Original Paper

Assessment of a Digital Symptom Checker Tool's Accuracy in Suggesting Reproductive Health Conditions: Clinical Vignettes Study

Kimberly Peven[1], BSN, MPH, PhD; Aidan P Wickham[1], MEng, MRes, PhD; Octavia Wilks[1], BSc, MPH, MBBCh; Yusuf C Kaplan[1], MD; Andrei Marhol[1], MD, PhD; Saddif Ahmed[1], BSc, MBBChB, MSc, DPhil; Ryan Bamford[1], BSc, MSc; Adam C Cunningham[1], PhD; Carley Prentice[1], BA; András Meczner[2], MD; Matthew Fenech[3], MD, PhD; Stephen Gilbert[4], PhD; Anna Klepchukova[1], MD; Sonia Ponzo[1*], BSc, MSc, PhD; Liudmila Zhaunova[1*], BSc, PhD

[1]Flo Health UK Limited, London, United Kingdom

[2]Your.MD Limited, London, United Kingdom

[3]Una Health GmbH, Hamburg, Germany

[4]Else Kröner Fresenius Center for Digital Health, TUD Dresden University of Technology, Dresden, Germany

[*]these authors contributed equally

Corresponding Author:
Aidan P Wickham, MEng, MRes, PhD
Flo Health UK Limited
27 Old Gloucester Street
London, WC1N 3AX
United Kingdom
Phone: 44 60396823
Email: a_wickham@flo.health

Abstract

Background: Reproductive health conditions such as endometriosis, uterine fibroids, and polycystic ovary syndrome (PCOS) affect a large proportion of women and people who menstruate worldwide. Prevalence estimates for these conditions range from 5% to 40% of women of reproductive age. Long diagnostic delays, up to 12 years, are common and contribute to health complications and increased health care costs. Symptom checker apps provide users with information and tools to better understand their symptoms and thus have the potential to reduce the time to diagnosis for reproductive health conditions.

Objective: This study aimed to evaluate the agreement between clinicians and 3 symptom checkers (developed by Flo Health UK Limited) in assessing symptoms of endometriosis, uterine fibroids, and PCOS using vignettes. We also aimed to present a robust example of vignette case creation, review, and classification in the context of predeployment testing and validation of digital health symptom checker tools.

Methods: Independent general practitioners were recruited to create clinical case vignettes of simulated users for the purpose of testing each condition symptom checker; vignettes created for each condition contained a mixture of condition-positive and condition-negative outcomes. A second panel of general practitioners then reviewed, approved, and modified (if necessary) each vignette. A third group of general practitioners reviewed each vignette case and designated a final classification. Vignettes were then entered into the symptom checkers by a fourth, different group of general practitioners. The outcomes of each symptom checker were then compared with the final classification of each vignette to produce accuracy metrics including percent agreement, sensitivity, specificity, positive predictive value, and negative predictive value.

Results: A total of 24 cases were created per condition. Overall, exact matches between the vignette general practitioner classification and the symptom checker outcome were 83% (n=20) for endometriosis, 83% (n=20) for uterine fibroids, and 88% (n=21) for PCOS. For each symptom checker, sensitivity was reported as 81.8% for endometriosis, 84.6% for uterine fibroids, and 100% for PCOS; specificity was reported as 84.6% for endometriosis, 81.8% for uterine fibroids, and 75% for PCOS; positive predictive value was reported as 81.8% for endometriosis, 84.6% for uterine fibroids, 80% for PCOS; and negative predictive value was reported as 84.6% for endometriosis, 81.8% for uterine fibroids, and 100% for PCOS.

Conclusions: The single-condition symptom checkers have high levels of agreement with general practitioner classification for endometriosis, uterine fibroids, and PCOS. Given long delays in diagnosis for many reproductive health conditions, which lead

to increased medical costs and potential health complications for individuals and health care providers, innovative health apps and symptom checkers hold the potential to improve care pathways.

(JMIR Mhealth Uhealth 2023;11:e46718) doi: 10.2196/46718

KEYWORDS

women's health; symptom checkers; symptom checker; digital health; chatbot; accuracy; eHealth apps; mobile phone; mobile health; mHealth; mobile health app; polycystic ovary syndrome; gynecology; digital health tool; endometriosis; uterus; uterine; uterine fibroids; vignettes; clinical vignettes

Introduction

Background

Millions of women and people who menstruate worldwide are affected by reproductive health conditions. Endometriosis, uterine fibroids, and polycystic ovary syndrome (PCOS) are among the most common with prevalences estimated at 10%-15%, 20%-40%, and 5%-20%, respectively [1-12]. All 3 conditions have been associated with fertility issues [12-14]. Endometriosis is a condition where endometrial tissue is found outside of the uterus and is typically characterized by painful periods, abnormal bleeding, and chronic pelvic pain, among other symptoms [5,14,15]. Uterine fibroids are benign uterine tumors that can cause a variety of debilitating symptoms, such as heavy menstrual bleeding, pain, and bladder or bowel dysfunction [12,16]. PCOS is a complex endocrine disorder characterized by a variety of symptoms, such as menstrual dysfunction and hirsutism, of differing severity and without a certain etiology [13]. These conditions can have similar presentations; for example, both pain and intermenstrual bleeding are symptoms of endometriosis and uterine fibroids, and additionally, these conditions can coexist in individuals at the same time.

Long diagnostic delays are common for endometriosis, uterine fibroids, and PCOS, with patients reporting receiving a diagnosis between 2 and 12 years from the onset of symptoms [17-22]. Controversy over diagnostic criteria may further complicate or delay final diagnosis [12,23-25]. Another contributing factor to diagnostic delays is a low level of knowledge on reproductive health as affected persons may believe symptoms are normal or hereditary, thus delaying in seeking medical input until symptoms worsen [26]. Delays in diagnosis can lead to worsening of symptoms, further health complications with fertility or psychiatric conditions, and a reduced quality of life [26-31]. Both endometriosis and uterine fibroids severely affect quality of life, everyday functioning, and workplace productivity [32-36]. A common sequela of PCOS is also a lowered quality of life [37] but also includes infertility, type 2 diabetes, and cardiovascular and psychiatric conditions (eg, hypertension, depression, and anxiety) [38].

In addition to risks of developing complications with fertility or psychiatric conditions [27-30], long diagnostic delays are associated with increased health care use and costs [39]. Endometriosis costs an average of US $27,855 per patient annually in the United States alone [40], while overall yearly expenditure for uterine fibroids is estimated to be US $34.4 billion [35]. Further, patients with long diagnostic delays for endometriosis have 60% higher mean all-cause costs compared to those with short delays [39]. Similarly, the economic costs of PCOS on individuals and health care systems are estimated to be US $8 billion per year [8]. As diagnostic costs represent a small proportion of the total economic burden of disease, particularly in light of long diagnostic delays, access to simpler screening processes may be a cost-effective strategy [41].

Innovations in health technologies and mobile apps have the potential to bridge this economic gap, deliver better health outcomes, and improve quality of life. Worldwide, there are more than 6 billion smartphone subscribers [42] and more than 350,000 health-related mobile apps [43]. As such, people increasingly turn to the internet for health information [44-46], with an increasing number of digital health interventions existing to assist with condition diagnosis (eg, check user symptoms against common condition symptoms) [47,48].

Despite the widespread availability and advantages of symptom checker apps, there remains a knowledge gap on the accuracy of many of these tools [49]. Researchers, clinicians, and patient groups are increasingly demanding more rigorous validation and evaluation of digital health solutions, with scientists highlighting the need for evidence generation [50-53]. Case vignette studies represent an established methodology for the evaluation of digital symptom checkers. In such studies, relevant fictitious patient cases are assessed by the symptom checker under investigation, and the output is compared to that of a human expert assessing the same case [54]. However, several scoping reviews have identified significant variability in study designs and reported quantitative measures when assessing symptom checkers, with about half reports describing app characteristics and half examining actual accuracy metrics, which were found to vary greatly [49,55]. A recent review of digital and web-based symptom checkers found diagnostic accuracy of the primary diagnosis varied from 19% to 38% and triage accuracy ranged from 49% to 90% [56]. Even though information on their development and validation is limited and its reliability is in question [47,49], trust in symptom checker apps is high among laypersons [57].

Flo App and Symptom Checker Development

Flo (by Flo Health UK Limited) is a health and well-being mobile app and period tracker for women and people who menstruate, with over 58 million monthly active users [58]. Flo allows users to track their symptoms throughout their menstrual cycle (eg, cramps, menstrual flow, and mood) or pregnancy and postpartum (eg, lochia), as well as general health information like contraceptive use, ovulation or pregnancy test results, water intake, and sleep. Additionally, the app offers personalized, evidence-based, and expert-reviewed content via an in-app

library. Further, digital health assistants (chatbots) provide users with information about a range of conditions.

Flo has developed 3 single-condition symptom checker "chatbots" to assess symptoms of reproductive health conditions (endometriosis, uterine fibroids, and PCOS). The decision to focus on these conditions was based on their prevalence, the feasibility of symptom assessment via an app, and the multifactorial impact that these conditions can have (eg, quality of life, productivity, cardiovascular diseases, mental health conditions, and fertility). The symptom checkers use symptom information gained through conversation-like questions and answers as well as symptom or menstrual cycle information previously entered into the app. Users with acute presentations are provided with a list of red flag symptoms (eg, nausea with vomiting, fever, and vaginal bleeding not related to the period) at the beginning of the conversation and are advised to discontinue the conversation with the symptom checker and seek urgent medical advice if their presence is confirmed by the user. After the conversation, the symptom checker gives the user one of two possible outcomes: (1) a strong match for the condition—"You're experiencing several symptoms typically associated with [condition]" or (2) a weak or no match for the condition—"While you may be experiencing some symptoms of [condition], your combination of symptoms does not strongly indicate it." An informative summary is available for the user that reiterates which of the user's symptoms match the presentation of a particular condition as described in the relevant clinical guidelines. This summary can then be used by the user to facilitate any subsequent conversations with their health care provider. The symptom checker is not intended as a diagnostic tool, does not provide medical advice, and users are advised to seek medical input to further investigate any concerns they have.

To ensure medical accuracy and safety during the development of symptom checkers, Flo uses a combination of an in-house medical team and external doctors specializing in the conditions of interest. The medical team builds the chat sequences considering the most relevant signs and symptoms based on the latest medical guidelines and evidence. The chat sequence is medically tested, reviewed, and adjusted in an iterative product development process.

The aim of this study was to determine the accuracy (agreement between clinician and symptom checker) of 3 symptom checkers for endometriosis, uterine fibroids, and PCOS developed using current medical guidelines (Monash, European Society of Human Reproduction and Embryology, and American Academy of Family Physicians) [24,59,60]. To this end, we devised a case vignette study whereby fictional patient cases were assessed for symptoms of the earlier-mentioned conditions by both symptom checkers and medical practitioners. We also aimed to provide a comprehensive illustration of how we created, reviewed, and categorized vignette cases in the predeployment testing and validation of digital health symptom checker tools.

Methods

Vignette Testing

Overview

Clinical case vignettes were created, reviewed, approved, classified, and entered into the symptom checkers by independent general practitioners recruited specifically for this study. Vignette cases needed to encompass presentations of not just endometriosis, fibroids, and PCOS but also other similarly presenting reproductive (eg, amenorrhea) and general health (eg, thyroid disorder) conditions. General practitioners have knowledge of a wide range of condition symptomatology and are typically the first point of contact for a patient in a health care system in the United Kingdom (where the study took place). Therefore, we reasoned that general practitioners were a more suitable choice for vignette creation, review, and classification instead of obstetricians and gynecologists.

All general practitioners were UK-based with an average of 12 years of clinical experience and were not previously affiliated with Flo. All general practitioners were remunerated for their time. No human subjects, interviews, or patient-doctor transcripts were used in the creation of vignettes; all case vignettes involved in this study are fictitious and were created from each general practitioner's experience of treating patients with these conditions.

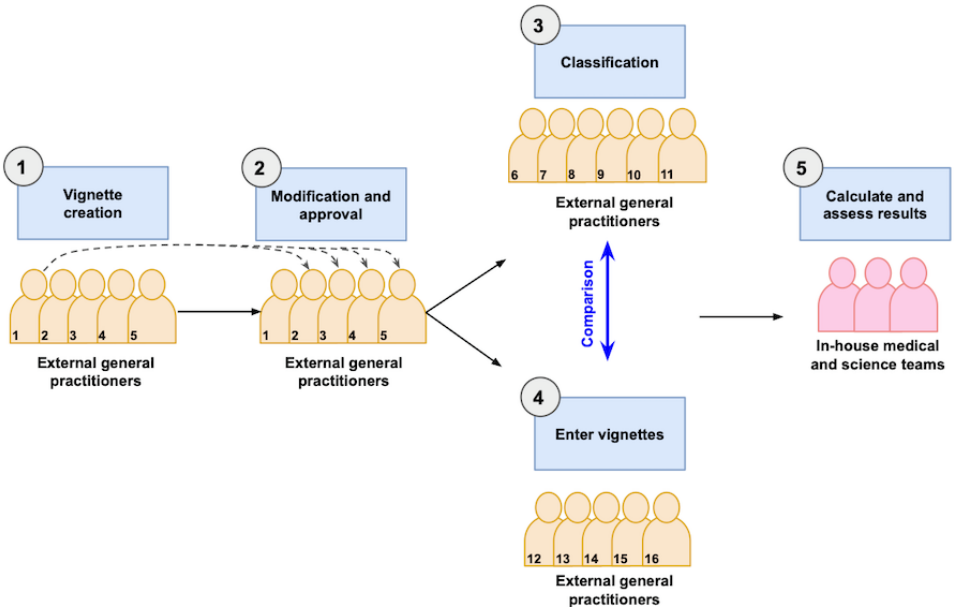
Vignette Creation, Review, and Approval

Five external general practitioners were recruited to independently create clinical case vignettes of simulated users (Figure 1, step 1). These simulated users would be presenting for the first time without any history of diagnosis or treatment for 1 of the 3 conditions of interest, namely, endometriosis, uterine fibroids, or PCOS. Cases were derived from the general practitioners' clinical experience and the literature. The general practitioners completed a template (Multimedia Appendix 1) for each vignette that contained information on the user's background, history of presenting condition, medical, surgical, and family history, as well as details on their menstrual cycle and other symptoms. The general practitioners were instructed to create a set number of cases for each of the 3 conditions and for each of the 3 possible outcomes to ensure a spread of severity and condition types: (A) "You're experiencing specific signs and symptoms commonly associated with [condition]", (B) "Although you're experiencing some of the potential signs and symptoms of [condition], they are not specific enough to indicate it strongly," and (C) "You're not experiencing any of the signs and symptoms commonly associated with [condition]." The general practitioners were instructed that "A" cases are those for which the user has specific features of the condition, and this differential diagnosis is the most likely cause of their symptoms. "B" and "C" cases are those which are considered to not have the condition. General practitioners were instructed that "B" cases represent users who show either too few or only some specific findings, and a clinician would not think of this condition as the most likely cause for these symptoms. "C" cases represent users who show either too few or nonspecific symptoms, and there would be other differential diagnoses that

are more likely to be the cause of the symptoms. Condition-negative cases had other diagnoses such as urinary tract infection, thrush, pregnancy, and functional constipation.

Each vignette was reviewed by a second general practitioner (Figure 1, step 2) who could either approve the vignette as-is or suggest changes to clarify the case. If changes were suggested, the case would then be reviewed, edited, and approved by a third general practitioner who would finalize the case. In total, 24 cases were created for each condition, in line with other single-condition or single-system symptom checker evaluations [61-64].

Figure 1. Vignette study procedure including (1) independent vignette creation by 5 external general practitioners; (2) vignette review, modification, and approval by a second and third general practitioner where required; (3) independent vignette classification by 6 external general practitioners not involved in other stages; (4) entry of vignettes into symptom checkers by 5 external general practitioners not involved in other stages; and (5) analysis of results.



Independent Classification of Vignettes

After vignette approval (Figure 1, step 2), all information related to the intended designation of each vignette was removed: the type of case (A, B, or C above) and any notes about the diagnosis the creator had in mind when creating the vignette were removed from the vignette template. To avoid bias from the case creator when setting the final classification, an additional independent panel of 6 additional external general practitioners not involved in previous steps of the vignette creation was recruited to classify the vignettes (Figure 1, step 3). The classifying general practitioners received a random selection of vignettes, each designated as either an endometriosis vignette, uterine fibroid vignette, or PCOS vignette. For each vignette, the general practitioners reviewed the case and designated the most likely outcome for the specified condition (endometriosis, uterine fibroids, or PCOS) matching the symptom checker wording: (1) a strong match for the condition—"You're experiencing several symptoms typically associated with [condition]" or (2) weak or no match for the condition—"While you may be experiencing some symptoms of [condition], your combination of symptoms does not strongly indicate it." During this step of classification, to ensure there was a shared agreement on the classification of each case, each vignette was reviewed independently by 3 general practitioners; the majority view (at least 2 out of 3) was taken as the "true value" or gold-standard classification for the vignette. While the vignettes were created with 3 levels of categorization for each condition, the classifying general practitioners were not aware of these levels and were asked to make a binary classification for each vignette.

Vignette Entry

An additional set of 5 external general practitioners (not involved in the other steps) were recruited to enter the vignette cases into a prototype of the symptom checkers (Figure 1, step 4). At this stage, the general practitioners were blinded to the condition assigned to the vignette, the classification, and the condition the symptom checker was assessing. If the symptom checker asked a question that was not contained in the vignette, general practitioners were instructed to follow a step-by-step protocol to determine the appropriate answer. First, if the symptom information requested by the symptom checker was specified in the vignette template but not included by the creator, a negative response should be selected (eg, the vignette template specifies pain symptoms should include whether the radiation of pain is present, but the vignette creator does not detail this in their description of pain, then pain radiation should be assumed to be absent). If the information was not part of the template, a neutral response (eg, "I don't know" and "I don't want to answer this question") should be selected. If no neutral response was available, a negative response should be selected. If no negative response was available, the answer mostly within normal limits should be selected (eg, the inputting general practitioner would select a period length of 2-7 days, as opposed to a period length of 1 day or less or a period length of 8 days or more).

Analysis

The final classification set by the independent general practitioner classifiers (Figure 1, step 3) was compared with the outcome of the symptom checker as tested in Figure 1, step 4. Outcomes were arranged in 2-way tables as shown in Figure 2. Accuracy statistics for percent agreement between general practitioner classification and symptom checker, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated using the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as detailed: accuracy (percent agreement): (TP+TN)/(TP+TN+FP+FN); sensitivity: TP/(TP+FN); specificity: TN/(FP+TN); PPV: TP/(TP+FP); and NPV: TN/(FN+TN) (Figure 2).

Figure 2. Two-way validation table demonstrating the true positive, true negative, false positive, and false negative cases produced when comparing the symptom checker output to the general practitioner gold standard.

		Symptom checker	
		Condition positive or strong match for the condition *"You're experiencing several symptoms typically associated with [condition]"*	Condition negative or weak match for the condition *"While you may be experiencing some symptoms of [condition], your combination of symptoms does not strongly indicate it."*
General practitioner (gold standard)	Condition positive or strong match for the condition *"You're experiencing several symptoms typically associated with [condition]"*	a) Both symptom checker and general practitioner designated strong match for the condition (exact match, true positive)	b) General practitioner designated strong match and symptom checker designated weak match (false negative)
	Condition negative or weak match for the condition *"While you may be experiencing some symptoms of [condition], your combination of symptoms does not strongly indicate it."*	c) General practitioner designated weak match and symptom checker designated strong match (false positive)	d) Both symptom checker and general practitioner designated weak match for the condition (exact match, true negative)

Results

Vignette Cases

Of the total of 24 cases that were created per condition (Table 1), 11-13 cases were classified as a strong match for the condition, and 11-13 cases were classified as a weak match for the condition after final classification by a panel (shown in Figure 1, step 3).

Table 1. Two-way validation table by condition (endometriosis [E], uterine fibroids [UF], and polycystic ovary syndrome [P]).

	Condition positive or strong match for the condition			Condition negative or weak match for the condition			Total		
	E, n	UF, n	P, n	E, n	UF, n	P, n	E, n	UF, n	P, n
General practitioner (gold standard)									
Condition positive or strong match for the condition	9	11	12	2	2	0	11	13	12
Condition negative or weak match for the condition	2	2	3	11	9	9	13	11	12
Total	11	13	15	13	11	9	24	24	24

Accuracy Metrics

Overall, exact matches (percent agreement) between the vignette classification and the symptom checker outcome ranged from 83% (20/24) for endometriosis and uterine fibroids to 88% (21/24) for PCOS (Figure 3 and Table 2). While there were no FN outcomes for PCOS, 8% (6/72) of all cases were falsely identified by the relevant symptom checker as negative for endometriosis and uterine fibroids. FP outcomes ranged from 8% (2/24) for endometriosis and uterine fibroids to 13% (3/24) of all cases for PCOS. An example vignette case showing a TP, TN, FP, and FN case (determined by agreement between general practitioner and symptom checker) is provided for each condition in Multimedia Appendix 2.

Figure 3. Overall symptom checker performance showing the proportion of false-positive outcomes, exact match outcomes, and false-negative outcomes by condition. PCOS: polycystic ovary syndrome.

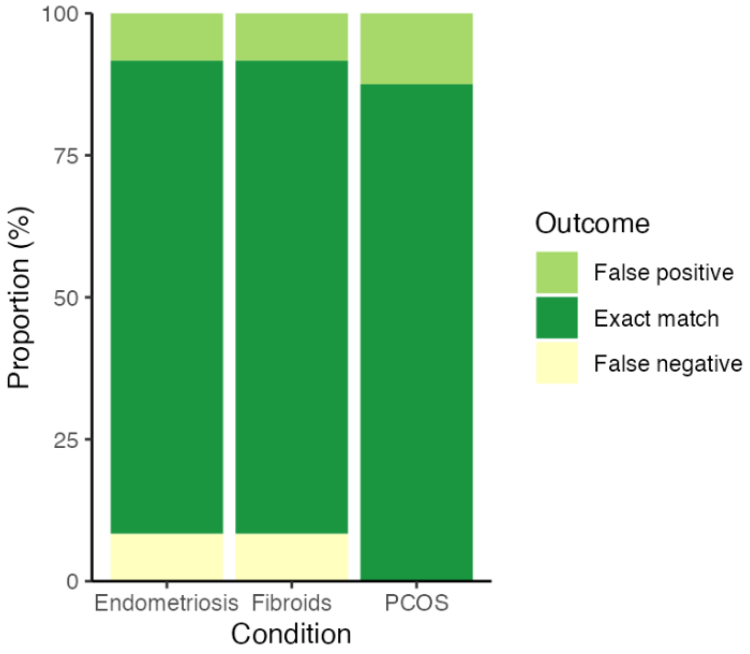


Table 2. Accuracy metrics for endometriosis, fibroids, and polycystic ovary syndrome (PCOS).

Condition	Values, n	Agreement (%)	Sensitivity (%)	Specificity (%)	PPV[a] (%)	NPV[b] (%)
Endometriosis	24	83.3	81.8	84.6	81.8	84.6
Fibroids	24	83.3	84.6	81.8	84.6	81.8
PCOS	24	87.5	100	75	80	100

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

While sensitivity was very high (100%) for PCOS (Table 2), specificity was high for all 3 conditions (>81%). PPV ranged from 80% for PCOS to 84.6% for uterine fibroids, while NPV ranged from 81.8% for uterine fibroids to 100% for PCOS.

Discussion

Summary

In this study, we provide an example methodology for the creation, review, and classification of vignette cases for the testing and validation of digital health symptom checker tools. The percent agreement between general practitioner-designated vignette cases and 3 single-condition symptom checkers for 3 reproductive health conditions (endometriosis, fibroids, and PCOS) was assessed. We found the designation given to case vignettes by the symptom checkers had high levels of agreement between general practitioner and symptom checker (83.3%-87.5%), sensitivity (81.8%-100%), specificity (75%-84.6%), PPV (80%-84.6%), and NPV (81.8%-100%) when compared to gold standard designation by general practitioners. Overall, these metrics show the high performance of the symptom checkers when tested on robustly designed clinical vignettes.

Comparison With Prior Work

This high accuracy of the identification of reproductive health conditions is particularly important as high rates of diagnostic error are reported by patients. A study of patients with self-reported surgically confirmed endometriosis found that 75.2% of patients reported being misdiagnosed with another physical health or mental health problem by their health care professional [65]. A similar study of patients diagnosed with PCOS found that 33.6% of women reported >2 years time to diagnosis, 47.1% visited ≥3 health professionals before a diagnosis was established, and 64.8% were dissatisfied with the diagnostic process [66]. The use of a tool like a symptom checker could give the user better knowledge and awareness of their symptoms in conversations they may have with their health care provider, leading to a more effective diagnostic pathway. We have shown in previous research that users agree Flo increases their knowledge of the menstrual cycle and facilitates easier conversations with their health care provider [67].

Other vignette studies of multicondition symptom checkers have shown mixed results for accuracy. A study by Gilbert et al [68] comparing urgency advice (ie, triage) from 7 multicondition symptom checker apps and 7 general practitioners to gold-standard vignettes found that the condition suggested first matched the gold standard (ie, M1 accuracy) for

71% of general practitioners and 26% of apps; when broadening to the condition suggested in the top 5 (ie, M5 accuracy), the accuracy of general practitioners rose to 83% and apps to 41%. Another study by Schmieding et al [69] comparing 22 symptom checkers using 45 vignettes found M1 accuracy of 46%, and M10 accuracy was 71%.

The multicondition symptom checkers evaluated by Gilbert et al [68] and Schmieding et al [69] were assessed using vignette cases that covered both common and less-common conditions seen in primary care practice, conditions that affect all body systems, and conditions that have a range of urgency levels. Further, these evaluated symptom checkers are designed to detect a wide range of conditions for a general population. In contrast, this study evaluated single-condition symptom checkers using vignettes specifically designed to represent presentations with specific symptoms of the condition (strong match or condition positive) and presentations with symptoms not specific to the condition (weak match or condition negative). This symptom checker design difference may explain the variation in accuracy found between our symptom checkers (single condition) and other studied symptom checkers (multicondition).

Evaluations of single-condition symptom checkers include a study of 12 web-based symptom checkers for COVID-19 [70] and a study of an app-based symptom checker for PCOS [62]. COVID-19 symptom checkers ranged widely in both sensitivity (14%-94%) and specificity (29%-100%), with only 4 symptom checkers having both sensitivity and specificity above 50% and 2 with both sensitivity and specificity above 75%. Sensitivity and specificity in our symptom checkers were between 75% and 100%. The PCOS symptom checker evaluated by Rodriguez et al [62] reported 12%-25% FP cases and no FN out of 8 cases tested. Our PCOS symptom checker had no FNs and 3 (13%) FP cases out of 24 cases tested. Our PPV and NPV values were 80%-100% for our 3 symptom checkers, suggesting relatively high chances that positively tested cases truly have the condition in question.

With the exception of COVID-19, which has a symptomatology and overall presentation that differs greatly from the reproductive health disorders assessed in this study, digital or app-based symptom checkers for a single condition are uncommon. Symptom-based patient-completed questionnaires and screening tools do exist, including for common reproductive health conditions such as endometriosis or PCOS. A patient self-assessment tool for endometriosis with 21 questions found sensitivity of 76% and specificity of 72%, PPV of 73%, and NPV of 75% [71]. Our endometriosis symptom checker had a similar but slightly higher sensitivity (81.8%), specificity (84.6%), PPV (81.8%), and NPV (84.6%). A 4-item questionnaire for use in the diagnosis of PCOS among women with a primary complaint of infertility had 77% sensitivity and 94% specificity, and a PPV and NPV calculated from their data as 87% and 88%, respectively [72]. Our PCOS symptom checker had higher sensitivity (100%), lower specificity (75%), higher NPV (100%), and lower PPV (80%), prioritizing the identification of cases. It should be noted, however, that our symptom checker is designed to be for a broader population than the 4-item clinical tool, including those who are not trying to get pregnant or experiencing fertility issues. Questionnaires such as these have some limitations. They may not be available to the public and additionally may be subject to more user error (eg, question skipping). App-based symptom checkers, on the other hand, can use historical data from users such as menstrual regularity to improve the accuracy of user answers. Additionally, users cannot accidentally skip questions, and the app will provide a detailed summary of results and recommendations.

It is not uncommon for variation of opinion between groups of general practitioners reviewing vignettes with El-Osta et al [73] reporting classification agreement between 3 general practitioners' primary diagnosis and the intention of the vignette being 32.4%. Each vignette in this study was reviewed independently by 3 different general practitioners, and in 71% (51/72) of cases, all 3 general practitioners agreed with the vignette assignment given at the vignette approval stage (Figure 1, step 2). However, it should be noted that El-Osta et al [73] provided a comparison between primary diagnosis of general practitioners and original vignette intention, whereas this analysis only concerns strong or weak match for known reproductive conditions. All 3 general practitioners agreed with each other (regardless of the vignette intention) for 81% (58/72) of vignette cases. This disagreement between general practitioners and some differences with the symptom checker results are to be expected, particularly when using symptom-based assessment for reproductive health conditions that can be complicated to diagnose, have overlapping symptomatology with other system conditions such as gastrointestinal and urinary conditions, and are often dismissed or considered to be "normal" variations in the menstrual cycle by some. These conditions have a notoriously prolonged time to diagnosis [16-19] and require investigations including imaging. Further, the sensitivity of different testing methods can vary. For example, physical examination for deep infiltrating endometriosis can have poor accuracy and requires imaging [74].

The possible applications of symptom checkers and health apps are far-reaching and could have benefits at the individual user level, health care professional level, and macro or health system level [63,75]. Especially for many reproductive health conditions where the time to diagnosis is currently long and contributes to high health care costs [17,26,66,76], an earlier diagnosis can lead to early treatment and thus decrease complications from untreated conditions and decrease health care costs of treating more advanced disease [27,28,39]. Menstrual cycle details such as cycle length, period length, or flow can be important information for health care providers when diagnosing patients. Health apps can help track cycle details over time and use these details when determining risk for conditions as well as in summary information for users to share with their health care providers (eg, the Flo app provides a "health report" where you can download a summary of symptoms over a period of time, average cycle length, and other details to share with a health care provider). Additionally, as people with symptoms such as heavy bleeding or menstrual pain may believe these are normal or hereditary [26], personalized assessment of symptoms and encouragement to seek further evaluation from a medical professional where appropriate may improve an individual's understanding of their symptoms and health status and decrease

time to diagnosis. Our prior research has demonstrated that 58% of Flo users report improvements in understanding the normality of certain cycle-related symptoms and recognizing the abnormal nature of others, while 1 in 3 Flo users reported that the use of the Flo app improved their communication with their health care provider [67]. Therefore, mobile apps with symptom checkers could identify users with risk factors for certain conditions, educate users about their symptoms, and further encourage conversation with their medical providers.

Strengths and Limitations

Strengths of this study include the use of different groups of independent, external general practitioners unfamiliar with the symptom checkers to create, enter, and classify case vignettes for symptom checker testing. Additionally, vignettes were created with a wide range of symptomatology to ensure the inclusion of borderline presentations as these are notoriously difficult to assess, even for doctors, although they represent a frequent reality as people do not often fit neatly into textbook case presentations. Further, each vignette case was reviewed by an independent, experienced general practitioner and classified by a separate panel viewing the vignettes for the first time. When generating vignette cases that represent typical presentations of a single condition as seen by a general practitioner, there will only be so many permutations of symptomatology that can be generated before repetitions of vignette cases occur; as a result, we created 72 vignette cases in total, 24 for each of our 3 conditions. The number of vignettes needed to evaluate symptom checkers is not well defined [54]. Other vignette symptom checker evaluations have used between 3 and 400 cases for testing, with single-condition or single-system evaluations (eg, mental health, ophthalmology, and PCOS) using fewer cases and multicondition evaluations using larger numbers of cases [61-63,77,78]. Among the 400 vignettes published by Hammoud et al [78], any single condition is only represented by at most 5 cases.

Limitations, however, should be noted. Vignette studies rely on clinical opinion of a small number of general practitioners. An audit study of clinical vignette benchmarking has shown significant variation between groups of general practitioners considering clinical vignettes [73]. To decrease bias from differences in clinical opinion, all cases were blindly reviewed by 3 general practitioners, one-third involved in cases of disagreement. We found agreement between all 3 general practitioners in 81% (58/72) of our cases. Vignettes also rely on the classical presentation of conditions that may present differently in real life or in patients with complex or atypical condition presentations. When creating vignettes, we recognize there is a possibility for bias, expected patterns, or entrenched unknowns in the understanding of each condition's symptomatology. Additionally, although we recognize that patients do not usually present to primary care practitioners with a prespecified suspected diagnosis and that therefore this aspect of the study design does not reflect usual medical practice, these chatbots are not meant to replace the interaction with primary care providers but rather to allow users to review their symptoms in advance of seeing a health care professional. As outlined in the medical guidelines, symptom severity, risk, and prevalence may vary across world regions and ethnicities. Neither the Flo app nor the symptom checker collect data on the user's race or ethnicity, so neither race nor ethnicity were included in the vignette creation process. In addition to this, the vignettes in this study were created by panels of general practitioners based in the United Kingdom and may not provide an accurate representation of symptoms for every cultural context. We recognize the inclusion of such information could help to identify at-risk individuals better.

While we found 100% sensitivity for our PCOS symptom checker, it is likely with a larger sample size and real-life cases, this level of perfect sensitivity will not be maintained. Other changes in accuracy statistics are likely to be seen in real-world use. Further, as real-world users may interpret their symptoms and the questions differently than doctors, future studies including the general population should be carried out to test each symptom checker's performance in the context of real-world deployment. The use of vignette-patient cases is an important part of predeployment algorithm testing for digital symptom checkers [54,79] and is the first stage of our symptom checker evaluation. The next stage in evaluating our symptom checkers will include the use of real-world data such as observational studies of condition diagnosed and undiagnosed people's symptoms and early field-testing of the in-app symptom checkers on users and comparing the output to an official diagnosis from a doctor. Evaluation of symptom checkers and digital health tools should follow multistage processes with increasing exposure to real environments exploring not only effectiveness but also usability and balance between probability of disease and risk of missing a diagnosis [79].

Conclusions

In conclusion, we have described a methodology for creating and classifying vignettes using multiple independent panels of general practitioners for the predeployment testing of digital health symptom checker tools. We found high levels of agreement between general practitioner classification and single-condition symptom checkers for 3 reproductive health conditions (endometriosis, fibroids, and PCOS). Given long delays in diagnosis for many reproductive health conditions, which lead to increased medical costs and potential health complications, innovative health apps and symptom checkers hold the potential to improve care pathways.

Acknowledgments

This study was funded by Flo Health UK Limited.

Conflicts of Interest

KP, APW, OW, YCK, A Marhoi, SA, ACC, AK, SP, and LZ were employees at Flo Health, Inc and have stock ownership in the company. RB, CP, A Meczner, MF, and SG are paid consultants for Flo Health, Inc. SG declares no nonfinancial interests but the following competing financial interests: he has or has had consulting relationships with Una Health GmbH, Lindus Health Ltd, Flo Health UK Limited, Thymia Ltd, and Ada Health GmbH and holds share options in Ada Health GmbH. MF declares no nonfinancial interests but the following competing financial interests: he has a consulting relationship with Flo Health UK Limited and holds share options in Una Health GmbH. A Meczner declares no nonfinancial interests but the following competing financial interests: he is an employee and shareholder at Healthily or Your.MD.

Multimedia Appendix 1

Vignette template.
[DOCX File , 19 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Example case vignettes with matching and mismatching classification between general practitioners and the symptom checker (SC).
[DOCX File , 62 KB-Multimedia Appendix 2]

References

1. Azziz R, Carmina E, Chen Z, Dunaif A, Laven JSE, Legro RS, et al. Polycystic ovary syndrome. Nat Rev Dis Primers. 2016 Aug 11;2:16057 [FREE Full text] [doi: 10.1038/nrdp.2016.57] [Medline: 27510637]
2. Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO. The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod. 2016;31(12):2841-2855 [FREE Full text] [doi: 10.1093/humrep/dew218] [Medline: 27664216]
3. Carmina E, Azziz R. Diagnosis, phenotype, and prevalence of polycystic ovary syndrome. Fertil Steril. 2006;86(Suppl 1):S7-S8 [FREE Full text] [doi: 10.1016/j.fertnstert.2006.03.012] [Medline: 16798288]
4. Deswal R, Narwal V, Dang A, Pundir CS. The prevalence of polycystic ovary syndrome: a brief systematic review. J Hum Reprod Sci. 2020;13(4):261-271 [FREE Full text] [doi: 10.4103/jhrs.JHRS_95_18] [Medline: 33627974]
5. Ellis K, Munro D, Clarke J. Endometriosis is undervalued: a call to action. Front Glob Womens Health. 2022;3:902371 [FREE Full text] [doi: 10.3389/fgwh.2022.902371] [Medline: 35620300]
6. Eskenazi B, Warner ML. Epidemiology of endometriosis. Obstet Gynecol Clin North Am. 1997;24(2):235-258 [FREE Full text] [doi: 10.1016/s0889-8545(05)70302-8] [Medline: 9163765]
7. Rawson JM. Prevalence of endometriosis in asymptomatic women. J Reprod Med. 1991;36(7):513-515 [Medline: 1834839]
8. Riestenberg C, Jagasia A, Markovic D, Buyalos RP, Azziz R. Health care-related economic burden of polycystic ovary syndrome in the United States: pregnancy-related and long-term health consequences. J Clin Endocrinol Metab. 2022;107(2):575-585 [FREE Full text] [doi: 10.1210/clinem/dgab613] [Medline: 34546364]
9. Teede H, Deeks A, Moran L. Polycystic ovary syndrome: a complex condition with psychological, reproductive and metabolic manifestations that impacts on health across the lifespan. BMC Med. 2010;8:41 [FREE Full text] [doi: 10.1186/1741-7015-8-41] [Medline: 20591140]
10. Waller KG, Lindsay P, Curtis P, Shaw RW. The prevalence of endometriosis in women with infertile partners. Eur J Obstet Gynecol Reprod Biol. 1993;48(2):135-139 [FREE Full text] [doi: 10.1016/0028-2243(93)90254-a] [Medline: 8491333]
11. Stewart EA. Uterine fibroids. Lancet. 2001;357(9252):293-298 [FREE Full text] [doi: 10.1016/S0140-6736(00)03622-9] [Medline: 11214143]
12. Khan AT, Shehmar M, Gupta JK. Uterine fibroids: current perspectives. Int J Womens Health. 2014;6:95-114 [FREE Full text] [doi: 10.2147/IJWH.S51083] [Medline: 24511243]
13. Azziz R, Woods KS, Reyna R, Key TJ, Knochenhauer ES, Yildiz BO. The prevalence and features of the polycystic ovary syndrome in an unselected population. J Clin Endocrinol Metab. 2004;89(6):2745-2749 [FREE Full text] [doi: 10.1210/jc.2003-032046] [Medline: 15181052]
14. Bulletti C, Coccia ME, Battistoni S, Borini A. Endometriosis and infertility. J Assist Reprod Genet. 2010;27(8):441-447 [FREE Full text] [doi: 10.1007/s10815-010-9436-1] [Medline: 20574791]
15. Kennedy S, Bergqvist A, Chapron C, D'Hooghe T, Dunselman G, Greb R, et al. ESHRE guideline for the diagnosis and treatment of endometriosis. Hum Reprod. 2005;20(10):2698-2704 [FREE Full text] [doi: 10.1093/humrep/dei135] [Medline: 15980014]
16. Stewart EA, Cookson CL, Gandolfo RA, Schulze-Rath R. Epidemiology of uterine fibroids: a systematic review. BJOG. 2017;124(10):1501-1512 [FREE Full text] [doi: 10.1111/1471-0528.14640] [Medline: 28296146]
17. Kiesel L, Sourouni M. Diagnosis of endometriosis in the 21st century. Climacteric. 2019;22(3):296-302 [FREE Full text] [doi: 10.1080/13697137.2019.1578743] [Medline: 30905186]

18. Hudelist G, Fritzer N, Thomas A, Niehues C, Oppelt P, Haas D, et al. Diagnostic delay for endometriosis in Austria and Germany: causes and possible consequences. Hum Reprod. 2012;27(12):3412-3416 [FREE Full text] [doi: 10.1093/humrep/des316] [Medline: 22990516]
19. Husby GK, Haugen RS, Moen MH. Diagnostic delay in women with pain and endometriosis. Acta Obstet Gynecol Scand. 2003;82(7):649-653 [FREE Full text] [doi: 10.1034/j.1600-0412.2003.00168.x] [Medline: 12790847]
20. Gibson-Helm ME, Lucas IM, Boyle JA, Teede HJ. Women's experiences of polycystic ovary syndrome diagnosis. Fam Pract. 2014;31(5):545-549 [FREE Full text] [doi: 10.1093/fampra/cmu028] [Medline: 24925927]
21. Stewart EA, Nicholson WK, Bradley L, Borah BJ. The burden of uterine fibroids for African-American women: results of a national survey. J Womens Health (Larchmt). 2013;22(10):807-816 [FREE Full text] [doi: 10.1089/jwh.2013.4334] [Medline: 24033092]
22. Borah BJ, Nicholson WK, Bradley L, Stewart EA. The impact of uterine leiomyomas: a national survey of affected women. Am J Obstet Gynecol. 2013;209(4):319.e1-319.e20 [FREE Full text] [doi: 10.1016/j.ajog.2013.07.017] [Medline: 23891629]
23. Spaczynski RZ, Duleba AJ. Diagnosis of endometriosis. Semin Reprod Med. 2003;21(2):193-208 [doi: 10.1055/s-2003-41326] [Medline: 12917789]
24. Becker CM, Bokor A, Heikinheimo O, Horne A, Jansen F, Kiesel L, et al. ESHRE guideline: endometriosis. Hum Reprod Open. 2022;2022(2):hoac009 [FREE Full text] [doi: 10.1093/hropen/hoac009] [Medline: 35350465]
25. Teede HJ, Misso ML, Costello MF, Dokras A, Laven J, Moran L, et al. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Hum Reprod. 2018;33(9):1602-1618 [FREE Full text] [doi: 10.1093/humrep/dey256] [Medline: 30052961]
26. Ghant MS, Sengoba KS, Vogelzang R, Lawson AK, Marsh EE. An altered perception of normal: understanding causes for treatment delay in women with symptomatic uterine fibroids. J Womens Health (Larchmt). 2016;25(8):846-852 [FREE Full text] [doi: 10.1089/jwh.2015.5531] [Medline: 27195902]
27. Ballweg ML. Impact of endometriosis on women's health: comparative historical data show that the earlier the onset, the more severe the disease. Best Pract Res Clin Obstet Gynaecol. 2004;18(2):201-218 [FREE Full text] [doi: 10.1016/j.bpobgyn.2004.01.003] [Medline: 15157638]
28. Peña AS, Witchel SF, Hoeger KM, Oberfield SE, Vogiatzi MG, Misso M, et al. Adolescent polycystic ovary syndrome according to the international evidence-based guideline. BMC Med. 2020;18(1):72 [FREE Full text] [doi: 10.1186/s12916-020-01516-x] [Medline: 32204714]
29. Matsuzaki S, Canis M, Pouly JL, Rabischong B, Botchorishvili R, Mage G. Relationship between delay of surgical diagnosis and severity of disease in patients with symptomatic deep infiltrating endometriosis. Fertil Steril. 2006;86(5):1314-1316; discussion 1317 [FREE Full text] [doi: 10.1016/j.fertnstert.2006.03.048] [Medline: 16978622]
30. Witchel SF, Teede HJ, Peña AS. Curtailing PCOS. Pediatr Res. 2020;87(2):353-361 [FREE Full text] [doi: 10.1038/s41390-019-0615-1] [Medline: 31627209]
31. Robinson KM, Christensen KB, Ottesen B, Krasnik A. Diagnostic delay, quality of life and patient satisfaction among women diagnosed with endometrial or ovarian cancer: a nationwide Danish study. Qual Life Res. 2012;21(9):1519-1525 [FREE Full text] [doi: 10.1007/s11136-011-0077-3] [Medline: 22138966]
32. Álvarez-Salvago F, Lara-Ramos A, Cantarero-Villanueva I, Mazheika M, Mundo-López A, Galiano-Castillo N, et al. Chronic fatigue, physical impairments and quality of life in women with endometriosis: a case-control study. Int J Environ Res Public Health. 2020;17(10):3610 [FREE Full text] [doi: 10.3390/ijerph17103610] [Medline: 32455618]
33. Della Corte L, Di Filippo C, Gabrielli O, Reppuccia S, La Rosa VL, Ragusa R, et al. The burden of endometriosis on women's lifespan: a narrative overview on quality of life and psychosocial wellbeing. Int J Environ Res Public Health. 2020;17(13):4683 [FREE Full text] [doi: 10.3390/ijerph17134683] [Medline: 32610665]
34. Marsh EE, Al-Hendy A, Kappus D, Galitsky A, Stewart EA, Kerolous M. Burden, prevalence, and treatment of uterine fibroids: a survey of U.S. Women. J Womens Health (Larchmt). 2018;27(11):1359-1367 [FREE Full text] [doi: 10.1089/jwh.2018.7076] [Medline: 30230950]
35. Al-Hendy A, Myers ER, Stewart E. Uterine fibroids: burden and unmet medical need. Semin Reprod Med. 2017;35(6):473-480 [FREE Full text] [doi: 10.1055/s-0037-1607264] [Medline: 29100234]
36. Simoens S, Dunselman G, Dirksen C, Hummelshoj L, Bokor A, Brandes I, et al. The burden of endometriosis: costs and quality of life of women with endometriosis and treated in referral centres. Hum Reprod. 2012;27(5):1292-1299 [FREE Full text] [doi: 10.1093/humrep/des073] [Medline: 22422778]
37. Barnard L, Ferriday D, Guenther N, Strauss B, Balen AH, Dye L. Quality of life and psychological well being in polycystic ovary syndrome. Hum Reprod. 2007;22(8):2279-2286 [FREE Full text] [doi: 10.1093/humrep/dem108] [Medline: 17537782]
38. Lizneva D, Suturina L, Walker W, Brakta S, Gavrilova-Jordan L, Azziz R. Criteria, prevalence, and phenotypes of polycystic ovary syndrome. Fertil Steril. 2016;106(1):6-15 [FREE Full text] [doi: 10.1016/j.fertnstert.2016.05.003] [Medline: 27233760]
39. Surrey E, Soliman AM, Trenz H, Blauer-Peterson C, Sluis A. Impact of endometriosis diagnostic delays on healthcare resource utilization and costs. Adv Ther. 2020;37(3):1087-1099 [FREE Full text] [doi: 10.1007/s12325-019-01215-x] [Medline: 31960340]
40. Soliman AM, Yang H, Du EX, Kelley C, Winkel C. The direct and indirect costs associated with endometriosis: a systematic literature review. Hum Reprod. 2016;31(4):712-722 [FREE Full text] [doi: 10.1093/humrep/dev335] [Medline: 26851604]

41. Azziz R. Overview of long-term morbidity and economic cost of the polycystic ovary syndrome. In: Azziz R, Nestler JE, Dewailly D, editors. Androgen Excess Disorders in Women Polycystic Ovary Syndrome and Other Disorders, Second Edition. Totowa, NJ. Humana Press; 2007:353-362
42. Ericsson mobility report. Ericsson. 2022. URL: https://www.ericsson.com/49d3a0/assets/local/reports-papers/mobility-report/documents/2022/ericsson-mobility-report-june-2022.pdf [accessed 2023-11-15]
43. Digital health trends 2021. IQVIA. 2021. URL: https://www.iqvia.com/insights/the-iqvia-institute/reports-and-publications/reports/digital-health-trends-2021 [accessed 2022-07-11]
44. Fox S, Duggan M. Health online 2013. Pew Research Center. 2013. URL: https://www.pewresearch.org/internet/2013/01/15/health-online-2013/ [accessed 2022-07-12]
45. Yigzaw KY, Wynn R, Marco-Ruiz L, Budrionis A, Oyeyemi SO, Fagerlund AJ, et al. The association between health information seeking on the internet and physician visits (The Seventh Tromsø Study—Part 4): population-based questionnaire study. J Med Internet Res. 2020;22(3):e13120 [FREE Full text] [doi: 10.2196/13120] [Medline: 32134387]
46. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: cross-sectional survey study. J Med Internet Res. 2020;22(1):e14679 [FREE Full text] [doi: 10.2196/14679] [Medline: 32012052]
47. Jutel A, Lupton D. Digitizing diagnosis: a review of mobile applications in the diagnostic process. Diagnosis (Berl). 2015;2(2):89-96 [FREE Full text] [doi: 10.1515/dx-2014-0068] [Medline: 29540025]
48. Wetzel AJ, Koch R, Preiser C, Müller R, Klemmt M, Ranisch R, et al. Ethical, legal, and social implications of symptom checker apps in primary health care (CHECK.APP): protocol for an interdisciplinary mixed methods study. JMIR Res Protoc. 2022;11(5):e34026 [FREE Full text] [doi: 10.2196/34026] [Medline: 35576570]
49. Millenson ML, Baldwin JL, Zipperer L, Singh H. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. Diagnosis (Berl). 2018;5(3):95-105 [FREE Full text] [doi: 10.1515/dx-2018-0009] [Medline: 30032130]
50. Kowatsch T, Otto L, Harperink S, Cotti A, Schlieter H. A design and evaluation framework for digital health interventions. It—Inf Technol. 2019;61(5-6):253-263 [FREE Full text] [doi: 10.1515/itit-2019-0019]
51. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med. 2020;3(1):110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
52. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med. 2019;2(1):38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
53. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med. 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
54. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. J Med Internet Res. 2022;24(10):e37408 [FREE Full text] [doi: 10.2196/37408] [Medline: 36287594]
55. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? Front Med (Lausanne). 2022;9:1040926 [FREE Full text] [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
56. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med. 2022;5(1):118 [FREE Full text] [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
57. Kopka M, Schmieding ML, Rieger T, Roesler E, Balzer F, Feufel MA. Determinants of laypersons' trust in medical decision aids: randomized controlled trial. JMIR Hum Factors. 2022;9(2):e35219 [FREE Full text] [doi: 10.2196/35219] [Medline: 35503248]
58. The world's most popular female health app. Flo Health. URL: https://flo.health/ [accessed 2023-11-23]
59. Teede H, Tay CT, Laven JSE, Dokras A, Moran LJ, Piltonen T, et al. International evidence-based guideline for the assessment and management of polycystic ovary syndrome 2023. Monash University. 2023. URL: https://bridges.monash.edu/articles/online_resource/International_Evidence-based_Guideline_for_the_Assessment_and_Management_of_Polycystic_Ovary_Syndrome_2023/24003834/1 [accessed 2023-11-15]
60. De La Cruz MSD, Buchanan EM. Uterine fibroids: diagnosis and treatment. Am Fam Physician. 2017;95(2):100-107 [FREE Full text] [Medline: 28084714]
61. Ćirković A. Evaluation of four artificial intelligence-assisted self-diagnosis apps on three diagnoses: two-year follow-up study. J Med Internet Res. 2020;22(12):e18097 [FREE Full text] [doi: 10.2196/18097] [Medline: 33275113]
62. Rodriguez EM, Thomas D, Druet A, Vlajic-Wheeler M, Lane KJ, Mahalingaiah S. Identifying women at risk for polycystic ovary syndrome using a mobile health app: virtual tool functionality assessment. JMIR Form Res. 2020;4(5):e15094 [FREE Full text] [doi: 10.2196/15094] [Medline: 32406861]
63. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR Form Res. 2019;3(4):e13863 [FREE Full text] [doi: 10.2196/13863] [Medline: 31663858]

64. Shen C, Nguyen M, Gregor A, Isaza G, Beattie A. Accuracy of a popular online symptom checker for ophthalmic diagnoses. JAMA Ophthalmol. 2019;137(6):690-692 [FREE Full text] [doi: 10.1001/jamaophthalmol.2019.0571] [Medline: 30973602]
65. Bontempo AC, Mikesell L. Patient perceptions of misdiagnosis of endometriosis: results from an online national survey. Diagnosis (Berl). 2020;7(2):97-106 [FREE Full text] [doi: 10.1515/dx-2019-0020] [Medline: 32007945]
66. Gibson-Helm M, Teede H, Dunaif A, Dokras A. Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome. J Clin Endocrinol Metab. 2017;102(2):604-612 [FREE Full text] [doi: 10.1210/jc.2016-2963] [Medline: 27906550]
67. Zhaunova L, Bamford R, Radovic T, Wickham A, Peven K, Croft J, et al. Characterization of self-reported improvements in knowledge and health among users of Flo period tracking app: cross-sectional survey. JMIR Mhealth Uhealth. 2023;11:e40427 [FREE Full text] [doi: 10.2196/40427] [Medline: 37099370]
68. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open. 2020;10(12):e040269 [FREE Full text] [doi: 10.1136/bmjopen-2020-040269] [Medline: 33328258]
69. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res. 2022;24(5):e31810 [FREE Full text] [doi: 10.2196/31810] [Medline: 35536633]
70. Munsch N, Martin A, Gruarin S, Nateqi J, Abdarahmane I, Weingartner-Ortner R, et al. Diagnostic accuracy of web-based COVID-19 symptom checkers: comparison study. J Med Internet Res. 2020;22(10):e21299 [FREE Full text] [doi: 10.2196/21299] [Medline: 33001828]
71. Cho HH, Yoon YS. Development of an endometriosis self-assessment tool for patient. Obstet Gynecol Sci. 2022;65(3):256-265 [FREE Full text] [doi: 10.5468/ogs.21252] [Medline: 35381626]
72. Pedersen SD, Brar S, Faris P, Corenblum B. Polycystic ovary syndrome: validated questionnaire for use in diagnosis. Can Fam Physician. 2007;53(6):1042-1047 [FREE Full text] [Medline: 17872783]
73. El-Osta A, Webber I, Alaa A, Bagkeris E, Mian S, Sharabiani MTA, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. BMJ Open. 2022;12(4):e053566 [FREE Full text] [doi: 10.1136/bmjopen-2021-053566] [Medline: 35477872]
74. Bazot M, Lafont C, Rouzier R, Roseau G, Thomassin-Naggara I, Daraï E. Diagnostic accuracy of physical examination, transvaginal sonography, rectal endoscopic sonography, and magnetic resonance imaging to diagnose deep infiltrating endometriosis. Fertil Steril. 2009;92(6):1825-1833 [FREE Full text] [doi: 10.1016/j.fertnstert.2008.09.005] [Medline: 19019357]
75. Critchley HOD, Babayev E, Bulun SE, Clark S, Garcia-Grau I, Gregersen PK, et al. Menstruation: science and society. Am J Obstet Gynecol. 2020;223(5):624-664 [FREE Full text] [doi: 10.1016/j.ajog.2020.06.004] [Medline: 32707266]
76. Hoeger KM, Dokras A, Piltonen T. Update on PCOS: consequences, challenges, and guiding treatment. J Clin Endocrinol Metab. 2021;106(3):e1071-e1083 [FREE Full text] [doi: 10.1210/clinem/dgaa839] [Medline: 33211867]
77. Fleming J, Jeannon JP. Head and neck cancer in the digital age: an evaluation of mobile health applications. BMJ Innov. 2020;6(1):13-17 [doi: 10.1136/bmjinnov-2019-000350]
78. Hammoud M, Douglas S, Darmach M, Alawneh S, Sanyal S, Kanbour Y. Avey: an accurate AI algorithm for self-diagnosis. medRxiv.. Preprint posted online on March 11, 2022 [FREE Full text] [doi: 10.1101/2022.03.08.22272076]
79. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet. 2018;392(10161):2263-2264 [FREE Full text] [doi: 10.1016/S0140-6736(18)32819-8] [Medline: 30413281]

Abbreviations

FN: false negative
FP: false positive
NPV: negative predictive value
PCOS: polycystic ovary syndrome
PPV: positive predictive value
TN: true negative
TP: true positive

Edited by L Buis, G Eysenbach; submitted 23.02.23; peer-reviewed by C Vorisek, E Baker, CI Sartorão Filho, S Dixon; comments to author 28.07.23; revised version received 06.09.23; accepted 07.11.23; published 05.12.23

Please cite as:
Peven K, Wickham AP, Wilks O, Kaplan YC, Marhol A, Ahmed S, Bamford R, Cunningham AC, Prentice C, Meczner A, Fenech M, Gilbert S, Klepchukova A, Ponzo S, Zhaunova L
Assessment of a Digital Symptom Checker Tool's Accuracy in Suggesting Reproductive Health Conditions: Clinical Vignettes Study
JMIR Mhealth Uhealth 2023;11:e46718
URL: https://mhealth.jmir.org/2023/1/e46718
doi: 10.2196/46718
PMID: