

Quality of condition suggestions and urgency advice provided by the Ada symptom assessment app evaluated with vignettes optimised for Australia*

Stephen Gilbert^{id A,B,C}, Matthew Fenech^A, Shubhanan Upadhyay^A,
Paul Wicks^A and Claire Novorol^A

^AAda Health GmbH, Karl-Liebknecht-Straße 1, 10178 Berlin, Germany.

^BEKFZ for Digital Health, University Hospital Carl Gustav Carus Dresden, Technische Universität Dresden, Dresden, Germany.

^CCorresponding author. Email: science@ada.com

Abstract. When people face a health problem, they often first ask, ‘Is there an app for that?’. We investigated the quality of advice provided by the Ada symptom assessment application to address the question, ‘How do I know the app on my phone is safe and provides good advice?’. The app was tested with 48 independently created vignettes developed for a previous study, including 18 specifically developed for the Australian setting, using an independently developed methodology to evaluate the accuracy of condition suggestions and urgency advice. The correct condition was listed first in 65% of vignettes, and in the Top 3 results in 83% of vignettes. The urgency advice in the app exactly matched the gold standard 63% of vignettes. The app’s accuracy of condition suggestion and urgency advice is higher than that of the best-performing symptom assessment app reported in a previous study (61%, 77% and 52% for conditions suggested in the Top 1, Top 3 and exactly matching urgency advice respectively). These results are relevant to the application of symptom assessment in primary and community health, where medical quality and safety should determine app choice.

Keywords: artificial intelligence, clinical decision support, health app governance, patient-centred care, self-evaluation in healthcare, triage.

Received 22 February 2021, accepted 11 May 2021, published online 14 October 2021

Introduction

A high proportion of Australian adults have access to the Internet and own smartphones (Hill *et al.* 2020), and approximately 80% of Australians report searching the Internet for health information (Cheng and Dunn 2015). Symptom assessment applications (SAAs) are algorithmic smartphone and Internet programs that ask patients questions about the problem concerning them, their demographics, relevant medical history and symptoms. The programs then use a range of algorithmic approaches to suggest one or more conditions that may best explain the symptoms. SAAs generally also suggest the next steps to patients, such as self-care at home versus seeking an urgent consultation, often alongside evidence-based condition information.

Hill *et al.* (2020) evaluated the most prominent freely available SAAs in Australia using clinical scenarios, known as vignettes, specifically adapted to those cases most typically encountered by Australian physicians, including shingles, heart attacks and viral upper respiratory infections. The authors selected the apps by identifying the most prominent in Internet search engines and app stores, using the terms ‘symptom

checker’ and ‘medical diagnosis’, ‘health symptom diagnosis’ and ‘symptom’ (Hill *et al.* 2020). They excluded apps that required account creation because they deemed these would limit user access in situations in which immediate advice was wanted (Hill *et al.* 2020, 2021; Gilbert *et al.* 2021). However, this exclusion criterion does not fairly reflect real-world SAA use, because many users create accounts through general interest or to address one issue, and then subsequently reuse the SAA for advice on later unrelated health problems. In addition, account creation for recent SAAs takes very little time to complete. The CE-marked Ada SAA was not identified or selected by Hill *et al.* (2020), despite being freely available in Australia since 2016 (Elder 2018) and despite being downloaded at least 200 times more in Australia during the period November 2018–January 2019 than the other apps included in that study, as addressed in Gilbert *et al.* (2021) and Hill *et al.* (2021). On the Ada SAA, 23 638 assessments were completed in Australia in the 12 months up to 11 April 2021, with 24% completed by (or for) men, 76% completed by women and 80% of assessments for users <35 years of age.

*A preprint version of this article is available at <https://www.medrxiv.org/content/10.1101/2020.06.16.20132845v1>.

In addition to other components of performance surveillance of the Ada app as described in Gilbert *et al.* (2020), Miller *et al.* (2020) and Morse *et al.* (2020), the aim of this study was to apply the methodology and vignettes of Hill *et al.* (2020) to provide a greater understanding of the results for the 36 free-to-Australian users SAAs evaluated by Hill *et al.* (2020).

Methods

One software/application was assessed in this study: the Ada symptom assessment application. Our searches in app stores and Internet search engines did not identify any other SAAs excluded, due to either the choice of search terms or the need for user account creation before use, from the study of Hill *et al.* (2020).

Patient vignettes

The vignettes used in this study were those created by Hill *et al.* (2020). Briefly, 30 patient vignettes from the well known study of Semigran *et al.* (2016) were selected and adapted by Hill *et al.* (2020), and supplemented with 18 new symptom-based scenarios, designed with reference to the scientific literature to include several Australia-specific conditions. The urgency advice of the vignettes was allocated into four triage categories: emergency, urgent, non-urgent and self-care. The Ada app's eight urgency advice levels can be mapped to these advice levels, as indicated in Table 1. This mapping is identical to the mapping used for the SAAs evaluated in the study of Hill *et al.* (2020). The gold standard vignette solutions (gold standard diagnosis and triage) are listed in Supplementary Table S1 and are the same as specified by Hill *et al.* (2020).

Among the conditions covered by the 48 vignettes, 85% are for common conditions and 15% are for uncommon conditions, which provides a good representation of Australian general practice presentations, where 10% of presentations relate to uncommon conditions (Cooke *et al.* 2013). The full vignettes are summarised in Table S1 and are listed in the data supplement of Hill *et al.* (2020).

The lay language summaries of the vignettes and the primary complaint identified were used to enter the vignettes into the Ada app. To maintain consistency, one investigator entered the information for each vignette (SG). This author has not been involved in the development of Ada's medical intelligence, question flow or interface design. The vignettes were entered between 12 and 14 June 2020 using an android smartphone and using version 3.5.0 of the Ada symptom assessment application because it is available on the Australian android (Google) app store. The Ada app has a broad coverage of user populations (e.g. childhood conditions, conditions in pregnant women), and it was therefore possible to enter all 48 vignettes.

Condition suggestion performance

The Ada app provides the user with between one and five condition suggestions at the end of each symptom assessment. 'Accurate condition suggestion' was defined as including the gold standard diagnosis as the top result (Top 1), or as being among the Top 3 or Top 10. In this paper, the Top 10 potential condition suggestions are listed to allow for easy comparison with the websites and apps evaluated in Hill *et al.* (2020). However, because the Ada app provides a maximum of five

Table 1. Mapping of the Ada app's levels of urgency advice to the gold standard triage categories from Hill *et al.* (2020)

Ada level	Ada advice level	Gold standard vignette triage advice
1	Call ambulance	Requires emergency care
2	Go to emergency department	Requires emergency care
3	See primary care within 4 h	Requires urgent care
4	See primary care same day	Requires urgent care
5	See primary care 2–3 days	Non-urgent care reasonable
6	See primary care 2–3 weeks	Non-urgent care reasonable
7	Do self-care/see pharmacist	Self-care reasonable
8	Do self-care	Self-care reasonable

condition suggestions, the Top 10 is always equal to the Top 5 suggestions. 'Incorrect condition suggestion' was defined as the correct condition not being included in the Top 5 results. The decision of whether or not the condition suggested by the Ada app was a match for the gold standard diagnosis was made by an author (SU) who has over 5 years of primary care and emergency department clinical experience. Strict matching criteria were applied: the condition provided by the Ada app must have fallen into the set listed for the gold standard diagnosis by Hill *et al.* (2020), with alternative medical names for the same condition being allowed. The outcome measures for condition suggestion accuracy used in the present study are exactly the same as 'diagnosis accuracy' used by Hill *et al.* (2020).

Urgency advice performance

Urgency advice accuracy was defined as the provision of a level of urgency advice that matched the gold standard vignette triage, as defined by Hill *et al.* (2020). The Ada app always provides a single overall urgency advice for the symptom assessment, so an unambiguous evaluation was possible for each vignette. The outcome measure for urgency advice accuracy used in the present study is the same as 'triage accuracy' used by Hill *et al.* (2020).

Data analysis

Simple descriptive statistical methods have been used to report the performance of the Ada app in the same format as the other apps assessed by Hill *et al.* (2020). For each vignette, the matching process determined whether there was a Top 1, Top 3 or Top 10 match, and the individual vignettes scores were expressed as a mean value for all 48 vignettes. The mean proportion of matching advice was reported for all vignettes, and for vignettes subdivided by gold standard triage advice level ('emergency care required', 'urgent care required', 'non-urgent care reasonable', 'self-care reasonable').

Results

Condition suggestion performance

The correct condition suggestion was listed first in 65% of vignettes and included among the first three results in 83% of vignettes. The condition suggestion results are summarised in Table 2 and the complete solution provided by the Ada app for the full set of vignettes is provided in Table S1.

Table 2. Accuracy of the condition suggestions provided by the Ada app

The data in this table can be compared with table S4 in Hill *et al.* (2020) to provide context to the SAAs evaluated in that study. Data show the number of vignettes listed first or in the Top 3 and Top 10 of 48 vignettes, with percentages in parentheses

Condition suggestion matching the vignette gold standard diagnosis		
Listed first	Listed in Top 3	Listed in Top 10
31 (65)	40 (83)	40 (83)

Table 3. Accuracy of the urgency advice provided by the Ada app

The data in this table can be compared with table S6 in Hill *et al.* (2020) to provide context to the SAAs evaluated in that study. Data are given as *n* (%)

Levels of Ada app urgency advice matching the vignette gold standard triage				
All cases (<i>n</i> = 48 vignettes)	Emergency care required (<i>n</i> = 13 vignettes)	Urgent care required (<i>n</i> = 14 vignettes)	Non-urgent care reasonable (<i>n</i> = 11 vignettes)	Self-care reasonable (<i>n</i> = 10 vignettes)
30 (63)	12 (92)	6 (43)	6 (55)	6 (60)

Urgency advice accuracy

Urgency advice exactly matched the vignette gold standard in 63% of cases, including just under 67% of emergency and urgent cases and 57% of less serious case vignettes. The urgency advice accuracy results are summarised in Table 3 and the complete solution provided by the Ada app for the full set of vignettes is provided in Table S1.

Discussion

Principal findings

It was possible to enter all study vignettes into the Ada app. The app provided the correct condition in the Top 1 in the case of 65% of vignettes, compared with 12–61% reported by Hill *et al.* (2020), and as a suggestion in the Top 3 in 83% of vignettes, compared with 23–77% in Hill *et al.* (2020). The urgency advice of the Ada app exactly matched the vignette gold standard in 63% of all cases. The range of performance of the apps evaluated by Hill *et al.* (2020), which provided any urgency advice, was 17–61%. The urgency advice of the Ada app exactly matched the vignette gold standard in 67% of emergency and urgent cases and in 57% of less serious case vignettes, compared with advice that matched exactly in 49% of all cases, including 60% of emergency and urgent cases but only 30–40% of less serious case vignettes for the apps evaluated by Hill *et al.* (2020). The global performance is reported above, and the performance of SAAs may not be uniform across all clinical areas. The number of vignettes in this study (48) is not sufficient to allow quantification of area-specific accuracy. For the 18 vignettes specifically developed for the Australian context, the Ada SAA listed the correct condition in 50% of vignettes, and in the Top 3 results in 67%, and 56% of urgency advice exactly matched the gold standard. Although performance is lower for the Ada SAA in these vignettes than in the non-Australian-specific vignettes,

this performance is still markedly superior to average SAA performance reported by Hill *et al.* (2020). Direct comparison of performance between the Ada SAA and the SAAs evaluated by Hill *et al.* (2020) is not possible because these data were not provided.

Comparisons with the wider literature on the performance of symptom assessment apps

The finding of this study, namely that the Ada app has relatively high accuracy with regard to condition suggestion and urgency compared with other available SAAs, reflects the findings of other studies (Nateqi *et al.* 2019; Cenev *et al.* 2020; Gilbert *et al.* 2020). The Ada app was recently compared with general practitioners (GPs) and competitor apps in a 200-vignette study (Gilbert *et al.* 2020). Many symptom assessment evaluations have focused on a single symptom assessment app or a specific medical subdiscipline/speciality with a small number of vignettes (Chambers *et al.* 2019), whereas the study by Gilbert *et al.* (2020) was a collaborative study with an independent university group. That study compared the condition coverage, accuracy and safety of eight popular symptom assessment apps to each other and to seven GPs. The performance of Ada app was closest to that of the human doctors; it offered 99% condition coverage for the vignettes, gave safe advice 97% of the time (the same performance as GPs) and provided the correct suggested conditions in its Top 3 approximately 70% of the time.

Relevance of the results to the Australian setting and implications for clinicians and policy makers

The results of Hill *et al.* (2020) show that the best-performing SAAs evaluated have good urgency advice accuracy but that many SAAs perform badly in terms of the accuracy of condition suggestion and urgency advice. It is possible that as a consequence of being one of the most widely used SAAs in Australia (Gilbert *et al.* 2021), and its optimisation based on user feedback, the Ada app performed better than some of the other apps assessed by Hill *et al.* (2020). In an editorial on the study of Hill *et al.* (2020), Dunn (2020) considered that 61% exact match to optimal advice insufficient. Nevertheless, it was also acknowledged that relatively conservative advice from SAAs is appropriate (Dunn 2020). Rørtveit *et al.* (2013) showed that GP telephone triage is often risk averse and moderate over-cautiousness is appropriate for safety, commenting that 'Pre-hospital triage of emergency patients is necessarily an inexact process and some degree of overtriage must generally be accepted'. Related to this, we showed that advice safety (defined as the proportion of urgency advice at gold standard level, more conservative or no more than one level less conservative on a six-level advice scale) of the best-performing SAAs was the same as GPs. Moreover, although not as accurate as GPs in Top 1 condition suggestion, the best apps are close to GP performance in providing the correct condition in their Top 3 and Top 5 condition suggestions (Gilbert *et al.* 2020).

Ethics, governance and international cooperation on validation

Dunn (2020) called for transparent surveillance of SAA performance to provide a firm basis for integrating symptom checkers

into the health system. Although more work needs to be done, there are significant advances towards such transparency: (1) the [World Economic Forum \(2020\)](#) is spearheading a collaborative framework exploring the governance of conversational artificial intelligence (AI) in health care with a particular focus on patient expectations, transparency, explainability, bias, fairness and data privacy/data rights issues; (2) the World Health Organization and International Telecommunications Union are developing an independent standard evaluation framework for benchmarking of AI algorithms, including SAAs, using confidential datasets ([Wiegand *et al.* 2019](#)); and (3) stronger European regulatory oversight of SAAs came into effect in May 2021 (<https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745>), including the requirement for proactive post-market surveillance (PMS) in all countries in which the SAAs are used. Ada is involved in the first two of these initiatives, along with a broad group of clinicians, researchers and industry collaborators.

GPs have a role in educating the public about app choice, and patients should be recommended apps with peer-reviewed clinical data, CE mark or equivalent regulatory approval or a minimum of a certified medical device quality management system ([International Organization for Standardization Technical Committee 210 \(ISO/TC 210\) 2016](#); [Leigh and Ashall-Payne 2019](#)).

Patient 'self-evaluation' and its role in community health

SAAs have millions of home users and their use is predicted to grow over the coming decades ([Hammond 2019](#)). Survey results show broad interest in the use of SAAs as 'at-home' tools to help users understand their symptoms and what next steps to take ([Meyer *et al.* 2020](#)). This offers convenience to users who would otherwise have to travel far or endure a long wait to see a physician, as well as to those who would like an additional perspective on their symptoms, or for whom primary care consultations are expensive.

'Patient-centred' linkages to the primary healthcare system and into secondary and tertiary care

Home SAA use has the potential to reduce the burden on primary and secondary care and to navigate patients to appropriate care ([Winn *et al.* 2019](#); [Miller *et al.* 2020](#); [Morse *et al.* 2020](#)). Confirmatory prospective studies of these findings are in progress. In a recent study, [Morse *et al.* \(2020\)](#) described their experience integrating an SAA in northern California for Sutter Health, in which 26 646 assessments were delivered in a broad patient population to provide assistance outside typical physician office hours. In that study, SAA triage recommendations were comparable to those of nurse-staffed telephone triage lines. Several studies have explored symptom assessment in diagnostic decision support (DDS; [Ramnarayan *et al.* 2007](#); [Ronicke *et al.* 2019](#)). In a retrospective study, [Ronicke *et al.* \(2019\)](#) explored the DDS potential of an Ada system, finding that the Ada system accelerated the diagnosis of rare diseases and provided correct disease suggestions earlier than the clinical diagnosis in 56% of cases, and that 33% of patients could have been identified as having a rare disease in the first documented clinical visit.

Study strengths and limitations

The vignettes in this study primarily described simple scenarios in which the patients did not have comorbidities. Clinical vignette-based studies are highly applicable to initial evaluation of SAAs in a specific clinical context, but further evaluation in the hands of real users is required to understand use, as well as effects on care delivery and patient safety ([Fraser *et al.* 2018](#)). The investigator who entered the vignettes in this study was familiar with the Ada symptom assessment app as a user/evaluator, but he was not involved in the development of the app's medical intelligence, question flow or user interface. Because only one investigator (SG) entered the vignettes into the SAA, there is no inter-rater reliability measure reported. The decision of whether the SAA suggested condition matched the gold standard diagnosis was made by a second coauthor (SU). The solution provided by the Ada app for the full set of vignettes can be referred to and verified in full, because this is provided in Table S1. The strengths of the study include independently created vignettes described in lay language and including conditions specific to Australia. A strength of vignettes studies is that they enable systematic comparisons without interfering with clinical care.

Conclusions

The Ada SAA has not been previously evaluated in an Australia-specific context, but has higher accuracy in condition suggestion and urgency advice than other SAAs thus far evaluated in this context using clinical vignettes. These results have relevance to the role of SAAs in primary and community health and to debates regarding the use of health apps, where medical quality and safety should determine app choice.

Conflicts of interest

Stephen Gilbert, Matthew Fenech and Shubhanan Upadhyay are employees or former employees of Ada Health GmbH; Claire Novorol is a cofounder and shareholder in Ada Health GmbH. The Ada Health research team has received research grant funding from Fondation Botnar and the Bill & Melinda Gates Foundation. Paul Wicks is an Employee of Wicks Digital health (WDH), a consultant to Ada, and has received speaker fees from Bayer and honoraria from Roche, the Fondazione Italiana di ricerca per la Sclerosi Laterale Amiotrofica (ARISLA), the American Medical Informatics Association (AMIA), the Innovative Medicines Initiative (IMI), Statisticians in the Pharmaceutical Industry (PSI) and the British Medical Journal (BMJ). Shubhanan Upadhyaya is a cochair of the Clinical Evaluation Working Group of the International Telecommunications Union/World Health Organization Focus Group on Artificial Intelligence for Health (FG-AI4H). Matthew Fenech is a contributor to the World Economic Forum Chatbots RESET Framework for governing responsible use of conversational AI in health care.

Declaration of funding

This study was funded by Ada Health GmbH.

References

- Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T (2020) Accuracy of online symptom checkers and the potential impact on service utilisation. *medRxiv* 2020.07.07.20147975. doi:[10.1101/2020.07.07.20147975](https://doi.org/10.1101/2020.07.07.20147975)

- Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, Turner J (2019) Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* **9**, e027743. doi:10.1136/bmjopen-2018-027743
- Cheng C, Dunn M (2015) Health literacy and the Internet: a study on the readability of Australian online health information. *Australian and New Zealand Journal of Public Health* **39**, 309–314. doi:10.1111/1753-6405.12341
- Cooke G, Valenti L, Glasziou P, Britt H (2013) Common general practice presentations and publication frequency. *Australian Family Physician* **42**, 65–68.
- Dunn AG (2020) Will online symptom checkers improve health care in Australia? *The Medical Journal of Australia*. doi:10.5694/mja2.50621
- Elder J (2018) The robot doctor will see you now. *The Sydney Morning Herald*. Available at <https://www.smh.com.au/lifestyle/health-and-wellness/the-robot-doctor-will-see-you-now-20180810-p4zwpv.html> [Verified 6 November 2020]
- Fraser H, Coiera E, Wong D (2018) Safety of patient-facing digital symptom checkers. *Lancet* **392**(10161), 2263–2264. doi:10.1016/S0140-6736(18)32819-8
- Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, Millen E, Montazeri M, Multmeier J, Pick F, Richter C, Türk E, Upadhyay S, Virani V, Vona N, Wicks P, Novorol C (2020) How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. *BMJ Open* **10**, e040269. doi:10.1136/bmjopen-2020-040269
- Gilbert S, Wicks P, Novorol C (2021) The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. *The Medical Journal of Australia* **214**, 143–143.e1. doi:10.5694/mja2.50917
- Hammond R (2019) The world in 2040. The future of healthcare, mobility, travel and the home. Future health, care and wellbeing. Allianz Partners. Available at <https://www.allianz-partners.com/content/dam/onemarketing/awp/azpartnerscom/italy/futurologo/en/Allianz-Partners-The-World-in-2040-Health-Care-Wellbeing-Report1.pdf> [Verified 2 May 2021]
- Hill MG, Sim M, Mills B (2020) The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. *The Medical Journal of Australia* **212**, 514–519. doi:10.5694/mja2.50600
- Hill MG, Sim M, Mills B (2021) The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. *The Medical Journal of Australia* **214**, 143–143.e1. doi:10.5694/mja2.50923
- International Organization for Standardization Technical Committee 210 (ISO/TC 210) (2016) ISO 13485:2016: medical devices – quality management systems – requirements for regulatory purposes. Available at <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/97/59752.html> [Verified 2 March 2021]
- Leigh S, Ashall-Payne L (2019) The role of health-care providers in mHealth adoption. *The Lancet Digital Health* **1**, e58–e59. doi:10.1016/S2589-7500(19)30025-1
- Meyer AN, Giardina TD, Spitzmueller C, Shahid U, Scott TM, Singh H (2020) Patient Perspectives on the Usefulness of an Artificial Intelligence–Assisted Symptom Checker: Cross-Sectional Survey Study. *Journal of Medical Internet Research* **22**, e14679.
- Miller S, Gilbert S, Virani V, Wicks P (2020) Patients' Utilization and Perception of an Artificial Intelligence–Based Symptom Assessment and Advice Technology in a British Primary Care Waiting Room: Exploratory Pilot Study. *JMIR Human Factors* **7**. doi:10.2196/19713
- Morse KE, Ostberg NP, Jones VG, Chan AS (2020) Use Characteristics and Triage Acuity of a Digital Symptom Checker in a Large Integrated Health System: Population-Based Descriptive Study. *Journal of Medical Internet Research* **22**, e20549. doi:10.2196/20549
- Nateqi J, Lin S, Krobath H, Gruarin S, Lutz T, Dvorak T, Gruschina A, Ortner R (2019) Vom symptom zur diagnose – tauglichkeit von symptom-checkern. *HNO* **67**, 334–342. doi:10.1007/s00106-019-0666-y
- Ramnarayan P, Cronje N, Brown R, Negus R, Coode B, Moss P, Hassan T, Hamer W, Britto J (2007) Validation of a diagnostic reminder system in emergency medicine: a multi-centre study. *Emergency Medicine Journal* **24**, 619–624. doi:10.1136/emj.2006.044107
- Ronicke S, Hirsch MC, Türk E, Larionov K, Tientcheu D, Wagner AD (2019) Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. *Orphanet Journal of Rare Diseases* **14**, 69. doi:10.1186/s13023-019-1040-6
- Rørtveit S, Meland E, Hunskaar S (2013) Changes of triage by GPs during the course of prehospital emergency situations in a Norwegian rural community. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* **21**, 89–89.e1. doi:10.1186/1757-7241-21-89
- Semigran HL, Levine DM, Nundy S, Mehrotra A (2016) Comparison of physician and computer diagnostic accuracy. *JAMA Internal Medicine* **176**, 1860–1861. doi:10.1001/jamainternmed.2016.6001
- Wiegand T, Krishnamurthy R, Kuglitsch M, Lee N, Pujari S, Salathé M, Wenzel M, Xu S (2019) WHO and ITU establish benchmarking process for artificial intelligence in health. *Lancet* **394**, 9–11. doi:10.1016/S0140-6736(19)30762-7
- Winn AN, Somai M, Fergestrom N, Crotty BH (2019) Association of Use of Online Symptom Checkers With Patients' Plans for Seeking Care. *JAMA Network Open* **2**, e1918561. doi:10.1001/jamanetworkopen.2019.18561
- World Economic Forum (WEF) (2020) Chatbots RESET: a framework for governing responsible use of conversational AI in healthcare. WEF. Available at <https://www.weforum.org/reports/chatbots-reset-a-framework-for-governing-responsible-use-of-conversational-ai-in-healthcare/> [Verified 2 March 2021]