Review

Michael L. Millenson*, Jessica L. Baldwin, Lorri Zipperer and Hardeep Singh

Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis

https://doi.org/10.1515/dx-2018-0009
Received April 1, 2018; accepted June 1, 2018; previously published online July 23, 2018

Abstract: Over a third of adults go online to diagnose their health condition. Direct-to-consumer (DTC), interactive, diagnostic apps with information personalization capabilities beyond those of static search engines are rapidly proliferating. While these apps promise faster, more convenient and more accurate information to improve diagnosis, little is known about the state of the evidence on their performance or the methods used to evaluate them. We conducted a scoping review of the peer-reviewed and gray literature for the period January 1, 2014–June 30, 2017. We found that the largest category of evaluations involved symptom checkers that applied algorithms to user-answered questions, followed by sensor-driven apps that applied algorithms to smartphone photos, with a handful of evaluations examining crowdsourcing. The most common clinical areas evaluated were dermatology and general diagnostic and triage advice for a range of conditions. Evaluations were highly variable in methodology and conclusions, with about half describing app characteristics and half examining actual performance. Apps were found to vary widely in functionality, accuracy, safety and effectiveness, although the usefulness of this evidence was limited by a frequent failure to provide results by named individual app. Overall, the current evidence base on DTC, interactive diagnostic apps is sparse in scope, uneven in the information provided and inconclusive with respect to safety and effectiveness, with no studies of clinical risks and benefits involving real-world consumer use. Given that DTC diagnostic apps are rapidly evolving, rigorous and standardized evaluations are essential to inform decisions by clinicians, patients, policymakers and other stakeholders.

Keywords: consumerism; crowdsourcing; diagnostic error; digital health; evidence-based medicine; health apps; health information technology; mHealth; patient engagement.

Introduction

The 2015 National Academy of Medicine (NAM) report Improving Diagnosis in Health Care concluded that most people will experience at least one diagnostic error in their lifetime [1]. The report, noting that over a third of adults go online to diagnose a health condition [2], urged professionals to direct patients to reliable online resources. How to determine the reliability of online resources, however, remains an unresolved question.

Currently available online resources have graduated beyond keyword searches on Google. Increasingly, they include sophisticated direct-to-consumer (DTC) diagnostic tools that use algorithms, sensors and "crowdsourcing" [3] to create Web 2.0 personalization and interactivity [4] for functions ranging from triage and differential diagnosis of common ailments to detecting skin changes suggestive of cancer.

With over a quarter million health apps available in major app stores [5], popular DTC diagnostic apps have been downloaded from tens of thousands to tens of millions of times [6]. Possible benefits include faster, more convenient and more targeted information to improve diagnosis [7] and reduction of unneeded visits and tests, but there is also the potential for unintended outcomes [8] such as inappropriate treatment and diagnostic error. The Food and Drug Administration (FDA) has long exempted "low risk" apps from its approval process [9], and the

***Corresponding author: Michael L. Millenson,** BA, Health Quality Advisors LLC, Highland Park, IL 60035, USA; and Northwestern University Feinberg School of Medicine, Department of General Internal Medicine and Geriatrics, Chicago, IL, USA, E-mail: michael@healthqualityadvisors.com. http://orcid.org/0000-0001-8364-1927
Jessica L. Baldwin and Hardeep Singh: Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey VA Medical Center, Houston, TX, USA; and Department of Medicine, Baylor College of Medicine, Houston, TX, USA, E-mail: jessica.baldwin@bcm.edu (J.L. Baldwin); hardeeps@bcm.edu (H. Singh)
Lorri Zipperer: Zipperer Project Management, Albuquerque, NM, USA, E-mail: lorri@zpm1.com

current FDA commissioner has said that apps helping consumers self-diagnose are an innovation that regulations should not impede [10]. Nonetheless, there are as yet no accepted vetting processes enabling clinicians or patients to distinguish between reliable apps and "digital snake oil" [11]. Diagnostic apps specifically have received scant attention in comparison to health management ones, even in overviews of the field [12, 13].

We conducted a scoping review to characterize the current state of evidence on how interactive, DTC diagnostic apps available to consumers perform and what methods are used to evaluate them.

Methods

Funding for our work was provided by the Gordon and Betty Moore Foundation; however, the foundation had no role in study design; collection, analysis and interpretation of data; or approval of final publication. Our scoping review used Arksey and O'Malley's five-stage methodological framework [14] summarized in Table 1.

Formulating research questions

An initial search in PubMed, Google Scholar, and the lay literature through General Reference Center Gold revealed a highly heterogeneous literature in which information of interest was often subsumed in broader examinations of diagnostic and/or health management apps. That search generated four research questions: what clinical conditions do these apps address? What functionality is involved in producing a tentative diagnosis? What methodologies are evaluators using to assess these apps? And what are the results of app evaluations, including evidence on risks and benefits? Our findings were intended to help guide medical practice, consumer choice and health policy by identifying the strengths and weaknesses of the evidence in the current literature and by highlighting evidence gaps.

Identification of relevant studies

With a medical librarian (LZ), we conducted a structured search of PubMed and Google Scholar for the period January 1, 2014–June 30, 2017, focusing on apps suggesting an initial diagnosis and marketed DTC without FDA approval. The timeframe was chosen in an attempt to minimize the inclusion of possibly technologically irrelevant evaluations of older apps. A lack of common keywords and inconsistent indexing made a structured and reproducible PubMed search difficult, leading to an iterative search process. Moreover, as no existing U.S. National Library of Medicine MeSH terms were closely related to our topic, we used broader, related terms such as "smartphone" and "diagnostic self-evaluation". In addition, we manually reviewed selected bibliographies, even if slightly outside the time frame. We also searched the lay literature through General Reference Center Gold and by looking more broadly at trade and general-interest publications, websites and reports from organizations active in this field [15]. We also interviewed physicians, researchers, digital health entrepreneurs and a venture capitalist.

Study selection

We included original research, descriptive studies and literature reviews related to diagnostic software applications consumers might commonly use, whether web-based or apps developed for a specific platform (e.g. iPhone) [16]. We excluded apps subject to FDA approval, those in a research phase, those using physical tests (e.g. Bluetooth-connected pregnancy tests) and static content (e.g. keyword searches).

Two authors (MLM and JLB) assessed full-text articles for relevance, given that an abstract might not accurately reflect whether an evaluation of a particular diagnostic app was performed. When there was a question about article inclusion, it was discussed with a third author (HS).

Data charting

Two authors (MLM and JLB) reviewed articles and organized information pertaining to type of digital platform(s), study design, app attributes, outcomes investigated and major findings [17].

Data summarization

Data was summarized according to app functionality; diseases evaluated; evaluation methodologies (including selection criteria, descriptions of app attributes and testing of diagnostic functionality); and study results.

Table 1: Steps involved in scoping review.

1.1 Formulated direct-to-consumer (DTC) diagnostic app-related research questions	
– Diagnoses/diseases evaluated?	
– App features and technologies evaluated?	
– Methodologies used in evaluations?	
– Evidence about app performance, risks and benefits?	
1.2 Identification of relevant studies	
Defined studies of interest	– English-language – Studies examining interactive apps suggesting provisional/initial diagnosis – Studies examining apps not subject to FDA approval
Initiated structured and iterative search	– Structured search terms, free text, and keywords – PubMed, Google Scholar, and General Reference Center databases from January 1, 2014 through June 30, 2017 – Iterative search with specific MeSH terms (e.g. mobile applications) and broader, related terms (e.g. "smartphone" and "diagnostic self-evaluation") – Manual bibliography search of selected articles, even if slightly outside time frame – Iterative Google keyword and bibliography-driven search of gray literature, including health care, informatics and general interest publications and websites – Interviews with physicians, entrepreneurs, and others
1.3 Study selection	
Inclusion and exclusion criteria applied	– Included original research, research letters, descriptive studies and literature reviews – "App" defined as interactive software, whether on web browser or mobile device, designed to perform a specific function directly for the user – "Direct-to-consumer" defined as marketed directly to individuals and not meant to primarily facilitate conversation with a clinician – "Diagnostic" defined as providing an initial or provisional diagnosis and not mainly providing additional information after initial diagnosis by a physician – "Interactive" defined as excluding physical tests, such as Bluetooth-connected pregnancy tests, and static content, such as search engine keyword searches – Abstracts and/or full text reviewed by two authors (MLM and JLB) to determine if criteria met
1.4 Charting the data	
Charting akin to narrative review, with general and specific information	– Selected articles reviewed by two investigators (MLM and JLB) as to study design, app characteristics, functional outcomes investigated, and other major findings – Reviewer analysis and notes recorded in a spreadsheet to facilitate final summarization using SPIDER format: Sample (clinical category and type of apps); Phenomena of Interest (attributes studied); Design; Evaluation (findings and discussion); and Research type[a]
1.5 Summarizing and reporting results	
Summarization with descriptive narrative aligned with primary research questions	Findings organized according to – Disease area – Technological functionality of apps – Methodology of evaluation – Results

[a]Ref. [17].

Results

Overview of selected studies

We identified 30 peer-reviewed articles and research letters (Tables 2 and 3) and six non-peer reviewed articles [47–52] meeting our definition. Although we focused on diagnostic apps, these were often described within broader studies evaluating medical apps.

Conditions evaluated

The greatest number of articles (10) focused on dermatology-related diagnostic apps, primarily conditions associated with malignancy [20, 25, 28, 34, 36, 39–41, 45, 46]. Next were eight articles on apps providing diagnostic and triage advice for a broad range of conditions [6, 19, 22, 24, 26, 27, 43, 44]. Other diagnostic areas included infectious disease [one article on acute infectious conditions; one article on sexually transmitted infections (STIs) [23, 38]; mental

Table 2: Peer-reviewed descriptive studies of direct-to-consumer (DTC) diagnostic apps.

Article type and descriptors				App attributes described by study				
Authors (year)	**App name(s) or description**	**Functionality of app(s)**	**Diseases or conditions**	**App content or features**	**User characteristics or behaviors**	**Usability or feasibility**	**Content validity/ quality control issues**	**Cost**
Bender et al. (2013) [18]	Smartphone apps for cancer (subset for "early detection")	Symptom checkers; sensors (image processing)	Cancer	X				
Bhattacharyya (2015) [19]	CrowdMed	Crowdsourcing	Various	X	X			
Brewer et al. (2013) [20]	Mobile apps for dermatology (subset for "self-surveillance/diagnosis")	Symptom checkers; sensors (image processing)	Skin diseases	X				X
Brouard et al. (2016) [21]	Mobile apps for cancer (subset for "screening")	Symptom checkers; sensors (image processing)	Cancer	X			X	X
Cheng et al. (2015) [22]	mTurk, oDesk, multiple web-based forums (e.g. Yahoo Answers, WebMD)	Crowdsourcing	Various	X		X		
Gibbs et al. (2017) [23]	Google Play and iTunes apps	Symptom checkers	Sexually transmitted infections (STI)	X		X	X	X
Jutel and Lupton (2015) [6]	Mobile apps to "assist in the process of diagnosis"	Symptom checkers	Various	X			X	
Juusola et al. (2016) [24]	CrowdMed[a]	Crowdsourcing	Various	X	X			
Kassianos et al. (2015) [25]	Smartphone apps for "melanoma detection"	Symptom checkers; sensors (image processing)	Melanoma	X				X
Lupton and Jutel (2015) [26]	Mobile apps for "self-diagnosis"	Symptom checkers	Various	X				X
Meyer et al. (2016) [27]	CrowdMed	Crowdsourcing	Various	X	X			X
Patel et al. (2015) [28]	Subset of apps for "self-surveillance/ diagnosis"	Symptom checkers; sensors (image processing)	Skin diseases	X				X
Pereira-Azevado et al. (2015) [29]	Urology-themed apps	Symptom checkers	Urology	X			X	X
Robillard et al. (2015) [30]	Apps to detect Alzheimer's disease	Symptom checkers	Alzheimer's disease	X		X	X	X
Rodin et al. (2017) [31]	Eye care apps (subset for "conducting self-tests")	Symptom checkers	Eye and vision care	X				
Shen et al. (2015) [32]	Mobile phone apps for depression	Symptom checkers	Depression	X			X	X

[a]Data gathered from survey of CrowdMed users.

Table 3: Peer-reviewed assessments of diagnostic performance of direct-to-consumer (DTC) diagnostic apps.

Author (year)	App(s) evaluated	Functionality of app(s)	Diseases or diagnoses	n	Source of comparison/ reference diagnosis	Source of case material	Method of assessing agreement	Major findings
Bisson et al. (2016) [33]	"Web-based symptom checker for knee pain"	Symptom checker (condition-specific)	Diagnoses related to knee pain	328	Clinical diagnosis from 7 board-certified sports medicine specialists	Self-reported symptoms from clinic patients	Sensitivity, specificity	Sensitivity 58%, specificity 48% for diagnoses selected by patients using app
Dorairaj et al. (2017) [34]	1 app marketed as a "skin cancer prevention tool"	Sensors (image processing)	Skin cancer	26	Clinical diagnosis from 3 plastic surgery residents; histological diagnosis	Photographs of skin lesions from clinic patients	Sensitivity, specificity, NPV, PPV, negative likelihood ratios	Sensitivity 80% (v. 100% for clinical dx); specificity was 9% (v. 55% for clinical dx)
Farmer et al. (2011) [35]	Boots WebMD symptom checker	Symptom checker (general)	Ear, nose and throat symptoms	61	Clinical diagnosis from 1 ENT specialist	Self-reported symptoms from clinic patients	Proportion of suggested differential diagnoses deemed accurate or appropriate	Symptom checker correctly diagnosed 70% of patients; however, top differential dx matched the clinical dx in only 16% of cases
Ferrero et al. (2013) [36]	Skin scan	Sensors (image processing)	Skin cancer	93	All cases of "biopsy-proven melanoma"	Photos from the National Cancer Institute and three medical reference sources	Percentage of photos classified as "high risk" lesions	Most lesions classified as "high" (10.8%) or "medium" (88.2%) risk
Hageman et al. (2015) [37]	WebMD symptom checker	Symptom checker (general)	"Hand and upper extremity conditions"	86	Clinical diagnosis from 1 of 3 orthopedic surgeons (hand specialists)	Self-reported symptoms from new patients	Pearson chi-square test	33% of diagnoses suggested by the app matched the "final diagnosis" of the surgeon
Luger et al. (2014) [38]	Google; WebMD symptom checker	Symptom checker	Acute infectious conditions	79	1 of 2 predefined conditions (vignette material taken from web-based medical reference sites)	Participants used vignettes to simulate self-report of symptoms while using 1 of the 2 tools	Percentage of patients who correctly identified the reference diagnosis	50% of patients using either tool reached correct diagnoses; qualitative findings from "think-aloud" interviews reported
Maier et al. (2015) [39]	SkinVision	Sensors (image processing)	Melanoma	195	Histological and clinical diagnoses (2 physicians)	Photographs of skin lesions taken using the app	Sensitivity, specificity, Overall accuracy	App achieved 81% accuracy versus 95% for clinicians (compared to histological diagnosis)
Nabil et al. (2017) [40]	SkinVision	Sensors (image processing)	Skin lesions	151	Clinical diagnosis (1 physician)	Photographs of skin lesions taken by researcher using the app	Weighted kappa	Inter-observer agreement between app and physician was "very low" (kappa = 0.073)
Ngoo et al. (2017) [41]	SkinVision; SpotMole; Dr. Mole	Sensors (image processing)	Skin lesions	57	Clinical diagnosis (2 physicians)	Photographs of the app taken "according to the instructions provided"	Sensitivity; specificity; kappa	Sensitivity ranged from "21% to 72%"; specificity ranged from "27% to 100%"; all apps had "low overall agreement" with physicians

Table 3 (continued)

Author (year)	App(s) evaluated	Functionality of app(s)	Diseases or diagnoses	n	Source of comparison/ reference diagnosis	Source of case material	Method of assessing agreement	Major findings
Powley et al. (2016) [42]	NHS and Boots WebMD symptom checkers	Symptom checkers	Inflammatory arthritis	34	All patients had a clinical diagnosis of inflammatory arthritis or inflammatory arthralgia (recruited from specialty care clinic)	Self-reported symptoms	Percentage of patients triaged appropriately and percentage of patients whose diagnoses were listed among the top differential diagnoses	56% of patients triaged to appropriate level of care; most patients' diagnoses appeared within top 5 differential diagnoses
Semigran et al. (2015) [43]	23 apps for diagnostic and triage purposes	Symptom checkers	Various	770	Predefined diagnoses described in 25 vignettes	Self-reported symptoms by standardized patients	Percentage of correct diagnoses listed first and in top 20; percentage of patients triaged appropriately	Correct diagnosis listed first in 34% of cases; and in top 3 in 51%; accurate triage advice given in 57% of cases
Semigran et al. (2016) [44]	Human Dx	Crowdsourcing technology, with individual physician responses compared to Semigran 2015	Various	45	Predefined diagnosis described in vignette	Clinician assessment of vignettes ranging in level of difficulty	Percent accuracy for first and top 3 diagnoses	72% of first diagnoses were accurate; in 84% of cases the correct diagnosis was included in top 3
Thissen et al. (2017) [45]	SkinVision	Sensors (image processing)	Skin lesions	108	Clinical and/or histological diagnosis	Lesions from dermatology clinic	Sensitivity and specificity	80% sensitivity and 78% specificity
Wolf et al. (2015) [46]	4 apps that assist users in determining whether a skin lesion may be malignant	Sensors (image processing)	Skin lesions	188	Histological diagnosis	Images taken during clinical care (retrieved from existing database)	Sensitivity, specificity, PPV, NPV	Sensitivity ranged from 6.8% to 98.4%; specificity ranged from 30.4% to 93.7%

"NHS" refers to the British National Health Service. Boots WebMD is branded version of the WebMD symptom checker in Britain.

health issues (one article on depression) [32]; neurology (one article on Alzheimer's disease) [30]; general oncology (two) [18, 21]; orthopedics (one on knee pain [33], one on hand surgery) [37]; eye and vision issues (one) [31]; otolaryngology (one general) [35]; rheumatology (one on inflammatory arthritis) [42]; and urology (one general) [29].

App functionality

The evaluations covered three broad functional categories of apps, with some articles including apps falling into more than one category. The largest category (20) involved medical symptom checkers that apply algorithms to user-answered questions to generate probable diagnoses and/or triage advice. The second most-common category (12) included apps that applied image processing technology and algorithms to smartphone photos. Articles we found were exclusively focused on conditions of the skin and eyes. Finally, five articles involved crowdsourcing using an online, distributed problem-solving model. (A prominent app in this category, CrowdMed, applies an algorithm to diagnostic suggestions submitted online by clinical and non-clinical "medical detectives" and then provides a second opinion.)

Evaluation methodologies

Most studies evaluated multiple apps. However, some focused on a specific app due to app developer funding [24], app prominence (e.g. WebMD's symptom checker) or a desire to show the need for greater regulation [36]. Selection criteria for which apps were included in evaluations appeared somewhat arbitrary. Some studies simply described the presence or absence of particular attributes, such as whether there was a disclosed privacy policy. App cost was not consistently addressed, nor did researchers consistently note that "free" apps may sell user data.

Assessment methodologies ranged from a structured rating grid completed by two expert panels to "think-aloud" feedback from consumers during use. User characteristics that were examined included age, gender, education, income, years of home ownership, health literacy, and computer literacy. As noted in Table 3, some studies engaged multiple experts to review app content and features, while others assessed performance directly by comparing an app's suggested diagnosis to a reference diagnosis from a clinician or other source, such as structured clinical vignettes. Although these apps are classified as low-risk devices by the FDA, it is important to note that we found no studies of accuracy or clinical risks and benefits based upon real-world use by consumers.

Quantitative studies of these apps' accuracy most often expressed their results in terms of percentage of true positives (or percent of responses correctly assigned to a specific category), sensitivity, and/or specificity for app-generated diagnoses when compared to diagnoses from a clinician or other reference source. Less commonly reported quantitative measures included positive predictive value, negative predictive value, and nonparametric statistics (e.g. kappa, chi-square, odds ratio) (Table 3).

Evaluation results

Potential privacy and security problems were highlighted by several studies; e.g. symptom checkers for STIs were rated as "poor to very poor" on informed consent, disclosure of privacy and confidentiality policies and possible conflicts of interest [23]. A similar conclusion was reached in a study of apps for detecting Alzheimer's disease [30].

Meanwhile, the cost of apps was difficult to ascertain. In the most comprehensive information we found, symptom checkers for both professionals and patients were said to range in price from "under $1 to $80 or more" [6]. In a study of dermatological diagnostic and management apps, app prices were given as ranging from 99 cents to $139.99 [20]. In neither study were prices for DTC diagnostic apps broken down separately. Only one of the three studies of the CrowdMed app mentioned its significant cost; i.e. users must offer a minimum $200 reward to the "crowd" of "medical detectives".

Actual diagnostic performance varied widely. A study of 23 general symptom checkers by Semigran et al. found an 80% rate of appropriate triage advice in emergent cases, but just 33% for appropriate self-care suggestions. Still, researchers judged these interactive apps preferable to a static Google search [43]. In a non-peer reviewed "contest", the Babylon Check symptom checker, which was not included in the Semigran study, was pitted against a junior doctor and experienced nurse using a standardized case and compared favorably [47]. A separate, non peer-reviewed article by the app's sponsor concluded that Babylon Check produced accurate triage advice in 88.2% of cases (based on pre-determined case vignettes), vs. 75.5% for doctors and 73.5% for nurses [50]. However, we also found articles calling into question some of the findings and asking for an independent evaluation and additional evidence for its accuracy [53].

Peer-reviewed results of general symptom checkers for particular diseases, rather than for general medical

and triage advice, showed few favorable results. In one study, the diagnosis suggested by a symptom checker matched a final diagnosis related to hand surgery just 33% of the time [37], while in another, a symptom checker provided "frequently inaccurate" advice related to inflammatory joint disease [42]. Specialty symptom checkers – like the general ones, based on answers to user questions – also fared poorly. An app for knee pain diagnoses had an accuracy rate of 58% [33]. Apps to screen for Alzheimer's disease were all rated "poor to very poor", and the authors noted that one tested app even concluded the user had the condition no matter what data were entered [30].

However, when specialty symptom checkers used data directly entered from sensors, they sometimes showed more promise, albeit with significant variability in the findings. For example, while one study warned of substantial potential for patient harm from a dermatology app's misleading results [36], another study of that same app using a different methodology 2 years later found an accuracy rate of 81% in detecting melanoma, a sensitivity of 73% and a specificity of 39.3% [39]. Meanwhile, vision diagnostic apps using sensors and directly targeting consumers received cautiously positive assessments in two non peer-reviewed articles [48, 51].

No studies examined actual patient outcomes. The closest approximation came in two studies of CrowdMed. In one study, patients said the app provided helpful guidance [27], while in another, users had fewer provider visits and lower utilization [24]. The patient's ultimate correct diagnosis was however, never confirmed. There were evaluations of consumer characteristics related to performance with varying results. Luger et al. found that individuals who diagnosed their symptoms more accurately using a symptom checker were slightly younger [38] while Powley et al. concluded that neither age nor gender had a significant impact on usability [42]. Hageman et al. identified more familiarity with the Internet as contributing to "optimal use and interpretation" [37].

Some study designs raised questions of evaluator bias against the interactive apps. Among the criticisms were whether a particular evaluation overweighed relatively rare diagnoses [54] or failed to compare app use for triage to a realistic consumer alternative, such as a telephone triage line [49]. Our scoping review raised similar concerns; e.g. studies in which an orthopedist assessed whether a symptom checker could "guess" the correct diagnosis [37], a dermatologist setting out to show the need for greater regulation [36] and an otolaryngologist comparing a symptom checker's diagnostic accuracy to his own [35]. This potential bias could be due to the tendency to judge algorithms differently than fellow humans [55].

Discussion

Patient diagnosis is evolving "from art to digital data-driven science", both within and outside the exam room [56]. DTC diagnostic technology is rapidly evolving: the second half of 2017, for example, witnessed the widespread online dissemination of a depression-assessment questionnaire [57], as well as with the debut of smartphone enhancements utilizing sensors and AI that target the same condition [58]. The pace of change should inspire urgency to improve the evidence base on app performance. However, most of the studies we identified simply described various apps' attributes, a finding similar to the conclusions of a broad systematic review of mHealth apps [59].

Our findings demonstrate the need to accelerate investments into evaluation and research related to consumer facing diagnostic apps. Conversely, there appears to be some progress in evaluating physician-facing diagnostic apps, such as determining accuracy of diagnosing complex cases by the Isabel clinical decision support system [60] and determining test ordering and diagnostic accuracy of an app for testing and diagnosis for certain hematologic conditions [61]. A recent systematic review and meta-analysis concluded that differential diagnosis generators (often used as apps) "have the potential to improve diagnostic practice among clinicians" [62]. Nevertheless, the review found many studies with poor methodological quality, in addition to high between-study heterogeneity [62].

Based on our review, we make three key recommendations to advance research, policy, and practice. First, researchers should consistently name all individual apps evaluated and provide all results by individual app. Apps are medical devices, and accurate and timely diagnosis is a significant public health issue. Given that some of these publicly available apps seemed to perform far better than others, identification is central to enabling the type of clinician-patient partnership recommended by NAM's Improving Diagnosis report, as well as the accountability that comes from policy oversight and replication of research findings. Since these products are aimed at consumers, price information should also routinely be included.

Second, evaluations of apps should explicitly address underlying technological and functional differences. These may or may not be tied to whether an app

is accessed via a web browser or is downloaded. Functionally, for example, an app relying on algorithmic analysis of answers to questions, even if it is downloaded to a mobile device, is very different than algorithmic analysis of data from that device's sensors. In turn, the technological basis of those algorithms – for example, the use of artificial intelligence (AI) – has substantial future implications. For example, current evidence suggests that the sensor-based diagnoses of DTC dermatology apps are approaching high reliability [40] and that general symptom checker accuracy might be significantly improved with AI [50]. These technological distinctions should be recognized by researchers and can inform evidence-based discussions about the clinical and economic impact of consumer use of DTC diagnostic apps and the appropriate public policy response.

Third, researchers should validate and standardize evaluation methodologies. The Standards for Universal reporting of patient Decision Aid Evaluation (SUNDAE) checklist for decision aids studies may serve as one example [63]. In addition to ensuring that evaluations name individual apps and identify their functionality appropriately, a methodology should include agreed-upon sampling and selection criteria; characteristics related to usability and performance; and standards for assessing sensitivity, specificity, and other measures of app accuracy. These actions will help avoid bias while also ensuring that the evidence base aligns with the varying needs of clinicians, patients, researchers, private-sector entrepreneurs, and policymakers.

Conclusions

Overall, the current evidence base on DTC, interactive diagnostic apps is sparse in scope, uneven in the information provided, and inconclusive with respect to safety and effectiveness, with no studies of clinical risks and benefits involving real-world consumer use. Although some studies we examined rigorously determined the sensitivity and specificity of app-generated diagnoses, methodologies varied considerably. Given that DTC diagnostic apps are rapidly evolving, more frequent and rigorous evaluations are essential to inform decisions by clinicians, patients, policymakers, and other stakeholders.

Acknowledgments: We thank Annie Bradford, PhD for help with the medical editing. We also thank Kathryn M. McDonald, MM, PhD and Daniel Yang, MD for their valuable comments on an earlier draft of this manuscript.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: This project was funded by the Gordon and Betty Moore Foundation, Funder ID: 10.13039/100000936, Grant number: 5492. Dr. Singh is additionally supported by the VA Health Services Research and Development Service (Presidential Early Career Award for Scientists and Engineers USA 14-274), the VA National Center for Patient Safety and the Agency for Healthcare Research and Quality (R01HS022087) and in part by the Houston VA HSR&D Center for Innovations in Quality, Effectiveness and Safety (CIN13-413).

Employment or leadership: None declared.

Honorarium: None declared.

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Improving Diagnosis in Health Care. National Academies of Sciences, Engineering and Medicine. 2015. Available at: http://iom.nationalacademies.org/Reports/2015/Improving-Diagnosis-in-Healthcare.aspx. Accessed: 14 Jun 2016.
2. Fox S, Duggan M. Health Online 2013. 2013. Available at: http://www.pewinternet.org/2013/01/15/health-online-2013/. Accessed: 12 Jul 2017.
3. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. Am J Prev Med 2014;46:179–87.
4. O'Reilly T. What is Web 2.0. 2005. Available at: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html. Accessed: 15 Dec 2017.
5. Research2guidance. The mHealth App Market is Getting Crowded. 2016. Available at: https://research2guidance.com/mhealth-app-market-getting-crowded-259000-mhealth-apps-now/. Accessed: 4 Sep 2017.
6. Jutel A, Lupton, D. Digitizing diagnosis: a review of mobile applications in the diagnostic process. Diagnosis 2015;2:89–96.
7. Singh H, Graber M, Onakpoya I, Schiff GD, Thompson MJ. The global burden of diagnostic errors in primary care. BMJ Qual Saf 2016;26:484–94.
8. McCartney M. How do we know whether medical apps work? Br Med J 2013;346:f1811.
9. Administration USFD. Mobile Medical Applications. Available at: https://www.fda.gov/medicaldevices/digitalhealth/mobilemedicalapplications/default.htm. Accessed: 15 Dec 2017.
10. Comstock J. In past editorials, Trump's FDA pick advocated hands-off approach for health apps. 2017. Available at: http://www.mobihealthnews.com/content/past-editorials-trumps-fda-pick-advocated-hands-approach-health-apps. Accessed: 15 Dec 2017.

11. AMA Wire. Medical innovations and digital snake oil: AMA CEO speaks out. 2016. Available at: https://wire.ama-assn.org/life-career/medical-innovation-and-digital-snake-oil-ama-ceo-speaks-out. Accessed: 15 Dec 2017.
12. Aitken M, Lyle J. Patient Adoption of mHealth: Use, Evidence and Remaining Barriers to Mainstream Acceptance. Parsippany, NY: IMS Institute for Healthcare Informatics. 2015. https://pascale-boyerbarresi.files.wordpress.com/2015/03/iihi_patient_adoption_of_mhealth.pdf. Accessed: 12 July 2017.
13. American Medical Association. Report 6 of the Council on Medical Service (I-16). Integration of mobile health applications and devices into practice. 2016. https://www.ama-assn.org/sites/default/files/media-browser/public/about-ama/councils/Council%20Reports/council-on-medical-service/interim-2016-council-on-medical-service-report-6.pdf. Accessed: 12 July 2017.
14. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Social Res Methodol 2005;8:19–32.
15. Fiordelli M, Diviani N, Schulz PJ. Mapping mHealth research: a decade of evolution. J Med Internet Res 2013;15:e95.
16. TechTarget. Computing Fundamentals. 2007. Available at: http://searchmobilecomputing.techtarget.com/definition/app. Accessed: 12 Jul 2017.
17. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res 2012;22:1435–43.
18. Bender JL, Yue RY, To MJ, Deacken L, Jadad AR. A lot of action, but not in the right direction: systematic review and content analysis of smartphone applications for the prevention, detection, and management of cancer. J Med Internet Res 2013;15:e287.
19. Bhattacharyya M. Studying the Reality of Crowd-Powered Healthcare. Paper presented at: AAAI HCOMP2015.
20. Brewer AC, Endly DC, Henley J, Amir M, Sampson BP, Moreau JF, et al. Mobile applications in dermatology. JAMA Dermatol 2013;149:1300–4.
21. Brouard B, Bardo P, Bonnet C, Mounier N, Vignot M, Vignot S. Mobile applications in oncology: is it possible for patients and healthcare professionals to easily identify relevant tools? Ann Med 2016;48:509–15.
22. Cheng J, Manoharan M, Lease M, Zhang Y. Is there a Doctor in the Crowd? Diagnosis Needed! (for less than $5). Paper presented at: iConference 2015.
23. Gibbs J, Gkatzidou V, Tickle L, Manning SR, Tilakkumar T, Hone K, et al. 'Can you recommend any good STI apps?' A review of content, accuracy and comprehensiveness of current mobile medical applications for STIs and related genital infections. Sex Transm Infect 2017;93:234–5.
24. Juusola JL, Quisel TR, Foschini L, Ladapo JA. The impact of an online crowdsourcing diagnostic tool on health care utilization: a case study using a novel approach to retrospective claims analysis. J Med Internet Res 2016;18:e127.
25. Kassianos AP, Emery JD, Murchie P, Walter FM. Smartphone applications for melanoma detection by community, patient and generalist clinician users: a review. Br J Dermatol 2015;172:1507–18.
26. Lupton D, Jutel A. 'It's like having a physician in your pocket!' A critical analysis of self-diagnosis smartphone apps. Soc Sci Med 2015;133:128–35.
27. Meyer AN, Longhurst CA, Singh H. Crowdsourcing diagnosis for patients with undiagnosed illnesses: an evaluation of crowdmed. J Med Internet Res 2016;18:e12.
28. Patel S, Madhu E, Boyers LN, Karimkhani C, Dellavalle R. Update on mobile applications in dermatology. Dermatol Online J 2015;21.
29. Pereira-Azevedo N, Carrasquinho E, Cardoso de Oliveira E, Cavadas V, Osório L, Fraga A, et al. mHealth in urology: a review of experts' involvement in app development. PLoS One 2015;10:e0125547.
30. Robillard JM, Illes J, Arcand M, Beattie BL, Hayden S, Lawrence P, et al. Scientific and ethical features of English-language online tests for Alzheimer's disease. Alzheimers Dement (Amst) 2015;1:281–8.
31. Rodin A, Shachak A, Miller A, Akopyan V, Semenova N. Mobile apps for eye care in Canada: an analysis of the iTunes store. JMIR Mhealth Uhealth 2017;5:e84.
32. Shen N, Levitan MJ, Johnson A, Bender JL, Hamilton-Page M, Jadad AA, et al. Finding a depression app: a review and content analysis of the depression app marketplace. JMIR Mhealth Uhealth 2015;3:e16.
33. Bisson LJ, Komm JT, Bernas GA, Fineberg MS, Marzo JM, Rauh MA, et al. How accurate are patients at diagnosing the cause of their knee pain with the help of a web-based symptom checker? Orthop J Sports Med 2016;4:2325967116630286.
34. Dorairaj JJ, Healy GM, McInerney A, Hussey AJ. Validation of a melanoma risk assessment smartphone application. Dermatol Surg 2017;43:299–302.
35. Farmer SE, Bernardotto M, Singh V. How good is Internet self-diagnosis of ENT symptoms using Boots WebMD symptom checker? Clin Otolaryngol 2011;36:517–8.
36. Ferrero NA, Morrell DS, Burkhart CN. Skin scan: a demonstration of the need for FDA regulation of medical apps on iPhone. J Am Acad Dermatol 2013;68:515–6.
37. Hageman MG, Anderson J, Blok R, Bossen JK, Ring D. Internet self-diagnosis in hand surgery. Hand (NY) 2015;10:565–9.
38. Luger TM, Houston TK, Suls J. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. J Med Internet Res 2014;16:e16.
39. Maier T, Kulichova D, Schotten K, Astrid R, Ruzicka T, Berking C, et al. Accuracy of a smartphone application using fractal image analysis of pigmented moles compared to clinical diagnosis and histological result. J Eur Acad Dermatol Venereol 2015;29:663–7.
40. Nabil R, Bergman W, KuKutsh NA. Poor agreemenet between a mobile phone application for the analysis of skin lesions and the clinical diagnosis of the dermatologist, a pilot study. Br J Dermatol 2017;177:583–4.
41. Ngoo A, Finnane A, McMeniman E, Tan JM, Janda M, Soyer HP. Efficacy of smartphone applications in high-risk pigmented lesions. Australas J Dermatol 2017;1–8. [Epub ahead of print].
42. Powley L, McIlroy G, Simons G, Raza K. Are online symptoms checkers useful for patients with inflammatory arthritis? BMC Musculoskelet Disord 2016;17:362.
43. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. Br Med J 2015;351:h3480.
44. Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of physician and computer diagnostic accuracy. JAMA Intern Med 2016;176:1860–61.

45. Thissen M, Udrea A, Hacking M, von Braunmuehl T, Ruzicka T. mHealth app for risk assessment of pigmented and nonpigmented skin lesions-a study on sensitivity and specificity in detecting malignancy. Telemed J E Health 2017;23:948–54.
46. Wolf JA, Moreau JF, Patton TJ, Winger DG, Ferris LK. Prevalence and impact of health-related internet and smartphone use among dermatology patients. Cutis 2015;95:323–8.
47. Chapman M. A health app's AI took on human doctors to triage patients. 2016. Available at: https://motherboard.vice.com/en_us/article/z43354/a-health-apps-ai-took-on-human-doctors-to-triage-patients. Accessed: 12 Jul 2017.
48. Shah V, Hemang K, Pandya MD. Smartphones for visual function testing. 2015. Available at: https://www.retinalphysician.com/issues/2015/may-2015/smartphones-for-visual-function-testing. Accessed: 18 Dec 2017.
49. Husain I. Self-diagnosis app study scrutinized the wrong way. 2015. Available at: https://www.imedicalapps.com/author/iltifat/#. Accessed: 12 Jul 2017.
50. Middleton K, Butt M, Hammerla N, Hamblin S, Mheta K, Parsa A. Sorting out symptoms: design and evaluation of the 'Babylon Check' automated triage system. 2016. Available at: https://arxiv.org/abs/1606.02041. Accessed: 12 Jul 2017.
51. Lee L. Portable vision testing kit puts an eye doctor in your smartphone. 2016. Available at: https://newatlas.com/eyeque-personal-vision-tracker/47148/. Accessed: 18 Dec 2017.
52. Hagan P. Can an app really help you spot a risky mole? SkinVision can help you 'be your own doctor' by finding irregularities. 2016. Available at: http://www.dailymail.co.uk/health/article-3845614/Can-app-really-help-spot-risky-mole-SkinVision-help-doctor-finding-irregularities.html. Accessed: 18 Dec 2017.
53. McCartney M. Margaret McCartney: innovation without sufficient evidence is a disservice to all. Br Med J 2017;358:j3980.
54. Fraser HS, Clamp S, Wilson CJ. Limitations of study on symptom checkers. JAMA Intern Med 2017;177:740–1.
55. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 2015;144:114–26.
56. Mandl KD, Bourgeois FT. The evolution of patient diagnosis: from art to digital data-driven science. J Am Med Assoc 2017;318:1859–60.
57. National Alliance on Mental Illness. Google partners with NAMI to shed light on clinical depression. 2017. Available at: https://www.nami.org/About-NAMI/NAMI-News/2017/Google-Partners-with-NAMI-to-Shed-Light-on-Clinica. Accessed: 12 Jul 2017.
58. Morse J. So how worried should we be about Apple's Face ID? 2017. Available at: http://mashable.com/2017/09/14/apple-faceid-privacy-concerns/#oL77nLsigiqV. Accessed: 8 Dec 2017.
59. Grundy QH, Wang Z, Bero LA. Challenges in assessing mobile health app quality: a systematic review of prevalent and innovative methods. Am J Prev Med 2016;51:1051–9.
60. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 2008;23(Suppl 1):37–40.
61. Meyer AN, Thompson PJ, Khanna A, Desai S, Mathews BK, Yousef E, et al. Evaluating a mobile application for improving clinical laboratory test ordering and diagnosis. J Am Med Inform Assoc 2018;25:841–7.
62. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The effectiveness of electronic differential diagnoses (ddx) generators: a systematic review and meta-analysis. PLoS One 2016;11:e0148991.
63. Sepucha KR, Abhyankar P, Hoffman AS, Bekker HL, LeBlanc A, Levin CA, et al. Standards for UNiversal reporting of patient Decision Aid Evaluation studies: the development of SUNDAE checklist. BMJ Qual Saf 2018;27:380–8.