Master Thesis 2019-2020

# eHealth for risk screening and early diagnosis: A scoping review on the accuracy and availability of online diagnostic tools

Author: Nikola Jovicic

First supervisor: Dr. Stans Drossaert

Second Supervisor: Jochem Goldberg

University of Twente

Faculty Behavioural, Management and Social Sciences, Psychology

Master Health Psychology & Technology

Workload: 35 EC

**Abstract**

**Introduction:** Health applications under the form of symptom checkers or other diagnostic tools have promising implications to not only reduce some of the burdens of the modern health care market, but also in empowering its users in becoming increasingly and more positively involved in their own health care. However, not much is known about the availability of such tools, their diagnostic capabilities and accuracy, and their impact on the behaviors and attitudes of its users.

**Objective:** To conduct a scoping review in order to explore these gaps and to report what is actually known about diagnostic tools across the literature by mapping all known tools and studies in that subject.

**Methods:** A search strategy was devised, and three databases were searched for potential studies: PubMed, Scopus, and PsycInfo. 330 studies were identified and were subject to full-text reviews. We included studies of any design as long as they appraised the accuracy of diagnostic tools available to the general public and provided information on the behavioral impact upon the use of such. Thus, 31 studies were selected for final review. Data was extracted in tables, where the characteristics of the tools and studies were summarized and presented.

**Results:** Three different types of diagnostic tools have been identified: *Diagnostic Symptom Checkers, Symptom Checkers with Triage Functions,* and *Risk Calculators.* 80 unique diagnostic tools have been identified, of which 46 had been covered in 3 studies alone. Most diagnostic tools operate on question-based algorithms; however, most studies did not report on either which algorithm it runs on, their validation status, or how many diagnoses they can actually produce. Overall, diagnostic and triage accuracy tends to be poor, with only a few studies showing good rates. However, their accuracy is also extremely variable, which is best demonstrated in studies assessing multiple diagnostic tools where they received the same input before attempting their diagnosis. We have identified a handful of studies on risk calculators, yet only one had its predictive capabilities tested on actual disease incidences. Lastly, the behavioral aspect on the use of diagnostic tools is critically underexplored, with very few studies reporting on changes in behavioral attitudes and actions, such as only half of users actually complying to the advice given by such tools.

**Conclusion:** Almost all diagnostic tools are outperformed by health professionals and their rates of accuracy remain sub-optimal. Considering not only the scarcity of relevant studies in the field and also the evidence, studies need to increase their efforts in clearly outlining the characteristics

of their diagnostic tools, assess a broader range of medical condition as well as examine already known tools to more medical conditions, and measure what people actually do with the health information given.

**Keywords**: Diagnostic tools, e-health, self-diagnosis, symptom checkers, diagnostic & triage accuracy

# Table of Contents

## 1. Introduction

The modern health care market faces many problems that endanger its integrity. These problems entail a large variety of factors, such as growing and increasingly older populations (Bloom et al., 2011), increases in global incidence rates of chronic conditions (Hajat & Stein, 2018), and the overall rising health care expenditures (Hughes, 2010; National Academy Press, 2001). For example, in the US alone, the number of people aged 65 or above has increased from 37.8 million to 50.9 million between 2007 and 2017 (ACL, 2018), and the health care spending has increased from 2.2 trillion dollars (Hartman et al., 2009), to 3.5 trillion dollars (Centers for Medicare & Medicaid Services, 2017) within the same timeframe. Technology may be an answer to alleviating the burdens of health care and to seeking and obtaining health information.

Access to health information used to be strictly limited to traditional media such as health professionals, books, or magazines (Diaz et al., 2002; Dutta-Burgman, 2009). However, an increasing number of people have started using electronic media for obtaining health information (Baker et al., 2003), such as self-screening/diagnosing themselves for a health concern (Fox & Duggan, 2013). modern media technologies have been rapidly advancing and contributing to the distribution of health information, with the Internet playing a focal role in this advancement (Hsieh et al, 2016). The Internet has emerged as an alternative compared to traditional health care media and has become one of the largest go-to sources for seeking health information for the global population (Chen et al., 2018). 75% of the global internet users have searched for health information online at some point (Doherty-Torstrick et al., 2016), while every third adult in the US regularly uses the Internet in the attempt to self-diagnose themselves (Fox & Duggan, 2013). Also, a 2015 UK survey asking individuals using the internet for health information, 73% indicated having attempted to search a symptom or to self-diagnose, 63% wanted to know more about managing a condition or illness, 39% wanted more information on improving their health and researching potential treatments, and 38% listed wanting to know more about risks associated to some procedures (Statista Research Department, 2015). While these figures might differ between different countries, they do indicate a noticeable trend between the use of the Internet and health information seeking behaviors. Furthermore, popular search engines such as Google or Yahoo are being commonly used for initial health queries (Wang et al., 2012). Estimates for the total number of health related searches can vary between different studies and search engines, but some engines like Google alone yields around 70000 health queries per

minute (Murphy, 2019). Due to the increased digitalization of health information and services and the growing interest in the exploration of such information, patients have become more active agents in their own health care instead of just being passive recipients (McMullan, 2006). That being said, our study is especially interested in tools or programs accessible to the general public via the internet, allowing to self-screen or self-diagnose oneself. Throughout this study, we will refer to them as diagnostic tools, which allow a user to diagnose or appraise their own medical condition or query.

*Rising popularity of e-health and diagnostic applications*

The trend towards making health related investigations on the internet and its growing interest has invariably contributed to the expansion of online health services and applications tailored towards a wide user-base, with e-Health emerging as a major discipline in that area. E-health can be understood as "*a variety of technologies or electronic services facilitating healthcare for patients, providers, and stakeholders as well as the distribution of health information and services*" (Sousa & Lopez, 2017, p.471), and is one of the most rapidly growing areas in the health care market (Srivasta et al., 2015). The number of health apps has grown explosively from 40000 (Boulos et al., 2014) to 318000 (IQVIA Institute, 2017) between 2012 and 2017 alone. In terms of global download rates of those applications, the numbers have doubled from 200 million unique downloads to 400 million between 2010 and 2018 (App Annie, 2019). Most of these health applications are centered around improving chronic care management, medications management, disease management, as well as increasing self-monitoring behaviors and health literacy (Silva et al., 2015; Sousa & Lopez, 2017).

A brief search resulted in us finding at least 3 types of such tools or/and applications, namely computerized diagnostic decision support systems or CDDSS (Nurek et al., 2015), crowdsourcing platforms (Meyer et al., 2016), and symptom checkers (Morita et al., 2017). CDDSS' are targeted towards physicians and other health professionals as a complimentary tool within their practices. Crowdsourcing platforms such as CrowdMed are websites where undiagnosed patients can submit their symptoms and other potentially relevant information, in which other people try to come up with a diagnosis (Meyer et al., 2016). Symptom checkers are algorithmic tools that allow to generate diagnoses based on the input by the users' perceived symptoms and questions asked by the program to assist in the diagnostic process (Morita et al.,

2017; Semigran et al., 2015). A symptom checker can also be classified as a tool providing diagnostic information based on user-entered symptoms (Kafle et al., 2018). These symptom checkers can also give guidance on the course of action their users should take upon interaction with the diagnostic tool and whether their health concern requires emergent care or not, which is often labelled as triage advice (Middleton et al., 2016). We want to highlight that there are no major distinctions between a ‘’symptom checker’’ and a ‘’diagnostic tool”, as both pertain to help in the early detection or diagnosis of an illness/condition. That being said, the use of such diagnostic tools is also very popular; fifty million people are reported to use self-triage through symptom checkers annually (Wyatt, 2015), while some popular applications have been downloaded anywhere between tens of thousands to tens of millions of times (Lupton & Jutel, 2015). However, what can be said about their effectiveness and the evidence in the literature?

*The advantageous, adverse, and ambiguous evidence for health applications*

E-Health applications hold great promise to improve quality and efficiency of care, energizing and engaging both health professionals and patients, reducing health care costs, reducing the complexity associated with delivering medical information, and is considered as the latest and most promising strategy to improve the current health care system (Elbert et al., 2014; Ossebaard & Gemert-Pijnen, 2016). Particularly, symptom checkers can reduce unnecessary emergency and doctor visits, prescription drugs, and empower patients to be more involved with their own health (Semigran et al., 2015). It is also promising that there are many studies which attempted to analyze which demographic or behavioral determinants potentially mediate positive and negative outcomes behind the use of health applications. Determinants such as age (Clarke et al., 2017), gender (Laz & Berenson, 2013), educational attainment (Ybarra & Suman, 2006), health literacy (Valizadeh-Haghi, & Rahmatizadeh, 2018) and race (Lewis, 2017) have been linked to various degrees of understanding health information, using health applications, and compliance to the advice given by said applications.

Nonetheless, the actual effectiveness of e-health related technologies is debatable. One exhaustive systematic review by Ekeland et al. (2014) found mixed results regarding the influence of eHealth applications on health care costs and health outcomes despite having examined a large body of literature. Some users experience increased levels of anxiety while searching for health information, amplifying the possibility of detrimental health outcomes

(Doherty-Torstrick et al., 2016. In some cases, the ready accessibility of health information has lead to the emergence of a novel phenomenon called ‘Cyberchondria’, which can be described as the obsessive and excessive use of the internet to find a disease matching real or even imagined symptoms (Bagaric & Jokic-Begic, 2019; Loos, 2013). It is also currently unknown whether the existing research around the demographic determinants to health applications for health outcomes would also extend to symptom checkers and other similar diagnostic tools. For example, does the use of such tools promote or hinder behavioral changes, such as actually seeking treatment from a doctor? It is also difficult to establish the number of available diagnostic tools, how they are being used, and their actual effectiveness in terms of accuracy. Due to the newness of eHealth regarding applications for self-diagnosis, self-screening, or early detection, not many studies have been performed that examined either the characteristics of such tools or their efficiency. For example, a study attempting to characterize the content and the functions of smartphone-applications for cancer found that only 34 out of 295 applications were able to provide assistance in the early detection for cancer, but no further analyses were made pertaining their accuracy (Bender et al., 2013). Another problem is that similar studies tend to refer to such diagnostic tools as health applications. Although they do classify as such, it is difficult to distinguish them from other health applications and to clearly identify their performance and effectiveness.

To our knowledge, only 4 studies have attempted to synthesize evidence concerning online diagnostic tools (Aboueid et al., 2019; Chambers et al., 2019; Millenson et al., 2018; Semigran et al., 2015). The presentation of the evidence includes-, but is not limited to, their accuracy, costs, regulations, user experiences, or clinical effectiveness. All studies agreed that the overall strength of evidence accuracy for symptom checkers/diagnostic tools remain inconclusive and weak; they tend to be inaccurate and too variable in their diagnostic performance and perform inferiorly when compared to health practitioners. Additional measures such as changes in behavioral attitudes, risk perceptions, or compliance to the advice given by such applications were also stated to be critically underexplored in the literature, and that a systematic review on this element of symptom checkers/diagnostic tools might not even worthwhile. However, some shortcomings were found in relation to the preemptive conclusions made in this matter. Two of those studies examined few symptom checkers/diagnostic tools,

where the assessment of accuracy was just one small component to other overarching themes/research questions (Aboueid et al., 2019; Chambers et al., 2019). In terms of the functionality of the diagnostic tool, one study did not include symptom checkers with question-based algorithms (Aboueid et al., 2019), one study excluded symptom checkers that only assessed specific medical conditions (Semigran et al., 2015, and the last study included studies assessing a diverse set of diagnostic tools such as CDDSS', crowdsourcing platforms, and symptom checkers in their assessment, but excluded those from before 2011 (Millenson et al., 2018).

*Summary and Objectives*

The growth of eHealth as a discipline and the number of available health applications indicates great promise towards the alleviation of health care burdens and the shift of the patient-role to that of a more active agent. Despite the fact that the number of available health applications is massive, it is currently unknown how many of those are specialized in the appraisal or early detection of diseases. The literature on the assessment of the accuracy and (behavioral) health outcomes of diagnostic tools is also scarce and the level of evidence is weak and unclear. The few studies that did attempt to synthesize evidence towards the use of such applications suffer from shortcomings in their study design and thus painted an incomplete picture in this field of study. The shortcomings entail either a low number of assessed diagnostic tools, the exclusion of tools using question-based algorithms, tools tailored towards specific medical conditions, and tools assessed in studies published before 2011.

In light of the knowledge gaps pertaining to the studies appraising diagnostic tools, we would like to map what kind of tools, programs, or applications for self-screening/diagnostic purposes have been examined in the literature, highlight the types of algorithms used among them, and on which medical conditions they have been tested on. We also want to assess their accuracy and distinguish between the performance of the different types of symptom checkers we might encounter in our study. Lastly, the behavioral impact of the use of such symptom checkers is barely explored, if at all, and little is known about determinants influencing the health decision making of appraising health information given by such tools. We want to explore the literature by assessing whether other studies examining the characteristics and accuracy of

such tools have included socio-demographic and behavioral explanations for their effect on health outcomes. Thus, our objectives are summarized in the following research questions:

1. What is known about the availability, characteristics, and functions of diagnostic tools accessible to the general public?

2. What is the accuracy of such diagnostic tools and what can be said about their predictive/diagnostic abilities?

3. What is the behavioral impact of the use of diagnostic tools on users/patients? Did studies attempt to describe health outcomes or interactions with diagnostic tools based on socio-demographic or other determinants?

## 2. Methods

Based on our research questions and the scarcity of evidence from systematic reviews on this topic, we will conduct a scoping review, which is described as the ''*ideal tool to determine the scope or coverage of a body of literature on a given topic*'' (Munn et al., 2018, p.2), as well as identifying research gaps in the existing literature (Arksey & Malley, 2005). Our scoping review will follow the methodological framework proposed by Arksey and O'Malley (2005). The methodology is based on five stages, namely: Identifying the research question(s), identifying relevant studies, study selection, charting the data, and collating/summarizing/and reporting of results. We have already identified our research questions, and we have also merged the latter two stages into one stage called data extraction and analysis.

### *2.1 Identification of relevant studies*

The literature search was conducted on 3 separate search engines, namely PubMed, Scopus, and PsycInfo, and lasted from May 1, 2019 to July 15, 2019. Due to the newness of symptom checkers in the field of study, we attempted to increase the sensitivity of our search engines towards finding relevant studies by applying many diverse search terms, in which every search engine received the same input. **Table 1** provides an overview on the various search terms used. Furthermore, we manually searched for additional studies in the hopes of covering as many promising sources as possible. This included looking at articles suggested by the respective search engines, examining citations within potential studies, and picking all studies included in

the aforementioned reviews of diagnostic tools for further appraisal. Thus, a total of 6536 studies have been generated, in which 119 had been handpicked.

**Table 1: Search Strategies**

| Search Terms | | |
|---|---|---|
| diagnose yourself AND health | ‘’Illness recognition ‘’ | (online risk assessment) AND health AND (diagnosis OR prediction) AND |
| ‘’Digital health intervention’’ | early help seeking AND online | (web based OR internet OR computer) |
| digital application AND health AND (online OR web OR web based OR internet OR computer) NOT diet NOT alcohol NOT smoking NOT exercise | health risk assessment AND internet-based NOT treatment | "online screening " |
| | health symptom checker | ‘’Preventive health behavior “ |
| ‘’Disorder recognition “ | mental illness OR mental disorder OR mental health OR psychological health | "self assessment" AND health AND (web based OR online) |
| ‘’Early disease diagnosis ‘’ | AND "online detection"OR "early detection" | self diagnosis tool AND health AND (internet OR web OR online OR web based) |
| ‘’Health risk Appraisal “ | ‘’NHS Checkers’’ | |
| Health assessment” AND online | ‘’self triage AND symptom checkers’’ | ‘’Symptom Appraisal “ |
| Health information seeking AND (symptoms OR family risk OR genetic predisposition*) AND (online OR web-based OR internet) | Self-diagnosis tools | ‘’Symptom Checker “ |
| | self-screening AND mental illness OR mental OR disorder OR health | health risk assessment AND internet-based NOT treatment |
| ‘’Health Risk assessment “AND web-based | | HRA AND (web-based OR web based OR internet OR online) |
| ‘’Health risk assessment” AND (online OR web OR web-based OR internet) | HRA AND (effectiveness OR efficiency) | |
| Symptom appraisal AND web use | ((web OR internet OR “search engine” OR google OR online OR “on line”) AND (“help seeking” OR “help-seeking” OR “information seeking” OR “information-seeking”) AND (symptom OR symptoms OR diagnoses OR diagnosis)) | |

*2.2 Study Selection*

The abstracts and citation titles of 6536 studies were screened, and relevant studies were downloaded into a folder. After the removal of 93 duplicates, full article examinations (n=330) were required to be performed for further eligibility. Both processes were performed by one researcher (N.J.). Due to the large number of required full-text screenings, concise inclusion criteria were required. Studies were **included** (1) if they assessed the accuracy of programs, tools, interventions, websites, or any other health application/diagnostic tool, capable of

appraising or diagnosing user-entered symptoms, (2) provided data or evidence concerning behavioral actions or attitudes upon the use of such tools, and (3) were available to the general public. All studies were considered, regardless of the study types/designs, and the year of publication. This would allow us to discover all potential studies in the field. Studies were **excluded** if they (1) did not assess the accuracy of diagnostic tools, (2) were exclusively available or tested on/by health professionals, (3) and did not mention or indicate availability to the general public. We also excluded the 3 reviews from *Aboueid, Chambers, Millenson* and their colleagues as their interpretation on the outcome measures of accuracy contained the aforementioned gaps in knowledge that we try to mediate in this study. *Figure 1* portrays a flow diagram describing the process of screening and selecting relevant articles.

*2.3 Data extraction & analysis*

After the final article selection (n=31), data was extracted and recorded in a Word document. The extracted data consisted of the names of the authors, date and country of publication, (1) the assessed medical condition (2), the name(s) of the diagnostic application(s) (3), the main objective of the study (4), the sample size (5), the study design/methodology (6), and the results (7) (**Table 2-5)**. Additional tables were created for the names of the diagnostic tools as well as the medical conditions they had been tested on (**Appendix**). This process would allow us to identify common themes and relevant information within the studies. After charting the data following this template while also becoming familiar with the content of the studies, we found that we could present the data findings according to 5 areas of focus pertinent to our research questions: Characteristics of online diagnostic tools (1), diagnostic accuracy (2), triage accuracy (3), risk estimation capacities (4), and behavioral actions and changes upon symptom checker use (5). *Characteristics of online diagnostic tools* allowed us to identify and categorize the various diagnostic tools used within the studies, which would give us answers to our first research question. Then, based on that categorization, we were able to distinguish between 3 types of tools (*Diagnostic symptom checkers, symptom checkers with triage functions, and risk calculators*). As our second research question pertains to appraising the accuracy of diagnostic tools, their accuracy could be interpreted and summarized based on the type they belong to (*Diagnostic accuracy, triage accuracy, and risk estimation capacities)*. Further analyses of accuracy were established based on the outcome measures used within the studies to test their

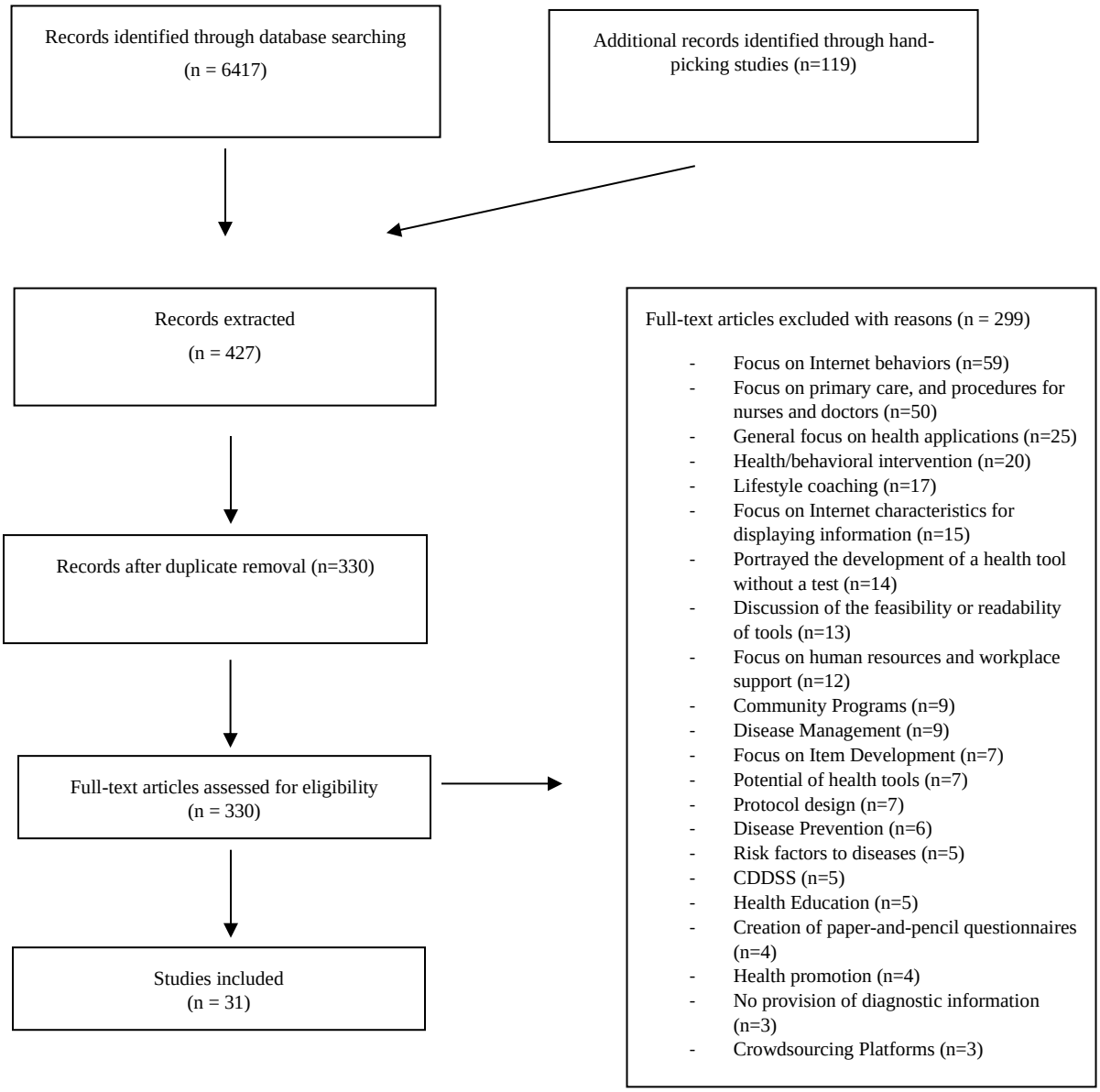


*Figure 1.* Flow diagram of article screening and selection process

accuracy. Lastly, *behavioral actions and changes upon symptom checker use* pertains to our last research question and allowed us to examine behavioral and demographical factors upon the use of diagnostic tools.

## 3. Results

20 studies examined the *diagnostic accuracy* of symptom checkers (**Table 2**) and 8 studies examined *triage accuracy* (**Table 3)**. 5 studies examined symptom checkers with *risk estimation capacities* (**Table 4**), and 5 studies examined *behavioral actions and changes* after the use of symptom checkers (**Table 5**). Further study characteristics and results will be described under each of the abovementioned categories.

### 3.1. Characteristics of online diagnostic tools

*3.1.a. Classification of diagnostic tools*

3 types of diagnostic tools have been identified. First, there are *diagnostic symptom checkers* which provide one or multiple diagnoses based on the symptoms entered by the user (19). Second, symptom checkers that provide the user with triage advice are designated as *symptom checkers with triage functions* (26). It is important to note that a diagnostic symptom checker can also have triage functions, so both types are not always mutually exclusive to each other. Third, we have symptom checkers called *risk calculators*, capable of providing risk estimations in which the user might develop specific disorders years into the future (16).

*3.1.b. Etymological use of diagnostic tools*

The etymological use of online diagnostic tools varies within all included studies. 12 studies are explicitly referring to them as *symptom checkers* (1; 4; 6; 9; 11; 12; 19; 23; 25; 26; 27; 28). Other diagnostic tools are being referred to as *online programs* (2; 3; 5; 7; 10; 13; 21; 22; 24), of which 4 are explicitly referred to as either *self-screeners*, *screening tools*, and *self-assessment tools* (5; 7; 10; 22). The remaining studies appraising diagnostic tools were dubbed *risk calculator websites* (10; 14; 15; 16; 17).

*3.1.c. Availability of diagnostic tools*

This review has identified 80 unique diagnostic tools (**Appendix)**, of which only 6 had their names not disclosed (2; 3; 22; 30). Some diagnostic tools have been appraised on multiple occasions across different studies. WebMD stands at the top with at least 7 studies using the symptom checker as either the primary point of interest, or as a comparator to another checker (1; 4; *6; 9; 23; 26; 27*). Other frequently studied symptom checkers are: SkinVision (18; 20; 29), Mayo Clinic (1; 12; 26), Isabel (1; 12; 26), Symcat (1; 12; 26), Symptomate (1; 26), OA RISK C

(16; 17), AskMD (12; 26), Healthline (4; 26), and 1 unnamed symptom checker for orthopedics (2; 3).

*3.1.d. Medical conditions assessed*

8 studies focused on a wide but unspecified range of disorders (*11; 12; 19; 21; 22; 25; 26; 28*). Even though some studies have given some examples such as common colds, acute liver failure, or migraines (11; 21; 25; 28), a full account of all tested medical conditions was not given. Specific illness conditions have been assessed on multiple occasions across different studies, with dermatological issues (8; 18; 20; 29; 30), arthritis (16; 17; 23), psychiatric disorders (5; 7; 31), and knee pain (2; 3), being the most commonly examined conditions. The remaining conditions are degenerative cervical myelopathy (4), hand pain (9), diabetes (10), ENT complaints (6), ophthalmology (27), influenza (24), breast cancer (14), cardiovascular disorders (15), HIV/Hep C (1), and chlamydia (13).

*3.1.e. Functionality and Validation of Diagnostic Tools/Symptom Checkers*

The diagnostic tools which we have identified above rely on algorithms commonly referred to as knowledge models, which also vary in sophistication and design, but ultimately enable the analysis of user-entered symptoms to generate diagnostic information. Most symptom checkers use *question-based algorithms* that rely on asking the user a series of questions about their symptoms (2; 3; 5; 7; 9; 10; 13-17; 21-24; 28; 31), and can be found for a diverse set of medical conditions. Typically, the answer to each question contributes to identifying the potential illness in which the symptom checker will attempt to return one or multiple diseases based on the answers. For example, one diagnostic symptom checker asks the user 15 questions covering the most common risk factors associated to specific psychological disorders, then provides a diagnosis based on the answers given (5). Risk calculators also usually work in this fashion; for example, a diabetes risk calculator asks the user a series of risk factors associated to developing it, such as blood pressure medication or having a family history with diabetes (10). Moreover, 4 studies describe the creation of novel symptom checkers which are even more sophisticated than other knowledge models (11; 12; 19; 25) but have only been applied to an unspecified broad range of conditions. The difference between the first mentioned algorithm and said complex one is that the latter does not ask the same questions, but rather adapts to the users answers by providing new questions and deprioritizing less relevant ones. For example, one

symptom checker with triage function with a sophisticated algorithm could ask the user whether he feels any pain, and if so whether it is localized in his head (19). If the user says yes to the pain but no to the ‘’head’’, the symptom checker will adapt by choosing a different set of follow-up questions that deprioritize the head region. The remaining group of known algorithms provide diagnostic information through *imaging and processing technologies* (8; 18; 20; 29; 30). The algorithm ranks each assessed picture by rank of probability for presenting evidence for having a specific disorder. This assessment is usually performed via methods such as fractal and imaging analysis, which attempts to recognize and interpret irregular shapes and patterns found on your skin (18). In our review, this type of algorithm was only found in smartphone applications for dermatological purposes.

In terms of validation, many studies have indicated the presence or lack of clinical validity of their own symptom checker or another one, or attempted to establish it in their own respective studies (2; 5; 7; 8; 10; 11; 12; 16; 19; 22; 24; 25). On the other hand, many studies have not made any references towards the validity or the algorithm of the examined symptom checker at all. Also, not much is known about the number of diagnoses a symptom checker can generate. Only 7 studies report the maximum amount of diagnoses a symptom checker is capable of listing and appraising (2; 3; 10; 13; 21; 26; 28).

In sum, despite the terminologies used to describe the tools, all of them can be considered symptom checkers or diagnostic tools as they all attempt to interpret the user’s symptoms and provide potential diagnostic information. Moreover, only few symptom checkers, like WebMD, have been assessed across multiple studies. Similarly, few medical conditions have been tested multiple times as well, with several studies not even stating their exact condition(s). Not all studies have described the functionality and validation status of their symptom checkers, but those that did frequently mention algorithms generating diagnoses based on user-answered questions.

### 3.2 Accuracy of Diagnostic Tools

This category refers to the evidence stated by studies (n=20) examining the accuracy of *diagnostic symptom checkers* (**Table 2**). Almost all studies examined specific illness conditions, except for 4 studies appraising a broad range of unspecified conditions (12; 25; 26; 28). 4 types of outcome measures have been identified which were used to appraise diagnostic accuracy. The

## Table 2: Assessment of studies regarding diagnostic accuracy

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author, Year, Country** | **Illness Condition** | **Methodology** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 1.Berry et al., 2019. USA | HIV, Hepatitis C | 5 Symptom checkers: Mayo Clinic, WebMD, Symptomate, Symcat, Isabel | 90 patients with HIV, 67 patients with hepatitis C, 11 patients with both HIV and hepatitis C. 106 male, 62 female | Analyze diagnostic and triage accuracy of a series of symptom checkers. | Compare output of different symptom checkers with the same input | Top 1 diagnosis for HIV, hepatitis C, and both respectively: 4.4%-7.8%; 3%-16.4%, and 6%-11.3%. Same diseases for being listed at all: 5.6%-42.4%;11.9%-37.3%; and 8.9–39.3%. Diagnostic accuracy vastly inferior to doctors |
| 2.Bisson et al., 2014. USA | Knee Pain (orthopedics) | Unspecified, web-based program | 537 participants. 272 male, 255 female | Design and evaluate the accuracy of an internet-based program generating differential diagnoses for knee pain | Assess the diagnostic overlap between the program and orthopedic surgeons | Sensitivity 89%, Specificity 27%. (for overall list of differentials including the doctor's diagnosis) |
| 3.Bisson et al., 2016. USA | Knee pain (orthopedics) | Unspecified, web-based program to generate differential diagnosis related to knee pain | 328 participants 163 male, 165 female | Evaluate the diagnostic accuracy of an internet-based program assessing knee pain as well as the user's ability to self-diagnose their knee pain from a list of possible diagnoses given by the same program | Assess the diagnostic overlap between the program and orthopedic surgeons. Observe the participants ability to correctly self-diagnose their knee pain based on the returned differentials by the program | Sensitivity 91%, Specificity 23% for diagnoses chosen by program. Sensitivity 58% and specificity 48% for diagnoses chosen by users. Program lists the correct diagnosis most of the time, but the user needs to find the proper diagnosis among listed differentials. |
| 4.Davies et al., 2018. UK | Degenerative Cervical Myelopathy (DCM) | 4 Symptom Checkers: WebMD, Healthline, Healthtools.AARP, and NetDoctor | 31 classical DCM symptoms | Evaluate whether online symptom checkers are able to correctly identify DCM when typical DCM symptoms are entered | Compare output of different checkers with the same input. Assess the extent to which the input classical DCM symptoms lead to DCM being listed as differential | 14/31 (45%) of typical DCM symptoms were listed as a potential DCM differential. 3/31 (10%) were listed in the top third of DCM differentials. Average rank of DCM diagnosis: WebMD (5.6), Healthline (12.9), Healthtools.AARP (15.5), and Netdoctor (no data due to low n). |
| 5.Donker et al., 2009. Netherlands | Psychiatric problems | 1 Online Program: Web Screening Questionnaire (WSQ) | 502 participants. 217 male, 285 female participants | Test and examine the validity and diagnostic accuracy of the WSQ | Performance of the WSQ is compared to other established questionnaires for mental disorders | Social phobia, panic disorder with agoraphobia, agoraphobia, OCD, and alcohol abuse/dependence: Sensitivity 72%-100%, Specificity 63%-80%. Depressive disorder, GAD, PTSD, specific phobia, and panic disorder: Sensitivity 80%-93%, Specificity 44%-51%. |

## Table 2 continued

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author, Year, Country** | **Illness Condition** | **Methodology** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 6. Farmer et al., 2011. UK | ENT complaints | 1 Symptom Checker: Boots WebMD | 61 patients 31 male, 30 female | Report the findings of a study examining the accuracy of Boots WebMD to diagnose ENT complaints | Compare the output of the symptom checker to that of a doctor and assess the diagnostic overlap | 70% of patients were correctly diagnosed by the symptom checker. First diagnosis matched clinical diagnosis in 16 % of patients. |
| 7.Farvolden et al., 2003. Canada | Psychiatric problems | 1 online program: WB-DAT | 183 participants. 79 male, 114 female | Validate and examine the accuracy of the novel web-based self-report screener WB-DAT | Diagnostic output of the WB-DAT was compared to Structured Clinical Interviews for Axis I Disorders from the DSM-IV, and assess the diagnostic overlap | Range of agreement (Cohen's kappa) ranged from 0.57-70. Sensitivity 71%-95%, Specificity 87%-97%. Acceptable to good agreement rates. |
| 8.Ferrero et al., 2013. USA | Dermatological issues | 1 dermatological app: Skin Scan | 93 photos of biopsy-proven melanoma | Investigate the ability of Skin Scan to detect skin lesions for high risk for melanomas | Compare the output of Skin Scan to 93 photos of biopsy-proven melanoma by its ability to correctly identify them as high-risk lesions | 10.8% (10/93) were identified as high-risk lesions, 88.3% (82/93) were labelled as medium risk, and 1.2% (1/93) were labelled as low risk. The performance is poor to acceptable as all lesions were biopsy-proven melanomas. |
| 9.Hageman et al., 2014. USA | Hand-issues | 1 online Symptom Checker: WebMD | 86 participants. 44 male, 42 female participants | Investigate diagnostic accuracy of WebMD for hand-related issues, as well as demographic factors potentially leading to the identification of the correct diagnosis | Output of WebMD was compared to the diagnoses of 3 hand surgeons. Demographic factors were examined to see whether the users affected WebMD's diagnosis accuracy | 33% of the diagnoses derived by the WebMD matched the final diagnosis of the hand surgeon. A multivariable model including sex (female), additional years of education, and prior use of the Internet to research their medical condition or symptoms explained 15% of the variation in correspondence of diagnosis. |
| 10.Heikes et al., 2007. USA | Diabetes and pre-diabetes | 1 online self-screening tool: DIABETES RISK CALCULATOR | 30,818 participants from NHANES III dataset. Unclear gender assignment. | Develop an online tool accessible to the general population for calculating their probability of undiagnosed diabetes or pre-diabetes, and demonstrate its diagnostic accuracy | 3 separate online tools were created in the process, in which DIABETES RISK CALCULATOR had the highest accuracy among them. Then, said tool was tested by receiving input from the NAHNES III dataset, which includes patient information from 30818 participants. Lastly, DIABETES RISK CALCULATOR output was compared to the patient's actual diagnosis. | Sensitivity and specificity for undiagnosed diabetes: 88% and 75% respectively Sensitivity and specificity for pre-diabetes and undiagnosed diabetes: 75% and 65% respectively. |

## Table 2 continued

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author (Year)** | **Illness Condition** | **Methodology** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 12.Kim & Lee, 2018. South Korea | Unspecified/General | 1 symptom checker: Human Disease Diagnosis Ontology (HDDO) | 6 other symptom checkers | Describe the creation and processing capabilities of a new symptom checker, while also testing its diagnostic capabilities | Diagnostic performance of HDDO was compared to 6 other symptom checkers | HDDO outperformed all Symptom Checkers in terms of listing the correct diagnosis in the Top 1, 3, and 20 diagnoses.<br>Mean Reciprocal Rank@1: 0.024 versus 0.125<br>Mean Reciprocal Rank@3: 0.074 versus 0.346<br>Mean Reciprocal Rank@20: 0.171 versus 0.527<br>HDDO outperforms other symptom checkers in diagnostic capabilities. |
| 18.Maier et al., 2015. Germany | Dermatological issues | 1 dermatological app: SkinVision | 195 melanocytic lesions | Evaluate the diagnostic performance of SkinVision | Compare the output of SkinVision to the diagnoses given by 2 dermatologists | SkinVision scored: Sensitivity 73%, the specificity 83%, overall accuracy 81%. The dermatologists outperformed SkinVision with a sensitivity of 88%, specificity 97%, and overall accuracy of 95%. |
| 20.Ngoo et al., 2017. Australia | Dermatological issues | 3 dermatological apps: SkinVision, SpotMole, Dr. Mole. | 57 pigmented lesions, of which 42 are clinically suspicious | Establish the extent to which dermatological apps can identify benign and high-risk lesions | Compare the output of the dermatological apps with the same input, and assess their overlap with diagnoses given by 2 dermatologists | The sensitivity of the Melanoma Apps ranged from 21.4% (Dr. Mole)-72.2%(SkinVision) and specificity from 27.3%(SkinVision) -100.0% (Dr. Mole). Inter-rater agreement between dermatologist and app was poor (Kappa=-0.01 SE 0.16; p=0.97) to slight (Kappa 0.16 SE 0.09; p=0.12) |
| 23.Powley et al., 2016. UK | Inflammatory arthritis | 2 symptom checkers: NHS and Boots WebMD | 21 patients with inflammatory arthritis. Unclear gender assignment. | Assess the extent to which the symptom checkers can correctly diagnose patients with inflammatory arthritis | Compare output of 2 symptom checkers to patients actual diagnoses | 4/21 patients with inflammatory arthritis were given the correct diagnosis on the first try. 69 % of RA patients and 75 % of PsA patients had their actual diagnosis listed amongst the top 5 differentials. |
| 25.Ruotsalo & Lipsanen, 2018. Finland | Unspecified/General | Interactive Symptom Elicitation (ISE) | 12 participants. 4 male, 8 female | Demonstrate a new symptom checker and assess its diagnostic capabilities | Participants entered details (symptoms) about a medical they have previously suffered from in the past. Then, the output of the ISE was compared to the actual diagnosis | Mean Reciprocal Rank@1 to 20: 0.362-0519.<br>Confidence of correct diagnosis from MRR@1 to 20: 13%-65%. |

## Table 2 continued

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author, Year, Country** | **Illness Condition** | **Methodology** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 26.Semigran et al., 2015. USA | Unspecified/general | 23 symptom checkers, of which 15 have triage options | 45 patient vignettes (15 emergent cases, 15 non-emergent cases-15 self-care cases) | Assess the diagnostic and triage accuracy of symptom checkers | Compared the output of the symptom checkers with the same input. | Average Top 1 correct diagnosis given in 34% of cases, and correct diagnosis was listed in a Top 20 differential in 58% of cases. |
| 27.Shen et al., 2019. USA | Ophthalmic issues | 1 online symptom checker: WebMD | 42 patient vignettes (18 emergent cases, 24 non-emergent cases) | Assess the diagnostic and triage accuracy of WebMD for ophthalmic diagnoses | Compared the output of WebMD to real diagnoses. | Top 1 diagnosis was correct in 11/42 cases (26%); Top 3 in 16/42 cases (40%), and correct diagnosis was not included in 18/42 (43%) cases. |
| 28. Sole et al., 2010. USA | Unspecified/General | 1 online symptom checker: 24/7 WebMed | 1290 participants (30% male, 70% female) | Describe the initiation of a new symptom checker and assess the symptom checker's congruence between its diagnosis, the users chief complaint, and the diagnosis given by the Student Health Services (SHS). Also, observe the proportion of users who requested an appointment with the SHS as well as compliance to triage advice given by 24/7 WebMD | Assessed the rate of agreement between the symptom checkers diagnosis with the SHS's actual diagnosis. | Agreement between the chief complaint and 24/7 WebMed classification ($\kappa\kappa$= .94) agreement between chief complaint and diagnosis in SHS ($\kappa\kappa$ = .91) and agreement between 24/7WebMed classification and SHS diagnosis ($\kappa\kappa$ = .89). |
| 29. Thissen et al., 2017. Netherlands | Dermatological issues | 1 dermatological application: SkinVision | 341 lesions. 239 malignant lesions and 102 benign lesions | Evaluate the ability of SkinVision to accurately identify malignant and non-malignant melanoma | Compared the number of correctly classified lesions by the dermatological app to actual diagnosis of all lesions | Sensitivity 80%, Specificity 78%. High number of correct diagnoses and low number of false positives. |
| 30. Wolf et al., 2013. USA | Dermatological issues | 4 unnamed dermatological applications | 188 lesions. 60 melanoma cases and 128 benign lesions | Assess the performance of 4 dermatological applications to evaluate and rate pictures of skin lesions to their likelihood of malignancy | Compared the output of all applications analyzing the same input, and evaluating their accuracy | Sensitivity ranged from 6.8%-98.4%; Specificity ranged from 30.4%-93.7%. |

first outcome measure consisted of determining the probability to which the correct diagnosis will be returned by the symptom checker based on a *rank* (1; 4; 6; 8; 9; 12; 23; 25; 26; 27), the second type consisted examining the *sensitivity* and *specificity* of the symptom checkers differentials (*2; 3; 5; 7; 10; 18; 20; 29; 30*), and the third measured the *rate of agreement between the diagnosis and other common diagnostic methods* (7; 20; 28). The last measure consisted of establishing the *concordance of the diagnosis between the symptom checker and health professional*s (2; 3; 6; 9; 18; 20). Each of these methods will be elaborated on in their respective paragraphs.

*3.2.a Diagnosis based on rank*

This outcome measure is characterized by examining how often a diagnosis is correctly identified as either the first diagnosis (Top 1) or is included at all in a list of differentials (quantified as Top 20 or as ‘’the diagnosis being mentioned at all’’). The interpretation of a Top 1 result would indicate the capability of a symptom checker to immediately recognize a particular condition, while a Top 20 (being listed at all) result would list the correct diagnosis among a list of differentials, but would require further human interaction in order to extrapolate the correct diagnosis. For example, Powley et al. (2016) observed how 21 patients with arthritis interacted with the symptom checker, and that only 19% (n=4) of them had their condition correctly listed on the first try, which implies the symptom checker being insufficiently sensitive towards this condition.

Overall, the top 1 diagnosis matched the correct diagnosis between 2.4% - 36.2% (12; *25*) of all cases. The correct diagnosis was included in the list of overall differentials between 5.6% - 72% (1; 23) of all cases. We see that the results were highly variable across the different ranks and studies. However, if the results are assessed by grouping them according to their illness conditions, then variability increases even further. Among specific illness conditions, values for top 1 diagnosis ranged from 3%-26% (1;27), while the range for the diagnosis being listed at all was 5.6% -72% (1; 23). Among broad illness conditions, this varied from 2.4% -36.2% (12; *25*) for top 1 diagnoses and 17.1% -58% (12; 26) for being listed at all. It appears that symptom checkers assessing specific medical conditions achieved on average the highest probability for the correct diagnosis being included in a list of returned differentials, while symptom checkers assessing broad conditions have the highest chance of their first returned diagnosis matching the

correct one. The gap for Top 1 diagnoses among specific conditions is smaller than for broad conditions (23% vs 33.8%), but for diagnoses being listed at all the variations become smaller for the broad conditions (30% vs 66.4%). Also, only one study reported data in which symptom checkers were not able to generate a correct diagnosis included in a list of differentials at all, listed at 40% of all cases (27).

*3.2.b. Sensitivity and Specificity*

Another form of measurement consisted of those of *sensitivity* and *specificity* (studies n=9). Parikh et al. (2008) have described what these scores represent. Sensitivity scores display the rate at which people with a medical condition are being correctly diagnosed, while specificity refers to people without a disease being correctly identified for not being ill. Low sensitivity scores translate to a high number of false negatives; the number of people incorrectly being declared disease-free. Low specificity translates to a high number of false positives; people being diagnosed with a disease although they have none. The validity of diagnostic tests is typically measured via those two components (Palinkas et al., 2016) and many studies were identified in our review which use this blueprint.

The range for *sensitivity* is 6.8% - 100% (28;5) and the range for specificity is 23% - 100% (3; 18). However, caution is advised for interpreting these results. For example, within the same study, the lowest recorded sensitivity score was 6.8% but the highest sensitivity was 98.4% (30). Similarly, another study's highest specificity score found was 100%, but the lowest possible score was 27.3% (20). All studies using these measurements consisted of symptom checkers assessing specific illness conditions. The sensitivity scores according to the medical condition ranged 6.8%-98.4% among dermatological applications (18; 20; 29; 30), 89%-91% among knee pain checkers (2; 3), 88% for diabetes (10),and 71%-100% for mental health checkers (5; 7). Knee pain checkers have the slightest gap of sensitivity scores (2%) while dermatological ones have the largest gap (91.6%). Concerning the dermatological sensitivity scores, they also vary independently between the respective studies assessing multiple tools. In study (20), the lowest/highest score is 21.4%-72.2% (50.8% gap), while in study (30) it is 6.8%-98.4% (91.6% gap). Although both have considerable variations in sensitivity scores, study (20) cuts the variation almost in half compared to study (30).

The variation of *specificity* scores according to the symptom checkers function was 27.3%-100% (20; 30) among dermatological checkers, 44%-97% among mental health checkers (5; 7), 23%-27% (2; 3) among knee pain checkers, and 75% for diabetes. Again, knee pain symptom checkers have the lowest gap in specificity scores (4%) while dermatological checkers have the largest gap (72.7%). For dermatological checkers, the specificity scores also differ between the respective studies, similarly to our findings on sensitivity. Study (20) reported results ranging from 27.3%-100% (72.7% gap) while study (30) had results ranging from 30.4%-93.7% (63.3% gap).

*3.2.c. Agreement between symptom checker and other diagnostic methods*

The third major method consisted of comparing the diagnostic performance of the symptom checker to traditional methods and tools which had been used in the past to perform the same task. For example, Farvolden et al. (2003) examined psychiatric disorders by pitting their symptom checker (WB-DAT) against traditional paper-and-pencil tests, which were commonly used to test the same disorders.

Only a few studies have been found that assessed the degree to which differentials returned by symptom checkers are compared to other common diagnostic methods (7; 20; 28). Every study had a different comparator when they examined the performance of their symptom checker. Study (7) compared the performance of the WB-DAT to other traditional paper-and-pencil tests to screen for depressive-, and anxiety disorders. Here, the agreement rates between the traditional tests and the symptom checker ranged from *acceptable to good*. In study (20), the comparator consisted of dermatologists analyzing the same input which was entered into different dermatological symptom checkers. There, the agreement between dermatologists and the application was *poor to slight*. In study (28), the comparator was the diagnosis of student health services. The agreement between the symptom checkers classification, the user's chief complaint, and the SHS diagnosis was *good*.

*3.2.d. Concordance between diagnosis given by symptom checkers and health professionals*

Finally, some studies examined the performance of symptom checkers in parallel to those of doctors by comparing their diagnostic correspondence to each other (2; 3; 6; 9; 18; 20). Usually, this meant that doctors received the same input as the symptom checkers did, and that the doctors generated a diagnosis after the symptom checker did. In order to interpret their

performance, studies have used at least one of the three abovementioned outcome measures in their assessment. We decided to choose this category because the within included studies were the only ones that have pitted diagnostic tools against health professionals/physicians in the pursuit of comparing and assessing the performance of both.

3 studies have compared both doctors and diagnostic symptom checkers performance by assessing their sensitivity and specificity scores (2; 3; 18). The symptom checkers in 2 of these studies assessed knee pain and had sensitivity scores of 89%-91% and specificity scores of 27%-23% respectively (2; 3). The third study attempts to diagnose malignant melanomas and showed sensitivity, specificity, and total accuracy scores of 73%, 83%, and 81% for the symptom checker, and 88%, 97%, and 95% for health professionals respectively (18). 2 other studies have examined the diagnostic overlap in top 1 or final diagnoses between symptom checkers and health professionals (6; 9). Study (9) examining hand-pain stated that only 33% of the symptom checker's diagnoses match with the doctors final diagnosis, while study (6) examining ENT complaints stated that the symptom checker's first diagnosis matches the doctors diagnosis 16% of the time, yet reached 70% when the differential list was expanded. The last study discovered poor inter-rater agreement between diagnoses made by health professionals and the symptom checkers in regards to diagnosing malignant melanomas (20).

In sum, although symptom checkers have the potential and occasionally prove fair to good rates of correctly identifying the correct diagnosis if they are allowed to return a list of differentials, they tend to have very poor to poor accuracy when only the first diagnosis was being considered. Some tools have excellent sensitivity and specificity scores, but others have very poor scores as well. The differences among said scores between different symptom checkers assessing the same medical condition can also be extremely variable. For example, although both dermatological and mental health symptom checkers reported the highest possible scores in sensitivity (100%) and specificity (100%) respectively, they have also proven to have the largest variations in those scores as well. There are some indications that the symptom checkers are capable of reliably identifying the medical condition in question, but their accuracy remains inferior to health professionals.

### 3.3. Accuracy of Triage Tools

The second category of tools refer to *symptom checkers with triage functions*. As opposed to diagnostic symptom checkers which attempt to provide a diagnosis, a *symptom checker with triage functions* guide the user to the most appropriate source of help for their health concern (19). These recommendations are typically referred to as *levels of triage urgency*, which pertain to the severity of the health concern and urgency in which medical treatment or attention should be sought. 8 studies examined triage accuracy (**Table 3**), where 4 contained solely *symptom checkers with triage functions* (11; 19; 22; 24), while the other 4 reported the use of *diagnostic symptom checkers* which also possessed triage functions. This reiterates the point that a diagnostic tool can be assigned to more than one type of function. 4 studies examined specific illness conditions (1; 23; 24; 27), while the remaining ones covered a broad area of possible conditions (11; 19; 22; 26). The outcome measures of most studies mainly consisted of analyzing the extent to which triage recommendations by symptom checkers reflected the actual *level of triage urgency* (1; 11; 19; 22; 26; 27). Lastly, 2 studies used *other methods* to measure triage accuracy (23; 24). Both will be explored below.

*3.3.a Triage accuracy for emergent-, non-emergent-, and self-care cases*

3 levels of triage urgency are frequently used, namely *emergent care* (in which immediate/emergent medical attention is required), *non-emergent care* (medical care is advised but not urgently), and *self-care* (professional medical care may be not required at all). The levels of triage urgency are usually represented in patient vignettes, where a patient vignette can be understood as the profile of an individual containing full descriptions of his medical condition(s), including his actual triage level. The symptom checker would receive the vignette/input, and then the researchers would compare its output to the actual triage level.

4 studies used the 3 abovementioned triage levels (11; 19; 22; 26), and 2 studies only emergent and non-emergent levels in their assessment (1; 27). However, some issues need to be clarified before the presentation of their results. 2 studies that used 3 triage levels did not specify the data for each specific level, but rather presented the appropriateness of triage advice given in general (19) or just for emergent cases (22). Study (19) reports that their symptom checker outperforms doctors and nurses overall (88.2% vs 75.5% vs 73.5% in all cases). Study (22) used

## Table 3: Assessment of studies regarding triage accuracy

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author, Year, Country** | **Illness Condition** | **Methodology** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 1.Berry et al., 2019. USA | HIV, Hepatitis C | 3 online symptom checkers: Symptomate, Symcat, Isabel | 90 patients with HIV, 67 patients with hepatitis C, 11 patients with both HIV and hepatitis C. | Analyze diagnostic and triage accuracy of a series of symptom checkers. | Compare output of different symptom checkers with the same input | 35.6% of HIV cases, 59.7% of hepatitis c, and 45.5% for both cases together were declared emergent by the symptom checkers.<br>Triage assessment is inferior to those of ED doctors. |
| 11.Kafle et al., 2018. Canada | Unspecified, General | 1 Symptom Checker: Knowledge Base (KB) | 45 patient vignettes (15 emergent cases, 15 non-emergent cases- 15 self-care cases) | Describe the creation and processing capabilities of a new personalized symptom checker, while also testing its diagnostic capabilities | Output of KB was measured in order to observe whether it could rank 45 patient vignettes according to three degrees of urgency | 20% for both emergent and non-emergent care cases, and 50% of self-care cases were correctly labeled as Top 1 diagnosis.<br>53.33% emergent, 66.67% non-emergent, and 78.57% self-care cases were correctly included in a Top 20 list of differentials.<br>Self-care cases have an acceptable to good range of being correctly labelled by the symptom checker, while emergent to non-emergent cases range from poor to acceptable |
| 19.Middleton et al., 2016. UK | Unspecified, General | 1 online symptom checker: Babylon Check | 102 patient vignettes, and unspecified number of actors to simulate symptoms of each vignette | Describe the development of the symptom checker Babylon Check and to test its diagnostic performance | 12 clinicians and 17 nurses independently diagnosed the actors simulating a patient vignette while Babylon Check attempted to diagnose each vignette. Then, the diagnostic overlap between the clinicians, nurses, and Babylon check was assessed | Babylon check outperforms both doctors and nurses: an accurate outcome) is produced in 88.2% of cases for Babylon check, in 75.5% of cases for doctors, and in 73.5% of cases for nurses. |
| 22.Poote et al., 2014. UK | Unspecified, General | Online Self-assessment system, unspecified name | 154 participants. 55 male, 99 female | Evaluate the triage advice given by the self-assessment system by rating congruence with diagnoses given by GP's. Also assess whether the level of agreement can be influenced by gender, age and nature of symptom | Triage output by the system was compared to triage advice given by 7 GP's. | Perfect agreement between system and GP's occurred in 39% of consultations.<br>Advice for more urgent level of care seeking was recommended in 86 consultations (56%) by the system, opposed to the 72 consultations (47%) by the GP's.<br>No significant association between the advice given by the self-assessment system and participants age, gender, and nature of symptom was found. |

## Table 3 continued

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author, Year, Country** | **Illness Condition** | **Methodology** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 23.Powley et al., 2016. UK | Inflammatory arthritis | 2 symptom checkers: NHS and Boots WebMD | 21 patients with inflammatory arthritis. Unclear gender assignment. | Assess the extent to which the symptom checkers can correctly diagnose patients with inflammatory arthritis | Compare output of 2 symptom checkers to patients actual diagnoses | 56 % of patients were given appropriate triage advice. |
| 24. Price et al., 2013. USA | Influenza-like illness | 1 web-based program: Strategy for Off-site Rapid Triage (SORT for Kids) | 294 participants. 152 male, 142 female | Assess the usability and triage accuracy given by SORT to diagnose influenza-like illnesses | Prior to completion of ED visit, participants filled out the SORT program. Then, the output of SORT was compared to the actual number of clinically necessary and non-necessary ED visits. | 14/15 ED visits were correctly identified as clinically necessary, while 28/271 of non-necessary visits were incorrectly labelled as necessary. Resulting in sensitivity 93.3% and specificity 12.9%. SORT provides many correct advices, but also many false positives. |
| 26.Semigran et al., 2015. USA | Unspecified/general | 23 symptom checkers, of which 15 have triage options | 45 patient vignettes (15 emergent cases, 15 non-emergent cases- 15 self-care cases) | Assess the diagnostic and triage accuracy of symptom checkers | Compared the output of the symptom checkers with the same input. | Triage advice was correct for 80% of emergent cases, 55% of non-emergent ones, and 33% for self-care cases. Appropriate triage advice was higher for uncommon diagnoses than for common diagnoses: 63% vs 52%. Correct triage advice varies across different levels of urgency and rarity of disease. |
| 27.Shen et al., 2019. USA | Ophthalmic issues | 1 online symptom checker: WebMD | 42 patient vignettes (18 emergent cases, 24 non-emergent cases) | Assess the diagnostic and triage accuracy of WebMD for ophthalmic diagnoses | Compared the output of WebMD to real diagnoses. | Triage urgency based on the top diagnosis was appropriate in 7/18 (39%) emergent cases and 21/ 24 (87.5%) of nonemergent cases. WebMD provides less accurate triage advice for emergent ophthalmic issues than for non-emergent ones. |

3 levels but only reported on the difference of recommended urgent care between doctors and the symptom checker (47% vs 56%).

Overall, appropriate *emergent* triage advice ranged between 20%-93.3% (11; 23), non-emergent cases are being correctly identified between 20% (11) and 87.5% (27) of times, and self-care cases vary between 33%-78.57% (26; 11). Results indicate that *emergent* triage recommendations tend to have the highest probabilities of being correctly recognized than non-emergent and self-care cases. However, one study (11) shows the inverse trend, where larger proportions of self-care cases tend to be more correct over non-emergent and emergent ones (78.57% vs 66.7% vs53.3%), while another study (27) showed higher accuracy for non-emergent cases than emergent ones (87.5% vs 39%). Only 2 studies showed good to very good rates of properly identifying emergent cases (23; 26), but only one study outperformed doctors in a clinical setting in this entire review (19). Nonetheless, these studies prove the exception to the norm as the remaining studies and reported triage levels tend to provide poor and unreliable rates of triage accuracy as well as high variability.

*3.3.b Other methods for triage accuracy*

Only two studies used a different method in order to appraise triage accuracy. 1 study assessing influenza-like illnesses among children reported that 14/15 emergency department visits were deemed clinically necessary while 28/271 of non-necessary visits were incorrectly labelled as necessary (24). The results indicate a near perfect sensitivity towards detecting sick patients, although the sample size is low. At the same time, a large number of false positives provide insights into a low specificity from the symptom checker. The remaining study (23) assessing inflammatory arthritis observed that 56% of the 21 patients suffering from the condition actually received the correct triage advice. Not only does the low sample size prevent a sound generalization from the results to the medical condition, but its accuracy is also low.

In summary, emergent cases tend to be correctly identified at an increased frequency than non-emergent and self-care cases, but several exceptions weaken the strength to this claim. Symptom checkers with triage options are often said to be good starting points for patients with health concerns, but still perform inferiorly to doctors. Overall, their results tend to be inaccurate and also extremely variable. Only 1 study (19) showed a superior performance by the symptom checker versus doctors and nurses alike.

### 3.4. Estimates given by risk calculators for developing X disorder

There are very few studies examining diagnostic tools-, or so called risk calculators, with the ability to provide risk estimations for developing specific disorders in the future (**Table 4).** The risk estimations usually follow a timeframe of developing a disease in the next 5 years, 10 years, 30 years, or in the entire lifetime. 2 studies use a variety of timeframes in order to observe differences in the risk of developing diseases (14; 17), while 3 studies focus solely on lifetime risk (10; 16) or 10 year risk (15). One study analyzed the results of 13 breast cancer risk calculator websites (14), another focused on 10 websites for CVD (15), and the diabetes study focused on one online self-screening tool (10). 2 studies solely focused on one risk calculator website for arthritis (16; 17).

*For breast cancer*, the 30-year risk for developing it varies between 3%-6.2% for low risk subjects, and 35%-64% for high risk subjects. However, the study used only 4 patient vignettes containing the profiles of one female patient each. *For CVD*, the 10-year risk varies between the different risk calculator websites: 3% to 25% for the profile of a 55-year old man, 0% to 4% for a45-year old woman). *For arthritis*, study (17) appraised changes in both 10-year risk and lifetime risk for OA before and after the use of the OA RISK C. The 10-year risk has dropped from 25.4% to 12.5% and from 47.6% to 28.1% in lifetime risk after the use of OA RISK C by the users. However, OA RISK C has calculated the average 10-year and lifetime risk at 3.6% and 25.3% respectively, indicating that participants tended to overestimate their risk. Those findings are similar to study (16), in which the average lifetime risk estimate given by the symptom checker was 25%, while users tended to overestimate their risk by up to 38%. *The last study* examined the risk estimation given by the risk calculator by using the medical profiles of the participants a priori to their diagnosis and compared the results to the actual diagnoses (10). This study is unique in regards of not using a timeframe for the probability of developing diabetes. The disease outcomes are already known, and the study simply observed the agreement between the risk calculators' diagnoses and the actual outcomes. The sensitivity and specificity scores for developing diabetes among the results were 88% and 75% respectively, indicating fairly good rates of accuracy for diagnosing diabetes.

In summary, it is difficult to interpret the risk estimations given by said risk calculators, as only one study compared risk estimations to actual disease incidence rates (10). All the other

## Table 4: Assessment of studies regarding risk estimations for developing certain disorders across different symptom checkers

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author, Year, Country** | **Illness Condition** | **Methodology** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 10.Heikes et al., 2007. **USA** | Diabetes and pre-diabetes | 1 online self-screening tool: DIABETES RISK CALCULATOR | 30,818 participants from NHANES III dataset. Unclear gender assignment. | Develop an online tool accessible to the general population for calculating their probability of undiagnosed diabetes or pre-diabetes, and demonstrate its diagnostic accuracy | 3 separate online tools were created in the process, in which DIABETES RISK CALCULATOR had the highest accuracy among them. Then, said tool was tested by receiving input from the NAHNES III dataset, which includes patient information from 30818 participants. Lastly, DIABETES RISK CALCULATOR output was compared to the patient's actual diagnosis. | Sensitivity and specificity for undiagnosed diabetes: 88% and 75% respectively<br>Sensitivity and specificity for pre-diabetes and undiagnosed diabetes: 75% and 65% respectively. |
| 14. Levy et al., 2007. USA | Breast cancer | 13 breast cancer risk calculator websites | 4 patient vignettes, all female, aged 21, 26, 42, and 48 respectively | Observe the range of breast cancer risk estimations given by various breast cancer risk calculator websites | Each breast cancer website received the same input. Then, their output was compared. | 30-year risk of breast cancer development among low risk subjects varied between 3%-6.2% across risk calculator websites, while high risk subjects had a variation between 35%-64%. 30-year risk of breast cancer development varied in 1 subject between 3.3%-30%. |
| 15.Lippi & Sanchis-Gomar (2018). USA | Cardiovascular disease (CVD) | 10 CVD risk calculator websites | 2 patient vignettes: a 55-year man with an intermediate CVD risk and a 45-year woman with a low risk | Observe the range of CVD risk estimations among and between popular internet-based CVD risk calculators | Each CVD website received the same patient vignettes and their output was compared. | 10-year CVD risk of the 55-year old man varied from 3% to over 25%, whereas that of the 45-year women varied between 0% and 4%. |
| 16. Losina et al., 2015. USA | Knee osteoarthritis (OA) | OA Risk C; risk calculator | 45 patients 6 male, 39 female | Describe the development of a novel risk calculator for knee osteoarthritis, demonstrate its capabilities to generate risk estimations, and to compare said risk estimations to those made by participants trying to guess their own risk | Participants completed the OA Risk C without knowing the risk estimations given by the program. Then, participants were instructed to give an estimation about their own risk for developing OA based on the information they have seen while using the program. | The average lifetime risk estimate by users was 38 %, while average lifetime risk determined by the calculator was 25 %. |

## Table 4 continued

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author, Year, Country** | **Illness Condition** | **Methodology** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 17.Losina et al., 2017. USA | OA | OA Risk C; risk calculator | 375 participants 195 male, 180 female | Measure changes in risk perception as well as their willingness to change risk behaviors related to OA before and after using OA RISK C | Both intervention and control group participants estimated their own OA risk prior to the trial. The intervention group viewed general information about OA and used OA Risk C while the control group only viewed general information on OA.<br>After the trial, they entered their own OA risk estimate again. Also, both group's willingness to change their behaviors before and after the intervention was measured. | OA Risk C risk estimation for both groups were on average: 3.6% for 10 years and 25.3% for lifetime risk.<br>Control group risk estimation changes before and after intervention: 26.9%-27.7% for 10 years; 48%-47.1% for lifetime.<br>Intervention group risk estimation changes before and after intervention: 25.4%-12.5% for 10 years; 47.6%-28.1% for lifetime. |

studies merely displayed the ability of listing risk estimations over specific time periods. Although it is useful to know that there are many risk calculator websites for breast cancer and CVD, the range of medical conditions as well as actual measures of accuracy in this field of literature are abysmal.

### 3.5. Behavioral actions and changes after the use of diagnostic tools

The last category pertains to the behaviors of people after the use of diagnostic tools and the changes in behaviors, attitudes, or actions that have been observed. Only 5 studies (**Table 5)** have been found that fit in this category. In terms of the types of diagnostic tools, we found 2 symptom checkers with triage functions (21; 28), 2 diagnostic symptom checkers (13; 31), and one risk calculator (17). Among those studies, only one (28) had measures of accuracy included. Only a few medical conditions are explicitly covered here, namely chlamydia, arthritis, and psychiatric disorders, while 2 studies have examined a broad but unspecified set of medical conditions in their assessment (21; 28). 2 main outcome measures have been identified, namely *adherence & compliance to the medical advice* and *changes in behavioral intentions*, which will be explored below.

#### *3.5.a. Adherence & compliance to the medical advice*

This outcome measure entails how many people actually complied or adhered to the medical advice or diagnosis given by a diagnostic tool. 4 studies fall under this category (13; 21; 28; 31). 1 study examined the proportion of participants who went to get themselves tested for chlamydia after using the symptom checker, as well as those who sought medical treatment after a positive result (*13*). 56% of the participants decided to get themselves tested after using the checker, of which 18% were tested positively for chlamydia. 50% of those who were tested positively actually underwent medical treatment within 7 days of confirmation. Another study measured the user's intentions to visit a GP for their complaint, and those who complied to triage recommendations given by the symptom checker (21). 73 patients (38%) intended to visit a GP for their complaint, while 20 patients (57%) of the follow-up actually followed the triage recommendations. The third study examined the number of requested appointments to student health services and actual visits after symptom checker use (28). One additional measurement in this study was compliance to triage advice after receiving a recommendation to seek help urgently. 143/1290 (11.08%) of participants who used the triage system requested an

## Table 5: Assessment of studies regarding behavioral actions following symptom checker use:

| | Characteristics of the tool | | Characteristics of Study | | | |
|---|---|---|---|---|---|---|
| **Author, Year, Country** | **Illness Condition** | **Tool/Program/Website used** | **Sample Size** | **Aim/ Main Objective(s)** | **Methodology/Study Design** | **Results** |
| 13. Kwan et al., 2012. Australia. | Chlamydia | OLC | 675 participants. 206 male, 171 female | Assess the number of pathology form downloads, confirmed diagnoses of chlamydia, and proportion of participants who underwent clinical control upon positive confirmation of chlamydia | Observed the number of people who got themselves tested after downloading the pathology form. Afterwards, examined the proportion of positively diagnosed participants which have undergone clinical treatment | 56% of pathology form downloads resulted in a test. Of those, 18% were diagnosed with chlamydia. 50% of those who had a positive result underwent clinical control within 7 days of confirmation. |
| 17.Losina et al., 2017. USA | OA | OA Risk C | 375 participants. 195 male, 180 female | Measure changes in risk perception as well as their willingness to change risk behaviors related to OA before and after using OA RISK C | Both intervention and control group participants estimated their own OA risk prior to the trial. The intervention group viewed general information about OA and used OA Risk C while the control group only viewed general information on OA.<br>After the trial, they entered their own OA risk estimate again. Also, both group's willingness to change their behaviors before and after the intervention was measured. | Both groups reported no changes in their willingness to move into an action phase regarding weight and diet controls. The intervention group reported an increase in participants willingness for exercise compared to the control group: 26.9% vs 13.6%. |
| 21**.** Nijland et al., 2010. Netherlands | Unspecified/General | 1 online triage program: http://www.dokterdokter.nl | 13133 people having accessed the program. 192 patients having completed the follow-up survey (65 male, 127 female).<br>35 who completed the follow-up questionnaire. | Assess the participants use of the online triage program, as well as their intention to visit a GP, and their compliance to the advice given by the program | Observed the total number of participants having entered a complaint and those who completed the triage process. Then, participants who have completed the optional survey at the end of the triage process were assessed in which they expressed their intentions to visit a GP as well as the reasons why they used the triage system.<br>Lastly, a follow up questionnaire on actual compliance was completed by a further subset of participants. Here, a regression analysis was performed in order to | Out of the 13133 participants, 6538 have entered a complaint and 3812 completed the triage process.<br>Out of 192 patients, 73 patients (38%) intended to visit a GP for their complaint prior to triage. Common reasons for using the triage system included gathering information about a health complaint (n=104; 38%) and deciding whether it would be necessary to contact a GP (n=72; 20%).<br>A follow-up questionnaire on actual compliance was completed by 35 patients. Among these, 20 (57%) had complied with the advice provided by the system. The regression analysis revealed a strong relation between intention to comply and actual compliance, while intention to comply was also strongly related to the |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | observe which factors could influence compliance to the advice given by the system | attitude towards the advice. In turn, attitude towards the advice was shaped by perceived effectiveness of the delivered advice and trust in the web-based triage. |
| 28. Sole et al., 2010. USA | Unspecified/General | 24/7 WebMed | 1290 participants (30% male, 70% female) | Describe the initiation of a new symptom checker in a college health setting, assess it's congruence between its diagnosis, the users chief complaint, and the diagnosis given by the Student Health Services (SHS). Also, observe the proportion of users who requested an appointment with the SHS as well as compliance to triage advice given. | Assessed the number of participants who used the system and the proportion of those who requested an appointment. Also, assessed the subset of participants who complied to the triage advice given. | Of the 1290 participants that used the triage system, 143 (11.08%) requested an appointment at SHS.<br>The system recommended to students who requested appointments electronically (59.3%) to seek care within 24 hours, of which 41.3% did (compliance). |
| 31.Van Ameringen et al., 2015. Canada | Psychiatric disorders | MACSCREEN | 770 participants, of which 103 completed the follow-up survey (24 male, 79 female). | Test hypothesis that access to a reliable and relevant self-report screening tool would encourage treatment-seeking behaviors among users | Prior to MACSCREEN completion, participants revealed the reasons to why they use the screening tool. After completion of MASCREEN, the study also measured the participants behavioral intentions for treatment seeking, their actual treatment seeking, and reasons for not seeking treatment. They also list reasons to which they will go seek treatment, and demographic factors influencing treatment seeking are also listed. | The most commonly mentioned reasons for using MASCREEN were: Concern for potentially having an anxiety problem (83.5%), a desire to confirm diagnosis given by a health professional (33%), and avoiding discomfort from talking to a health professional (34.4%).<br><br>Most frequently mentioned desired course of action include: seeking further help from a health professional (85.4%), looking for more health information online (34%), and talking with a family member (25.2%).<br><br>From this follow-up survey, 53/103 sought treatment while 50/103 didn't. Reasons for not seeking treatment include fear/lack of desire to take psychiatric medications (57%), sensing discomfort discussing their anxiety with a physician (28%), and anxiety not being severe enough to warrant treatment (28%).<br><br>Reasons for seeking treatment include: Symptoms need to get worse (44%), acquiring the financial means for medication or treatment (42%), and requiring to be convinced of treatment effectiveness (34%). |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | Disorder symptom severity is not suggested to not accurately influencing treatment seeking, but functional impairment and disease burden are. Being married also positively affects treatment seeking while being single has a higher likelihood of not engaging in treatment seeking.<br>Lastly, being female has a small but significant trend towards seeking treatment whereas no significant relation was found with educational attainment levels. |

appointment at the SHS. From this proportion, the triage system recommended 59.3% to seek care within 24 hours, of which 41.3% did. The last study (31) measured that 51% of participants (total n=103) sought treatment after using MACSCREEN, even though 85.4% indicated that they would seek treatment after MACSCREEN completion.

*3.5.b. Changes in behavioral intentions*

*Changes in behavioral intentions* refers to factors potentially influencing a person's willingness and attitudes towards taking actions for improving their health and/or seeking medical help. 3 studies have explored parts of this domain (17; 21; 31). Willingness to exercise-, which is considered as a health promoting activity, has increased among OA RISK C users to 26.9% (17). A strong relation was found between intention to comply and actual compliance, with intention to comply being also strongly related to the attitude towards the advice (21). In turn, attitude towards the advice was shaped by perceived effectiveness of the delivered advice and trust in the web-based triage. The last study measured the reasons listed by the participants for not seeking treatment even after having received a recommendation in doing so. Reasons such as symptoms requiring getting worse, as well as demographic factors such as being married, functional impairment and disease burden were linked being significantly related to treatment seeking, but not for educational attainment.

In sum, the results indicate that only a fraction of the users follow the triage advice, and that even fewer undergo clinical treatment in the case of confirmed diagnosis. The studies also show a severe lack in the literature examining behavioral intentions and actual treatment seeking after symptom checker use. Lastly, almost no studies attempt to draw connections between demographic factors and key behaviors influencing the use of symptom checkers, albeit an abundance of research doing so for other aspects health information seeking and e-health applications.

## 4.Discussion

This scoping review was conducted for three main purposes. We wanted to (1) examine which types of online diagnostic tools were available to the general public and describe their characteristics. We also wanted to (2) review the accuracy of such tools across the literature. Lastly, we were interested in (3) the behavioral impact said tools can have on its users. In doing so, we have mapped both the extent to which the scientific literature is known in that field, as

well as all appraised diagnostic tools. We have described the characteristics and different functionalities of all tools, the results on their diagnostic and predictive abilities, what users have done with the information they received from them, and listed the changes in behaviors or attitudes concerning adhering to the information they received and actually seeking medical help.

Diagnostic tools can be differentiated in terms of the terminology used to describe them (e.g. self-screener) and according to the function(s) (e.g. diagnostic symptom checker) they serve. However, despite the fact that every diagnostic tool we have examined can be classified as a symptom checker, a lot of confusion could occur when different designations and terms are being used to refer to their use. This resulted in our efforts of mapping and extracting evidence related to symptom checkers being very difficult. Search terms such as ‘’screeners’ or ‘’checkers’’ could either yield very few or very many search results, with the majority of them not even referring to diagnostic performance tests, which was evident during our article selection process. Additionally, the findings indicate a prospective future for knowledge models with sophisticated algorithms, capable of autonomously complementing and expanding their database based on the previous interactions with user-entered symptoms in order to facilitate future diagnoses. The potential ramifications of sophisticated algorithms could be immeasurable for the health care model, as they could theoretically achieve (near) perfect accuracy scores over time. Thus, their progress and advancements should be tracked in the future. Nonetheless, many studies in this review have failed to mention the algorithms of their diagnostic tool, their validation status, the number of possible detectable conditions, and only one study has actually listed the frequency in which the diagnostic tool was not able to provide a diagnosis at all (Shen et al., 2019). While the provision of these technical information may not change the viability of such tools in terms of accuracy, they would allow for more concise observations of their technical prowess and limitations as well as facilitate our efforts in grasping their full potential.

There are few studies which attempted to identify the number of health applications for self-diagnostic purposes available in the market, with even fewer ones examining their characteristics and accuracy. This review has identified an abundance of diagnostic tools and related studies tailored towards the general public. Although, 46/80 diagnostic tools were examined in just 3 studies (14; 15; 26), while very few tools have been examined across multiple studies and medical conditions alike. These disproportions highlight how some tools have

detailed accounts of their accuracy and have been tested on multiple medical conditions, while others are only explored on a surface level, rendering the generalizability of the results difficult. Furthermore, while we were able to report under which circumstances a specific diagnostic tool or a general/broad diagnostic tool tend to have better rates of accuracy as well as denote differences in accuracy within the same medical condition, we couldn't find or make any claims concerning both types of tools attempting to diagnose the same medical condition. This is due to the absence of studies having assessed the same medical condition across the different types of diagnostic tools and also some known broad diagnostic tools not having been tested on multiple conditions consistently. This is unfortunate because it could have provided us some insights into what type of tool and which tools exactly tend to perform better for which medical conditions. Other potential ramifications could include which type of tools would be safer or more reliable to use by the general public. Future studies should start filling these gaps by testing broad diagnostic tools to the medical conditions that have been already assessed among studies appraising specific tools. Only then can we make further definitive and comparative accounts on the accuracy and reliability of diagnostic tools.

Our overall assessment encompassing diagnostic and triage accuracy is that their accuracy tends to be poor and extremely variable. Not only does their overall accuracy tend to be low, but their diagnostic performance also varies drastically between different medical conditions as well as in those assessing the same condition. However, the accuracy of most symptom checkers dramatically increases whenever they are allowed to provide multiple diagnoses and the correct diagnosis is included in that list. A key problem is that even though symptom checkers are proficient in including the correct diagnosis in the differential process, that the extraction of the correct diagnosis depends on the user himself and is rarely expanded upon. Among triage and diagnostic accuracy studies, only one reported data on users trying to choose the correct diagnosis (Bisson et al., 2016), where the users frequently chose the wrong diagnosis. These findings are very intriguing because not only do symptom checkers tend to be inaccurate, but users also have difficulties to choose the correct diagnosis even among sound symptom checkers. Thus, future studies should also at least assess the accuracy in which the user is able to extract the correct diagnosis from said lists. Also, we also want to point out that not diagnostic tools perform poorly. There are a couple of studies which can boast with acceptable to good rates of accuracy (2; 3; 5;7; 10; 18; 19; 29), and if the results were to be interpreted in

isolation (not taking into account how accurate other traditional diagnostic methods are), then our overall judgement of diagnostic tools would have been more favorable. That being said, the larger picture needs to be taken into consideration because even the good results are inferior to the source of comparison in their respective studies.

Almost every diagnostic tool performed inferiorly when compared to doctors. Some exceptions were found where studies reported acceptable to good accuracy rates, and one symptom checker even outperforming doctors and nurses for providing appropriate triage advice (19). Nonetheless, with the latter study being the exception, the overall trend indicates that the diagnoses and advice on recommended triage care given by doctors remain the more consistent, accurate, and safer option. However, there needs to be more clarity in terms of what acceptable diagnosis rates constitute for both doctors and symptom checkers. While the performance of symptom checkers is frequently compared to diagnoses given by doctors, it would be wrong to assume that doctors are always accurate. Indeed, overdiagnoses are common across the entire medical field (Jenniskens et al., 2017). Diagnostic errors occurred in about 12 million patients alone in the US (Singh et al., 2014), and the gravity of such errors can range from insignificant to major (Schiff et al., 2009). This is why some studies indicate that interpretations of symptom checkers accuracy are incomplete because the comparisons between both parties tend to be inconsistent, lack a standardized evaluation method (Morita et al., 2017), and thus cannot be held to the same standard. Nonetheless, until these weaknesses are mediated in future studies, patients/users should use both diagnostic symptom checkers and those with triage functions with a lot of caution and rely on health information given by health professionals.

Concerning risk calculators, only few studies were found which appraised and described their potential predictive abilities. Even then, not much can be said about their accuracy as only one study compared risk estimations to actual disease incidence rates (10). Most studies merely reported the risk estimations given by the risk calculators without linking them to the actual incidences. Another interesting observation is that users tended to overestimate their lifetime risk for developing a certain disorder when compared to prognoses given by the risk calculators. So not only do users tend to have difficulties choosing the correct diagnosis, it also seems that they overestimate their own risk or vulnerability to a health concern. This exhibits a noticeable trend where users tend to have difficulties interacting with health information online. It is essential to

conduct studies where diagnoses given by symptom checkers are compared to the actual incidence rates over time; only then can we truly interpret their predictive abilities and their reliability. Until then, risk calculators should not substitute the opinion and appraisal of a physician, but rather serve as a supplementary source of information due to their lack of actual diagnostic abilities.

During our assessment of diagnostic tools, we have found several outcome measures used for accuracy. Yet, the study designs did not inherently differ significantly between the different types of studies examining said tools. Every study that assessed accuracy compared the performance of the symptom checker at question to one or multiple comparators. While outcome measures such as sensitivity and specificity and range of agreement with the diagnosis given by doctors are very suitable measures for accuracy, we do have some qualms about rank-based measures. While Top-1 measures are also indicative of accuracy (did the diagnostic tool provide the correct diagnosis immediately, Yes/No), a ‘’the correct diagnosis included in the list of differentials’’ measure is imprecise because the correct diagnosis has never been chosen by the tool at all. For example, if the tool lists 20 potential diagnoses, the user can choose the wrong diagnosis up to 19 times. Its only redeeming quality consists of observing whether users can extract the correct diagnosis, but we have already made this point earlier. Ultimately, we do not recommend the use of rank based measures, but their weaknesses can be mediated if study designs are tailored towards them. Further inconsistencies include some studies not listing all results despite having explicitly stated all outcome measures used. For example, 2 studies have mentioned the use of 3 triage levels in their assessment, but only provided data for either emergent cases (Poote et al., 2014) or triage appropriateness in general (Middleton et al., 2016). Diligent data reporting is required as this field of study still remains underexplored compared to other e-health fields, and where our knowledge is limited. Considering that there are many studies which have demonstrated the accuracy of various diagnostic tools, and it being highly probable that their accuracy will only improve slowly over the next years, efforts should be consolidated towards maximizing the collection of all additional data instead. This includes the technical characteristics of diagnostic tools, more rigorous examinations in terms of overlapping medical conditions across different types of tools, conscientious usage of outcome measures, and additional user-centric observations (choosing the correct diagnosis among a list). This would

allow pathways of recognizing limitations and weaknesses of diagnostic tools to become more manageable, while also accelerating efforts in modernizing and improving existing tools.

Little evidence was found in regards to the behavioral impact of symptom checkers on users. This was surprising due to the fact that not only does the literature have an abundance of studies assessing the number of available medical applications in the world as well as their download rates (IQVIA Institute, 2017), the user's interest in searching the internet for health information (Berry, 2018), and demographic variables influencing the use of m-health resources (Yun et al., 2017), not much is known about what users do after consulting with a symptom checker. Of the few studies that did, we discovered how about half of most target populations actually comply with the advice given by symptom checkers. While one study was capable of providing an array of potentially interrelated determinants influencing each other for triage compliance, changes in behavioral actions or attitudes upon symptom checker use remains underexplored. There are separate studies outside of this review, such as the one by Luger and colleagues (2014), which examined how users interpret their own symptom and how they navigate the Internet for self-diagnosis, or the study by Lupton & Jutel (2015) investigating the potential doctor-patient relationship after having consulted the Internet. While those studies provide insights into other key aspects on the use of (online) diagnostic tools, we're still missing data to the most pertinent aspects to health outcomes such as triage compliance, actual treatment seeking behavior, and intentions(as well as actually moving to an action phase!) to change health averse behaviors. Thus, studies examining symptom checkers need to increase their efforts in measuring and reporting baseline attitudes and actions taken by users in order to expand our understanding of the consequences of symptom checkers on the users.

**Strengths and limitations**

To our knowledge, our scoping review provides the most exhaustive overview on symptom checkers available to the general public up to date. We have established the classifications and functionalities of symptom checkers, mapped the medical conditions they have been assessed on, and presented the most prominent studies in that domain. We also pointed out several weaknesses related to research practices within the studies, such as lacking descriptions or mentions of the symptom checker's validity or algorithmic functions. Lastly, this

review attempts to provide a starting point for future studies as evidence in this domain of study is scarce and difficult to establish.

There are a few limitations which need to be acknowledged. First, we do not exclude the possibility that some articles might have been missed because our search terms might have not captured all relevant studies. This is partially due to the numerous terminologies used to describe symptom checkers, but also possibly due to our strict inclusion criteria. Our inclusion criteria necessitated that studies needed to mention the availability of symptom checkers to the general public, or at the very least indicate that the symptom checker at question could also be used by said population. One might create the argument that studies not indicating the targeted user-group could instantiate a flaw in the study design and would absolve us from this weakness. Nevertheless, we could have missed some studies due to their availability not being clearly stated. Second, the list of symptom checkers and included studies is not representative of all available online diagnostic tools. For example, one study evaluated an online oral health and risk assessment tool designed for the use of dentists (Busby et al., 2013), and another examined symptom checkers that send input from users to real doctors in order to generate a diagnosis (Meyer et al., 2016). Those types of studies were excluded as they did not pertain to our research questions, however, we might have missed further relevant studies due to this choice. Third, despite having attempted to counteract such weaknesses by hand searching further articles under relevant studies, some articles were not able to be extracted due to lack of access. We have examined the few existing studies which appraised accuracy such as those mentioned in Millenson et al. (2018) but were unable to access some despite having contacted the respective authors, thus lowering the completeness of our review.

## 5. Conclusion

The potential that symptom checkers hold for alleviating the burdens of the global health care system in terms of reducing costs, unnecessary hospitalizations, or patient anxiety are enormous. However, these are currently completely out shadowed by generally low and variable rates of accuracy and inconsistent research methodologies. Despite some symptom checkers having excellent scores and even outperforming doctors in one instance, more resources need to be invested in increasing their accuracy to warrant their use as reliable tools. The understanding of the behavioral impact of symptom checkers on users is vastly underdeveloped and

underexplored. Future studies would need to start with thorough and rigorous descriptions of applied research methodologies, the characteristics of the symptom checkers, and optimize outcome measures for accuracy, such as measuring the accuracy of user-selected diagnoses and expanding the assessment of different medical conditions.

**REFERENCES**

(References with an Asterix(*) and a number were part of the review)


Aboueid, S., Liu, R. H., Desta, B. N., Chaurasia, A., & Ebrahim, S. (2019). The Use of Artificially Intelligent Self-Diagnosing Digital Platforms by the General Public: Scoping Review. *JMIR Medical Informatics*, *7*(2). doi: 10.2196/13445

Administration for Community Living (2018). Profile of Older Americans. *Retrieved from:* https://acl.gov/aging-and-disability-in-america/data-and-research/profile-older-americans

App Annie. The State of Mobile 2019. Retrieved from: https://www.appannie.com/de/insights/market-data/the-state-of-mobile-2019/

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), p.19-32. doi:10.1080/1364557032000119616

Bagaric, B., & Jokic-Begic, N. (2019). Cyberchondria – Health Anxiety Related to Internet Searching. *Socijalna Psihijatrija, 47*(1), p.28-50. doi:10.24869/spsih.2019.28

Baker, L., Wagner, T. H., Singer, S., & Bundorf, M. K. (2003). Use of the Internet and E-mail for Health Care Information. *Jama, 289*(18), p.2400-2406. doi:10.1001/jama.289.18.2400

Bender, J. L., Yue, R. Y. K., To, M. J., Deacken, L., & Jadad, A. R. (2013). A Lot of Action, But Not in the Right Direction: Systematic Review and Content Analysis of Smartphone Applications for the Prevention, Detection, and Management of Cancer. *Journal of Medical Internet Research*, *15*(12). doi: 10.2196/jmir.2661

*1. Berry, A., Cash, B., Wang, B., Mulekar, M., Haneghan, A. V., Yuquimpo, K., Swaney, A., Marshall, C. M., & Green, W. (2019). Online symptom checker diagnostic and triage accuracy for HIV and hepatitis C. *Epidemiology and Infection, 147*. doi:10.1017/s0950268819000268

*2. Bisson, L. J., Komm, J. T., Bernas, G. A., Fineberg, M. S., Marzo, J. M., Rauh, M. A., Smolinski, J. R., & Wind, W. M. (2014). Accuracy of a Computer-Based Diagnostic Program for Ambulatory Patients With Knee Pain. *The American Journal of Sports Medicine, 42*(10), p.2371-2376. doi:10.1177/0363546514541654

*3. Bisson, L. J., Komm, J. T., Bernas, G. A., Fineberg, M. S., Marzo, J. M., Rauh, M. A., Smolinski, J. R., & Wind, W. M. (2016). 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, 4*(2), p.1-5. 232596711663028. doi:10.1177/2325967116630286

Bloom, D.E., Cafiero, E.T., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L.R., Fathima, S., Feigl, A.B., Gaziano, T., Mowafi, M., Pandya, A., Prettner, K., Rosenberg, L., Seligman, B., Stein, A., & Weinstein, C. (2011). The Global Economic Burden of Non-communicable Diseases. *Geneva: World Economic Forum*. Retrieved from: https://www.world-heart-federation.org/wp-content/uploads/2017/05/WEF_Harvard_HE_GlobalEconom icBurdenNonCommunicableDiseases_2011.pdf

Boogerd, A. E., Arts, T., Engelen, JLPG. L., & Van de Belt, H. T. (2015). "What Is eHealth": Time for An Update? *Journal of Medical Internet Research, 4*(1), p.1-3. doi: 10.2196/resprot.4065

Boulos, M. N. K., Brewer, A. C., Karimkhani, C., Buller, D. B., & Dellavalle, R. P. (2014). Mobile medical and health apps: state of the art, concerns, regulatory control and certification. *Online Journal of Public Health Informatics, 5*(3). doi: 10.5210/ojphi.v5i3.4814

Broderick, J., Devine, T., Langhans, E., Lemerise, J. A., Lier, S., & Harris, L. (2014). Designing Health Literate Mobile Apps. *NAM Perspectives*. National Academy of Medicine, Washington, DC. doi.org/10.31478/201401a

Busby, M., Chapple, E., Matthews, R., & Chapple, I. L. C. (2013). Practitioner evaluation of a novel onlineintegrated oral health and risk assessment tool: a practice pilot. *British Dental Journal, 215*(3), p.115–120. doi: 10.1038/sj.bdj.2013.738

Centers for Medicare & Medicaid Services. National Health Expenditures 2017 Highlights. Retrieved from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData

Chambers, D., Cantrell, A. J., Johnson, M., Preston, L., Baxter, S. K., Booth, A., & Turner, J. (2019). Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open, 9*(8). doi: 10.1136/bmjopen-2018-027743

Chen, Y., Li, C., Liang, J., & Tsai, C. (2018). Health Information Obtained From the Internet and Changes in Medical Decision Making: Questionnaire Development and Cross-Sectional Survey. *Journal of Medical Internet Research,20*(2). doi:10.2196/jmir.9370

Clarke, C. S., Round, J., Morris, S., Kharicha, K., Ford, J., Manthorpe, J., Iliffe, S., Goodman, C., & Walters, K. (2017). Exploring the relationship between frequent internet use and health and social care resource use in a community-based cohort of older adults: an observational study in primary care. *BMJ Open*, *7*(7). doi: 10.1136/bmjopen-2017-015839

*4. Davies, B. M., Munro, C. F., & Kotter, M. R. (2019). A Novel Insight Into the Challenges of Diagnosing Degenerative Cervical Myelopathy Using Web-Based Symptom Checkers. *Journal of Medical Internet Research, 21*(1). doi:10.2196/10868

Diaz, A. J., Griffith, A. R., Ng, J. J., Reinert, E. S., Friedmann, D. P., & Moulton, W. M. (2002). Patients' Use of the Internet for Medical Information. *Journal of General Internal Medicine, 17*(3)., p.180-185. doi: 10.1046/j.1525-1497.2002.10603x

Doherty-Torstrick, E. R., Walton, K. E., & Fallon, B. A. (2016). Cyberchondria: Parsing Health Anxiety From Online Behavior. *Psychosomatics, 57*(4), 390–400. doi: 10.1016/j.psym.2016.02.002

*5. Donker, T., Straten, A. V., Marks, I., & Cuijpers, P. (2009). A Brief Web-Based Screening Questionnaire for Common Mental Disorders: Development and Validation. *Journal of Medical Internet Research, 11*(3), p1-12. doi:10.2196/jmir.1134

Dutta-Bergman, M. J. (2004). Primary Sources of Health Information: Comparisons in the Domain of Health Attitudes, Health Cognitions, and Health Behaviors. *Health Communication, 16*(3), p.273-288. doi:10.1207/s15327027hc1603_1

Ekeland, A. G., Bowes, A., & Flottorp, S. (2012). Methodologies for assessing telemedicine: A systematic review of reviews. *International Journal of Medical Informatics, 81*(1), p.1-11. doi:10.1016/j.ijmedinf.2011.10.009

Elbert, N. J., Os-Medendorp, H. V., Renselaar, W. V., Ekeland, A. G., Roijen, L. H., Raat, H., Nijsten, E. C. T., & Pasmans, G. M. A. S. (2014). Effectiveness and Cost-Effectiveness of eHealth Interventions in Somatic Diseases: A Systematic Review of Systematic Reviews and Meta-Analyses. *Journal of Medical Internet Research, 16*(4), p.1-23. doi:10.2196/jmir.2790

*6. Farmer, S., Bernardotto, M., & Singh, V. (2011). How good is Internet self-diagnosis of ENT symptoms using Boots WebMD symptom checker? *Clinical Otolaryngology, 36*(5), p.517-518. doi:10.1111/j.1749-4486.2011.02375.x

*7. Farvolden, P., Mcbride, C., Bagby, R. M., & Ravitz, P. (2003). A Web-Based Screening Instrument for Depression and Anxiety Disorders in Primary Care. *Journal of Medical Internet Research, 5*(3), p.1-7. doi:10.2196/jmir.5.3.e23

*8. Ferrero, N. A., Morrell, D. S., & Burkhart, C. N. (2013). Skin scan: A demonstration of the need for FDA regulation of medical apps on iPhone. *Journal of the American Academy of Dermatology, 68*(3), p.515-516. doi:10.1016/j.jaad.2012.10.045

Fox, S., & Duggan, M. (2013). Health online 2013. Retrieved from: https://www.pewresearch.org/internet/2013/01/15/health-online-2013/

Gesser-Edelsburg, A., Shadbari, E. A. N., Cohen, R., Halavi, M. A., Hijazi, R., Paz-Yaakobovitch, & Birman, Y. (2019). Differences in Perceptions of Health Information Between the Public and Health Care Professionals: Nonprobability Sampling Questionnaire Survey. *Journal of Medical Internet Research, 21*(7). doi: 10.2196/14105

Gill, K. H., Gill, N., & Young., D. S. (2013). Online Technologies for Health Information and Education: A literature review, *J Consum Health Internet, 17*(2), p.139-150. doi:10.1080/15398285.2013.780542.

*9. Hageman, M. G., Anderson, J., Blok, R., Bossen, J. K., & Ring, D. (2014). Internet Self-Diagnosis in Hand Surgery. *Hand, 10*(3), p.565-569. doi:10.1007/s11552-014-9707-x

Hajat C, & Stein E. (2018). The global burden of multiple chronic conditions: A narrative review. *Preventitive Medicine Reports, 12*, p.284-293. doi.org/10.1016/j.pmedr.2018.10.008

Hartman, M., Martin, A., Mcdonnell, P., & Catlin, A. (2009). National Health Spending In 2007: Slower Drug Spending Contributes To Lowest Rate Of Overall Growth Since 1998. *Health Affairs*, *28*(1), 246–261. doi: 10.1377/hlthaff.28.1.246

*10. Heikes, K. E., Eddy, D. M., Arondekar, B., & Schlessinger, L. (2007). Diabetes Risk Calculator: A simple tool for detecting undiagnosed diabetes and pre-diabetes. *Diabetes Care, 31*(5), p.1040-1045. doi:10.2337/dc07-1150

Hughes, D. A. (2010). From NCE to NICE: the role of pharmacoeconomics. *British Journal of Clinical Pharmacology, 70*(3), 317–319. doi: 10.1111/j.1365-2125.2010.03708.x

Hsieh, R. W., Chen, L., Chen, T., Liang, J., Lin, T., Chen, Y. (2016). The Association Between Internet Use and Ambulatory Care-Seeking Behaviors in Taiwan: A Cross-Sectional Study. *Journal of Medical Internet Research, 8*(12). DOI: 10.2196/jmir.5498

IQVIA Institute for Human Data Science (November 2017). The Growing Value of Digital Health. *Evidence and Impact on Human Health and Healthcare System*. Retrieved from: https://www.iqvia.com/institute/reports/the-growing-value-of-digital-health

Jenniskens, K., Groot, J. A. H. D., Reitsma, J. B., Moons, K. G. M., Hooft, L., & Naaktgeboren, C. A. (2017). Overdiagnosis across medical disciplines: a scoping review. *BMJ Open, 7*(12). doi: 10.1136/bmjopen-2017-018448

*11.Kafle, S., Pan. P., Torkamani, A., Halley, S., Powers, J., & Kardes, H. (2018). Personalized symptom checker using medical claims. *Health Recommender Systems*.

*12. Kim, G., & Lee, D. (2019). Intelligent Health Diagnosis Technique Exploiting Automatic Ontology Generation and Web-Based Personal Health Record Services. *IEEE Access, 7*, p.9419-9444. doi:10.1109/access.2019.2891710

Krey, M. (2018). MHealth Apps: Potentials for the Patient – Physician Relationship. *Journal of Advances in Information Technology,9*(4), 102-109. doi:10.12720/jait.9.4.102-109

Kostopoulou, O., Rosen, A., Round, T., Wright, E., Douiri, A., & Delaney, B. (2014). Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. *British Journal of General Practice*, *65*(630). doi: 10.3399/bjgp15x683161

*13. Kwan, K. S., Jachimowicz, E. A., Bastian, L., Marshall, L., & Mak, D. B. (2012). Online chlamydia testing: An innovative approach that appeals to young people. *Medical Journal of Australia, 197*(5), p.287-290. doi:10.5694/mja11.11517

Laz, H. T., & Berenson, B. A. (2013). Racial and ethnic disparities in internet use for seeking health information among young women. *Journal of Health Communication, 18*(2). p.250-260. doi: 10.1080/10810730.2012.707292

*14. Levy, A. G., Sonnad, S. S., Kurichi, J. E., Sherman, M., & Armstrong, K. (2008). Making Sense of Cancer Risk Calculators on the Web. *Journal of General Internal Medicine, 23*(3), p.229-235. doi:10.1007/s11606-007-0484-x

Lewis, M. J. (2017). INTERNET ACCESS AND RACIAL/ETHNIC DISPARITIES IN USING INTERNET HEALTH RESOURCES. *SEHSD Working Paper. U.S.Census Bureau.* Retrieved from: https://www.census.gov/content/dam/Census/library/working-papers/2017/demo/SEHSD-WP2017-31paper.pdf

*15. Lippi, G., & Sanchis-Gomar, F. (2018). The 'lottery' of cardiovascular risk estimation with Internet-based risk calculators. *Journal of Medical Systems, 42*(4), p1-5. doi:10.1007/s10916- 018-0925-6

Loos, A. (2013). Cyberchondria: Too Much Information for the Health Anxious Patient? *Journal of Consumer Health On the Internet, 17*(4), p.439-445. doi:10.1080/15398285.2013.833452

*16. Losina, E., Klara, K., Michl, G. L., Collins, J. E., & Katz, J. N. (2015). Development and feasibility of a personalized, interactive risk calculator for knee osteoarthritis. *BMC Musculoskeletal Disorders, 16*(1). doi:10.1186/s12891-015-0771-3

*17. Losina, E., Michl, G. L., Smith, K. C., & Katz, J. N. (2017). Randomized Controlled Trial of an Educational Intervention Using an Online Risk Calculator for Knee Osteoarthritis:

Effect on Risk Perception. *Arthritis Care & Research, 69*(8), p.1164-1170. doi:10.1002/acr.23136

Luger, T. M., Houston, T.K., & Suls, J. (2014). Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. *Journal of Medical Internet Research, 16*(16), p.1-13. doi:10.2196/jmir.2924

Lupton, D., & Jutel, A. (2015). 'Its like having a physician in your pocket!' A critical analysis of self-diagnosis smartphone apps. *Social Science & Medicine, 133*, 128–135. doi: 10.1016/j.socscimed.2015.04.004

*18. Maier, T., Kulichova, D., Schotten, K., Astrid, R., Ruzicka, T., Berking, C., & Udrea, A. (2014). Accuracy of a smartphone application using fractal image analysis of pigmented moles compared to clinical diagnosis and histological result. *Journal of the European Academy of Dermatology and Venereology, 29*(4), P.663-667. doi:10.1111/jdv.12648

Mcmullan, M. (2006). Patients using the Internet to obtain health information: How this affects the patient–health professional relationship. *Patient Education and Counseling, 63*(1-2), p.24-28. doi:10.1016/j.pec.2005.10.006

Meyer, A. N., Longhurst, C. A., & Singh, H. (2016). Crowdsourcing Diagnosis for Patients With Undiagnosed Illnesses: An Evaluation of CrowdMed. *Journal of Medical Internet Research, 18*(1). doi: 10.2196/jmir.4887

*19. Middleton, K., Butt, M., Hammerla, N., Hamblin, S., Mehta, K., & Parsa, A. (2016). Sorting out symptoms: design and evaluation of the 'babylon check' automated triage system. Available at: https://arxiv.org/abs/1606.02041.

Millenson, L. M., Baldwin, L. J., Zipperer, L., & Singh, H. (2018). Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis.*Diagnosis, 5*(3), p.95-105. doi.org/10.1515/dx-2018-0009

Morita, T., Rahman, A., Hasegawa, T., Ozaki, A., & Tanimoto, T. (2017). The Potential Possibility of Symptom Checker. *International Journal of Health Policy and Management, 6*(10), 615–616. doi: 10.15171/ijhpm.2017.41

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., Mcarthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*(1), p1.-7. doi:10.1186/s12874-018-0611-x

Murphy, M. (2019, March 10). Dr Google will see you now: Search giant wants to cash in on your medical queries. *The Telegraph.* Retrieved from: https://www.telegraph.co.uk/technology/2019/03/10/google-sifting-one-billion- health-questions-day/

National Academy Press. (2001). *Preparing for an aging world: the case for cross-national research*. DOI: 10.17226/10120

*20. Ngoo, A., Finnane, A., Mcmeniman, E., Tan, J., Janda, M., & Soyer, H. P. (2017). Efficacy of smartphone applications in high-risk pigmented lesions. *Australasian Journal of Dermatology, 59*(3), p.1-18. doi:10.1111/ajd.12599

*21. Nijland, N., Cranen, K., Boer, H., Julia E W C Van Gemert-Pijnen, & Seydel, E. R. (2010). Patient use and compliance with medical advice delivered by a web-based triage system in primary care. *Journal of Telemedicine and Telecare, 16*(1), p.8-11. doi:10.1258/jtt.2009.001004

Nurek, M., Kostopoulou, O., Delaney, B. C., & Esmail, A. (2015). Reducing diagnostic errors in primary care. A systematic meta-review of computerized diagnostic decision support systems by the LINNEAUS collaboration on patient safety in primary care. *European Journal of General Practice,21*(1), p.8-13. doi:10.3109/13814788.2015.1043123

Ossebaard, H. C., & Gemert-Pijnen, L. V. (2016). eHealth and quality in health care: implementation time. *International Journal for Quality in Health Care*, *28*(3), 415–419. doi: 10.1093/intqhc/mzw032

Palinkas, M., Canto, G. D. L., Rodrigues, L. A. M., Bataglion, C., Siéssere, S., Semprini, M., & Regalo, S. C. H. (2016). The Real Role of Sensitivity, Specificity and Predictive Values in the Clinical Assessment. *Journal of Clinical Sleep Medicine*, *12*(02), 279–280. doi: 10.5664/jcsm.5506

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*(1), 45. doi: 10.4103/0301-4738.37595

*22. Poote, A. E., French, D. P., Dale, J., & Powell, J. (2014). A study of automated self-assessment in a primary care student health centre setting. *Journal of Telemedicine and Telecare, 20*(3), p.123-127. doi:10.1177/1357633x14529246

*23. Powley, L., Mcilroy, G., Simons, G., & Raza, K. (2016). Are online symptoms checkers useful for patients with inflammatory arthritis? *BMC Musculoskeletal Disorders, 17*(1). doi:10.1186/s12891-016-1189-2

*24. Price, R. A., Fagbuyi, D., Harris, R., Hanfling, D., Place, F., Taylor, T. B., & Kellermann, A. L. (2013). Feasibility of Web-Based Self-Triage by Parents of Children With Influenza-Like Illness. *JAMA Pediatrics, 167*(2), 112. doi:10.1001/jamapediatrics.2013.1573

*25. Ruotsalo, T., & Lipsanen, A. (2018). Interactive Symptom Elicitation for Diagnostic Information Retrieval. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. doi:10.1145/3209978.3210172

Schiff, G. D. (2009). Diagnostic Error in Medicine. *Archives of Internal Medicine, 169*(20), 1881. doi: 10.1001/archinternmed.2009.333

*26. Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self-diagnosis and triage: audit study. *Bmj*. doi: 10.1136/bmj.h3480

*27. Shen, C., Nguyen, M., Gregor, A., Isaza, G., & Beattie, A. (2019). Accuracy of a Popular Online Symptom Checker for Ophthalmic Diagnoses. *JAMA Ophthalmology, 137*(6), doi:10.1001/jamaophthalmol.2019.0571

Silva, B. M., Rodrigues, J. J., De la Torre Diez, I., Lopez-Coronado, M., & Saleem, K. (2015). Mobile-health: A review of current state in 2015. *Journal of Biomedical Informatics,56*, p.265-272. http://dx.doi.org/10.1016/j.jbi.2015.06.003

Singh, H., Meyer, D. N. A., & Thomas, J. E. (2014). The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ Quality & Safety, 23*(9), p.727-731. doi: 10.1136/bmjqs-2013-002627

Singh, H., Schiff, G. D., Graber, M. L., Onakpoya, I., & Thompson, M. J. (2016). The global burden of diagnostic errors in primary care. *BMJ Quality & Safety*, *26*(6), 484–494. doi: 10.1136/bmjqs-2016-005401

*28. Sole, M. L., Stuart, P. L., & Deichen, M. (2006). Web-Based Triage in a College Health Setting. *Journal of American College Health, 54*(5), p.289-294. doi:10.3200/jach.54.5.289-294

Sousa, E. C. V & Lopez, D. K. (2017). Towards Usable E-Health, A Systematic Review of Usability Questionnaires. *Applied Clinical Informatics, 8*(2), p.470-490. https://doi.org/10.4338/ACI-2016-10-R-0170

Srivastava, S., Pant, M., Abraham, A., & Agrawal, N. (2015). The Technological Growth in eHealth Services. *Computational and Mathematical Methods in Medicine*, *2015*, 1–18. doi:10.1155/2015/894171

Statista Research Department. 2015. Share of individuals who have used the internet to search for health care information in the United Kingdom (UK) in 2015. Retrieved from: https://www.statista.com/statistics/505053/individual-use-internet-for-health-information-search-united-kingdom-uk/

Svenstrup, D., Jørgensen, L. H., & Winther, O. (2015). Rare disease diagnosis: A review of web search, social media and large-scale data-mining approaches. *Rare Diseases, 3*(1), p.1-7.

*29. Thissen, M., Udrea, A., Hacking, M., Braunmuehl, T. V., & Ruzicka, T. (2017). MHealth App for Risk Assessment of Pigmented and Nonpigmented Skin Lesions—A Study on Sensitivity and Specificity in Detecting Malignancy. *Telemedicine and E-Health, 23*(12), p.948-954. doi:10.1089/tmj.2016.0259

Valizadeh-Haghi, S., & Rahmatizadeh, S. (2018). eHealth Literacy and General Interest in Using Online Health Information: A Survey Among Patients with Dental Diseases. *Online Journal of Public Health Informatics. 10*(3). doi:10.5210/ojphi.v10i3.9487

*31. Van Ameringen, M., Simpson, W., Patterson, B., & Turna, J. (2015). Internet screening for anxiety disorders: Treatment-seeking outcomes in a three-month follow-up study. *Psychiatry Research, 230*(2), p.689-694. doi:10.1016/j.psychres.2015.10.031

Wang, L., Wang, J., Wang, M., Li, Y., Liang, Y., & Xu, D. (2012). Using Internet Search Engines to Obtain Medical Information: A Comparative Study. *Journal of Medical Internet Research*, *14*(3). doi: 10.2196/jmir.1943

*30. Wolf, J. A., Moreau, J., Akilov, O., Patton, T., English, C. J., Ho, J., & Ferris, K. L. (2013). Diagnostic Inaccuracy of Smart Phone Applications for Melanoma Detection. *JAMA Dermatology, 149*(4), p.422-426. doi:10.1001/jamadermatol.2013.2382.

Wyatt, J. C. (2015). Fifty million people use computerised self triage. *British Medical Association.* doi:10.1136/bmj.h3727

Ybarra, M.L., Suman, M. (2006). Help seeking behavior and the Internet: a national survey. *International Journal of Medical Informatics, 75*(1), p.29-41. DOI: 10.1016/j.ijmedinf.2005.07.029

# Appendix

**Overview on diagnostic tools & medical conditions assessed**

| Illness condition | General and Unspecified | Dermatological | Arthritis | Psychiatric disorders | Orthopedic (Knee Pain) | DCM | Hand pain | Diabetes | ENT | Ophthalmic | Influenza | Breast Cancer | HIV /Hep C | CVD | Chlamydia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Diagnostic Tool | | | | | | | | | | | | | | | |
| ACC/AHA 2013 (15) | | | | | | | | | | | | | | ✓ | |
| ACC/AHA ASCVD (15) | | | | | | | | | | | | | | ✓ | |
| ASCVD (15) | | | | | | | | | | | | | | ✓ | |
| AskMD (12; 26) | ✓ | | | | | | | | | | | | | | |
| AstraZeneca (14) | | | | | | | | | | | | ✓ | | | |
| Australian absolute cardiovascular disease risk calculator (15) | | | | | | | | | | | | | | ✓ | |
| Babylon Check (19) | ✓ | | | | | | | | | | | | | | |
| BetterMedicine (26) | ✓ | | | | | | | | | | | | | | |
| Breastcancerquiz.com (14) | | | | | | | | | | | | ✓ | | | |
| Cardiovascular Risk Calculator (15) | | | | | | | | | | | | | | ✓ | |
| CVD Risk Check (15) | | | | | | | | | | | | | | | |
| DIABETES RISK CALCULATOR (10) | | | | | | | | ✓ | | | | | | | |
| DocResponse (26) | ✓ | | | | | | | | | | | | | | |
| DoctorResponse (26) | ✓ | | | | | | | | | | | | | | |
| Dokterdokter.nl (21) | ✓ | | | | | | | | | | | | | | |
| Drugs.com (26) | ✓ | | | | | | | | | | | | | | |
| Dr. Mole (20) | | ✓ | | | | | | | | | | | | | |
| EarlyDoc (26) | ✓ | | | | | | | | | | | | | | |
| Esagil (26) | ✓ | | | | | | | | | | | | | | |
| Family Doctor (26) | | | | | | | | | | | | | | | |
| FreeMD (26) | ✓ | | | | | | | | | | | | | | |
| GenneX (14) | | | | | | | | | | | | ✓ | | | |
| Halls, MD (14) | | | | | | | | | | | | ✓ | | | |
| Harvard Center for Cancer Prevention (14) | | | | | | | | | | | | ✓ | | | |
| Harvard Medical School Family Health Guide (26) | ✓ | | | | | | | | | | | | | | |
| Healthline (4; 26) | ✓ | | | | | ✓ | | | | | | | | | |
| Healthwise (26) | ✓ | | | | | | | | | | | | | | |
| Healthy Children (26) | ✓ | | | | | | | | | | | | | | |
| Human Disease Diagnosis Ontology (HDDO) (12) | ✓ | | | | | | | | | | | | | | |

**Overview on diagnostic tools & medical conditions assessed**

| Illness condition | General and Unspecified | Dermatological | Arthritis | Psychiatric disorders | Orthopedic (Knee Pain) | DCM | Hand pain | Diabetes | ENT | Ophthalmic | Influenza | Breast Cancer | HIV /Hep C | CVD | Chlamydia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Diagnostic Tool | | | | | | | | | | | | | | | |
| Harvard Center for Cancer Prevention (14) | | | | | | | | | | | | ✓ | | | |
| Harvard Medical School Family Health Guide (26) | ✓ | | | | | | | | | | | | | | |
| Healthline (4; 26) | ✓ | | | | | ✓ | | | | | | | | | |
| Healthwise (26) | ✓ | | | | | | | | | | | | | | |
| Healthy Children (26) | ✓ | | | | | | | | | | | | | | |
| Human Disease Diagnosis Ontology (HDDO) (12) | ✓ | | | | | | | | | | | | | | |
| Healthtools AARP (4) | | | | | | ✓ | | | | | | | | | |
| Heartscore (15) | | | | | | | | | | | | | | ✓ | |
| Interactive Symptom Elicitation (ISE) (25) | ✓ | | | | | | | | | | | | | | |
| Isabel (1; 12; 26) | ✓ | | | | | | | | | | | | ✓ | | |
| iTriage (26) | ✓ | | | | | | | | | | | | | | |
| Knowledge Base (KB) (11) | ✓ | | | | | | | | | | | | | | |
| MACSCREEN (31) | | | | ✓ | | | | | | | | | | | |
| Mary Bird Perkins (14) | | | | | | | | | | | | ✓ | | | |
| Mayoclinic (1; 12; 26) | ✓ | | | | | | | | | | | | ✓ | | |
| Mayoclinic Heart Disease Risk Calculator (15) | | | | | | | | | | | | | | ✓ | |
| MD+ CALC (15) | | | | | | | | | | | | | | ✓ | |
| MEDoctor (26) | ✓ | | | | | | | | | | | | | | |
| National Surgical Breast and Bowel Project (14) | | | | | | | | | | | | ✓ | | | |
| NCI (14) | | | | | | | | | | | | ✓ | | | |
| NetDoctor (4) | ✓ | | | | | ✓ | | | | | | | | | |
| NHS (23; 26) | ✓ | | ✓ | | | | | | | | | | | | |
| OA RISK C (16; 17) | | | ✓ | | | | | | | | | | | | |
| National Surgical Breast and Bowel Project (14) | | | | | | | | | | | | ✓ | | | |
| NCI (14) | | | | | | | | | | | | ✓ | | | |
| NetDoctor (4) | ✓ | | | | | ✓ | | | | | | | | | |
| NHS (23; 26) | ✓ | | ✓ | | | | | | | | | | | | |
| OA RISK C (16; 17) | | | ✓ | | | | | | | | | | | | |
| OLC (13) | | | | | | | | | | | | | | | ✓ |

**Overview on diagnostic tools & medical conditions assessed**

| Illness condition | General and Unspecified | Dermatological | Arthritis | Psychiatric disorders | Orthopedic (Knee Pain) | DCM | Hand pain | Diabetes | ENT | Ophthalmic | Influenza | Breast Cancer | HIV/Hep C | CVD | Chlamydia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Diagnostic Tool | | | | | | | | | | | | | | | |
| QRISK3–2017 (15) | | | | | | | | | | | | | | ✓ | |
| RealAge (14) | | | | | | | | | | | | ✓ | | | |
| Framingham Risk Score (ATP-III) (15) | | | | | | | | | | | | | | ✓ | |
| Skin Scan (8) | | ✓ | | | | | | | | | | | | | |
| SkinVision (18; 20; 29) | | ✓ | | | | | | | | | | | | | |
| SORT for kids (25) | | | | | | | | | | | ✓ | | | | |
| SpotMole (20) | | ✓ | | | | | | | | | | | | | |
| Steps2Care (26) | ✓ | | | | | | | | | | | | | | |
| Symcat (1; 12; 26) | ✓ | | | | | | | | | | | | ✓ | | |
| Symptify (26) | ✓ | | | | | | | | | | | | | | |
| Symptomate (1; 26) | ✓ | | | | | | | | | | | | ✓ | | |
| Unspecified dermatological applications (30) | | ✓ | | | | | | | | | | | | | |
| Unspecified online self-assessment system (22) | ✓ | | | | | | | | | | | | | | |
| Unspecified web-based program (2; 3) | | | | | ✓ | | | | | | | | | | |
| WB-DAT (7) | | | | ✓ | | | | | | | | | | | |
| WebMD (1; 4; 12; 9; 23; 26; 27) | ✓ | | ✓ | | | ✓ | | | ✓ | ✓ | | | ✓ | | |
| Women’s Cancer Network (14) | | | | | | | | | | | | ✓ | | | |
| WSQ (5) | | | | ✓ | | | | | | | | | | | |
| 24/7 WebMed (28) | ✓ | | | | | | | | | | | | | | |