LSE INTERNATIONAL
DEVELOPMENT REVIEW

Deamant, E. (2025). Health and Economic Impacts of AI Symptom Checkers: A Narrative Literature Review. LSE International Development Review, 4(1), pp. 99 – 110. https://doi.org/10.82191/lseidr.100

# Health and Economic Impacts of AI Symptom Checkers: A Narrative Literature Review

By Lane Deamant

## ABSTRACT

**Background:** Symptom checkers, digital tools used for health assessments without a medical professional's direct input, have seen increasing integration of artificial intelligence (AI). Substantial research has explored the hypothetical diagnostic accuracy of AI symptom checkers, but relatively little has looked at the practical implications of these tools for users and wider health systems.

**Objectives/Methods:** This narrative review aims to evaluate the state of the literature on the health and economic impacts of AI symptom checkers. 129 articles were considered for inclusion from seven databases, alongside a hand search of reference lists and Google Scholar. Nine articles were selected.

**Results:** A conceptual framework was developed to summarise the main possible health-related outcomes of AI symptom checkers. Depending on the user's needs and the symptom checker's recommendations, there is some limited evidence to suggest that symptom checkers may reduce unnecessary usage of healthcare. However, other – also limited – research suggests that symptom checkers may recommend unneeded care ('overtriage') or, with greater risks to patient health, not make recommendations for needed care ('undertriage'). Other outcomes (e.g., quality of care, mental stress, cost savings) were minimally reported.

**Discussion and Conclusion:** Several key research gaps were found, including a need for more real-world and follow-up evidence, explicit consideration of (recent) AI, and economic evaluations. Each of these research directions would aid decisionmakers in determining the impacts of AI symptom checkers, as the current research base points to both positive and negative potentialities.

**Keywords:** artificial intelligence, symptom checker, chatbot, narrative review, health, economics

## INTRODUCTION

Digital tools known as 'symptom checkers' promise quick, reliable and cheap assessments of users' symptoms, lists of potential diagnoses, and recommendations for treatment, all without the inconvenience of seeing a general practitioner (GP) (Aboueid et al. 2019). Physicians have long been aware, and largely discouraging, of patients searching their symptoms on 'Dr Google' (Reardon 2023). The emergence of more formalised symptom checkers, especially artificially intelligent ones, is expected to bring new changes to how patients seek healthcare information, although the exact nature of these changes remains to be demonstrated (Farnood, Johnston & Mair 2020).

### Background

Symptom checkers are a wide set of tools that gather information on users' symptoms and provide advice (You & Gui 2021). For some scholars, symptom checkers are seen as a challenge or risk for healthcare systems, such as 'cyberchondriac' patients excessively entering symptoms, triggering new anxieties, and potentially visiting health clinics when they have no real need to (Eichenberg & Schott 2019: 3). Scholars also worry about overly-trusting users who may avoid in-person care following the advice of a symptom checker (Radionova et al. 2023). For the optimists, however, symptom checkers are seen as a near-panacea for a host of healthcare goals from more informed patients to more correct diagnoses and reduced healthcare costs through patient triage (Meyer et al. 2020). Any impacts of these tools depend on alignment between the tool's programming for what recommendations it makes, the integration of artificial intelligence (AI), the patient's health needs at the time of use, and the patient's behaviour following use of the tool (Radionova et al. 2023; Meyer et al. 2020).

AI has numerous definitions but is typically understood as computer systems that 'think or act like humans', including behaviours of integrating information and learning (Morrow et al. 2023: 1). While AI trained on large datasets are often seen as advanced or promising, there are also risks of inaccuracies or even falsities ('hallucinations') (Hatem, Simmons & Thornton 2023). Symptom checkers are increasingly utilising AI (called 'AI symptom checkers' in this review) through specialised chatbots which simulate conversations with users or through other AI-powered questionnaires (Semigran et al. 2015). Broader tools such as OpenAI's ChatGPT can also be considered to have 'symptom-checking capabilities' when users  enter symptoms and request health-related advice (Chen, Chen & Tian 2023: 1).

Most symptom checkers follow a process of establishing a patient history, evaluating symptoms, giving an initial list of possible diagnoses, and making recommendations that range from self-treatment to seeing a GP or urgent care within a specified time frame (You & Gui 2021: 1356). Given that most symptom checkers are intended for personal rather than clinical use, they typically provide disclaimers about their diagnoses being non-definitive and not supplied from medical professionals (Ubie 2023). These disclaimers allow the tools to be considered 'low-risk' by regulatory regimes in the US, the UK and other countries, a designation that carries lower requirements to demonstrate benefit prior to and following market authorisation (O'Reilly-Jacob et al. 2021). This classification exists despite concerns about symptom checkers' risks for health-seeking behaviour (Iacobucci 2020), as this review will explore. As a partial consequence of the regulatory landscape, data are not widely available on patient behaviour following use of AI symptom checkers, nor on the consequences of these actions for individual and aggregate health (Chambers et al. 2019).

At the same time, there is a relatively wide research base on the diagnostic accuracy of AI symptom checkers (Radionova et al. 2023). These studies are often based on hypothetical, researcher-invented patient cases ('vignettes') (Lyons et al. 2024). This research presents a mixed picture on accuracy (Radionova et al. 2023). Some vignette studies find these tools to be as accurate or better than medical professionals (Gräf et al. 2022; Lyons et al. 2024), while others report worse accuracy (Ceney et al. 2021; Ćirković 2020). Accuracy is an important component of evaluations of symptom checkers because users who follow inaccurate advice could experience adverse health consequences, although this remains to be proven in the literature (Pairon, Philips & Verhoeven 2023). A modest research base also suggests that users generally intend to follow the advice of symptom checkers, although subsequent behaviours are underexplored in the literature (Pairon, Philips & Verhoeven 2023). The literature further lacks evidence on the costs and/or savings these tools might bring to the healthcare system and to individual patients (Pairon, Philips & Verhoeven 2023).

**Research Objectives**

In sum, a limited body of research has raised conflicting and disparate sets of claims about the possible effects of AI symptom checkers, primarily based on hypothetical patient cases (Radionova et al. 2023). Empirical health and economic impacts are left uncertain. The scholarly context may also be affected by the rapidly changing nature of symptom checkers under investigation through constant algorithmic upgrades and the steady launching of new tools (Pairon, Philips & Verhoeven 2023; Saenger et al. 2024). A narrative literature review was therefore merited to map the existing evidence on the health and economic impacts of AI symptom checkers, consolidate understanding through the development of a conceptual framework, and elaborate the research gaps. The conceptual framework complements ongoing research and policy goals of ensuring that the growing market for digital health technology is supportive of population health (Chambers et al. 2019). These objectives led to the following research questions:

1. What is the state of the literature on the impacts of AI symptom checkers for users and health systems?
2. What are the possible health impacts identified by the literature?
3. What key research gaps exist?

## METHODS

### Design

A narrative literature review was conducted on the health and economic outcomes of AI symptom checkers. A systematic review was deemed inappropriate given that research on these tools has been minimal and methodologically varied, albeit with enough literature to form a review (Aboueid et al. 2019). At the time of writing (June 2024), the author could find no other systematic or narrative literature review that focussed exclusively on the empirical health and economic impacts of AI symptom checkers. Using a rigorous search strategy, this narrative review sought to synthesise as much of the existing research as possible and to identify future research trajectories (Paré & Kitsiou 2017).

The two main concepts guiding the literature search were artificial intelligence and symptom checkers. An initial search of the literature was conducted to document as many synonyms of these concepts as possible, which were then included as search terms. The final search strategy (Appendix A) was developed with assistance from the London School of Economics (LSE) Department of Health Policy librarian, Heather Dawson. Articles were searched for 'AI' [or related terms] within 6 words of 'symptom checker' [or related terms]. This strategy helped reduce spurious matches by ensuring the concepts were mentioned in relation to each other. Searching title, keywords and abstracts rather than 'all fields' also helped ensure the relevance and specificity of articles generated.

### Inclusion/Exclusion Criteria

Studies were included if they evaluated the health and/or economic impacts of AI symptom checkers using data from real users. Given the nascency of the research (Radionova et al. 2023), the review did not specify the research methodologies of included articles, the native languages of participants, or the type of outcomes under investigation (e.g., individual or aggregate). Studies were excluded that focused solely on (1) the diagnostic accuracy of the tool(s); (2) user experience or app design; (3) descriptions of user characteristics; (4) tools used by medical professionals; (5) non-artificially intelligent tools; (6) hypothetical or theoretical perspectives on symptom checkers (including vignettes); and/or (7) were not published in English.

As a quality measure, the review excluded conference abstracts, manuscripts, non-peer-reviewed articles and other reports. Such grey literature was initially considered for inclusion but was found by the author to have a high degree of sponsorship or direct production by companies who make symptom checkers, increasing the risk of biassed results (Resnik 2019: 3). Literature was restricted to 2017–2024 to improve the likelihood of picking up genuine AI tools.

### Publication Screening

The narrative review searched the following databases: Embase, Medline (OVID), CINAHL Plus, APA PsycInfo, Journals@OVID, EBSCO EconLit, and Health Management Information Consortium (HMIC), at which point saturation was reached with no new articles emerging. The searches were finalised on 14 June, 2024. In line with other relevant narrative reviews (Radionova et al. 2023), the search process has been summarised in a PRISMA-style flow diagram (Figure 1). 129 articles were found, of which 64 were duplicates. Given the high number of duplicates, the author also searched Google Scholar using key terms and hand-searched reference lists from database-identified sources, a 'snowball' method that has been shown to increase the identification of relevant studies for literature reviews (Wohlin et al. 2022).

Following removal of duplicates, the remaining titles and abstracts were screened by hand according to the inclusion/exclusion criteria. In a Google Sheet, the author marked articles with 'Exclude' (with reasons), 'Include' or 'Maybe', with the latter two undergoing full-text review. To help account for the absence of another reviewer, the author discussed 'Maybe' articles with a secondary supervisor at LSE. Following this screening process, nine articles were selected for inclusion in the final review.

### Quality Assessment

The included articles were appraised using Sirriyeh and colleagues' Quality Assessment Tool for Studies with Diverse Designs (QATSDD) (Sirriyeh et al. 2012). QATSDD was developed out of a systematic review of existing quality assessment tools, followed by a piloted and peer-reviewed process to design the final tool (Sirriyeh et al. 2012: 748). Using input from health researchers, the authors established face and content validity and interrater

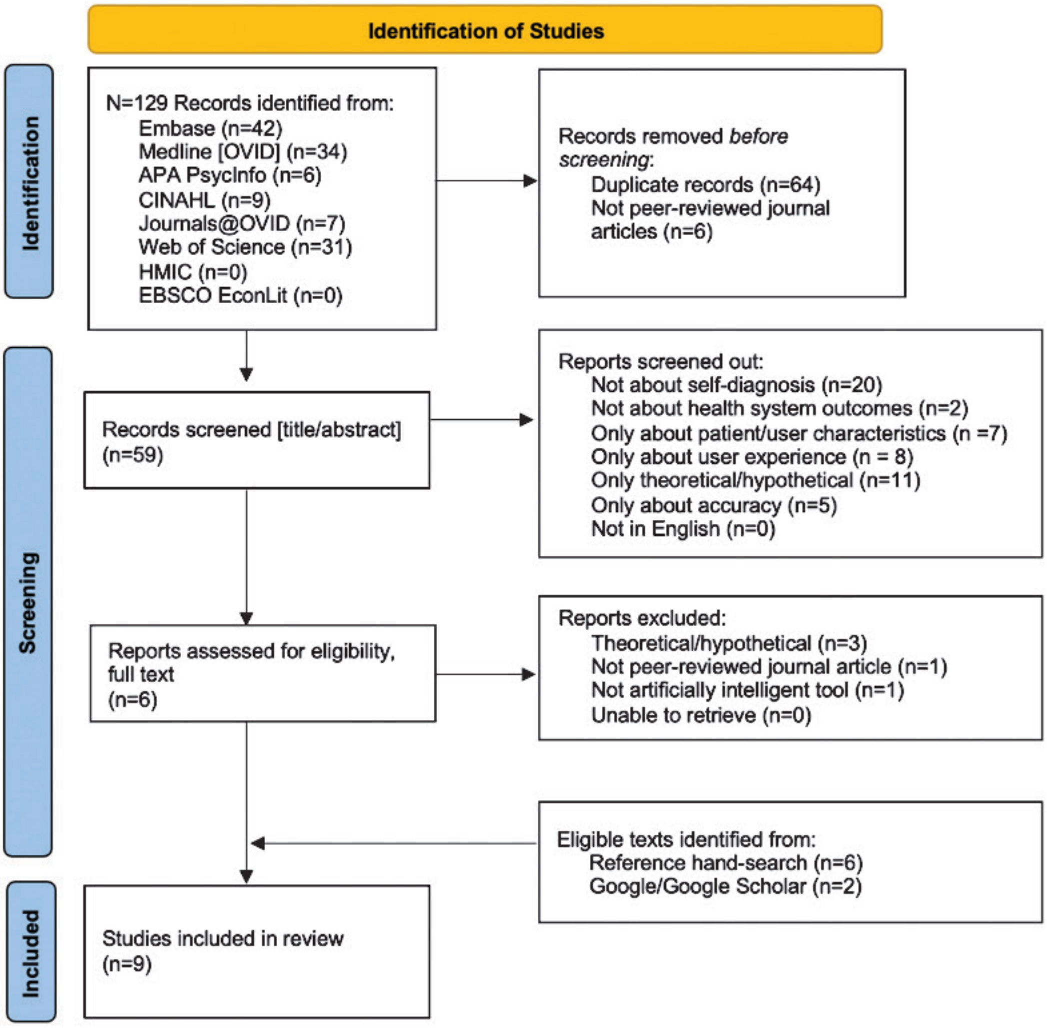
Identification of Studies
Identification
N=129 Records identified from:
Embase (n=42)
Medline [OVID] (n=34)
APA PsycInfo (n=6)
CINAHL (n=9)
Journals@OVID (n=7)
Web of Science (n=31)
HMIC (n=0)
EBSCO EconLit (n=0)
Records removed before screening:
Duplicate records (n=64)
Not peer-reviewed journal articles (n=6)
Screening
Records screened [title/abstract] (n=59)
Reports screened out:
Not about self-diagnosis (n=20)
Not about health system outcomes (n=2)
Only about patient/user characteristics (n =7)
Only about user experience (n = 8)
Only theoretical/hypothetical (n=11)
Only about accuracy (n=5)
Not in English (n=0)
Reports assessed for eligibility, full text (n=6)
Reports excluded:
Theoretical/hypothetical (n=3)
Not peer-reviewed journal article (n=1)
Not artificially intelligent tool (n=1)
Unable to retrieve (n=0)
Eligible texts identified from:
Reference hand-search (n=6)
Google/Google Scholar (n=2)
Included
Studies included in review (n=9)

**Figure 1:** Flowchart for Literature Review Search
(Diagram template source: Page et al. 2021)

reliability (Sirriyeh et al. 2012). QATSDD assesses each study on well-established research practices with scores ranging from 0 (absent) to 3 (complete), which can be summed and turned into percentages (Sirriyeh et al. 2012: 749). The total QATSDD scores for each study in this review are available in Appendix B. Shortcomings revealed by these quality scores are explored in the discussion section.

### Data Collection

Information for extraction was determined by a high-level scan of the articles as well as reference to other literature reviews for consistency (Pawson 2002). Data management and organisation was done on Google Sheets. The following key pieces of information were collected from the articles: Year of publication, country, AI symptom checker used, research methodology, target population, sample size and relevant sample characteristics, primary health outcomes under investigation, any secondary health outcomes, economic outcomes, strengths, limitations, main findings, and reported author affiliations. A summary of the extracted data is available in Appendix C.

### Evidence Synthesis

The evidence collected during the review was synthesised through a qualitative, iterative thematic analysis approach (Dixon-Woods et al. 2005: 47). This involved moving back and forth between the data collection sheet and the source materials to identify categories and themes on symptom checkers' health and economic outcomes.

The synthesis approach was qualitative for several reasons: the descriptive nature of several included studies, the limited comparability of numerical figures due to different methods and sample populations and the small number of articles available (Sukhera 2022: 416). While qualitative themes are used to organise the results section, numerical figures from the articles are also provided to contextualise the themes and avoid overgeneralisations. Overall, the aim was a rich summary of the current scholarly understanding of the impacts of AI symptom checkers. This included the development of a conceptual framework (Jabareen 2009), which identifies key health outcomes following the use of symptom checkers. Conceptual frameworks are particularly important for nascent fields of research (Yigitcanlar et al. 2021), such as AI symptom checkers.

### Ethics

As this narrative review utilised only secondary, publicly available and anonymised data, ethics approval was not required in line with LSE guidelines. The ethics submission was reviewed by Professor Joan Costa-i-Font (LSE Department of Health Policy) on 12 June 2024, submission #395087.

## RESULTS

### Overview

Nine studies were included in the review in line with the stated inclusion/exclusion criteria (see section 2.2). A summary of the studies' characteristics can be found in Appendix C. Quality assessment scores were mixed, as seen in Appendix B. Total quality scores ranged from 16.7–77.1%, with most studies being 'good quality' (Klingenberg et al. 2020).

### Conceptual Framework

The conceptual framework (Figure 2) charts the main types of potential health impacts of symptom checker usage, developed from literature focusing on AI tools that interact with patients and provide advice. The vertical axis shows how patients use the tool with varying health needs, which they may or may not know, and which they attempt to communicate to the symptom checker. On the horizontal axis, the tool subsequently makes health-related recommendations of varying intensities (e.g., from self-treatment to emergency care). If the user's needs and the tool's advice match, then ideal outcomes can be achieved. However, if user needs and symptom checker advice do not align, negative outcomes may occur. Users may seek unnecessary care or, more seriously, users may not seek necessary care. Each quadrant has different amounts of research substantiating its likelihood, as reported in the sections below. Cost implications of the various outcomes were not reported in the included studies but could be explored in future research.

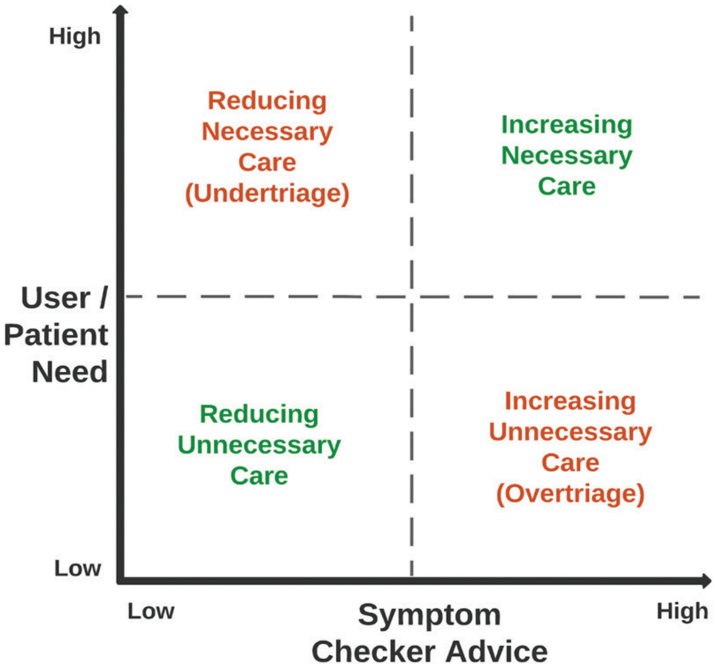
High
Reducing Necessary Care (Undertriage)
Increasing Necessary Care
User / Patient Need
Reducing Unnecessary Care
Increasing Unnecessary Care (Overtriage)
Low
Low
Symptom Checker Advice
High

**Figure 2:** Conceptual Framework of Health Impacts of Symptom Checker Usage (Source: Author's creation)

## Main Health Outcomes

### *1. Reducing unnecessary care (lower left quadrant)*

The most predominant health outcome in the included studies was reducing unnecessary use of healthcare services. Patients may at times seek care that is more intense than their actual need, such as seeking primary care when self-care is suitable, or urgent care when primary care is suitable (O'Cathain et al. 2020). Symptom checkers have the potential to reduce this usage by recommending care in line with patient need.

There is limited evidence to suggest that patients may take up lower-intensity symptom checker advice. In a study of at-home users of the Buoy AI symptom checker (location not specified), 32% reported reductions in the intensity of care they intended to seek compared to their intentions before usage (Winn et al. 2019). Conversely, 91.6% of respondents in a study of English patients who were prompted to use the Ada AI symptom checker while in their GP waiting room reported that they would still have come to the GP even if they had used a symptom checker beforehand (Miller et al. 2020: 6). A qualitative study found that German Ada users also generally intended for a physician to confirm or reject the results, emphasising the importance of a physical examination over the tool's symptom assessment (Müller et al. 2024: 9). Providers have indicated similar views. In a post-study survey of Italian GPs whose patients were invited to use an unspecified AI symptom checker while in the waiting room, only about one-third of the GPs anticipated that symptom checkers would be useful in reducing future 'unnecessary visits' (Mahlknecht et al. 2023: 13).

Providing further insights into the health needs of patients who intend to seek (or have already sought) care, researchers found that nearly half (43.4%) of patients who were prompted to use Ada in a German emergency department (ED) waiting room were considered 'non-emergencies' by Ada (Cotte et al. 2022: 8). If those patients had used Ada at home and subsequently sought a GP appointment rather than emergency care, unnecessary care may have been avoided. Such behaviours may be likely given that, in another study, Buoy users were found to be most likely to follow lower-intensity symptom checker advice such as 'wait and watch', self-treatment or primary care in the next several days, compared to higher-intensity advice to seek care in the next several hours (Carmona et al. 2022: 12). These findings suggest that symptom checkers may reduce overuse of care if patients who do not need urgent care are redirected towards less urgent pathways.

### *2. Increasing unnecessary care ('overtriage', lower right quadrant)*

While the previous section explored symptom checkers' recommendations for patients to appropriately reduce the intensity of their care, the opposite problem may also occur wherein patients are recommended higher levels of care than they need. This is referred to as 'overtriage' and may occur as programmers prioritise patient safety (Schmieding et al. 2021: 5).

The included studies raised similar concerns about AI symptom checkers providing results that are safe but excessive (Fraser et al. 2022). German Ada users reported anecdotes of the tool providing overly urgent advice at times, such as recommending emergency care for a respondent's tonsillitis when the person 'knew it wasn't necessary' as they could get antibiotics from their GP instead (Müller et al. 2024: 5). Two studies asked expert physician panels to assess Ada's recommendations for emergency room patients, with the panels finding overtriages by Ada in one-quarter to one-half of all cases (Cotte et al. 2022; Fraser et al. 2022). To test a wider range of symptom checkers, Fraser and colleagues (2023) ran patient data (originally entered by patients themselves into Ada) through the following tools: OpenAI's ChatGPT version 3.5, ChatGPT version 4.0 (an updated version), and WebMD Symptom Checker (AI status not stated). ChatGPT performed best in reducing overtriages, with 0% of ChatGPT v3.5's triages, 3% of ChatGPT v4.0's triages and 11% of WebMD's triages being classified by the authors as 'too cautious' (Fraser et al. 2023: 5). However, the converse result may be higher rates of potentially unsafe triages, discussed in the following section.

### *3. Reducing necessary care ('undertriage', upper left quadrant)*

Recommending less urgent care than a patient needs is called 'undertriage' and is considered the greatest risk to patient health and safety arising from symptom checker usage. The same studies of patient-entered data in emergency departments found that small minorities (8.9-14%) of Ada's recommendations were considered potential undertriages, as assessed by at least two out of three independent physicians (Cotte et al. 2022: 7; Fraser et al. 2022: 5). Meanwhile, ChatGPT v3.5 had 41% undertriages and the updated ChatGPT v4.0 had 22%

undertriages (Fraser et al. 2023: 5). Patient follow-up provided mixed results. Of Ada's 20 potential undertriages of patients in a German emergency waiting room, 15 actually required emergency care, indicating that those patients would have experienced a health risk if they had followed Ada's advice while at home (Cotte et al. 2022: 8). However, the patient-users were recruited while already seeking emergency care, so these effects could not be assessed. Another study found no 'adverse' outcomes for the five English emergency department patients identified as potential Ada undertriages, but the sample size was not powered to detect uncommon events (Fraser et al. 2022: 11).

Compared to the previous studies that used data entered by patients while they were in clinical waiting rooms, Saenger and colleagues (2024) described a case study of real at-home symptom checker usage. The subject was a 63-year-old male patient who experienced vision problems post-surgery. The patient entered his symptoms into ChatGPT and received a non-emergency treatment advice, contributing to his decision to delay seeking care for 24 hours before he called an ambulance. According to the emergency department report, he was having a suspected transient ischemic attack (mini-stroke), which may have merited seeking care earlier (Saenger et al. 2024: 237). The authors feel the case study indicates the risk that patients, engaging in 'unmoderated' symptom-checking conversations with chatbots, may not seek medical interventions when necessary (Saenger et al. 2024: 238). Several symptom checker users in Germany also expressed worries that other users might follow non-urgent symptom checker recommendations even if the situation was 'life-threatening', although no users reported undertriage experiences themselves (Müller et al. 2024: 13).

#### *4. Increasing necessary care (upper right quadrant)*

Several studies described potential benefits of symptom checkers increasing or encouraging usage of necessary care, particularly care that is considered more sensitive. German respondents reported qualitatively that Ada made it easier to raise uncomfortable topics such as body weight and sexually transmitted diseases, compared with a doctor (Müller et al. 2024: 9). In a survey of at-home Buoy users, 11% said they used the tool for symptoms they would be 'too embarrassed' to bring up with their primary care provider, with gynaecological issues being the second-most reported symptom (Carmona et al. 2022: 14). Moderately positive results also indicated that users felt 'encouraged to seek help' after using Buoy, although it is not known if those were the same users who reported using the tool for gynaecological issues (Carmona et al. 2022: 12). The authors nonetheless suggest that symptom checkers might be 'particularly useful for users affected by conditions considered personal, embarrassing, stigmatising… or requiring potentially uncomfortable or psychologically stressful physical examinations (such as pelvic examinations),' (Carmona et al. 2022: 14). Empirical outcomes for these patients or conditions were not described.

### Other Health Outcomes

#### *1. Quality of care*

Two studies investigated whether symptom checker usage might affect the quality of a subsequent doctor's visit. In a study where patients were invited to use an unnamed symptom checker while in an Italian GP waiting room, about one-quarter of patients and physicians reported the symptom checker having a positive impact on the quality of the medical visit, while the large majority were neutral (Mahlknecht et al. 2023: 11). In the same study, results were inconclusive on the effects the symptom checker had on visit duration, with equal proportions of patients reporting visits being either longer or shorter than expected (approx. 11.5% for each) (Mahlknecht et al. 2023: 11). In a qualitative study with German Ada users, several participants reported discussing the tool's results with their physicians; the overall impression was 'neutral' on how this affected the visit (Müller et al. 2024: 11). One participant reported positive views from their doctor about the person's symptom checker usage, but another participant reported not telling their physician about their usage due to fears of being perceived as challenging the physician's authority (Müller et al. 2024: 11).

#### *2. Reducing anxiety*

A final indirect health outcome of symptom checkers was the potential impact on user mental health. In a study of US Buoy users, reports were weak but positive on users feeling 'less anxious' post-use (Carmona et al. 2022: 12). Conversely, some German Ada users reported post-use experiences of 'mental stress' (Müller et al. 2024: 8). Further details were not provided in either study.

### Economic Outcomes

The literature search found very little evidence on the economic impacts of AI symptom checkers, with no studies focusing solely on economics or finances. Some German Ada users anticipated 'personal savings' from usage, including avoiding long travels to a clinic (13 out of 15 respondents were located rurally) (Müller et al. 2024: 12). Evidence of achieving these savings in practice was not elaborated. In the US context, Buoy users with insurance were found to be 2.21 times more likely to have intentions of seeking medical care after consultation, compared with uninsured individuals (Carmona et al. 2022: 12). Again, the implications of these differential intentions were not explored further by the authors. System-wide economic impacts were not investigated in any of the studies.

## DISCUSSION

### Overview

Symptom checkers are a rapidly growing consumer item with increasing integration of AI, yet their impacts on patient-users and health systems are largely unknown. Following an assessment of the empirical literature base on AI symptom checkers, this review contributes a conceptual framework, aiming to consolidate knowledge and guide future research on the direct health impacts of these tools. The conceptual framework maps four broad health outcomes occurring across two axes that both range from low to high: the patient's health needs, and the intensity of the symptom checker's advice. Depending on these configurations, the framework shows that symptom checkers can potentially increase or decrease both necessary and unnecessary care. Different symptom checkers, which vary in their programming (Fraser et al. 2023), produce different results on this matrix. The following general trends of AI symptom checkers were observed: positive implications for reductions in unnecessary care if patient needs are low in practice, but also relatively high rates of recommendations for unnecessary care (overtriage), lower but not insignificant rates of recommendations against necessary care (undertriage), and minimal evidence on encouraging necessary care. Indirect outcomes such as costs and anxiety are also relevant but have a similarly minimal research base.

The literature review identified nine publications that met the inclusion criteria. The small number of studies reflects the state of the literature on empirical outcomes of AI symptom checkers, yet was still sufficient to reach saturation of knowledge and generate a conceptual framework. The conceptual framework was furthered by the well-developed theoretical backings of the included studies, as noted in the quality assessment table (Appendix B). However, the overall quality of the included evidence was found to be inconsistent both across and within studies. Methodological concerns of the included studies are raised in subsequent sections to contextualise the conclusions that can be drawn from this review.

### Review Strengths and Limitations

AI symptom checkers are an emerging and heterogeneous area of scholarly investigation, adding complexity to a literature review. Steps were taken to help ensure rigour, such as using a variety of synonyms for each search term and running multiple pilot searches to see if relevant texts were picked up. The search strategy was also tested and refined with the help of an LSE librarian. To note, following full-text review, only one article resulted from the database search while eight articles were identified through the snowball method. The snowball strategy was useful in identifying relevant sources, although the technique made the search process less replicable (Radionova et al. 2023: 12). Excluding grey literature further restricted the scope. This decision was taken to enhance the review's quality and reliability, but industry influence may still have been at play as several of the included studies had author affiliations with the symptom checker under investigation. One study received funding from Buoy (Carmona et al. 2022) and three had authors employed by Ada (Cotte et al. 2022; Miller et al. 2020) or Buoy (Winn et al. 2019).

### Limitations of Included Studies

Several limitations of the included studies are apparent from the quality assessment scores (Appendix B). First, most studies used symptom checkers as a primary data collection tool, but few justified or assessed those data collection tools and methods of analysis. The studies subsequently lacked assessments of whether patients would enter their health information in the same way every time they used the symptom checker (reliability), as well as whether patient-entered information is a genuine representation of their health status and healthcare intentions (validity) (Winecoff & Bogen 2024). This is concerning because it has been shown that slight variations in patient-entered language can lead to different results being shown by chatbots (Saenger et al. 2024). With symptom checkers being sensitive to wording, further research would benefit from formal assessment of reliability and validity when using symptom checkers as a form of data collection (Mansab, Bhatti & Goyal 2021).

The QATSDD assessment also showed that many of the studies' samples were not fully representative of their target populations due to size and sampling methodology. Six out of nine studies conducted convenience samples of patients in clinical waiting rooms rather than randomly sampling at-home symptom checker users (two groups which may have substantially varying preferences). In particular, patients who are already physically at a clinic may be less able or willing to change their care intentions compared to at-home users who would be deciding whether to seek care in the first place (Miller et al. 2020: 7). This is indicated by the review's findings that many respondents still intended to see a physician after using a symptom checker in their own GP waiting room. Still, scholars have argued that a 'real-life setting' such as a GP office is superior to a simulated lab environment which may be even further removed from a user's healthcare experiences (Mahlknecht et al. 2023: 16).

### Research Gaps and Future Directions

A primary research gap is noted in the lack of studies designed to assess true patient health needs and behaviours following use of AI symptom checkers, the two dimensions necessary to quantify health outcomes (Pairon, Philips & Verhoeven 2023). While the included studies utilised real user data, few gathered follow-up data and none comprehensively. Further comparisons were lacking with non-users. The research also lacked granular data about the needs of different population groups such as older adults and minority language speakers, and any differential health outcomes from their use of symptom checkers (Savolainen & Kujala 2024: 8). It therefore remains to be demonstrated in the literature whether the potential benefits and harms of symptom checker usage will be distributed (un)equally within the overall populations of both users and non-users (Kopka et al. 2023).

One way to achieve more real-world patient data is to link symptom checker data to a patient's electronic health record (EHR) (Winn et al. 2019: 42). This may be more feasible in settings where the same healthcare provider administers both the EHR and the symptom checker (Liu et al. 2022). The EHR provides valuable and large-scale data on patient diagnoses and behaviours over time, but as a source of sensitive information, careful data management and privacy will be top priorities for any such future studies (Wood et al. 2021).

Additionally, this literature review aimed to assess artificially intelligent symptom checkers, given the existing literature's proposition that AI tools are distinct from non-AI ones in functionality and ethics (Meyer et al. 2020; Zawati & Lang 2024). With continuous learning from user inputs, AI has the potential to bring faster, more personalised health-related advice (within legal bounds) than previous symptom checkers (Wimbarti, Kairupan & Tallei 2024: 345). However, literature has also found that consumers have mixed views, including reports of distrust and fear, towards AI tools in general (Cicek, Gursoy & Lu 2024), and in medical uses specifically (Khullar et al. 2022). As this review has demonstrated, research is still highly limited on how AI symptom checkers impact health in practice. While the included studies all mentioned to varying extents the artificially intelligent nature of the symptom checkers under investigation, none directly assessed how the presence of AI might have affected the outcomes.

A final limitation of the review's findings is the fact that most studies gathered their data during 2018 or 2019. While these studies still investigated artificially intelligent symptom checkers, AI has developed substantially in the last several years, including being trained on larger datasets and user inputs (Saenger et al. 2024: 238). A quick web search reveals a plethora of AI symptom checkers that are available for personal use, many without peer-reviewed assessments (Wood et al. 2021). The rapid pace of new AI tools may be difficult for scholars to keep up with, leading to less rigorous but quick-to-publish studies such as Saenger and colleagues' (2024) case study on ChatGPT use, included in this review. Previous scholars on the topic of symptom checkers (not AI-specific) have noted that literature reviews will 'require regular updating to keep track of new studies' (Chambers et al. 2019: 11). This review aimed to provide such an update, with the addition of explicit consideration for AI, an area that has attracted the attention of the public as well as policymakers and regulators (Iacobucci 2020).

A final research gap is found in the lack of formal, high-quality economic evaluations of AI symptom checkers. Such evidence is increasingly important as decisionmakers in health systems look for quick and easy-to-implement solutions to ease budgetary pressure (Wetzel et al. 2024: 10). Rigorous economic evidence will be needed to assess whether AI symptom checkers are financially sustainable and whether health providers would, for example, save money by promoting or purchasing these tools for their constituents (Iacobucci 2020: 2). Without such evidence, health systems may face financial losses, especially if new AI businesses underperform or even go bankrupt (Oliver 2019).

## CONCLUSION

This narrative literature review has aimed to consolidate the nascent body of evidence on the health and economic impacts of AI symptom checkers for non-hypothetical users. The narrative synthesis approach allowed for the development of a conceptual framework that can be used as a roadmap to guide further assessments on how symptom checkers might impact individual and aggregate health. While other authors have variously discussed overtriage and undertriage, no other conceptual framework has, to the best of this author's knowledge, charted the possible health impacts of symptom checker usage as shown in Figure 2.

The framework describes how symptom checkers may alternately reduce or increase necessary or unnecessary care, depending on the level of user need at the outset and the intensity of advice given by the tool. The greatest amount of evidence was available for the quadrant of 'reducing unnecessary care', as substantial percentages (but still a minority) reported intentions to reduce the intensity of care sought following their use of AI symptom checkers. Other quadrants of the framework received less scholarly attention, indicating the need for research that covers all possible health impacts from these digital tools. While the framework provides a general overview of symptom checkers' potential health impacts, specific conclusions could not be verified due to the minimal data and low or mixed quality of the available studies. In particular, most studies used unvalidated data collection tools on small samples of patients who may not represent at-home symptom checker users.

Several key research gaps were identified: a need for more real-world and follow-up data on user health needs and behaviours, data on AI specifically (including highly recent AI and comparisons to non-AI), and economic evidence on how these tools might impact individual and health system costs. In the context of these research gaps, decision-makers face substantial uncertainty about AI symptom checkers and how they might impact health and finances. The concern is that AI symptom checkers may deliver users advice that is unhelpful and wasting resources (O'Reilly-Jacob et al. 2021), or at worst unsafe and endangering (Iacobucci 2020). With health outcomes at the centre of the conceptual framework, the review has aimed to highlight the imperative for AI symptom checker advice to be fully aligned with patient needs – not just in hypothetical models but in practice.

## REFERENCES

- Aboueid, S., Liu, RH., Desta, BN., Chaurasia, A. and Ebrahim, S. 2019. The use of artificially intelligent self-diagnosing digital platforms by the general public: A scoping review. JMIR Medical Informatics, 7(2): p.e13445. doi:10.2196/13445.
- Carmona, KA., Chittamuru, D., Kravitz, RL., Ramondt, S. and Ramírez, AS. 2022. Health information seeking from an intelligent web-based symptom checker: Cross-sectional questionnaire study. Journal of Medical Internet Research, 24(8): e36322. DOI: 10.2196/36322.
- Ceney, A., Tolond, S., Glowinski, A., Marks, B., Swift, S. and Palser, T. 2021. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS ONE, 16(7): e0254088. DOI: 10.1371/journal.pone.0254088.
- Chambers, D., Cantrell, AJ., Johnson, M., Preston, L., Baxter, SK., Booth, A. and Turner, J. 2019. Digital and online symptom checkers and health assessment/triage services for urgent health problems: A systematic review. BMJ Open, 9(8): e027743. DOI: 10.1136/bmjopen-2018-027743.
- Chen, A., Chen, DO. and Tian, L. 2023. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. Journal of the American Medical Informatics Association. Available ahead of print. DOI: 10.1093/jamia/ocad245.
- Cicek, M., Gursoy, D. and Lu, L. 2024. Adverse impacts of revealing the presence of “Artificial Intelligence (AI)” technology in product and service descriptions on purchase intentions: The mediating role of emotional trust and the moderating role of perceived risk. Journal of Hospitality Marketing & Management, 34(1): 1–23. DOI: 10.1080/19368623.2024.2368040.
- Ćirković, A. 2020. Evaluation of four artificial intelligence-assisted self-diagnosis apps on three diagnoses: Two-year follow-up study. Journal of Medical Internet Research, 22(12): e18097. DOI: 10.2196/18097.
- Cotte, F., Mueller, T., Gilbert, S., Blümke, B., Multmeier, J., Hirsch, MC., Wicks, P., Wolanski, J., Tutschkow, D., Schade Brittinger, C., Timmermann, L., and Jerrentrup, A. 2022. Safety of triage self-assessment using a symptom assessment app for walk-in patients in the emergency care setting: Observational prospective cross-sectional study. JMIR mHealth and uHealth, 10(3): e32340. DOI: 10.2196/32340.
- Dixon-Woods, M., Agarwal, S., Jones, D., Young, B. and Sutton, A. 2005. Synthesising qualitative and quantitative evidence: A review of possible methods. Journal of Health Services Research and Policy, 10(1): 45–53. DOI: 10.1177/135581960501000110.
- Dunn, A.G. 2020. Will online symptom checkers improve health care in Australia? Medical Journal of Australia, 212(11): 512–513. DOI: 10.5694/mja2.50621.
- Eichenberg, C. and Schott, M. 2019. Use of web-based health services in individuals with and without symptoms of hypochondria: Survey study. Journal of Medical Internet Research, 21(6): e10980. DOI: 10.2196/10980.
- Farnood, A., Johnston, B. and Mair, F.S. 2020. A mixed methods systematic review of the effects of patient online self-diagnosing in the “smart-phone society” on the healthcare professional-patient relationship and medical authority. BMC Medical Informatics and Decision Making, 20(1): 253. DOI: 10.1186/s12911-020-01243-6.
- Fraser, HSF., Cohan, G., Koehler, C., Anderson, J., Lawrence, A., Pateña, J., Bacher, I., and Ranney ML. 2022. Evaluation of diagnostic and triage accuracy and usability of a symptom checker in an emergency department: Observational study. JMIR mHealth and uHealth, 10(9): e38364. DOI: 10.2196/38364.
- Fraser, HSF., Crossland, D., Bacher, I., Ranney, M., Madsen, T. and Hilliard, R. 2023. Comparison of diagnostic and triage accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and physicians for patients in an emergency department: Clinical data analysis study. JMIR mHealth and uHealth, 11: e49995. DOI: 10.2196/49995.
- Gräf, M., Knitza, J., Leipe, J., Krusche, M., Welcker, M., Kuhn, S., Mucke, J., Hueber, A. J., Hornig, J., Klemm, P., Kleinert, S., Aries, P., Vuillerme, N., Simon, D., Kleyer, A., Schett, G., and Callhoff, J. 2022. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatology International, 42(12): 2167–2176. DOI: 10.1007/s00296-022-05202-4.
- Hatem, R., Simmons, B. and Thornton, JE. 2023. A call to address AI “hallucinations” and how healthcare professionals can mitigate their risks. Cureus, 15(9): e44720. DOI: 10.7759/cureus.44720.
- Iacobucci, G. 2020. Row over Babylon’s chatbot shows lack of regulation. BMJ, 368: m815. DOI: 10.1136/bmj.m815.
- Jabareen, Y. 2009. Building a conceptual framework: philosophy, definitions, and procedure. International Journal of Qualitative Methods, 8(4): 49–62. DOI: 10.1177/160940690900800406.
- Khullar, D., Casalino, LP., Qian, Y., Lu, Y., Krumholz, HM. and Aneja, S. 2022. Perspectives of patients about artificial intelligence in health care. JAMA Network Open, 5(5): e2210309. DOI: 10.1001/jamanetworkopen.2022.10309.
- Klingenberg, OG., Holkesvik, AH. and Augestad, LB. 2020. Digital learning in mathematics for students with severe visual impairment: A systematic review. British Journal of Visual Impairment, 38(1): 38–57. DOI: 10.1177/0264619619876975.
- Kopka, M., Scatturin, L., Napierala, H., Fürstenau, D., Feufel, MA., Balzer, F. and Schmieding, ML. 2023. Characteristics of users and nonusers of symptom checkers in Germany: Cross-sectional survey study. Journal of Medical Internet Research, 25: e46231. DOI: 10.2196/46231.
- Liu, AW., Odisho, AY., Brown III, W., Gonzales, R., Neinstein, AB. and Judson, TJ. 2022. Patient experience and feedback after using an Electronic Health Record-integrated COVID-19 symptom checker: Survey study. JMIR Human Factors, 9(3): e40064. DOI: 10.2196/40064.
- Lyons, RJ., Arepalli, SR., Fromal, O., Choi, JD. and Jain, N. 2024. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Canadian Journal of Ophthalmology, 59(4): e301–e308. DOI: 10.1016/j.jcjo.2023.07.016.
- Mahlknecht, A., Engl, A., Piccoliori, G. and Wiedermann, CJ. 2023. Supporting primary care through symptom checking artificial intelligence: A study of patient and physician attitudes in Italian general practice. BMC Primary Care, 24(1): 174. DOI: 10.1186/s12875-023-02143-0.
- Mansab, F., Bhatti, S. and Goyal, D. 2021. Reliability of COVID-19 symptom checkers as national triage tools: An international case comparison study. BMJ Health Care Informatics, 28(1): e100448. DOI: 10.1136/bmjhci-2021-100448. PMID: 34663637; PMCID: PMC8523957.
- Meyer, AND., Giardina, TD., Spitzmueller, C., Shahid, U., Scott, TMT. and Singh, H. 2020. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: Cross-sectional survey study. Journal of Medical Internet Research, 22(1): e14679. DOI: 10.2196/14679.
- Miller, S., Gilbert, S., Virani, V. and Wicks, P. 2020. 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 Human Factors, 7(3): e19713. DOI: 10.2196/19713.
- Morrow, E., Zidaru, T., Ross, F., Mason, C., Patel, KD., Ream, M. and Stockley, R. 2023. Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. Frontiers in Psychology, 13: 971044. DOI: 10.3389/fpsyg.2022.971044.

- Müller, R., Klemmt, M., Koch, R., Ehni, HJ., Henking, T., Langmann, E., Wiesing, U., and Ranisch, R. 2024. “That’s just Future Medicine” - a qualitative study on users’ experiences of symptom checker apps. BMC Medical Ethics, 25(1): 17. DOI: 10.1186/s12910-024-01011-5.
- O’Cathain, A., Connell, J., Long, J. and Coster, J. 2020. “Clinically unnecessary” use of emergency and urgent care: A realist review of patients’ decision making. Health Expectations, 23(1): 19–40. DOI: 10.1111/hex.12995.
- Oliver, D. 2019. David Oliver: Lessons from the Babylon Health saga. BMJ, 365: l2387. DOI: 10.1136/bmj.l2387.
- O’Reilly-Jacob, M., Mohr, P., Ellen, M., Petersen, C., Sarkisian, C., Attipoe, S., and Rich, E. 2021. Digital health & low-value care. Healthcare (Amsterdam), 9(2): 100533. DOI: 10.1016/j.hjdsi.2021.100533.
- Page, MJ., McKenzie, JE., Bossuyt, PM., Boutron, I., Hoffmann, TC., Mulrow, CD., Shamseer, L., Tetzlaff, JM., Akl, EA., Brennan, SE., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, MM., Li, T., Loder, EW., Mayo-Wilson, E., McDonald, S., McGuinness, LA., Stewart, EW., Thomas, J., Tricco, AC., Welch VA., Whiting P., and Moher, D. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372: n71. DOI: 10.1136/bmj.n71.
- Pairon, A., Philips, H. and Verhoeven, V. 2023. A scoping review on the use and usefulness of online symptom checkers and triage systems: How to proceed? Frontiers in Medicine (Lausanne), 9: 1040926. DOI: 10.3389/fmed.2022.1040926.
- Paré, G. and Kitsiou, S. 2017. Methods for literature reviews. In: Lau, F. and Kuziemsky, C. (eds.). Handbook of eHealth evaluation: An evidence-based approach [Internet]. Victoria (BC): University of Victoria. Chapter 9. Available at: https://www.ncbi.nlm.nih.gov/books/NBK481583/ [Last accessed 20 February 2025].
- Pawson, R. 2002. Evidence-based policy: In search of a method. Evaluation, 8(2): 157–181. DOI: 10.1177/1358902002008002512.
- Radionova, N., Ög, E., Wetzel, AJ., Rieger, MA. and Preiser, C. 2023. Impacts of symptom checkers for laypersons’ self-diagnosis on physicians in primary care: Ccoping review. Journal of Medical Internet Research, 25: e39219. DOI: 10.2196/39219.
- Reardon, S. 2023. AI chatbots can diagnose medical conditions at home. How good are they? Scientific American, 31 March. Available at: https://www.scientificamerican.com/article/ai-chatbots-can-diagnose-medical-conditions-at-home-how-good-are-they/ [Last accessed 15 February 2025].
- Resnik, DB. 2019. Institutional conflicts of interest in academic research. Science and Engineering Ethics, 25(6): 1661–1669. DOI: 10.1007/s11948-015-9702-9.
- Saenger, J.A., Hunger, J., Boss, A. and Richter, J. 2024. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. Wiener Klinische Wochenschrift, 136(7–8): 236–238. DOI: 10.1007/s00508-024-02329-1.
- Savolainen, K. and Kujala, S. 2024. Testing two online symptom checkers with vulnerable groups: Usability study to improve cognitive accessibility of eHealth services. JMIR Human Factors, 11: e45275. DOI: 10.2196/45275.
- Schmieding, ML., Mörgeli, R., Schmieding, MAL., Feufel, MA. and Balzer, F. 2021. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. Journal of Medical Internet Research, 23(3): e24475. DOI: 10.2196/24475. Erratum in: Journal of Medical Internet Research, 23(5): e30215. DOI: 10.2196/30215.
- Semigran, HL., Linder, JA., Gidengil, C. and Mehrotra, A. 2015. Evaluation of symptom checkers for self-diagnosis and triage: Audit study. BMJ, 351: h3480. DOI: 10.1136/bmj.h3480.
- Shahsavar, Y. and Choudhury, A. 2023. User intentions to use ChatGPT for self-diagnosis and health-related purposes: Cross-sectional survey study. JMIR Human Factors, 10: e47564. DOI: 10.2196/47564.
- Sirriyeh, R., Lawton, R., Gardner, P. and Armitage, G. 2012. Reviewing studies with diverse designs: The development and evaluation of a new tool. Journal of Evaluation in Clinical Practice, 18(4): 746–752. DOI: 10.1111/j.1365-2753.2011.01662.x.
- Sukhera, J. 2022. Narrative reviews: Flexible, rigorous, and practical. Journal of Graduate Medical Education, 14(4): 414–417. DOI: 10.4300/JGME-D-22-00480.1.
- Ubie. 2023. Ubie user terms of use. Last updated 30 October 2023. Available at: https://ubiehealth.com/terms [Last accessed 19 February 2025].
- Wallace, W., Chan, C., Chidambaram, S., Hanna, L., Iqbal, F.M., Acharya, A., Normahani, P., Ashrafian, H., Markar S.R., Sounderajah V., and Darzi A. 2022. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digital Medicine, 5(1): 118. DOI: 10.1038/s41746-022-00667-w.
- Wetzel, AJ., Koch, R., Koch, N., Klemmt, M., Müller, R., Preiser, C., Rieger, M., Rösel, I., Ranisch, R., Ehni, HJ., and Joos, S. 2024. Better see a doctor? Status quo of symptom checker apps in Germany: A cross-sectional survey with a mixed-methods design (CHECK.APP). Digital Health, 10: 20552076241231555. DOI: 10.1177/20552076241231555.
- Wimbarti, S., Kairupan, BHR. and Tallei, TE. 2024. Critical review of self-diagnosis of mental health conditions using artificial intelligence. International Journal of Mental Health Nursing, 33(2): 344–358. DOI: 10.1111/inm.13303.
- Winecoff, A. and Bogen, M. 2024. Trustworthy AI needs trustworthy measurements. Center for Democracy and Technology. Available at: https://cdt.org/insights/trustworthy-ai-needs-trustworthy-measurements/#:~:text=Validity%20refers%20to%20how%20well,foundational%20in%20the%20social%20sciences [Last accessed 3 August 2024].
- Winn, A.N., Somai, M., Fergestrom, N. and Crotty, B.H. 2019. Association of use of online symptom checkers with patients’ plans for seeking care. JAMA Network Open, 2(12): e1918561. DOI: 10.1001/jamanetworkopen.2019.18561.
- Wohlin, C., Kalinowski, M., Felizardo, K.R. and Mendes, E. 2022. Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. Information and Software Technology, 147: 106908. DOI: 10.1016/j.infsof.2022.106908.
- Wood, A., Denholm, R., Hollings, S., Cooper, J., Ip, S., Walker, V., Denaxas, S., Akbari, A., Banerjee, A., Whiteley, W., Lai, A., Sterne, J., Sudlow, C., and CVD-COVID-UK consortium 2021. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: Data resource. BMJ, 373: n826. DOI: 10.1136/bmj.n826.
- Yigitcanlar, T., Corchado, JM., Mehmood, R., Li, RYM., Mossberger, K. and Desouza, KC. 2021. Responsible urban innovation with local government artificial intelligence (AI): A conceptual framework and research agenda. Journal of Open Innovation: Technology, Market, and Complexity, 7(1): 1–16. DOI: 10.3390/joitmc7010071.
- You, Y. and Gui, X. 2021. Self-Diagnosis through AI-enabled chatbot-based symptom checkers: User experiences and design considerations. AMIA Annual Symposium Proceedings, 2020: 1354-1363.
- Zawati, MH. and Lang, M. 2024. Does an app a day keep the doctor away? AI symptom checker applications, entrenched bias, and professional responsibility. Journal of Medical Internet Research, 26: e50344. DOI: 10.2196/50344.