	60.2 (7.2)	58.2-61.7	5.5 (1.1-11.0)
Symptomate[g], n=14	9 (64)	60.9 (11.6)	58.6-63.2	26.4 (17.6-35.2)
Drugs.com, n=42	25 (59)	60.6 (6.5)	59.3-61.9	51.6 (41.8-61.5)
FreeMD, n=44	26 (59)	60.2 (6.7)	58.9-61.6	56.0 (45.1-65.9)
Doctor Diagnose, n=16	10 (62)	69.5 (10.9)	67.3-71.7	63.7 (53.8-73.6)
Family Doctor, n=41	22 (53)	58.1 (7.0)	56.7-59.6	68.1 (58.2-78.0)
Early Doc, n=17	9 (52)	63.4 (11.4)	61.1-65.7	76.9 (68.1-85.7)
Isabel[g], n=45	23 (51)	60.9 (6.8)	59.4-62.2	89 (82.4-94.5)
NHS[h], n=44	23 (52)	62.0 (6.9)	60.9-63.4	89 (82.4-94.5)
Symcat[g], n=45	20 (44)	60.9 (6.8)	59.5-62.2	97.8 (94.5-100)
Healthwise, n=44	19 (43)	61.2 (7)	59.7-62.6	98.9 (96.7-100)
iTriage[h,i], n=43	14 (32)	60.5 (6.9)	59.1-61.9	100 (100-100)

[a]SC: symptom checkers

[b]SCs are listed in order by the proportion of participants outperforming them.

[c]Most SCs did not evaluate every case vignette. Their accuracy is given as the proportion of correctly solved vignettes of the total vignettes that they evaluated.

[d]The participants' accuracy is based on their assessment of the same case vignettes assessed by the respective SC.

[e]For the participant sample, 95% CIs were calculated using bootstrapping.

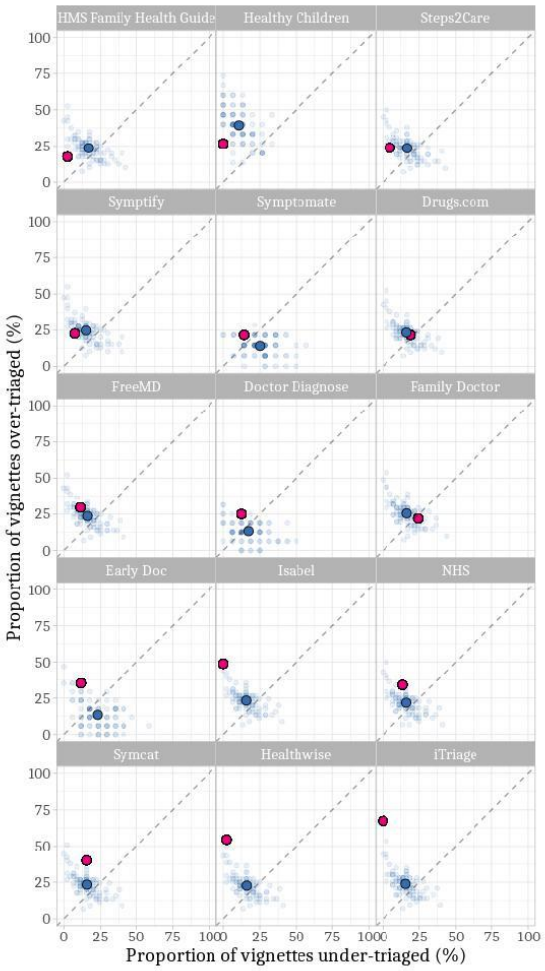
[f]HMS: Harvard Medical School.

[g]Four SCs were apparently designed never to recommend self-care.

[h]NHS: National Health Service.

[i]One SC advised seeking emergency care for all case vignettes.

Figure 4. Comparison of the overtriage inclination of symptom checkers (SCs) and participants. The dashed line shows where proportions of over and undertriaged errors are equal. Proximity to the left lower corner indicates a high triage accuracy. The red dot marks the respective symptom checker. The faded blue dots refer to the performance of individual participants. The larger blue dot marks their average performance. The SCs are ordered from left to right and top to bottom by the proportion of participants outperforming them, with the lowest proportional difference at the top left and the highest proportional difference on the bottom right.



Discussion

Principal Findings

Our study suggests that an average SC has no greater overall triage accuracy than an average user. However, this does not imply that SCs are not useful. Specifically, our data confirm a prior study showing that the lay population has difficulties reliably identifying medical emergencies [25]. On average, participants failed to identify every third emergency, and 12% (11/91) of our participants identified emergencies less reliably than the worst-performing SC.

Most SCs tended to overtriage. From a clinical and legal perspective, it can make sense to accept the resulting inflated cost of false alarms to avoid potentially missing an emergency (*defensive decision making*). In contrast, false alarms raised by SCs can functionally exacerbate overcrowding in health care services. In fact, the ability of some SCs to reliably detect emergencies can be partially attributed to their general tendency—by design—to recommend emergency care even for self-care cases (the least urgent triage level) where no medical care is warranted. This trade-off must be considered before recommending their use.

Studies on the effects of SC advice on users are scarce. Therefore, general recommendations on whether laypersons should use SCs cannot be formulated as yet. On the basis of a detailed analysis of the performance variation among SCs and human decision makers, we showed that the five best SCs that Semigran et al [11] included in their sample outperformed almost all our participants and thus could be seen as beneficial to users. In contrast, SCs mistake self-care cases for emergencies a substantial number of times. This hints at SCs being better suited to help users who are looking for an answer on where they should seek professional help (ie, by discriminating between emergency and nonemergency cases) rather than on whether they should seek medical care at all (ie, by discriminating between self-care and non–self-care cases).

Finally, SCs and participants struggled with different kinds of case vignettes, that is, SCs performed poorly in some clinical situations, whereas in others, their performance was superior to that of their users. For example, the 15 pediatric cases evaluated by the SC *Healthy Children* appear to have been more challenging for participants (mean accuracy of 49.9%) than the 30 nonpediatric cases (mean accuracy of 66.3%). To provide a more differentiated picture of SC triage performance, further analyses should also investigate performance differences with respect to different types of cases.

Limitations

Compared with the general population of the United States [26], our participants were better educated and included more men than women. The median and mean ages were similar to those of the general US population. One study suggests that the groups most likely to seek health information online are younger White females from high-income households, most with a bachelor's degree or higher [1]. Most participants in a survey among users of a specific SC (Isabel) were female and White but older than the average population [27]. Despite the fact that our sample's demographic distribution did not fully resemble the US population or, presumably, the population of SC users, we consider our findings to have at least some external validity for these populations, as demographic variables showed no significant influence on triage accuracy.

The data on SCs date back to a study published in 2015 [11], where the specific versions of the SCs assessed were not specified. Therefore, changes in performance due to possible upgrades were not considered. Such upgrades are likely, and new SCs have since entered the market. Other SCs included in the Semigran et al sample [11] are no longer available online, including the best-performing SC (*HMS Family Health Guide*). This speaks to the general problem that future research evaluating the performance of SCs will have to address the rapidly changing markets and technological developments.

As we built our study on the materials of the Semigran et al study [11], we also inherited their limitations: the chosen 45 case vignettes do not cover the entire spectrum of prehospital case presentations, especially with the omission of mental health–related scenarios. In addition, some case vignettes lacked a proper diagnosis and stated only the presenting complaints (eg, "Vomiting" for vignette 45, "Constipation" for vignette 40, "Back pain" for vignette 20). This prevented a plausibility check of the gold standard triage level that should be assigned to each vignette.

In general, assessing triage capability with case vignettes has limited validity. This limitation is arguably greater for human participants than for SCs. Although SCs assess a case with a set algorithm and are therefore dependent only on input, contextual (social, emotional, etc) factors play a greater role in human decision making. In a real-life setting, humans might also notice and process more or less information than presented in a case vignette. In addition, reading "back pain" in a dry case vignette is surely a different matter than experiencing it. Thus, our results might be more valid for situations where SC users utilize the tool to triage someone other than themselves. Research shows that this is common practice, as up to 50% of online health information seekers do so on behalf of someone else [1].

Conclusions

Prior publications have emphasized the need for a framework within which the safety and usefulness of SCs should be analyzed. Assessing the average performance of SCs, as has often been done, fosters few actionable recommendations. Given the high-performance variability among SCs, we consider benchmarking with case vignettes as a valuable first step in identifying the best SCs, which could then be tested extensively against relevant competitors.

Although comparing SCs' triage capability against that of health care professionals is certainly useful [28], this approach implicitly asks whether the former could replace the latter, rather than assessing whether and under which circumstances a user should rely on an SC or refrain from using it. Similar to the common practice of testing a new medicine against a placebo, we suggest that SCs should be benchmarked against a realistic alternative, for example, an SC user relying on his own appraisal (stand-alone triage capability).

Following this approach, our study suggests that the lay population would benefit from some SCs to some extent. Although SCs detect emergencies more reliably than the average user, they are more risk averse than the general population and recommend emergency care more often than is actually necessary. This is a cause for concern, as it might unnecessarily increase the burden on already overwhelmed health care services. Thus, advice on when not to seek emergency care would be the most useful feature of SCs, but it is precisely in that situation that they performed the worst. Further research should investigate which user groups benefit the most from using SCs and whether it is possible to identify the characteristics of scenarios where laypersons are superior to SCs in assessing triage levels. The detailed analyses presented in this paper provide a first step toward a framework for comparatively assessing the respective weaknesses and strengths of both SCs and human decision makers to be able to determine when humans should rely on SCs rather than on their gut feeling and vice versa.

Acknowledgments

The authors express their gratitude to the participants, to Felix Grün for his support in designing the questionnaire and for his valuable feedback, to Eike Richter for his advice on statistical methods, and to Frances Lorié for proofreading the manuscript. The project was funded by the home institutions of the previous authors (MF and FB). No external funding was required for this study. The authors acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Fund of Charité—Universitätsmedizin Berlin.

Authors' Contributions

MS conceived the study, created the questionnaire, designed and conducted the analyses, and wrote the first draft of the paper. MALS assisted with case vignette adaptations. RM assisted with case vignette adaptations and manuscript development. FB and MF provided critical input and advised on the study and questionnaire design, analysis methods, and drafts of the paper. FB and MF contributed equally and share the last authorship. All authors accept full responsibility for the final version of the paper. The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Conflicts of Interest

All authors have completed the International Committee of Medical Journal Editors uniform disclosure form and declare no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; and no other relationships or activities that could appear to have influenced the submitted work.

Multimedia Appendix 1

Adapted case vignettes and case difficulty level.
[DOCX File , 40 KB-Multimedia Appendix 1]

References

1. Fox S, Duggan M. Health online 2013. Pew Research Center. 2013. URL: https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/PIP_HealthOnline.pdf [accessed 2021-02-18]
2. Martin SS, Quaye E, Schultz S, Fashanu OE, Wang J, Saheed MO, et al. A randomized controlled trial of online symptom searching to inform patient generated differential diagnoses. NPJ Digit Med 2019 Nov 11;2(1):110 [FREE Full text] [doi: 10.1038/s41746-019-0183-0] [Medline: 31728417]
3. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 01;9(8) [FREE Full text] [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
4. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. The Lancet 2018 Nov;392(10161):2263-2264. [doi: 10.1016/s0140-6736(18)32819-8]
5. Bavdekar SB, Pawar M. Evaluation of an Internet-Delivered Pediatric Diagnosis Support System (ISABEL®) in a Tertiary Care Center in India. Indian Pediatrics. 2005. URL: http://www.indianpediatrics.net/nov2005/1086.pdf [accessed 2021-02-18]
6. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of Four Computer-Based Diagnostic Systems. N Engl J Med 1994 Jun 23;330(25):1792-1796. [doi: 10.1056/nejm199406233302506]
7. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med 2012 Feb 26;27(2):213-219 [FREE Full text] [doi: 10.1007/s11606-011-1804-8] [Medline: 21789717]
8. Farmer N. An update and further testing of a knowledge-based diagnostic clinical decision support system for musculoskeletal disorders of the shoulder for use in a primary care setting. J Eval Clin Pract 2014 Oct 15;20(5):589-595. [doi: 10.1111/jep.12153] [Medline: 24828447]
9. Farmer N, Schilstra MJ. A Knowledge-based Diagnostic Clinical Decision Support System for Musculoskeletal Disorders of the Shoulder for Use in a Primary Care Setting. Shoulder & Elbow 2017 Feb 06;4(2):141-151. [doi: 10.1111/j.1758-5740.2011.00165.x]
10. 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 2021 Feb 09;214(3):143. [doi: 10.5694/mja2.50923] [Medline: 33423279]
11. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. Br Med J 2015 Jul 08;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
12. Gottliebsen K, Petersson G. Limited evidence of benefits of patient operated intelligent primary care triage tools: findings of a literature review. BMJ Health Care Inform 2020 May 07;27(1):e100114. [doi: 10.1136/bmjhci-2019-100114]

13. Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Sci Adv 2018 Jan 17;4(1):eaao5580 [FREE Full text] [doi: 10.1126/sciadv.aao5580] [Medline: 29376122]
14. Anhang Price R, Fagbuyi D, Harris R, Hanfling D, Place F, Taylor TB, et al. Feasibility of web-based self-triage by parents of children with influenza-like illness: a cautionary tale. JAMA Pediatr 2013 Feb 01;167(2):112-118. [doi: 10.1001/jamapediatrics.2013.1573] [Medline: 23254373]
15. Unipark. QuestBack. URL: https://www.unipark.com/ [accessed 2021-02-18]
16. Mortensen K, Hughes TL. Comparing Amazon's Mechanical Turk Platform to Conventional Data Collection Methods in the Health and Medical Research Literature. J Gen Intern Med 2018 Apr 4;33(4):533-538 [FREE Full text] [doi: 10.1007/s11606-017-4246-0] [Medline: 29302882]
17. Rouse SV. A reliability analysis of Mechanical Turk data. Computers in Human Behavior 2015 Feb;43:304-307. [doi: 10.1016/j.chb.2014.11.004]
18. Casey LS, Chandler J, Levine AS, Proctor A, Strolovitch DZ. Intertemporal differences among MTurk workers: time-based sample variations and implications for online data collection. SAGE Open 2017 Jun 14;7(2). [doi: 10.1177/2158244017712774]
19. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing URL: https://www.R-project.org/ [accessed 2021-02-18]
20. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw 2019 Nov;4(43):1686. [doi: 10.21105/joss.01686]
21. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using. J Stat Soft 2015;67(1). [doi: 10.18637/jss.v067.i01]
22. Bray A, Ismay C, Chasnovski E, Baumer B, Cetinkaya-Rundel M. infer: tidy statistical inference. 2020. URL: https://CRAN.R-project.org/package=infer [accessed 2021-02-18]
23. Wickham H. ggplot2. Elegant graphics for data analysis. 2016. URL: https://link.springer.com/book/10.1007/978-3-319-24277-4 [accessed 2021-02-18]
24. Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. J Med Internet Res 2021:1-29. [doi: 10.2196/24475]
25. Mills B, Hill M, Buck J, Walter E, Howard K, Raisinger A, et al. What constitutes an emergency ambulance call? Australasian J Paramed 2019 Mar 22;16. [doi: 10.33151/ajp.16.626]
26. Age and sex composition in the United States. United States Census Bureau. 2019. URL: https://www.census.gov/content/census/en/data/tables/2019/demo/age-and-sex/2019-age-sex-composition.html [accessed 2020-08-10]
27. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient Perspectives on the Usefulness of an Artificial Intelligence-Assisted Symptom Checker: Cross-Sectional Survey Study. J Med Internet Res 2020 Jan 30;22(1):e14679 [FREE Full text] [doi: 10.2196/14679] [Medline: 32012052]
28. Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of Physician and Computer Diagnostic Accuracy. JAMA Intern Med 2016 Dec 01;176(12):1860-1861. [doi: 10.1001/jamainternmed.2016.6001] [Medline: 27723877]

Abbreviations

HMS: Harvard Medical School
MTurk: Mechanical Turk
PDT: Pacific Daylight Time
SC: symptom checker

Edited by G Eysenbach; submitted 24.09.20; peer-reviewed by M Hill, E Berner, J Knitza; comments to author 04.10.20; revised version received 22.10.20; accepted 18.01.21; published 10.03.21

Please cite as:
Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F
Benchmarking Triage Capability of Symptom Checkers Against That of Medical Laypersons: Survey Study
J Med Internet Res 2021;23(3):e24475
URL: https://www.jmir.org/2021/3/e24475
doi: 10.2196/24475
PMID:

8 Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

9 Publikationsliste

- Karduck L, Behnke AL, Gabrysch C, Kasper A, Lennartz N, von Philipsborn P, Poppinga SK, Schmidt D, Schmidt M, Schmieding ML, Schulz L, Schürmann C, Speer L, Strube S.
  Assessing universities' impact on global health: a comparative study of 36 German universities.
  European Journal of Public Health, Volume 25, Issue suppl_3, October 2015, ckv175.149.

- Poncette A, Mosch L, Spies C, Schmieding M, Schiefenhövel F, Krampe H, Balzer F.
  Improvements in Patient Monitoring in the Intensive Care Unit: Survey Study.
  J Med Internet Res 2020;22(6):e19091.
  (Impact Factor: 5.03)

- Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F. *Benchmarking Triage Capability of Symptom Checkers Against That of Medical Laypersons: Survey Study.*
  J Med Internet Res 2021;23(3):e24475.
  (Impact Factor: 5.03)

10 Danksagung

Der erste Dank gilt meinen Betreuern, Markus Feufel und Felix Balzer: Danke für stete Ermunterung und Unterstützung. Ich danke den Ko-Autoren Rudolf Mörgeli und Maike Schmieding für ihre Mitwirkung.

Auch denen möchte ich Dank aussprechen, deren Arbeit mich zu dieser Promotion inspiriert hat: Malte Joswig, Eta S. Berner, Julia Dressel und Hany Farid, sowie Robert S. Ledley und Lee B. Lusted.

Ferner danke ich Eike Richter und Felix Grün für ihren Rat zur statistischen Auswertung bzw. zur Rekrutierung der Probanden, sowie Fridtjof Schiefenhövel und Patrick Heeren für die Einarbeitung in das Handwerk der Medical Data Science.

Der letzte Dank gilt meinen Eltern, Simone und Holger Schmieding, und den Freunden, die mich bestärkt haben, diese Arbeit zu beginnen, weiterzumachen und abzuschließen: Nicola Bendzko, Nico und Minh Meißner, Laura Hahn, Luisa von Albedyll, Franziska Hörth, Sylvia Hartmann & Katharina Kanthak, Nadja Henningsen, Marie-Therese Holzer, Louisa Morrison, Clemens Olesch, Kamilla Toewe, Jakob Voran und Solveig Mosthaf.