Evaluating the Accuracy of a New Artificial Intelligence Based Symptom Checker: A Clinical Vignette Study

Mohammad Hammoud*, Shahd Douglas, Mohamad Darmach, Sara Alawneh, Swapnendu Sanyal, and Youssef Kanbour

Rimads QSTP-LLC, Qatar Science and Technology Park, Doha, Qatar

mhh@rimads.ai, shahd@rimads.ai, mdarmach@rimads.ai, salawneh@rimads.ai, swapnendu@rimads.ai, youssef@rimads.ai

ABSTRACT

Objectives To evaluate the accuracy of a new Artificial Intelligence (AI) based symptom checker and compare it against that of some popular symptom checkers and seasoned primary care physicians.

Design Vignette study.

Setting 400 gold-standard primary care vignettes.

Intervention/Comparator We propose a 4-stage comprehensive experimentation methodology that capitalizes on the standard clinical vignette approach to evaluate 6 symptom checkers. To this end, we developed and peer-reviewed 400 vignettes, each approved by at least 5 out of 7 independent and experienced general practitioners. To establish a frame of reference and interpret the results of symptom checkers accordingly, we further compared the best-performing symptom checker against 3 primary care physicians with an average experience of 16.6 years.

Primary Outcome Measures We thoroughly studied the diagnostic accuracies of symptom checkers and physicians from 7 standard performance angles, including (a) *M1* as a measure of a symptom checker's or a physician's ability to return a vignette's main diagnosis at the top of their differential list, (b) F1-measure as a trade-off score between sensitivity and precision, and (c) NDCG as a measure of a differential list's ranking quality, among others.

Results The new AI-based symptom checker, namely, Avey significantly outperformed 5 popular symptom checkers, namely, Ada, WebMD, K-Health, Buoy, and Babylon by averages of 24.5%, 175.5%, 142.8%, 159.6%, 2968.1% using *M1*; 8.7%, 88.9%, 66.4%, 88.9%, 2084% using F1-measure; and 21.2%, 93.4%, 113.3%, 136.4%, 3091.6% using NDCG, respectively. In contrast, physicians slightly outpaced Avey by an average of 1.2% using F1-measure, while Avey exceeded them by averages of 10.2% and 25.1% using *M1* and NDCG, respectively.

Conclusions Avey demonstrated a superior performance against current symptom checkers and compared favorably to physicians.

Strengths and Limitations of this Study

- This study investigated thoroughly the performance of 6 symptom checkers and a panel of experienced physicians from 7 different accuracy dimensions, one of which was explored for the first time in literature.
- To the best of our knowledge, the study developed and peer-reviewed the largest benchmark vignette suite in the domain thus far.
- To minimize bias, the symptom checkers were only tested by independent primary care physicians and using only gold-standard vignettes.
- To establish a standard of full transparency and facilitate the reproducibility of the study, all the peer-reviewed vignettes and results (i.e., 45 sets of experiments) were made publicly available.
- The study lacks an evaluation on real patients and a rigorous process to choose symptom checkers.

1. INTRODUCTION

Digital health has become ubiquitous. Every day millions of people turn to the Internet for health information and treatment advice [1, 2]. For instance, in Australia, around 80% of people search the Internet for health information, and nearly 40% seek guidance online for self-treatment [3, 4]. In the US, almost two-thirds of adults search the Web for health information and one-third utilize

*Correspondence to Dr. Mohammad Hammoud; Rimads QSTP-LLC and Carnegie Mellon University; mhh@rimads.ai

it for *self-diagnosis*, trying to discover by themselves the underlying causes of their health symptoms [5]. A recent study showed that half of the patients investigated their symptoms on search engines before visiting emergency departments [6, 7].

While search engines like Google and Bing are exceptional tools for educating people on almost any matter, they may facilitate misdiagnosis and induce risks stemming from unrelated health content [5]. This is because Web search entails sifting through an ocean of results, which could emanate from all sorts of sources, and making personal judgments on which data to unveil. Some governments have even launched "Don't Google It" advertising campaigns to urge their residents to avoid assessing their health using search engines [8, 9]. In fact, search engines are not medical diagnostic tools, and laymen are typically not equipped to exploit them for self-diagnosis.

In contrast to search engines, symptom checkers are patient-facing medical diagnostic tools that emulate clinical reasoning[1], especially if they use Artificial Intelligence (AI) [4, 10]. They are trained to make medical expert-like judgments on behalf of patients. More precisely, a patient can start a consultation session with a symptom checker by inputting a chief complaint (in terms of one or more symptoms). Afterwards, the symptom checker asks questions to the patient and collects answers from them. Eventually, the symptom checker generates a differential diagnosis (i.e., a ranked list of possible diseases) that explains the causes of the patient's symptoms.

Symptom checkers are increasingly becoming an integral part of digital health, with more than 15 million users per month [11] that are likely to keep growing [12]. A UK-based study that engaged 1,071 patients found that more than 70% of individuals between the ages of 18 and 39 years would use a symptom checker [13]. A recent study examining a specific symptom checker found that over 80% of patients perceived it to be useful and more than 90% indicated that they would use it again [14]. Various credible healthcare institutions and entities such as the UK National Health Service (NHS) [15] and the government of Australia [16] have officially adopted symptom checkers for self-diagnosis and referrals.

Symptom checkers are inherently scalable (i.e., they can assess millions of people instantly and concurrently) and universally available. Besides, they promise to provide patients with necessary high-quality, evidence-based information [17], reduce unnecessary medical visits [18, 19, 20, 21], alleviate the pressure on healthcare systems [22], improve accessibility to timely diagnosis [18], and guide patients to the most appropriate care pathways [12], to mention just a few.

Nevertheless, the utility and promise of symptom checkers cannot be materialized if they do not prove to be accurate [10]. To elaborate, a recent study has shown that most patients (more than 76%) use symptom checkers solely for self-diagnosis [14]. As such, if symptom checkers are not meticulously engineered and rigorously evaluated on their diagnostic capabilities, they may put these patients at risk [23, 24, 25]. To this end, this paper comprehensively investigates the diagnostic performance of symptom checkers via measuring the accuracy of a few popular symptom checkers and a new AI-based one. In addition, it compares the accuracy of the best-performing symptom checker against that of a panel of experienced physicians to put things in perspective and interpret results accordingly.

To begin with, we shed some light on the new AI-based symptom checker, namely, Avey, which was extensively researched, designed, developed, and tested in-house for around 4 years before it was launched. Avey uses an intelligent inference engine with three major components: (1) a diagnostic algorithm, (2) a recommendation system, and (3) a ranking model. The inference engine taps into a probabilistic graphical model, namely, a Bayesian network (see Figure 1 for an actual visualization of this network). During a session with Avey, the engine's diagnostic algorithm operationalizes the Bayesian network and generates after every patient's answer a probability for each modeled disease, conditional on the *findings*[2] that have been discovered or inferred thus far.

Questions are asked during a patient's session with Avey via its recommendation system, which predicts the future impact of every finding that has not yet been asked and recommends the one that exhibits the highest impact on the current diagnostic hypothesis of the algorithm. At the end of the session, the ranking model ranks all the possible diseases and outputs them as a differential diagnosis to the patient.

To evaluate Avey and related symptom checkers, we propose a comprehensive scientific methodology that capitalizes on the standard clinical vignette approach. Delivering on this methodology, we compiled and peer-reviewed 400 vignettes with seven external medical doctors using a super-majority voting scheme. To the best of our knowledge, this yielded the largest benchmark

[1] Clinical reasoning is the reasoning process that leads to a medical diagnosis. It is inferential by nature where a doctor starts from an initial hypothesis based on a patient's complaint, gathers relevant information, makes several inferences, and produces a possible medical diagnosis that may explain the cause of the patient's symptoms.

[2] A finding is defined as a symptom, an attribute, or an etiology. An attribute is a feature of a symptom or an etiology (e.g., in "severe chest pain", "severe" is an attribute and "chest pain" is a symptom).

vignette suite in the domain thus far. Furthermore, we defined and utilized seven standard accuracy metrics, one of which measures for the first time in the field the ranking qualities of the differential diagnoses of symptom checkers and doctors.

We leveraged our benchmark vignette suite and accuracy metrics to study the performance of Avey and five other major symptom checkers, namely, Ada [26], K-Health [27], Buoy [28], Babylon [29], and WebMD [30]. Results show that Avey significantly outperforms the five popular symptom checkers. For instance, Avey outpaced Ada, K-Health, Buoy, Babylon, and WebMD by averages of 24.5%, 142.8%, 159.6%, 2968.1%, and 175.5%, respectively in generating the vignettes' main diagnoses at the top of their differential lists.

Moreover, we compared Avey's performance against three highly seasoned primary care physicians with an average experience of 16.6 years. Results show that Avey compares favorably to the physicians and even outperforms them under some accuracy metrics, including the ability to rank diseases correctly within their generated differential lists, among others.

Finally, to facilitate the reproducibility of our study and support future related studies, we made our benchmark vignette suite publicly and freely available at [31]. Besides, we posted all the results of the symptom checkers and physicians at [31] to establish a standard of full transparency and allow the community to cross-validate the results, a step much needed in health informatics [32].

2. METHODS

2.1 Stages

Building on prior related work [4, 5, 11, 12, 33, 34], we adopted a clinical vignette approach to measure the performance of Avey and several other symptom checkers. A seminal work at Harvard Medical School has established the value of this approach in validating the accuracy of symptom checkers [11, 34], especially that it has been also a common approach for testing physicians on their diagnostic abilities [34].

To this end, we concretely defined our experimentation methodology in terms of 4 stages, namely, *vignette creation*, *vignette standardization*, *vignette testing on symptom checkers*, and *vignette testing on doctors*. The 4 stages are demonstrated in Figure 2.

In the vignette creation stage, an internal team of medical doctors rigorously compiled a set of vignettes from October 10, 2021 until November 29, 2021. All the vignettes were drawn from reputable medical websites and training material for health care professionals, including USMLE Step 2 CK, MRCP Part 1 Self-Assessment, American Board of Family Medicine, and American Board of Pediatrics, among others [35, 36, 37, 38, 39, 40, 41, 42]. In addition, the internal medical team supplemented the vignettes with information that might be 'asked' by symptom checkers and physicians in stages 3 and 4. The vignettes involved 14 body systems and encompassed common and less-common conditions relevant to primary care practice (see Table 1). They fairly represent real-life and/or practical cases in which patients might seek primary care advice from physicians or symptom checkers.

Table 1:The body systems and numbers of common and less-common diseases covered in our benchmark vignette suite.

Body System	# of Disease	% of Common Diseases	% of Less-Common Diseases
Hematology	23	8.69	91.30
Cardiovascular	46	58.69	41.30
Neurology	22	40.90	59.09
Endocrine	20	65	35
ENT	23	69.56	30.43
GI	44	47.72	53.27
Obs/Gyn	54	59.25	40.74
Infectious	23	26.08	73.91
Respiratory	37	70.27	29.72
Orthopedics & Rheumatology	32	65.62	34.37
Ophthalmology	18	83.33	16.66
Dermatology	12	75	25
Urology	14	57.14	42.85
Nephrology	32	53.12	46.87

The internal medical team constructed each vignette with eight major components: (*i*) the age and sex of the assumed patient, (*ii*) a maximum of three chief complaints, (*iii*) the history of the suggested illness associated with details on the chief complaints and other present and relevant findings, (*iv*) absent findings, including ones that are expected to be solicited by symptom checkers and physicians in stages 3 and 4, (*v*) basic findings that pertain to physical examinations that can still be exploited by symptom checkers, (*vi*) past medical and surgical history, (*vii*) family history, and (*viii*) the most appropriate main and differential diagnoses.

The output of the vignette creation stage (i.e., stage 1) is a set of vignettes that serves as an input to the *vignette standardization* stage (i.e., stage 2). Seven external medical doctors from four specialties, namely, Family Medicine, General Medicine, Emergency Medicine, and Internal Medicine, with an average experience of 8.4 years were recruited from the professional networks of SD, SA, and MD to review the vignettes in this stage. None of these doctors had any involvement with Avey's project and they were all entirely unaware of it before they were recruited.

We designed and developed a full-fledged web portal to streamline the process of reviewing and standardizing the vignettes. To elaborate, the portal allows the internal medical team to upload the vignettes to a web page that is shared across the seven external recruited doctors. Each doctor can access the vignettes and review them independently, without seeing the reviews of other doctors.

After reviewing a vignette, a doctor can reject or accept it. Upon rejecting a vignette, a doctor can propose changes to improve its quality and/or clarity. The internal medical team reviews the suggested changes and updates the vignette accordingly, before re-uploading it to the portal for a new round of peer reviews[3]. Multiple review rounds can take place before a vignette is rendered gold-standard. A vignette becomes gold-standard only if it is accepted by at least five out of the seven (i.e., super-majority) external doctors. Once a vignette is standardized, the portal moves it automatically to stages 3 and 4.

Stage 2 started on October 17, 2021 and ended on December 4, 2021. As an outcome, 400 vignettes were produced and standardized. To allow for external validation, we made all the vignettes publicly available at [31]. Lastly, we note that none of the 400 vignettes were used in Avey's development.

The output of stage 2 serves as an input to stage 3, namely, *vignette testing on symptom checkers*. For this sake, we recruited three independent primary care physicians from two specialties, namely, Family Medicine and General Medicine, with an average experience of 4.2 years from the professional networks of SD and MD. None of these physicians had any involvement with the development of Avey and they were all completely unaware of it before they were recruited. Furthermore, two of them were not among the seven doctors who reviewed the vignettes in stage 2. These doctors were recruited solely to test the gold-standard vignettes on Avey and other related symptom checkers.

The approach of having primary care physicians test symptom checkers has been shown recently to be more *reliable* than having laypeople do it [33, 43, 44]. This is because the standardized vignettes act as *proxies* for patients, while testers act as only *data extractors* from the vignettes and *data feeders* to the symptom checkers. Consequently, the better the testers are in extracting and feeding data, the more reliable the clinical vignette approach becomes. In fact, a symptom checker cannot be judged on its accuracy if the answers to its questions do not precisely align with the vignettes. To this end, physicians are deemed more capable of playing the role of testers than laypeople, especially that AI-based symptom checkers may often ask questions that have no answers in the vignettes, even if the vignettes are quite comprehensive. Clearly, when these questions are asked, laypeople will not be able to answer them properly, diminishing thereby the reliability of the clinical vignette approach and the significance of the reported results. In contrast, physicians will judiciously answer these questions in alignment with the vignettes and capably figure out whether the symptom checkers are able to virtually 'diagnose' them (i.e., produce the correct differential diagnoses in the vignettes). We elaborate further on the rationale behind using physicians as testers in Section 4.2.

Besides vignettes, we chose five symptom checkers, namely, Ada [26], Babylon [29], Buoy [28], K-Health [27], and WebMD [30] to test and compare them against Avey. Four of these symptom checkers (i.e., Ada, Buoy, K-Health, and WebMD) were selected because of their superior performance in [33] and one (i.e., Babylon) due to its popularity. We tested the vignettes on the most up-to-date versions of these symptom checkers that were available on Google Play, App Store, or websites (e.g., Buoy) between the dates of November 7, 2021 and January 31, 2022.

The six symptom checkers (Avey and the five competitors) were tested through their normal question-answer flows. As in [33], each of the external physicians in stage 3 randomly pulled vignettes from the gold-standard pool and tested them on each of the six symptom checkers (see Figure 2). By the end of stage 3, each physician tested a total of 133 gold-standard vignettes on each symptom checker, except one physician who tested 1 extra vignette to complete the 400 vignettes. Each physician saved a screenshot of each symptom checker's output for each vignette to allow for results' verification, extraction[4], and analysis. We posted all these screenshots online at [31] to establish a standard of full transparency and allow for external cross-validation and study-replication.

[3] That is, the internal medical team always asks the seven external doctors to review again, and accept or reject every updated vignette, irrespective of how big the update is.

[4] Different symptom checkers and doctors can refer to the same disease differently. As such, the internal medical team considered an output disease by a symptom checker (in stage 3) or a doctor (in stage 4) as a reasonable match to a disease in the gold-standard vignette if it was an alternative name, an umbrella name, or a directly related disease to it.

In stage 4, we recruited three more independent and experienced primary care physicians with an average experience of 16.6 years from the professional networks of SD, SA, and MD. One of those physicians is a Family Medicine doctor with 30+ years of experience. The other two are also Family Medicine doctors, each with 10+ years of experience. None of these physicians had any involvement with the development of Avey and were completely unaware of it before they were recruited. Furthermore, none of them were among the seven or three doctors of stages 2 or 3, respectively and were only recruited for pursuing stage 4.

The sole aim of stage 4 is to compare the accuracy of the winning symptom checker against that of experienced primary care physicians. Hence and akin to [11], we concealed the main and differential diagnoses of the 400 gold-standard vignettes from the three recruited doctors and exposed the remaining information through our web portal. The doctors were granted access to the portal and asked to provide their main and differential diagnoses for each vignette without checking any reference, mimicking as closely as possible the way they conduct real-world sessions with patients. As an outcome, each vignette was 'diagnosed' by each of the three doctors. Again, the results of the doctors were posted online at [31] to allow for external cross-validation.

2.2 Accuracy Metrics

To evaluate the performance of symptom checkers and doctors in stages 3 and 4, we utilize seven standard accuracy metrics. As in [33, 45], for every tested gold-standard vignette, we use the matching-1 (***M1***), matching-3 (***M3***), and matching-5 (***M5***) criteria to measure if a symptom checker or a doctor is able to output the vignette's main diagnosis at the top (i.e., $M1$), among the first 3 diseases (i.e., $M3$), or among the first 5 diseases (i.e., $M5$) of their differential list. For each symptom checker and doctor, we report the percentages of vignettes that fulfil $M1$, $M3$, and $M5$. The mathematical definitions of $M1$, $M3$, and $M5$ are given in Table 2.

Table 2: The descriptions and mathematical definitions of the seven accuracy metrics used in our study.

Metric	Description	Mathematical Definition
***M1*%**	The percentage of vignettes where the gold-standard main diagnosis is returned at the top of a symptom checker's or a doctor's differential list	$\frac{\sum_{v=1}^{N} i_v}{N}$, where N is the number of vignettes and i_v is 1 if the symptom checker or doctor returns the gold-standard main diagnosis within vignette v at the top of their differential list; and 0 otherwise
***M3*%**	The percentage of vignettes where the gold-standard main diagnosis is returned among the first 3 diseases of a symptom checker's or a doctor's differential list	$\frac{\sum_{v=1}^{N} i_v}{N}$, where N is the number of vignettes and i_v is 1 if the symptom checker or doctor returns the gold-standard main diagnosis within vignette v among the top 3 diseases of their differential list; and 0 otherwise
***M5*%**	The percentage of vignettes where the gold-standard main diagnosis is returned among the first 5 diseases of a symptom checker's or a doctor's differential list	$\frac{\sum_{v=1}^{N} i_v}{N}$, where N is the number of vignettes and i_v is 1 if the symptom checker or doctor returns the gold-standard main diagnosis within vignette v among the top 5 diseases of their differential list; and 0 otherwise
Average Recall	Recall is the proportion of diseases that are in the gold-standard differential list and are generated by a symptom checker or a doctor. The average recall is taken across all vignettes for each symptom checker and doctor	$\frac{\sum_{v=1}^{N} r_v}{N}$, where N is the number of vignettes and $r_v = \frac{True\ Positive}{True\ Positive+False\ Negative}$ of the symptom checker or doctor for vignette v
Average Precision	Precision is the proportion of diseases in the symptom checker's or doctor's differential list that are also in the gold-standard differential list. The average precision is taken across all vignettes for each symptom checker and doctor	$\frac{\sum_{v=1}^{N} p_v}{N}$, where N is the number of vignettes and $p_v = \frac{True\ Positive}{True\ Positive+False\ Positive}$ of the symptom checker or doctor for vignette v
Average F1-measure	F1-measure captures the trade-off between precision and recall. The average F1-measure is taken across all vignettes for each symptom checker and doctor	$\frac{\sum_{v=1}^{N} \frac{2(p_v+r_v)}{p_v+r_v}}{N}$, where N is the number of vignettes and r_v and p_v are as defined at column 3 in rows 5 and 6 above, respectively
Average NDCG	Normalized Discounted Cumulative Gain (NDCG) is a measure of ranking quality. The average NDCG is taken across all vignettes for each symptom checker and doctor	$\frac{\sum_{v=1}^{N} \frac{DCG_v}{gold\ DCG_v}}{N}$, assuming N vignettes, n number of diseases in a gold-standard vignette v, and $relevance_i$ for the disease at position i in v's differential list. $DCG_v = \sum_{i=1}^{n} \frac{2^{relevance_i}-1}{\log_2(i+1)}$, which is computed over the differential list of a doctor or a symptom checker for v. $gold\ DCG_v$ is defined exactly as DCG_v, but is computed over the gold-standard differential list of v

Besides, as in [33, 43, 46], for each tested gold-standard vignette, we use ***recall*** (or sensitivity in medical parlance) as a measure of the percentage of relevant diseases that are returned in the symptom checker's or doctor's differential list. Moreover, we utilize ***precision*** as a measure of the percentage of diseases in the symptom checker's or doctor's differential list that are relevant. For each symptom checker and doctor, we report the average recall and average precision across all vignettes. The average recall and average precision are defined mathematically in Table 2.

Typically, there is a trade-off between recall and precision (the higher the recall, the lower the precision, and vice versa). Thus, in accordance with the standard practice in information retrieval[5], we further use the ***F1-measure*** that combines the trade-off between recall and precision in one easily interpretable score. The mathematical definition of the F1-measure is provided in Table 2. The higher the F1-measure of a symptom checker or a doctor, the better.

Finally, we measure the ranking qualities of each symptom checker and doctor using the Normalized Discounted Cumulative Gain (***NDCG***) [47] metric that is widely used in practice [48]. To begin with, each disease at position i in the differential list of a gold-standard vignette is assigned $relevance_i$. The higher the rank of a disease in the differential list, the higher the relevance of that disease to the correct diagnosis. Next, Discounted Cumulative Gain (DCG) is defined mathematically as $\sum_{i=1}^{n} \frac{2^{relevance_i}-1}{\log_2(i+1)}$, assuming n diseases in a vignette's differential list (see Table 2). As such, DCG penalizes a symptom checker or a doctor if they rank a disease lower in their output differential list than the gold-standard list. Capitalizing on DCG, Normalized DCG (NDCG) is the ratio of a symptom checker's or a doctor's DCG divided by the corresponding gold-standard DCG. Table 2 provides the mathematical definition of NDCG.

2.3 Patient and Public Involvement

No patients were involved in any part of this study, but rather vignettes that acted as proxies for patients during testing with symptom checkers and physicians.

3. RESULTS

3.1 Avey versus Symptom checkers

In this section, we present our findings of stage 3. As indicated in Section 2.1, the 400 gold-standard vignettes were tested over six symptom checkers, namely, Avey, Ada, WebMD, K-Health, Buoy, and Babylon. Not every vignette was successfully diagnosed by every symptom checker. For instance, 18 vignettes failed on K-Health because their constituent chief complaints were not available in K-Health's search engine, hence, the sessions could not be initiated. Moreover, 35 vignettes failed on K-Health because of an age limitation, whereby only vignettes that encompassed ages of 18 years or more were accepted.

Besides search and age limitations, some symptom checkers (in particular, Buoy) crashed while diagnosing certain vignettes, even after trying multiple times. In addition, many symptom checkers did not produce differential diagnoses for some vignettes albeit concluding the diagnostic sessions. For example, Babylon did not generate differential diagnoses for 351 vignettes. The reason of why some symptom checkers could not produce diagnoses for some vignettes is uncertain, but we conjecture that it might relate to either not modelling the needed diseases or falling short to recall such diseases despite being modelled. Table 3 summarizes the failure rates and reasons across the examined symptom checkers. Alongside, the table shows the average number of questions asked by each symptom checker upon successfully diagnosing vignettes.

Table 3: Failure reasons, failure counts, success counts, and average number of questions across the six tested symptom checkers (DDx = Differential Diagnosis; Qs = Questions).

	Failure Reasons & Counts			Success Counts		Avg. # of Qs
	Search Limitations	Age Limitations	Crashed	No DDx Generated	DDx Generated	
Avey	0	0	0	2	398	24.3
Ada	0	0	0	0	400	29.4
WebMD	2	1	0	3	394	2.64
K-Health	18	35	0	2	345	25.3
Buoy	2	3	5	74	316	25.6
Babylon	15	0	0	351	34	5.9

Figure 3 demonstrates the accuracy results of all the symptom checkers over the 400 vignettes, irrespective of whether they failed or not during some diagnostic sessions[6]. As depicted, Avey outperformed Ada, WebMD, K-Health, Buoy, and Babylon by averages of 24.5%, 175.5%, 142.8%, 159.6%, 2968.1% using $M1$; 22.4%, 114.5%, 123.8%,118.2%, 3392% using $M3$; 18.1%, 79.2%, 116.8%, 125%, 3114.2% using $M5$; 25.2%, 65.6%, 109.4%, 154%, 3545% using recall; 8.7%, 88.9%,66.4%, 88.9%, 2084% using F1-measure; and 21.2%, 93.4%, 113.3%,136.4%, 3091.6% using NDCG. Ada was able to surpass Avey by an average of 0.9% using precision, although Avey significantly outpaced it across all the remaining metrics, even with asking an average of 17.2% lesser number of questions (see Table 3). As shown in Figure 3, Avey also outperformed WebMD, K-Health, Buoy, and Babylon by averages of 103.2%, 40.9%, 49.6%, 1148.5% using precision, respectively.

[5] Information retrieval is a field in computer science, wherein the differential diagnosis problem lies partially under.

[6] In this set of results, a symptom checker is penalized if it fails to start a session, crashes, or does not produce a differential diagnosis albeit concluding a session.

Figure 4 illustrates the accuracy results of all the symptom checkers across only the vignettes that were successful. In other words, symptom checkers were not penalized if they failed to start sessions or crashed during sessions. Nonetheless, Avey still outperformed Ada, WebMD, K-Health, Buoy, and Babylon by averages of 24.5%, 173.2%, 110.9%, 152.8%, 2834.7% using $M1$; 22.4%, 112.4%, 94%, 112.9%, 3257.6% using $M3$; 18.1%, 77.8%, 88.2%, 119.5%, 3003.4% using $M5$; 25.2%, 64.5%, 81.8%, 147.1%, 3371.4% using recall; 8.7%, 87.6%, 44.4%, 83.8%, 1922.2% using F1-measure; and 21.2%, 91.9%, 85%, 130.7%, 2964% using NDCG. Under average precision, Ada outpaced Avey by an average of 0.9%, while Avey surpassed WebMD, K-Health, Buoy, and Babylon by averages of 101.3%, 22%, 45.6%, and 1113.8%, respectively.

Finally, Figure 5 (a) shows the accuracy results of all the symptom checkers over only the vignettes that resulted in differential diagnoses on every symptom checker (i.e., the intersection of successful vignettes with differential diagnoses across all symptom checkers). In this set of results, we excluded Babylon since it failed to produce differential diagnoses for 351 out of the 400 vignettes. As demonstrated in the figure, Avey still outperformed Ada, WebMD, K-Health, and Buoy by averages of 28.1%, 186.9%, 91.5%, 89.3% using $M1$; 22.4%, 116.3%, 85.6%, 59.2% using $M3$; 18%, 80.1%, 85.7%, 65.5% using $M5$; 23%, 64.9%, 78.5%, 97.1% using recall; 7.2%, 92.7%, 42.2%, 47.1% using F1-measure; and 21%, 93.6%, 77.4%, 76.6% using NDCG. Under average precision, Ada surpassed Avey by an average of 2.4%, while Avey outpaced WebMD, K-Health, and Buoy by averages of 109.5%, 20.4%, and 16.9%, respectively.

All the combinations of all the results (i.e., 45 sets of results), including a breakdown between common and less-common diseases, can be found at [49]. In general, Avey demonstrates a superior performance against all the competitor symptom checkers, irrespective of the combination of results.

3.2 Avey versus Human Doctors

In this section, we present our findings of stage 4. As discussed in Section 2.1, we tested the 400 gold-standard vignettes on three doctors with an average clinical experience of 16.6 years. Table 4 shows the results of the doctors across all our accuracy metrics. In addition, Figure 5(b) depicts the results of Avey against *Average MD*, which is the average performance of the three medical doctors. As shown, the human doctors provided average $M1$, $M3$, $M5$, recall, precision, F1-measure, and NDCG of 61.2%, 72.5%, 72.9%, 46.6%, 69.5%, 55.3%, 61.2%, respectively. In contrast, Avey demonstrated average $M1$, $M3$, $M5$, recall, precision, F1-measure, and NDCG of 67.5%, 87.3%, 90%, 72.9%, 43.7%, 54.6%, 76.6%, respectively.

Table 4: Accuracy results (in %) of three medical doctors, MD_1, MD_2, and MD_3, with an average experience of 16.6 years.

	M1	*M3*	*M5*	Recall	Precision	F1-Measure	NDCG
MD_1	49.7	62	62.7	41.2	58.6	48.4	52.2
MD_2	61.3	67.2	67.5	41.2	78.1	53.9	58
MD_3	72.5	88.2	88.5	57.3	71.7	63.7	73.5

To this end, Avey compares favorably to the considered highly experienced doctors, yielding inferior performance in terms of precision and F1-measure, but superior performance in terms of $M1$, $M3$, $M5$, and NDCG. More precisely, the doctors outperformed Avey by averages of 37.1% and 1.2% using precision and F1-measure, while Avey outpaced them by averages of 10.2%, 20.4%, 23.4%, 56.4%, and 25.1% using $M1$, $M3$, $M5$, recall, and NDCG, respectively.

3.3 Ordering of Symptom Checkers and Doctors

We now demonstrate the order of the six considered symptom checkers and three physicians (referred to as MD_1, MD_2, and MD_3) from best-performing to worst-performing under each accuracy metric. Alongside, we report the resultant statistical ranges and standard deviations. Table 5 shows all the results.

4. DISCUSSION

4.1 Principal Findings

In this paper, we capitalized on the standard clinical vignette approach to assess the accuracies of Avey, five popular symptom checkers, and three primary care physicians with an average experience of 16.6 years. We found that Avey significantly outperforms the five symptom checkers and compares favorably to the three physicians. For instance, under $M1$, Avey outperforms the next best-performing symptom checker, namely, Ada, by 24.5% and the worst-performing symptom checker, namely, Babylon, by 2968.2%. On average, Avey outperforms the five symptom checkers by 694.1% using $M1$. In contrast, under $M1$, Avey underperforms the best-performing physician by 6.9% and outperforms the worst-performing one by 35.8%. On average, Avey outperforms the three physicians by 13% using $M1$.

Table 5: Ordering of symptom checkers and physicians (denoted MD_1, MD_2, and MD_3) from best-performing to worst-performing.

Metric	Descending Order (*best to worst*)	Symptom Checkers		Doctors	
		Range (%)	Standard Deviation (%)	Range (%)	Standard Deviation (%)
M1%	MD_3, Avey, MD_2, Ada, MD_1, K Health, Buoy, WebMD, and Babylon	65.3	21	22.8	9
M3%	MD_3, Avey, Ada, MD_2, MD_1, WebMD, Buoy, K Health, and Babylon	84.8	27	26.2	11
M5%	Avey, MD_3, Ada, MD_2, MD_1, WebMD, K Health, Buoy, and Babylon	87.2	27	25.8	11
Average Recall	Avey, Ada, MD_3, WebMD, MD_1 & MD_2 (a tie), K Health, Buoy, and Babylon	70.9	22	16.1	8
Average Precision	MD_3, MD_2, MD_1, Ada, Avey, K Health, Buoy, WebMD, and Babylon	40.6	13	19.5	8
Average F1-Measure	MD_3, Avey, MD_2, Ada, MD_1, K Health, Buoy & WebMD (a tie), and Babylon	32.9	16	15.3	6
Average NDCG	Avey, MD_3, Ada, MD_2, MD_1, WebMD, K Health, Buoy, and Babylon	74.2	23	21.3	9

4.2 Strengths and Limitations

This paper proposed a comprehensive and rigorous experimentation methodology that taps into the standard clinical vignette approach to evaluate symptom checkers and primary care physicians. Based on this methodology, we developed and peer-reviewed the largest benchmark vignette suite in the domain thus far. A recent study utilized 200 vignettes and was deemed one of the most comprehensive to date [33]. The seminal work of [34] utilized 45 vignettes and many studies followed suit [4, 7, 12, 43].

Using this standardized suite, we evaluated the performance of a new AI symptom checker, namely, Avey, five popular symptom checkers, namely, Ada, WebMD, K-Health, Buoy, and Babylon, and a panel of experienced physicians to put things in perspective and interpret results accordingly. To measure accuracy, we used seven standard metrics, one of which was leveraged for the first time in the field to quantify the ranking qualities of symptom checkers' and physicians' differential diagnoses. To minimize bias, the six symptom checkers were tested by only independent primary care physicians and using only peer-reviewed vignettes.

To facilitate the reproducibility of this study and support future related studies, we made all our peer-reviewed vignettes publicly and freely available at [31]. In addition, we posted online all our reported results (e.g., the screenshots of the sessions with symptom checkers and the answers of physicians) at [31, 49] to establish a standard of full transparency and allow for external cross-validation.

This study, however, lacks an evaluation with real patients and covers only 14 body systems with a limited range of conditions. As pointed out in Section 2.1, in the clinical vignette approach, vignettes act as proxies for real patients. The first step in this approach is to standardize these vignettes, which would necessitate an assembly of independent and experienced physicians to review and approve them. Likewise, upon replacing vignettes with real patients, a group of physicians (say, seven, as is the case in this study) is needed to check each patient at the same time and agree by a super-majority vote on their differential diagnosis. This corresponds to standardizing the diagnosis of the patient before they are asked to self-diagnose with each symptom checker. Afterwards, the diagnoses of the symptom checkers can be matched against the patient's standardized diagnosis and accuracy results can be reported accordingly.

Albeit appealing, the above approach differs from the standard clinical vignette approach (no vignettes are involved anymore but actual patients) and is arguably less practical, especially that it suggests checking and diagnosing a vast number of patients before testing on symptom checkers. In addition, the cases of the patients should cover enough diseases (e.g., as in Table 1), which could drastically increase the pool of the patients that need to be diagnosed by physicians before identifying a representative sample. This may explain why this alternative approach has not been utilized in any of the accuracy studies of symptom checkers thus far, granted that the clinical vignette approach is a standard one and further commonly used for testing the diagnostic abilities of physicians [34].

In any of these approaches, it is important to distinguish between *testers* and *subjects*. For instance, in the above alternative approach, the patients are the testers of the symptom checkers *and* the subjects by which the symptom checkers are tested. In contrast, in the clinical vignette approach, the testers are either physicians or laypeople, while the subjects are the standardized vignettes. As discussed in Section 2.1, employing physicians as testers serves in making the clinical vignette approach more reliable. This is because symptom checkers may ask questions that hold no answers in the standardized vignettes, making it difficult for laypeople to answer them appropriately and hard for the community to trust the reported results accordingly.

To this end, two methodologies have been pursued in literature. One is to *dry run* a-priori by a physician every gold-standard vignette on every considered symptom checker and identify every finding (i.e., symptom, etiology, or attribute) that could be asked by these symptom checkers. Subsequently, the physician supplements each vignette with more findings to ensure that laypeople can properly answer any question asked during actual testing. This is the methodology that was used in the seminal work of [11, 34].

The second methodology is not to dry run each vignette beforehand on each symptom checker, especially that it might not be possible to fully know what an AI-based symptom checker will ask during actual testing[7]. On the contrary, the second methodology suggests standardizing the vignettes in a way that precisely reflects real-life patient cases. Afterwards, *multiple* (to address bias and ensure reliability) independent physicians test the vignettes on each symptom checker. These physicians will then reliably answer any questions about any data not held in the vignettes, thus ensuring the correctness of the approach. This methodology has been shown to be more reliable for conducting accuracy studies [33, 43, 44]. As such, it was adopted in the most recent state-of-the-art papers (e.g., [4, 33]) and, consequently, in ours.

Aside from studying the accuracy of symptom checkers, real patients can be involved in testing the *usability* of such tools (e.g., by using a self-completed questionnaire after self-diagnosing with symptom checkers as in [78]). Clearly, this sort of studies is orthogonal to accuracy ones and lies outside the scope of this paper. We plan to conduct a usability study on Avey as discussed in Section 4.5.

Finally, we note that the physicians that were compared against the symptom checkers in stage 4 (i.e., vignette testing on doctors) may not be a representative sample of primary care physicians. Furthermore, our study did not follow a rigorous process to choose symptom checkers and considered only a few of them, which were either popular (i.e., Babylon) or performed superiorly in related recent studies (i.e., Ada, K-Health, Buoy, and WebMD).

4.3 Comparison to the Wider Literature

Much work, especially recently, has been done to study symptom checkers from different perspectives. It is not possible to do justice to this large body of work in this short article. As such, we briefly describe some of the most closely related ones, which focus primarily on the accuracy of self-diagnosis.

Semigran *et al*. [34] were the first to study the performance of many symptom checkers across a range of conditions in 2015. They tested 45 vignettes over 23 symptom checkers and discovered that they vary considerably in terms of accuracy, with $M1$ ranging from 5% to 50% and $M20$ (which measures if a symptom checker returns the gold-standard main diagnosis among its top 20 suggested conditions) ranging from 34% to 84%.

Semigran *et al*. published a follow-up paper [11] in 2016 that compared the diagnostic accuracy of physicians against symptom checkers using the same vignettes in [34]. Results showed that, on average, physicians outperformed symptom checkers (72.1% vs 34.0% along $M1$, and 84.3% vs 51.2% along $M3$). However, symptom checkers were more likely to output the gold-standard main diagnosis at the top of their differentials for low-acuity and common vignettes, while physicians were more likely to do it for high-acuity and uncommon vignettes.

The two studies of Semigran *et al*. [11, 34] provided useful insights into the first generation of symptom checkers. However, much has changed since 2015-2016. To exemplify, Gilbert *et al*. [33] recently compiled, peer-reviewed, and tested 200 vignettes over 8 popular symptom checkers and 7 General Practitioners (GPs). As in [34], they found a significant variance in the performance of symptom checkers, but a promise in the accuracy of a new symptom checker named Ada [26]. Ada exhibited accuracies of 49%, 70.5%, and 78% for $M1$, $M3$, and $M5$, respectively.

None of the symptom checkers in [33] outperformed GPs, but Ada came close, especially in $M3$ and $M5$. The authors of [33] pointed out that the nature of iterative improvements in software suggests an expected increase in the future performance of symptom checkers, which may at a point in time exceed that of GPs. As illustrated in Figure 3, we found that Ada is still largely ahead of the conventional symptom checkers, but Avey outperforms it. Furthermore, Avey surpasses physicians under various accuracy metrics as depicted in Figure 5 (b).

Hill *et al*. [4] evaluated 36 symptom checkers, 8 of which use AI, over 48 vignettes. They showed that accuracy varies considerably across symptom checkers, ranging from 12% to 61% using $M1$ and from 30% to 81% using $M10$ (where the correct diagnosis appears among the top 10 conditions). They also observed that AI-based symptom checkers outperform rule-based ones

[7] Some symptom checkers demonstrate *nondeterministic* behaviors, thus might ask some different questions at different times.

(i.e., symptom checkers that do not use AI). Akin to Hill *et al.* [4], Ceney *et al*. [12] detected a significant variation in accuracy across 12 symptom checkers, ranging from 22.2% (CAIDR [50]) to 72% (Ada) using *M*5.

Many other studies focused on the diagnostic performance of symptom checkers, but only across a limited set of diagnoses [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73]. For instance, Shen *et al.* [72] evaluated the accuracy of WebMD for ophthalmic diagnoses. Hennemann *et al.* [67] investigated the diagnostic performance of Ada for mental disorders. Nateqi *et al.* [70] validated the accuracies of Symptoma [74], Ada, FindZebra [75], Mediktor [76], Babylon, and Isabel [77] for ENT conditions. Lastly, Munsch *et al.* [69] assessed the accuracies of 10 web-based COVID- 19 symptom checkers.

Miller *et al.* [78] presented a real-world usability study of Ada over 523 participants (patients) in a South London primary care clinic over a period of 3 months. Nearly all patients (i.e., 97.8%) found Ada very easy to use. In addition, 22% of patients between ages of 18 and 24 suggested that using Ada before coming to the clinic would have changed their minds in terms of what care to consider next. Studies of other symptom checkers like Buoy and Isabel reported high degrees of utility as well [24, 79].

Some work has also explored the triage capabilities of symptom checkers [7, 43, 79, 80]. Studying the utility and triage capabilities of symptom checkers are beyond the scope of this paper and have been set as future work in Section 4.5.

Early AI models for medical diagnosis adopted expert systems [46, 51, 52, 53, 54]. Subsequent models employed probabilistic formulations to account for uncertainty in the diagnostic process [55] and focused on approximate probabilistic inference to optimize for efficiency [56, 57, 58].

With the increasing availability of Electronic Medical Records (EMRs), Rotmensch *et al.* [59] utilized Logistic regression (LR), naive Bayes (NB), and Bayesian networks with noisy OR gates (noisy OR) on EMRs to automatically construct medical knowledge graphs. Miotto *et al.* [60] proposed an EMR-based unsupervised deep learning approach to derive a general-purpose patient representation and facilitate clinical predictive modelling. Ling *et al.* [61] modeled the problem as a sequential decision-making process using deep reinforcement learning. Kannan *et al.* [46] showed that multiclass logistic regression and deep learning models can be effective in generalizing to new patient cases, but with an accuracy caveat concerning the number of diseases that can be incorporated.

4.4 Implications for Clinicians and Policymakers

As pointed out in Section 1, a UK-based study that engaged 1,071 patients found that more than 70% of individuals between the ages of 18 and 39 years would use a symptom checker [13]. This study was influential in the UK health policy circles, whereby it received press attention and prompted responses from NHS England and NHSX, a UK government policy unit that develops best practices and national policies for technology in health [78, 81]. Given that symptom checkers vary considerably in performance (as demonstrated in Section 3.1), this paper serves in scientifically informing patients, clinicians, and policymakers about the current accuracies of some of these symptom checkers.

Besides, this study advocates for extensive internal testing and rigorous external validation by independent physicians for any symptom checker before it is publicly launched. The work on Avey took around 4 years and was conducted by a professional team of medical doctors and computer scientists. Avey was launched only after it was verified and tested in-house over thousands of medical cases.

Lastly, this study suggests that any external scientific validation of any AI-based medical diagnostic algorithm should be fully transparent and eligible for replication. As a direct translation to this suggestion, we posted all the results of the tested symptom checkers and physicians online as a proof-of-work and to allow for cross-verification and study-replication. Moreover, we made all our peer-reviewed vignettes publicly and freely available. This will not only enable reproducing this study, but further supporting future related studies in academia and industry alike.

4.5 Unanswered Questions and Future Research

This paper focused solely on studying the diagnostic accuracies of symptom checkers. As such, we set forth 3 immediate and complementary future directions, namely, *usability*, *utility*, and *extendibility* ones. To elaborate, we will first study the usability and acceptability of Avey with real patients. In particular, we will investigate how patients will perceive Avey and interact with it. During this study, we will observe and identify any barrier in Avey's UX/UI and language aspects. Afterwards, we will incorporate necessary changes to make Avey's interface more friendly (e.g., through using sound rather than only text). Second, we will examine how patients will respond to Avey's output and gauge its influence on their subsequent choices for care. Finally, we will extend Avey's AI model to involve triage and measure its economic impact on patients and healthcare systems.

5. CONCLUSIONS

AI-based symptom checkers that undergo rigorous development and testing have the potential to become useful tools for timely, accurate, and instant self-diagnosis. In this paper, we introduced Avey, a new AI-based symptom checker that was extensively researched, designed, developed, and tested for around 4 years before it was launched. We further proposed an experimentation methodology to evaluate Avey against popular symptom checkers and seasoned primary care physicians. Results showed that Avey significantly outperforms the considered symptom checkers and compares favorably to physicians. In the future, we will extend Avey's AI model to involve triage and study its usability for real patients and utility for healthcare systems.

Acknowledgements: Vignette setting was carried out with the help of the following independent and experienced physicians: Dr. Azmi Qudsi, Dr. Doaa Eisa, and Dr. Muna Yousif. Vignette review (i.e., vignette standardization, or stage 2 of our experimentation methodology) was carried out by the following independent and experienced physicians: Dr. Zaid Abu Saleh, Dr. Odai Al-Batsh, Dr. Ahmad Alowaidat, Dr. Tamara Altawara, Dr. Arwa Khashan, Dr. Muna Darmach, and Dr. Nour Essale. Vignette testing on symptom checkers (i.e., stage 3 of our experimentation methodology) was carried out by the following independent and experienced physicians: Dr. Maram Alsmairat, Dr. Muna Darmach, and Dr. Ahmad Kakakan. Vignette testing on doctors (i.e., stage 4 of our experimentation methodology) was carried out by the following independent and experienced physicians: Dr. Mohmmad Almadani, Dr. Tala Hamouri, and Dr. Noor Jodeh.

Contributors: MH conceived the study, designed the experimentation methodology, and supervised the project. SD coordinated the work within and across the project stages (e.g., coordination of vignette creation, vignette standardization, vignette testing on symptom checkers, and vignette testing on doctors). MH conducted the literature review and documentation. SD, MD, and SA created the vignettes and verified the testing results. MD and SS carried out results compilation and summarization. MH and SS carried out data analysis and interpretation. YK developed the web portal for streamlining the processes of reviewing, standardizing, and testing the vignettes. SS maintained Avey's software and provided technical support. MH wrote the paper. All authors reviewed and commented on drafts of the paper. MH provided administrative support. MH is the guarantor for this work.

Funding: There are no funders to report for this submission.

Competing interests: All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/disclosure-of-interest/. All authors are employees of Rimads QSTP-LLC, which is the manufacturer of Avey (see authors' affiliations). MH is the founder and CEO of Rimads QSTP-LLC and holds equity in Rimads QSTP-LLC. The authors have no support from any organization for the submitted work; no financial relationships with any organizations that might have interests in the submitted work; and no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: No patients were involved in any part of this study, but rather vignettes that acted as proxies for patients. In addition, doctors were not subjects in stage 4 of the study (or any stage as a matter of fact), but rather the vignettes themselves. As such, no IRB approval is required.

Patient consent for publication: Not required.

Data sharing: All our gold-standard vignettes are made publicly and freely available at https://doi.org/10.6084/m9.figshare.21221036.v4 to enable the reproducibility of this work. In addition, all the outputs of the symptom checkers and physicians are posted at the same site to allow for external cross-validation. Lastly, the results of all our 45 sets of experiments are published at https://doi.org/10.6084/m9.figshare.21221006.v5 to establish a standard of full transparency.

REFERENCES

[1] Janet M Morahan-Martin. 2004. How internet users find, evaluate, and use online health information: a cross-cultural review. CyberPsychology & Behavior 7, 5 (2004), 497–510.

[2] Jeremy C Wyatt. 2015. Fifty million people use computerised self triage.

[3] Christina Cheng and Matthew Dunn. 2015. Health literacy and the Internet: a study on the readability of Australian online health information. Australian and New Zealand journal of public health 39, 4 (2015), 309–314.

[4] Michella G Hill, Moira Sim, and Brennen Mills. 2020. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Medical Journal of Australia 212, 11 (2020), 514–519.

[5] David M Levine and Ateev Mehrotra. 2021. Assessment of Diagnosis and Triage in Validated Case Vignettes Among Nonphysicians Before and After Internet Search. JAMA network open 4, 3 (2021), e213287–e213287.

[6] Seth S Martin, Emmanuel Quaye, Sarah Schultz, Oluwaseun E Fashanu, Jane Wang, Mustapha O Saheed, Prem Ramaswami, Hermes de Freitas, Berthier Ribeiro-Neto, and Kapil Parakh. 2019. A randomized controlled trial of on- line symptom searching to inform patient generated differential diagnoses. NPJ digital medicine 2, 1 (2019), 1–6.

[7] Malte L Schmieding, Rudolf Mörgeli, Maike AL Schmieding, Markus A Feufel, and Felix Balzer. 2021. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. Journal of medical Internet research 23, 3 (2021), e24475.

[8] Norman Bates. 2014. Don't google it. https://vimeo.com/115097884. [Online; accessed 08-Jan-2022].

[9] Sarah Larimer. 2014. Can this ad campaign get people in Belgium to stop Googling their symptoms? https://www.washingtonpost.com/news/to-your-health/wp/2014/11/11/can-this-ad-campaign-get-people-in-belgium-to-stop- googling-their-symptoms/. [Online; accessed 08-Jan-2022].

[10] Stephanie Aboueid, Samantha Meyer, James R Wallace, Shreya Mahajan, and Ashok Chaurasia. 2021. Young Adults' Perspectives on the Use of Symptom Checkers for Self-Triage and Self-Diagnosis: Qualitative Study. JMIR Public Health and Surveillance 7, 1 (2021), e22637.
[11] Hannah L Semigran, David M Levine, Shantanu Nundy, and Ateev Mehrotra. 2016. Comparison of physician and computer diagnostic accuracy. JAMA internal medicine 176, 12 (2016), 1860–1861.
[12] Adam Ceney, Stephanie Tolond, Andrzej Glowinski, Ben Marks, Simon Swift, and Tom Palser. 2021. Accuracy of online symptom checkers and the potential impact on service utilisation. Plos one 16, 7 (2021), e0254088.
[13] Healthwatch Enfield. 2019. Using technology to ease the burden on primary care. https://www.healthwatch.co.uk/reports-library/using-technology-ease- burden-primary-care. [Online; accessed 08-Jan-2022].
[14] Ashley ND Meyer, Traber D Giardina, Christiane Spitzmueller, Umber Shahid, Taylor MT Scott, and Hardeep Singh. 2020. Patient Perspectives on the Usefulness of an Artificial Intelligence–Assisted Symptom Checker: Cross-Sectional Survey Study. Journal of medical Internet research 22, 1 (2020), e14679.
[15] NHS Services. 2019. Babylon GP at hand. https://www.gpathand.nhs.uk/our- nhs-service. [Online; accessed 08-Jan-2022].
[16] Government of Australia. 2019. Healthdirect Symptom Checker. https://www. healthdirect.gov.au/symptom-checker. [Online; accessed 08-Jan-2022].
[17] Wouter A Spoelman, Tobias N Bonten, Margot WM de Waal, Ton Drenthen, Ivo JM Smeele, Markus MJ Nielen, and Niels H Chavannes. 2016. Effect of an evidence-based website on healthcare usage: an interrupted time-series study. BMJ open 6, 11 (2016), e013166.
[18] Stephanie Aboueid, Rebecca H Liu, Binyam Negussie Desta, Ashok Chaurasia, and Shanil Ebrahim. 2019. The use of artificially intelligent Self-Diagnosing digital platforms by the general public: Scoping review. JMIR medical informatics 7, 2 (2019), e13445.
[19] Shannon Brownlee, Kalipso Chalkidou, Jenny Doust, Adam G Elshaug, Paul Glasziou, Iona Heath, Somil Nagpal, Vikas Saini, Divya Srivastava, Kelsey Chalmers, et al. 2017. Evidence for overuse of medical services around the world. The Lancet 390, 10090 (2017), 156–168.
[20] Daniel J Morgan, Sanket S Dhruva, Scott M Wright, and Deborah Korenstein. 2016. 2016 update on medical overuse: a systematic review. JAMA internal medicine 176, 11 (2016), 1687–1692.
[21] Canadian Institute of Health Information. 2017. Unnecessary Care in Canada. https://www.cihi.ca/sites/default/files/document/choosing-wisely- baseline-report-en-web.pdf. [Online; accessed 08-Jan-2022].
[22] Stephanie Aboueid, Samantha B Meyer, James R Wallace, Shreya Mahajan, Teeyaa Nur, and Ashok Chaurasia. 2021. Use of symptom checkers for COVID-19-related symptoms among university students: a qualitative study. BMJ Innovations 7, 2 (2021).
[23] Saba Akbar, Enrico Coiera, and Farah Magrabi. 2020. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. Journal of the American Medical Informatics Association 27, 2 (2020), 330–340.
[24] Hamish Fraser, Enrico Coiera, and David Wong. 2018. Safety of patient-facing digital symptom checkers. The Lancet 392, 10161 (2018), 2263–2264.
[25] Marise J Kasteleyn, Anke Versluis, Petra van Peet, Ulrik Bak Kirk, Jens van Dalfsen, Eline Meijer, Persijn Honkoop, Kendall Ho, Niels H Chavannes, and Esther PWA Talboom-Kamp. 2021. SERIES: eHealth in primary care. Part 5: A critical appraisal of five widely used eHealth applications for primary care– opportunities and challenges. European Journal of General Practice 27, 1 (2021), 248–256.
[26] Ada Health. 2011. Symptom Checker. https://ada.com/. [Online; accessed 07-Jan-2022].
[27] K Health. 2016. Symptom Checker. https://khealth.com/. [Online; accessed 07-Jan-2022].
[28] Buoy Health. 2014. Symptom Checker. https://www.buoyhealth.com/. [Online; accessed 07-Jan-2022].
[29] Babylon Health. 2013. Symptom Checker. https://www.babylonhealth.com/. [Online; accessed 07-Jan-2022].
[30] WebMD. 1996. Symptom Checker. https://www.webmd.com/. [Online; accessed 07-Jan-2022].
[dataset][31] Hammoud, Mohammad; Douglas, Shahd; Darmach, Mohamad; Sanyal, Swapnendu; Alawneh, Sara; Kanbour, Youssef (2022): Evaluating the Accuracy of a New Artificial Intelligence Based Symptom Checker: A Clinical Vignette Study: Vignette Suite and Screenshots. figshare. Dataset. https://doi.org/10.6084/m9.figshare.21221036.v4
[32] Enrico Coiera, Elske Ammenwerth, Andrew Georgiou, and Farah Magrabi. 2018. Does health informatics have a replication crisis? Journal of the American Medical Informatics Association 25, 8 (2018), 963–968.
[33] Stephen Gilbert, Alicia Mehl, Adel Baluch, Caoimhe Cawley, Jean Challiner, Hamish Fraser, Elizabeth Millen, Maryam Montazeri, Jan Multmeier, Fiona Pick, et al. 2020. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ open 10, 12 (2020), e040269.
[34] Hannah L Semigran, Jeffrey A Linder, Courtney Gidengil, and Ateev Mehrotra. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. bmj 351 (2015).
[35] United States Medical Licensing Examination. 2019-2020. USMLE Step 2 CK. https://www.usmle.org/step-exams/step-2-ck. [Accessed 05-Feb-2022].
[36] John D Firth and Ian Gilmore. 2008. MRCP Part 1 Self-Assessment: Medical Masterclass Questions and Explanatory Answers. Radcliffe Publishing.
[37] Doug Knutson. 2018. Family Medicine PreTest Self-Assessment And Review. Mc- Graw Hill Professional.
[38] American Board of Family Medicine. 2018. In-Training Examination. https:
[39] American Academy of Pediatrics. 2020. 2021 PREP Self-Assessment. https://www.aap.org/. [Accessed 05-Feb-2022].
[40] CRC Press. 2011-2020. 100 Cases Book Series. https://www.routledge.com/100- Cases/book-series/CRCONEHUNCAS. [Accessed 05-Feb-2022].
[41] Alfred F Tallia, Joseph E Scherger, and Nancy Dickey. 2017. Swanson's Family Medicine Review E-Book. Elsevier Health Sciences.
[42] Ian Wilkinson, Ian Boden Wilkinson, Tim Raine, Kate Wiles, Anna Goodhart, Catriona Hall, and Harriet O'Neill. 2017. Oxford handbook of clinical medicine. Oxford university press.
[43] Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Mobasher Butt, Arnold DoRosario, and Saurabh Johri. 2020. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. Frontiers in artificial intelligence 3 (2020), 100.
[44] Stefanie Maria Jungmann, Timo Klan, Sebastian Kuhn, and Florian Jungmann. 2019. Accuracy of a Chatbot (ADA) in the diagnosis of mental disorders: compar- ative case study with lay and expert users. JMIR formative research 3, 4 (2019), e13863.
[45] Pueyo Ferrer, Martín Baranera, et al. 2017. A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application. Emergencias: revista de la Sociedad Espanola de Medicina de Emergencias 29, 6 (2017), 391–396.
[46] Anitha Kannan, Jason Alan Fries, Eric Kramer, Jen Jen Chen, Nigam Shah, and Xavier Amatriain. 2020. The accuracy vs. coverage trade-off in patient-facing diagnosis models. AMIA Summits on Translational Science Proceedings 2020 (2020), 298.
[47] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 20, 4 (2002), 422–446.
[48] Xue Zhao. 2013. A Theoretical Analysis of NDCG Ranking Measures. (2013).
[dataset][49] Hammoud, Mohammad; Douglas, Shahd; Darmach, Mohamad; Alawneh, Sara; Sanyal, Swapnendu; Kanbour, Youssef (2022): Evaluating the Accuracy of a New Artificial Intelligence Based Symptom Checker: A Clinical Vignette Study: results document. figshare. Dataset. https://doi.org/10.6084/m9.figshare.21221006.v5
[50] Caidr. 2006. Symptom Checker. https://caidr.squarespace.com/. [Online; accessed 08-Jan-2022].
[51] G Octo Barnett, James J Cimino, Jon A Hupp, and Edward P Hoffer. 1987. DXplain: an evolving diagnostic decision-support system. Jama 258, 1 (1987), 67–74.
[52] Bruce G Buchanan and Edward H Shortliffe. 1984. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. (1984).
[53] Tommi S. Jaakkola and Michael I. Jordan. 1999. Variational Probabilistic Inference and the QMR-DT Network. J. Artif. Int. Res. 10, 1 (may 1999), 291–322.
[54] S. L. Lauritzen and D. J. Spiegelhalter. 1988. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. Journal of the Royal Statistical Society. Series B (Methodological) 50, 2 (1988), 157–224. http:
[55] Randolph A Miller, Harry E Pople, and Jack D Myers. 1985. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. In Computer-assisted medical decision making. Springer, 139–158.
[56] Quaid Morris. 2001. Recognition Networks for Approximate Inference in BN20 Networks (UAI'01). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 370–377.
[57] Anne Marie Rassinoux, RA Miller, RH Baud, and JR Scherrer. 1996. Modeling principles for QMR medical findings.. In Proceedings of the AMIA Annual Fall Symposium. American Medical Informatics Association, 264.

[58] Michael Shwe and Gregory Cooper. 1991. An empirical analysis of likelihood- weighting simulation on a large, multiply connected medical belief network. Computers and Biomedical Research 24, 5 (1991), 453–475. https://doi.org/10. 1016/0010-4809(91)90020-W
[59] Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. 2017. Learning a health knowledge graph from electronic medical records. Scientific reports 7, 1 (2017), 1–11.
[60] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports 6, 1 (2016), 1–10.
[61] Yuan Ling, Sadid A. Hasan, Vivek Datla, Ashequl Qadir, Kathy Lee, Joey Liu, and Oladimeji Farri. 2017. Diagnostic Inferencing via Improving Clinical Con- cept Extraction with Deep Reinforcement Learning: A Preliminary Study. In Proceedings of the 2nd Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research), Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (Eds.), Vol. 68. PMLR, 271–285. https://proceedings.mlr.press/v68/ling17a.html
[62] Andrew C Berry, Nicholas A Berry, Bin Wang, Madhuri Mulekar, Anne Melvin, Richard J Battiola, Frederick K Bulacan, and Bruce B Berry. 2018. Use of online symptom checkers to delineate the ever-elusive GERD versus non-GERD cough. The clinical respiratory journal 12, 12 (2018), 2683.
[63] Andrew C Berry, Nicholas A Berry, Bin Wang, Madhuri S Mulekar, Anne Melvin, Richard J Battiola, Frederick K Bulacan, and Bruce B Berry. 2020. Symptom checkers versus doctors: A prospective, head-to-head comparison for cough. The clinical respiratory journal 14, 4 (2020), 413–415.
[64] Leslie J Bisson, Jorden T Komm, Geoffrey A Bernas, Marc S Fineberg, John M Marzo, Michael A Rauh, Robert J Smolinski, and William M Wind. 2014. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. The American journal of sports medicine 42, 10 (2014), 2371–2376.
[65] Aleksandar Ćirković. 2020. Evaluation of Four Artificial Intelligence–Assisted Self-Diagnosis Apps on Three Diagnoses: Two-Year Follow-Up Study. Journal of medical Internet research 22, 12 (2020), e18097.
[66] SE Farmer, Matteo Bernardotto, and V Singh. 2011. How good is Internet self- diagnosis of ENT symptoms using Boots WebMD symptom checker? Clinical otolaryngology: official journal of ENT-UK; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial Surgery 36, 5 (2011), 517–518.
[67] Severin Hennemann, Sebastian Kuhn, Michael Witthöft, Stefanie M Jungmann, et al. 2022. Diagnostic Performance of an App-Based Symptom Checker in Mental Disorders: Comparative Study in Psychotherapy Outpatients. JMIR Mental Health 9, 1 (2022), e32832.
[68] Tana M Luger, Thomas K Houston, and Jerry Suls. 2014. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. Journal of medical Internet research 16, 1 (2014), e2924.
[69] Nicolas Munsch, Alistair Martin, Stefanie Gruarin, Jama Nateqi, Isselmou Ab- darahmane, Rafael Weingartner-Ortner, Bernhard Knapp, et al. 2020. Diagnostic accuracy of web-based COVID-19 symptom checkers: comparison study. Journal of medical Internet research 22, 10 (2020), e21299.
[70] Krobath H Gruarin S Lutz T Dvorak T Gruschina A Ortner R. Nateqi J, Lin S. 2019. From symptom to diagnosis-symptom checkers re-evaluated : Are symp- tom checkers finally sufficient and accurate to use? An update from the ENT perspective. HNO (2019).
[71] Aimee E Poote, David P French, Jeremy Dale, and John Powell. 2014. A study of automated self-assessment in a primary care student health centre setting. Journal of telemedicine and telecare 20, 3 (2014), 123–127.
[72] Carl Shen, Michael Nguyen, Alexander Gregor, Gloria Isaza, and Anne Beattie. 2019. Accuracy of a popular online symptom checker for ophthalmic diagnoses. JAMA ophthalmology 137, 6 (2019), 690–692.
[73] Yuya Yoshida and Glenn Thomas Clark. 2021. Accuracy of online symptom checkers for diagnosis of orofacial pain and oral medicine disease. Journal of Prosthodontic Research 65, 2 (2021), 186–190.
[74] Symptoma. 2009. Symptom Checker. https://www.symptoma.com/. [Online; accessed 19-March-2022].
[75] FindZebra. 2013. Symptom Checker. https://www.findzebra.com/. [Online; accessed 19-March-2022].
[76] Mediktor. 2011. Symptom Checker. https://www.mediktor.com/en-us. [Online; accessed 19-March-2022].
[77] Isabel. 1999. Symptom Checker. https://symptomchecker.isabelhealthcare.com/. [Online; accessed 08-Jan-2022].
[78] Stephen Miller, Stephen Gilbert, Vishaal Virani, and Paul Wicks. 2020. Patients' utilization and perception of an artificial intelligence–based symptom assessment and advice technology in a British primary care waiting room: exploratory pilot study. JMIR human factors 7, 3 (2020), e19713.
[79] Aaron N Winn, Melek Somai, Nicole Fergestrom, and Bradley H Crotty. 2019. Association of use of online symptom checkers with patients' plans for seeking care. JAMA network open 2, 12 (2019), e1918561–e1918561.
[80] Fatma Mansab, Sohail Bhatti, and Daniel Goyal. 2021. Reliability of COVID-19 symptom checkers as national triage tools: an international case comparison study. BMJ Health & Care Informatics 28, 1 (2021).
[81] Ingrid Torjesen. 2019. Patients find GP online services "cumbersome," survey finds.

Figure 1: An actual visualization of Avey's *brain* (i.e., a probabilistic graphical model). At a high level, the nodes can be thought of representing diseases and findings, while the edges can be viewed as encompassing conditional independence assumptions and modelling clinical reasoning metrics.

Figure 2: Our 4-stage experimentation methodology (V_i = Vignette i, assuming n vignettes and $1 \leq i \leq n$; D_j = Doctor j, assuming 7 doctors and $1 \leq j \leq 7$; MD_k = Medical Doctor k, assuming 3 doctors and $1 \leq k \leq 3$; R_i = Result of vignette V_i as generated by a symptom checker or an MD). In the "vignette creation" stage, the vignettes are compiled from reputable medical sources by an internal team of medical doctors. In the "vignette standardization" stage, the vignettes are reviewed and approved by a panel of experienced and independent physicians. In the "vignette testing on symptom checkers" stage, the vignettes are tested on symptom checkers by a different panel of experienced and independent physicians. In the "vignette testing on doctors" stage, the vignettes are tested on a yet different panel of experienced and independent physicians.

Figure 3: Accuracy results considering for each symptom checker all the succeeded and failed vignettes.

Figure 4: Accuracy results considering for each symptom checker only the succeeded vignettes, with or without differential diagnoses.

Figure 5: (a) Accuracy results considering only the succeeded vignettes with differential diagnoses across all symptom checkers, and (b) accuracy results of Avey versus three medical doctors, on average (i.e., Average MD).