BMJ Health & Care Informatics

Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records

Videha Sharma ,[1,2] Ibrahim Ali,[3] Sabine van der Veer,[1] Glen Martin,[1] John Ainsworth,[1] Titus Augustine[2]

To cite: Sharma V, Ali I, van der Veer S, *et al.* Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *BMJ Health Care Inform* 2021;**28**:e100253. doi:10.1136/bmjhci-2020-100253

Received 09 October 2020
Accepted 03 February 2021

[1]Centre for Health Informatics, The University of Manchester, Manchester, UK
[2]Department of Renal and Pancreatic Transplantation, Manchester University NHS Foundation Trust, Manchester, UK
[3]Department of Renal Medicine, Salford Royal NHS Foundation Trust, Salford, UK

Correspondence to
Dr Videha Sharma;
videha.sharma@postgrad.manchester.ac.uk

INTRODUCTION

Prognostic risk prediction models aim to estimate the risk of a future outcome based on available clinical parameters.[1] There has been an increase in the development of such models given the move towards personalised and precision medicine, since they provide individualised risks for patients.[2 3] They can help convey risks and benefits more succinctly and promote shared decision-making. Despite their benefits, risk prediction models in front-line clinical practice remain underutilised and their potential impact on care outcomes has not been fullfilled.[4] A recent systematic review of clinical decision support systems by Kwan *et al* published in the *BMJ* demonstrated only a poor to moderate improvement of care and highlighted the importance of designing models and tools that critically consider care processes and patient outcomes.[5] Ongoing challenges include poor methodological development and lack of external validation of models.[6 7] However, where robust models have been externally validated, an underappreciated barrier to their adoption in clinical practice is the lack of integration with electronic health records (EHRs).

LACK OF INTEGRATION AS A BARRIER TO USE

Clinical risk prediction models have clear potential to influence clinical decision-making and enhance the quality of care delivered to patients.[8 9] However, developing a successful model is a rigorous process with many pitfalls, such as incomplete training data, risk of bias and failure to address clinical need. There are further challenges to externally validate and calibrate a model across different patient groups before being accepted for clinical use.[10] As a result, although there is a large body of literature on the development of risk prediction models, the evidence of successful clinical adoption and impact on care outcomes is largely absent.[11]

Risk prediction models are primarily developed using routinely collected clinical data, increasingly retrieved from EHRs.[12 13] Thus, the variables selected and assessed during model development are those available in electronic data repositories, such as demographics, diagnostic results, medical history or drug history. Some models that were robustly validated and gained international recognition were converted to online tools and made available through web-interfaces or mobile applications. An example of such a model is the CHA(2)DS(2)-VASc score, which is used to predict the risk of stroke in patients with atrial fibrillation (AF) and thus guide anticoagulation.[14] It has successfully achieved clinical impact and is the gold-standard risk prediction model for AF management as recommended by the National Institute of Health and Care Excellence.[15] To use the model however, a healthcare professional is required to access a website or open an app, manually complete data fields with the patients' details and receive a risk score to guide clinical decision-making. Though this task may seem trivial compared with the potential added benefit of greater quality decision-making, the practicalities and time constraints of clinical practice form a significant barrier to usage. This is compounded with the potential of manual transcription errors, which form a hazard of receiving incorrect results.

Another example of this is the kidney failure risk equation (KFRE) developed by Tangri *et al*, which is similarly available as an online tool.[16] This model uses routinely collected clinical data including patient age, gender, estimated glomerular filtration rate and urinary albumin:creatinine ratio, to provide a 2-year and 5-year risk of progression to kidney failure for patients with chronic kidney disease. The KFRE has been validated internationally, and is generally reviewed positively.[17] However, its widespread adoption is limited by the dependence on externally

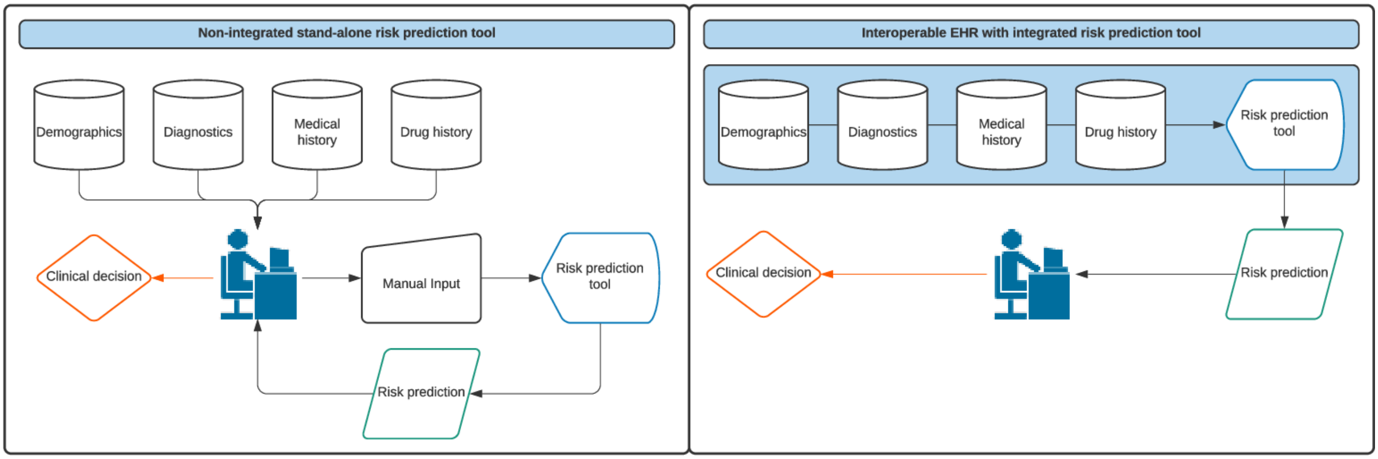

Figure 1 Current and proposed use of risk prediction tools. EHRs, electronic health records.

accessing the tool online and manually transcribing the variables into data fields before a risk score is presented. This impractical process has been shown to contribute as a barrier to clinical impact in primary care settings.[18]

A number of initiatives have attempted to increase the usability of risk prediction tools by developing user-friendly interfaces. An example of this is MDCalc(c), which is a medical calculator available via a website and a mobile application. By making the content easy to navigate and using an intuitive visual design it aims to enhance the user experience. However, the fundamental barrier of accessing the interface as an external application and manually completing fields is yet to be overcome. This is particularly a challenge as many healthcare institutions still lack interoperable EHRs and store clinical data across multiple digital systems. This means that a healthcare professional wishing to use a risk prediction tool may have to access multiple electronic sources to gather the required data to complete the fields and obtain a risk.

As healthcare providers increasingly turn to unified EHRs, the success of risk prediction models will be dependent on the integration of tools within these systems. Usability barriers may be mitigated if clinicians can access risk prediction tools, pertinent to their practice, within their local EHR and have a risk score presented automatically as fields are populated with relevant data from within the system. This intuitively simple concept would create a paradigm shift for the practical daily use of such tools and translate to patient benefit (figure 1). Risks may be presented graphically over a period of time to illustrate the impact of risk-addressing therapies and thus promote compliance. By improving accessibility in this way, it will also have an impact on future academic research evaluating these tools' performance and impact on clinical outcomes. Currently, research into usability of risk prediction tools, as standalone interfaces, or within EHRs is largely absent. User experience is a significant part of successful product development in areas outside of healthcare and formal methodology for evaluation in other fields has been established. Recognising the importance of this as part of model development is crucial to achieve value out of future solutions.[19]

FUTURE CONCEPTS

For risk prediction models that have undergone rigorous validation and assessment of clinical impact, integrating tools into EHRs will likely overcome a major barrier to use. Unfortunately, this practical implication has not been widely explored and new tools continue to appear as web-interface solutions risking non-adoption and thus failure to impact care. An example of such a recent model is the iPREDICTLIVING (2019) developed to predict risks around kidney donation to better inform renal transplant decision-making.[20] In the context of a sensitive and complex clinical decision as kidney donation, detracting the clinician from the human interaction by a time-consuming on-screen process will likely impact the patient experience. Digital health interventions should be centred around improving the quality of care delivered to patients, which includes better decisions, but also enhancing the patient–clinician relationship by providing clinicians the time to consult patients.

To realise a more streamlined workflow, a change in how we think about clinical risk prediction models is required. Frontline usability should be part of the initial exploration of the proposed model. This means involving clinicians (end-users) at the outset as part of research projects to understand how the tool would be practically used and the impact it would have on clinical encounters. The usability of such interventions plays a crucial role in preventing clinician fatigue and improving uptake.[21]

Technical challenges revolve around non-standardised coding of health data across EHR providers.[22] This means that even if a risk prediction model is made available as a standalone software, which can be integrated, misaligned clinical terminology may limit implementation. Involving EHR vendors early in the development of risk prediction models and imploring greater alignment across the industry will mitigate barriers to implementation and subsequent scale-up of novel solutions. An example of successful tool integration is QRISK, which has been embedded within a number of primary care clinical management systems.[23] The tool calculates individual cardiovascular risk and generates a score based on existing data. Not only has this impacted positively on front-line practice, regular use provides evolving data quality and completeness reflecting the changing population characteristics over time. This has allowed researchers to update and calibrate the tool for long-term accuracy.[24] Similar implementation through hospital EHR vendors may bring such models into routine secondary care settings unifying and standardising practice. Another relevant example was the PREDICT software used in general practice in New Zealand, which automatically recorded patients' risk profiles for cardiovascular disease and prospectively linked this to coded hospital and mortality databases. This allowed a risk prediction model to be developed that took in to account an area-based deprivation index and self-reported ethnicity alongside clinical parameters, resulting in greater personalised risk profiles for individual patients. The strength of this study was its prospective nature and ability to seamlessly collect healthcare data without additional intervention by clinicians delivering care.[25]

The tremendous potential of clinical risk prediction models mandates policy-makers to establish regulations to standardise the integration of tools into EHRs. Strategies to achieve this may be through EHR vendors working directly with data scientists to incorporate statistical models within their user interface, or alternatively provide non-proprietary application programming interfaces for third party developers to seamlessly integrate with. The potential success of this however, heavily relies on the engagement of front-line healthcare professionals who can provide the clinical context and workflow that a risk prediction model is intending to influence. Encouraging multidisciplinary research and development teams, which can appreciate the different facets of clinical context, statistical modelling and implementation science, supported by EHR vendors working to unified standards has the potential to bridge the current bench-to-bedside gap for clinical risk prediction models.

Twitter Videha Sharma @VidehaSharma

Contributors VS conceptualised the manuscript. VS and IA reviewed the literature and wrote the manuscript. GM reviewed the literature and edited the manuscript. SV, JA and TA reviewed and edited the manuscript. VS created the manuscript figures.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; internally peer reviewed.

ORCID iD

Videha Sharma http://orcid.org/0000-0001-7640-1239

REFERENCES

1 Steyerberg EW. *Clinical prediction models*. Springer, 2019.
2 Banning M. A review of clinical decision making: models and current research. *J Clin Nurs* 2008;17:187–95.
3 Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Acad Emerg Med* 2002;9:1184–204.
4 Ahmed I, Debray TPA, Moons KGM, *et al*. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Med Res Methodol* 2014;14:3.
5 Kwan JL, Lo L, Ferguson J, *et al*. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. *BMJ* 2020;370:m3216.
6 Wynants L, Van Calster B, Collins GS, Bonten MM, *et al*. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
7 Collins GS, de Groot JA, Dutton S, *et al*. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.
8 Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014;2:3.
9 Topol E. The Topol review: preparing the healthcare workforce to deliver the digital future. *Health Educ J* 2019.
10 Chen L. Overview of clinical prediction models. *Ann Transl Med* 2020;8:71.
11 Dekker FW, Ramspek CL, van Diepen M. Con: most clinical risk scores are useless. *Nephrol Dial Transplant* 2017;32:752–5.
12 Goldstein BA, Navar AM, Pencina MJ, *et al*. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24:198–208.
13 Rothman B, Leonard JC, Vigoda MM. Future of electronic health records: implications for decision support. *Mt Sinai J Med* 2012;79:757–68.
14 Lip GYH, Nieuwlaat R, Pisters R, *et al*. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro heart survey on atrial fibrillation. *Chest* 2010;137:263–72.
15 National Institute of Health and Care Excellence. *Clinical guideline 180 (CG180): atrial fibrillation: management*, 2014.
16 Tangri N, Stevens LA, Griffith J, *et al*. A predictive model for progression of chronic kidney disease to kidney failure. *JAMA* 2011;305:1553–9.
17 Peeters MJ, van Zuilen AD, van den Brand JAJG, *et al*. Validation of the kidney failure risk equation in European CKD patients. *Nephrol Dial Transplant* 2013;28:1773–9.
18 Major RW, Shepherd D, Medcalf JF, *et al*. The kidney failure risk equation for prediction of end stage renal disease in UK primary care: an external validation and clinical impact projection cohort study. *PLoS Med* 2019;16:e1002955.
19 Vermeeren AP, Roto V. User experience evaluation methods: current state and development needs. *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*, 2010:521–30.
20 Haller MC, Wallisch C, Mjøen G, *et al*. Predicting donor, recipient and graft survival in living donor kidney transplantation to inform pretransplant counselling: the donor and recipient linked iPREDICTLIVING tool - a retrospective study. *Transpl Int* 2020;33:729–39.
21 Sutton RT, Pincock D, Baumgart DC, *et al*. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020;3:1–10.
22 D'Amore JD, Mandel JC, Kreda DA, *et al*. Are meaningful use stage 2 certified EHRs ready for interoperability? findings from the smart C-CDA collaborative. *J Am Med Inform Assoc* 2014;21:1060–8.
23 Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.
24 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099.
25 Pylypchuk R, Wells S, Kerr A, *et al*. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. *Lancet* 2018;391:1897–907.

BMJ Health & Care Informatics

Mothers intention and preference to use mobile phone text message reminders for child vaccination in Northwest Ethiopia

Zeleke Abebaw Mekonnen ,[1] Kassahun Alemu Gelaye,[2] Martin C. Were,[3] Binyam Tilahun[1]

To cite: Mekonnen ZA, Gelaye KA, Were MC, *et al*. Mothers intention and preference to use mobile phone text message reminders for child vaccination in Northwest Ethiopia. *BMJ Health Care Inform* 2021;**28**:e100193. doi:10.1136/bmjhci-2020-100193

Received 11 June 2020
Revised 29 December 2020
Accepted 03 February 2021

[1]Department of Health Informatics, Institute of Public Health,College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
[2]Department of Epidemiology and Biostatistics, Institute of Public Health,College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Correspondence to
Zeleke Abebaw Mekonnen;
Zelekeabebaw7@gmail.com

ABSTRACT

Objectives With the unprecedented penetration of mobile devices in the developing world, mHealth applications are being leveraged for different health domains. Among the different factors that affect the use of mHealth interventions is the intention and preference of end-users to use the system. This study aimed to assess mother's intention and preference to use text message reminders for vaccination in Ethiopia.

Methods A cross-sectional study was conducted among 460 mothers selected through a systematic random sampling technique. Initially, descriptive statistics were computed. Binary logistic regression analysis was also used to assess factors associated with the outcome variable.

Results In this study, of the 456 mothers included for analysis, 360 (78.9%) of mothers have intention to use text message reminders for vaccination. Of these, 270 (75%) wanted to receive the reminders a day before the vaccination due date. Mothers aged 35 years or more (AOR=0.35; 95% CI: 0.15 to 0.83), secondary education and above (AOR=4.43; 95% CI: 2.05 to 9.58), duration of mobile phone use (AOR=3.63; 95% CI: 1.66 to 7.94), perceived usefulness (AOR=6.37; 95% CI: 3.13 to 12.98) and perceived ease of use (AOR=3.85; 95% CI: 2.06 to 7.18) were predictors of intention to use text messages for vaccination.

Conclusion In conclusion, majority of mothers have the intention to use text message reminders for child vaccination. Mother's age, education, duration of mobile phone use, perceived usefulness and perceived ease of use were associated with intention of mothers to use text messages for vaccination. Considering these predictors and user's preferences before developing and testing text message reminder systems is recommended.

Summary

What is already known?

- The immunisation programme in Ethiopia is challenged with a lack of effective methods to track vaccination schedules.
- Mobile phone short message service is a widely applicable appointment reminder intervention to improve healthcare.
- For the development and effective implementation of mHealth interventions, considering contextual differences and user preferences are crucial.

What does this paper add?

- This study determined users intention and preferences in a resource-limited setting which helped to design a locally tailored automated text message reminder system for the immunisation programme in Ethiopia.
- The results of this study were used as an input to design and test the effectiveness of a locally developed automated text message reminder system for the immunisation programme in Ethiopia.

BACKGROUND

Timely completion and uptake of the childhood vaccination is key to reducing the high morbidity and mortality of vaccine-preventable diseases (VPDs) among infants globally. Maintaining reductions in mortality from VPD relies on continued immunisation uptake that is reliant on parental decision-making and subsequent attendance at health facilities.[1] However, many children still miss scheduled vaccines in the extended programme of immunisation or are being vaccinated after the recommended ages.[1 2]

Adherence to childhood vaccination schedules is a function of various factors including the information gaps both from the service supply and demand sides.[3] The immunisation programme is also challenged with a lack of effective methods to track vaccination schedules.[4] Immunisation programmes usually involve the use of the child health card as a tool for reminding caregivers of children of the dates of their next vaccination.[3] However, it was observed that the majority of the mothers who missed their vaccination appointment were due to forgetfulness and difficulty in tracking vaccination schedules indicating a need to identify more innovative approaches. This necessitates the establishment of an appropriate and uninterrupted vaccine delivery strategy with more focus on demand-side interventions.[5–7]

To date, there is a continuous growth of mobile network coverage and unprecedented penetration of mobile devices globally.[8] By the end of 2018, 5.1 billion people around the world subscribed to mobile services, accounting for 67% of the global population.[9] In the same year, mobile subscribers in Ethiopia reached 44%.[10] A study conducted on mobile phone access in Gondar city among pregnant women reported that 76.7% of mothers owned a mobile phone. Among those women who had mobile phones, 90% were able to read text messages using their mobile phones.[11] With these advancements, leveraging mobile health (mHealth) applications in the health sector is becoming popular.[8]

mHealth is the use of mobile phone technology to deliver healthcare.[12] According to the WHO, mHealth has the ability to transform the delivery of healthcare and bring a paradigm shift in healthcare delivery processes all over the world.[13] mHealth is now extensively used in healthcare and there is a growing global trend in harnessing this technology for behaviour change, disease surveillance, prevention and control of various health problems and enhancing attendance for health services. Hence, the field of mHealth has been proposed as a potential solution to many of the challenges that developing countries face.[13–21]

mHealth applications and programmes make use of several aspects of mobile technology such as text messaging, voice and video services.[12] The WHO reported that short message service (SMS) is the the most common mobile phone features used for appointment reminders.[13] It is widely applicable appointment reminder intervention to improve healthcare-seeking behaviours considering participant characteristics such as forgetfulness and lack of knowledge.[22–24] Mobile phone-based text messaging demonstrates strong potential as a tool for healthcare improvement for several reasons; applicability on almost every model of mobile phone, relatively low cost and widely applicable to a variety of health behaviours and conditions.[25 26]

Implementing new technologies is inherently challenging. According to the non-adoption, abandonment, scale-up, spread and sustainability framework, the condition, the technology, the value proposition, the adopter system (comprising professional staff and clients), the organisational infrastructure, the context and the interaction between all these domains determine effective implementation of new technological innovations.[27] Evidence also indicated that mHealth initiative success is based on the accessibility, acceptance, effective adaptation to local contexts and strong stakeholder collaboration.[28–30] It is also important to take into account the diverse environment with cultural and contextual differences to adopt new technological interventions.[31–37]

Among the various factors contributing for the successful implementation of mHealth interventions, end-users perception and value propositions to the new system are crucial worth considering before actual implementation.[27 35 38 39] According to the theory of reasoned action, the adoption of new intervention is dependent on the behavioural intention of users. Effective technology use is also the result of an intention in making the behaviour, and this intention is influenced by the perceived ease of use and perceived usefulness including user's preference.[37 39 40]

The programme theory for this study is that clients will use the proposed SMS-based mHealth intervention if they have intention to use the system, the system is designed based on their preference and they believe that it will provide positive results.[37] Hence, investigating the user's intention and preference is crucial to design and implement more effective mHealth interventions in developing countries including Ethiopia.[28 31 41–47] Therefore, this study aimed to assess the intention and preference of mothers to use mobile phone text message reminders for child vaccination in northwest Ethiopia.

METHODS

Study design and setting

A health-facility based cross-sectional study was conducted from 1 October to 26 October 2018 in Gondar city administration, northwest Ethiopia. Gondar city administration has a total of 24 Kebele's (the smallest administrative unit in Ethiopia). From the total kebeles, 13 are urban and 11 are rural kebeles. The city administration had an estimated total population of 390 644. Of these, 12 149 were under 1 year of age. The city administration has also a total of 23 public health facilities.[48]

Source and study populations

The source population consisted of mothers paired with infants attending the vaccination units at health facilities. The study population included those eligible mother–infant pairs who visited the selected health facilities during the study period.

Inclusion and exclusion criteria

Those mothers of infants who visited vaccination units of health facilities and remaining with at least one vaccination appointment were included. Mothers who resided in the study area for at least 6 months prior to the study period and who owned a mobile phone were included for this particular study. Mothers whose infants had already received the last doses of vaccines were excluded from the study.

Sample size determination and sampling procedures

We could not find any study conducted in Ethiopia to determine the intention of mothers to use text message reminders for routine vaccination. Therefore, we did a pilot study to determine the proportion of those mothers who have the intention to use the text message reminders and it was found to be 77.6%. Finally, the sample size required for this study was determined by considering the following assumptions; proportion of intention to use text message reminder for child vaccination as 77.6% (from pilot study), 95% CI and 4% margin of error. With these assumptions, the sample size was 418. Taking a 10% non-response rate, the final sample size was 460.

All the eight health centres and the comprehensive specialised hospital in Gondar city were included in this study. The sample in each health facility was allocated proportionally to the number of clients who vaccinated their infants in the same period of the previous year. A systematic random sampling technique was applied to select the study participants. To select 460 study participants from the 2058 eligible participants, the sampling interval was calculated to be 4.4 which is rounded off to the nearest whole number 4. Accordingly, every fourth client who presented to the selected health facilities for their infant's vaccination were included in this study.

Study variables

The outcome variable was the intention to use text message reminders for vaccination. Based on the technology acceptance model (TAM), perceived ease of use and perceived usefulness were considered as predictor variables for this study.[49] Additionally, the sociodemographic characteristics of mothers were included as predictors for the outcome of interest.

Intention to use mobile text message reminders was defined as the user's likelihood to use mobile phone text message reminders for child vaccination.[31 43 50–52] Perceived ease of use was defined as the extent to which a person believes that using a particular system (in this case the text message reminder) would be free from effort.[31 43 50 51] Perceived usefulness was defined as the degree to which a person believes that using a particular system (in this case the text message reminder) would enhance his or her task (in this case timely vaccination of children).[31 43 50 51]

Items for the composite variables were measured on a 5-point Likert-type scale ranging from 'strongly disagree' (score 1) to 'strongly agree' (score 5). Item scores for each composite variable were added and divided by the number of items to create a composite variable scale (ranging from score 1 to 5) for data analysis.[53 54] Finally, the composite variable score was dichotomised as 'Yes' or 'No' based on the final score. Accordingly, final score of above three (agree and strongly agree) were categorised as 'Yes' while those final scores of three or below (strongly disagree, disagree and neutral) were categorised as 'No'.[55 56]

The household wealth index was created by principal components analysis, including items on asset ownership, housing characteristics and ownership of animals and farming. The household wealth index was calculated separately for urban and rural residents.

Data collection tools and procedures

The data collection instrument for this study was adapted from the scales used in the TAM which has four major variables: perceived usefulness, perceived ease of use, behavioural intention and actual use. The scales perceived usefulness and perceived ease of use were adapted from Davis's study[51] and the scale intention to use was adapted from Venkatesh *et al*'s study[55] to fit the study context. During adaption, the data collection instrument underwent forward and backward translation. Face and content validity of the data collection instrument was assessed by six experts and the proposed changes from the expert panels were considered for refinement of the data collection instrument.

Then, before the actual data collection, a pilot study was done out of the study area, in the health facilities of Bahir Dar city administration with a sample size of 100. The results of the pilot study were used to assess the validity and reliability of the data collection instrument. The internal consistency for each dimension of the data collection instrument was checked using Cronbach's alpha and scores on perceived usefulness (Cronbach alpha=0.95), perceived ease of use (Cronbach alpha=0.91) and intention to use text message reminders (Cronbach alpha=0.93) were deemed acceptable. Finally, nine data collectors and three supervisors were recruited for the actual data collection. Face to face interview technique was used to collect data from eligible study participants using the validated data collection instrument.

Statistical analysis

The data were checked for completeness, entered into Epi-data V.3.1 and exported to STATA V.14 software for analysis. Descriptive statistics on frequencies and percentages were computed and have been presented using graphs and tables. A binary logistic regression analysis model was used to identify the predictor variables for intention to use text message reminders for child vaccination. Finally, the results were reported as adjusted odds ratio (AOR) with their 95% CIs.

Multicollinearity and model fit statistics

The presence of multicollinearity was checked among independent variables using variance inflation factor (VIF) at a cut-off point of 10. Finally, the Hosmer and Lemeshow goodness of fit test was used to check the model fit.

RESULTS

Sociodemographic characteristics

In this study, a total of 460 study participants were included with a response rate of 99.1%. The mean (SD) age of mothers was 27.2 (4.9) years. As shown in table 1, 260 (57%) mothers belonged to an age group of 25–34 years. The majority of the mothers were currently married (90.8%), orthodox by religion (87.5%) and more than

Table 1 Sociodemographic characteristics of mothers who vaccinated their infants in health facilities of in Gondar city administration, northwest Ethiopia, 2018 (n=456)

Characteristics	Total (%)
Age of mother	
≤24	138 (30.3)
25–34	260 (57.0)
≥35	58 (12.7)
Marital status	
Currently married	414 (90.8)
Currently not married	42 (9.2)
Religion	
Orthodox	399 (87.5)
Muslim	45 (9.9)
Others	12 (2.6)
Mother's education	
No formal education	62 (13.6)
Primary	144 (31.6)
Secondary and above	250 (54.8)
Mother's occupation	
Housewife	263 (57.7)
Employed	60 (13.2)
Merchant	89 (19.5)
Others	44 (9.6)
Residence	
Rural	28 (6.1)
Urban	428 (93.9)
Family size	
<5	304 (66.7)
≥5	152 (33.3)
Household wealth index	
Poor	153 (33.6)
Middle	152 (33.3)
Rich	151 (33.1)
Distance to health facility (in minutes)	
<15 min	192 (42.1)
15–30 min	213 (46.7)
>30 min	51 (11.2)

half (54.8%) of the mothers attained secondary education and above (table 1).

Pertaining to the occupation of the mothers, the largest 263 (57.7%) of the total mothers were housewives followed by merchants 89 (19.5%). The study also indicated that the vast majority (93.9%) of the mothers have resided in urban kebeles (table 1).

Mobile phone utilisation

Three hundred and twenty-four (71.1%) of the mothers have been using mobile phones for more than two years and 232 (50.9%) of the mothers were using regular(standard) mobile phones.

Table 2 Mobile phone utilisation of mothers who vaccinated their infants in health facilities of Gondar city administration, northwest Ethiopia, 2018

Characteristics	Total (%)
Duration of mobile phone use	
<1 year	54 (11.8)
1–2 years	78 (17.1)
>2 years	324 (71.1)
Type of current mobile phone	
Regular/Standard	232 (50.9)
Smart	224 (49.1)
Changed phone number in the last 12 months	
No	425 (93.2)
Yes	31 (6.8)
Have additional phone number	
No	411 (90.1)
Yes	45 (9.9)
Usually experienced mobile network challenges	
No	415 (91.1)
Yes	41 (8.9)
Problem keeping a mobile phone charged	
No	420 (92.1)
Yes	36 (7.9)
Switch off mobile phone during day time	
No	427 (93.6)
Yes	29 (6.4)
Can read mobile text message	
No	41 (9)
Yes	415 (91)
Can send mobile text message	
No	58 (12.7)
Yes	398 (87.3)
Shared mobile phone with others in the house	
No	400 (87.7)
Yes	56 (12.3)

Mobile phone network challenges were usually encountered by 41 (8.9%) of the participants and 36 (7.9%) encountered a problem to keep their mobile phones charged. Regarding text message use, 415 (91%) and 398 (87.3%) of the mothers can read and send mobile text messages, respectively. About phone sharing, 56 (12.3%) shared their mobile phones to household members (table 2).

Intention to use text message reminders for child vaccination

In this study, 360 (78.9%) with 95% CI (74.9% to 82.4%) of mothers intended to use text message reminders for child vaccination, if offered the opportunity. Three hundred and eighty-eight (85.1%) of the mothers perceived the

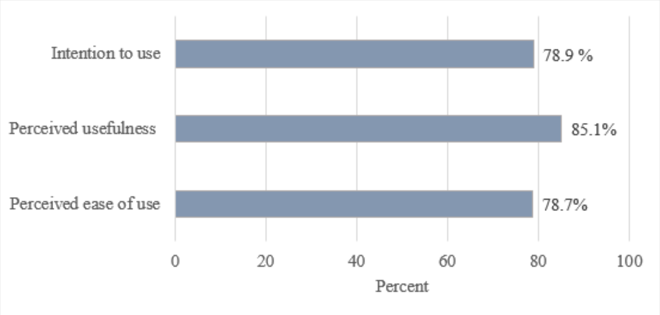

Figure 1 Intention, perceived usefulness and perceived ease of use in using text message reminders for vaccination, Gondar city administration, northwest Ethiopia (n=456).

mobile phone-based text message reminders as useful to child vaccination. Similarly, 359 (78.7%) of the mothers perceived the mobile phone-based text message reminders for child vaccination as easy to use (figure 1).

Preference of mothers to receive text message reminders for child vaccination

Most of the mothers preferred receiving text message reminders starting the first vaccination appointment (47.8%) followed by a second vaccination appointment (31.1%). Regarding the number of text messages, around two-thirds (64.2%) preferred to receive one text message reminder per each vaccination appointment. The study also indicated that three-fourths (75%) and 131 (36.4%) of the mothers wanted to receive the text message reminder a day before the due date and on the due date of the vaccination appointment, respectively. Regarding language preference, more than half (58.1%) of the mothers would prefer to receive text messages in Amharic (national) language while 38.9% preferred to receive the text message in both Amharic and English languages (table 3).

Predictors of intention to use text message reminders for child vaccination

Bivariable and multivariable binary logistic regression analyses were done to determine the association between the intention to use text message reminders and covariates. Accordingly, those variables which had a p value of less than 0.2 in the bivariable regression analysis (mothers age, mother's educational status, mother occupation, marital status, household wealth index, place of residence, duration of mobile phone use, mobile phone type, perceived ease of use and perceived usefulness) were considered for the multivariable regression analysis.

In the final multivariable logistic regression model, the variables mother age, mother education, duration of mobile phone use, perceived ease of use and perceived usefulness were found to have a significant association with intention of mothers to use text message reminders for vaccination.

As shown in table 4, mothers above 35 years of age were 65% less likely (AOR=0.35; 95% CI: 0.15 to 0.83) to have the intention to use text message reminders for child vaccination than those who are less than 25 years of age after controlling for other variables. Mothers who had primary education were 2.7 times more likely (AOR=2.75; 95% CI: 1.25 to 6.05) and who had secondary education and above were 4.4 times more likely (AOR=4.43; 95% CI: 2.05 to 9.58) to have intention to use text message reminders for child vaccination than those who had no formal education.

The study also indicated that perceived ease of use has a positive and significant effect on the mother's intention to use text message reminders for child vaccination. Keeping other factors constant, those who perceived the text message reminder as easy to use were 3.8 times more likely (AOR=3.85; 95% CI: 2.06 to 7.18) to have intention to use text message reminders for child vaccination as compared with their counterparts. Similarly, mothers who perceived the text message reminder as useful were 6.3 times more likely (AOR=6.37; 95% CI: 3.13 to 12.98) to have intention to use text message reminders for child vaccination as compared with their counterparts.

In the final multivariable model, marital status, occupation, household wealth index and the type of mobile phone mothers are currently using did not have a

Table 3 Preference of mothers to receive text message reminders for child vaccination in Gondar city administration, northwest Ethiopia, 2018 (n=360)

Characteristics	Total (%)
Preferred appointment to begin receiving reminders	
First appointment	172 (47.8)
Second	112 (31.1)
Third	47 (13.1)
Fourth	29 (8.1)
Preferred number of text messages per visit	
One	231 (64.2)
Two	115 (31.9)
Three	14 (3.9)
Preferred date to receive text message reminders	
On due date	131 (36.4)
A day before due date	270 (75)
Two days before due date	82 (22.8)
Three days before due date	16 (4.4)
Others	4 (1.1)
Preferred time of the day for receiving text message reminders	
Morning (06:01–before 12:00)	86 (23.9)
Afternoon (12:00–18:00)	147 (40.8)
Evening (18:01–00:00)	30 (8.3)
Any time	97 (26.9)
Preferred language	
Amharic only	209 (58.1)
Both Amharic and English	140 (38.9)
English only	11 (3.1)

Table 4 Bivariable and multivariable binary logistic regression analysis of factors associated with intention to use text message reminders for child vaccination in Gondar city, northwest Ethiopia, 2018

	Intention to use (n)			
Characteristics	No	Yes	COR (95% CI)	AOR (95% CI)
Age of mother				
≤24	23	115	1	1
25–34	48	212	0.88 (0.51 to 1.53)	0.77 (0.38 to 1.55)
≥35	25	33	0.26 (0.13 to 0.52)	0.35 (0.15 to 0.83)
Mother’s education				
No formal education	33	29	1	1
Primary	29	115	4.51 (2.37 to 8.59)	2.75 (1.25 to 6.05)
Secondary and above	34	216	7.23 (3.90 to 13.39)	4.43 (2.05 to 9.58)
Marital status				
Currently married	81	333	1	1
Currently not married	15	27	0.44 (0.22 to 0.86)	0.63 (0.24 to 1.64)
Mother’s occupation				
Housewife	60	203	1	1
Employed	11	49	1.32 (0.64 to 2.69)	0.94 (0.37 to 2.39)
Merchant	11	78	2.09 (1.05 to 4.19)	1.19 (0.50 to 2.83)
Others	14	30	0.63 (0.32 to 1.27)	0.91 (0.36 to 2.31)
Household wealth index				
Poor	50	103	1	1
Middle	24	128	2.58 (1.49 to 4.49)	1.27 (0.60 to 2.68)
Rich	22	129	2.85 (1.62 to 5.01)	1.15 (0.50 to 2.61)
Duration of mobile use				
<1 year	21	33	1	1
1–2 years	31	47	0.96 (0.47 to 1.96)	1.08 (0.45 to 2.61)
>2 years	44	280	4.05 (2.15 to 7.62)	3.63 (1.66 to 7.94)
Type of current mobile phone				
Regular/Standard	66	166	1	1
Smart	30	194	2.57 (1.59 to 4.15)	1.40 (0.69 to 2.88)
Perceived ease of use				
Not easy	50	47	1	1
Easy	46	313	7.24 (4.37 to 11.99)	3.85 (2.06 to 7.18)
Perceived usefulness				
Not useful	37	31	1	1
Useful	59	329	6.65 (3.83 to 11.56)	6.37 (3.13 to 12.98)

AOR, adjusted odds ratio; COR, crude odds ratio.

significant association with intention of mothers to use mobile phone text message reminders for child vaccination.

Multicollinearity and model fitness

A multicollinearity test was performed for the variables included in the final multivariable model. Hence, the variable place of residence had a VIF value of 12.3 and was removed from the final model due to its multicollinearity effect. The final model fitness was also assessed using Hosmer and Lemeshow test. The Hosmer and Lemeshow test showed that the model fits the data well (p value of 0.905).

DISCUSSION

The findings of this study showed that mothers have a high intention to use mobile phone text message reminders for their child’s vaccination. Mother’s age, educational status, duration of mobile phone use, perceived ease of use and perceived usefulness were significantly associated with intention of mothers to use mobile phone-based text message reminders for vaccination. Mothers preferred to receive mobile phone-based text messages one day before the due date of vaccination and in Amharic (national) language.

This study indicated that the majority of mothers have intentions to use text message reminders for child vaccination. This finding corroborated findings from a study in Lagos Nigeria.[57] A willingness study on pregnant women in Gondar city also reported consistent findings where around three-fourths of women were willing to receive text messages.[11] However, this finding was slightly higher than a finding from another study in Nigeria.[41] On the other hand, this finding is lower than a study finding from Kenya.[8] The difference might be due to the difference in the information communication technology infrastructure and investment in digitalisation across countries.

In this study, the educational status of mothers was positively associated with their intention to use mobile phone-based text message reminders for child vaccination. This finding is in accordance with other studies.[5 44 57–59] This may be explained by the fact that educated women are likely to be aware of incoming text messages and are likely to read and act on the received messages promptly. Evidence also showed that literacy status was shown to be an issue in text message reminder system implementation that has to be addressed when text message reminder system is being planned for implementation.[25] A potential drawback to implementing a mobile-phone-based text messaging intervention is that it requires the recipient to have a mobile phone and an adequate level of literacy, marginalising some population groups who could potentially benefit from the mHealth intervention.[21] In our study population, this could affect around 14% of women having no formal education.

The study also found that perceived usefulness has a positive significant association with intention of mothers to use text message reminders for child vaccination. This finding is consistent with other studies where users who did not believe in the possible advantages of e-Health were less inclined to use e-Health.[58–61] End-users need to perceive the system as being useful or they will not attempt to use it regardless of how easy or difficult it is to use. Therefore, during system development, there is a need to ensure that the system will improve the intended health outcomes.[25 62]

The findings also showed that perceived ease of use was positively associated with intention of mothers to use text message reminders for child vaccination. This finding corroborates with the findings of other studies.[54 58–60] When users have no or little previous experience of using a system, they usually pay more attention to the system's ease of use. This implies that users would be unwilling to use a new mHealth service regardless of how useful the system would be if they perceive it to be difficult to use. Research also showed that users will stop using mHealth interventions that are not user friendly.[54] Difficulty in using a new system could be solved if the user thinks that the system will be useful to them. One study reported that training users on the new mobile health technology improves perceived ease of use and, thereby, increases intention to use the actual system.[54] Hence, deployment of mHealth initiatives may require extra guidance on how to operate and use the new system for improved implementation.[25 62]

Mobile services are mainly designed for individual users, who may have different expectations and needs in accordance with their preferences. To develop an effective text message reminder system for vaccination, parental preferences must be fully understood and taken into consideration before deployment.[34] In this study, more than half of mothers would like to receive the text message reminders in Amharic (national) language. This finding is consistent with evidence from India.[26] On the contrary, from studies in Nigeria[7 44 57] majority of the mothers preferred English language for reminders on their mobile phones which could be attributed to their high literacy levels.

For successful implementation of mHealth programmes, clients should be able to choose when and how frequently they would receive text messages.[21] The findings of this study indicated that the majority of mothers preferred to receive one text message reminder per each vaccination appointment. The optimal timing most preferred by mothers for receiving the text message reminders is the day preceding the vaccination appointment date which corroborates the findings in other studies.[5 41 57] This might be because sending text messages to mothers 1 day before their vaccination appointments will increase the chances of the messages being seen and help them to get prepared for their child vaccination appointments ahead of time.

This study also showed that marital status, mother's occupation, household wealth index and type of current mobile phone were not found to have a significant association with intention of mothers' to use mobile text messages for child vaccination. In another study, it was also reported that the type of mobile phone did not have a significant association with intention to use SMS reminders.[11] Thus, the type of mobile phone the mother had and the differences in economic status would not be a major challenge for implementing text message reminder interventions for child vaccination. Though it did not have a significant effect in another study,[11] the variable place of residence has been removed from the final model due to its multicollinearity effect.

Implications for practice and research

This study has practical implications in particular for immunisation programme managers. Given the high proportion of mothers who had intention to use mobile phone-based text message reminders for vaccination, incorporating mobile text messages is a promising avenue to strengthen the routine immunisation programme in Ethiopia. If designed appropriately by considering user's preference in terms of frequency, timing and language;

text message-based mHealth interventions may be an innovative way for engaging users in care for improved child vaccination outcomes. The study also provides a basis for further interventional studies that can develop and assess the effectiveness of mobile text messaging interventions as a tool to improve the routine immunisation programme in Ethiopia.

Limitations

The findings of this study should be interpreted in light of some limitations. First, we studied intention for text message-based appointment reminder for those who already had a mobile phone and visiting vaccination units of health facilities in northwest Ethiopia. So, the findings may not be generalisable to the population of the entire country particularly for those residing in rural areas.

As most mHealth programmes focus on those with mobile phones, a potential drawback to the use of mobile phone-based text-message-reminders is the potential marginalisation of certain populations, such as those that do not have a mobile phone. However, these limitations may be reduced as mobile technology advances and mobile subscriptions grow in developing countries. This study also did not address the ecological and systemic barriers to implementation beyond user's intention to use the technology which demand further research.

CONCLUSION

In this study, we found that majority of mothers have intention to use mobile phone text message reminders for child vaccination. Most of the mothers also would like to receive the text message reminders in Amharic language one day before the vaccination due date. Predictors of mothers' intention to use mobile phone text message reminders include mother's age, mother's education, duration of mobile phone use, perceived ease of use and perceived usefulness of the proposed system.

Considering these predictors and user's preferences indicated in this study, the development of an automated mobile phone-based text message reminder system and testing its effectiveness is recommended for the immunisation programme in Ethiopia.

Acknowledgements The authors would like to thank the German Academic Exchange Service (DAAD) and Doris Dukes Charitable Foundation (DDCF) for supporting this study through the University of Gondar. Also, the authors are grateful for the study participants, data collectors, supervisors and stakeholders involved in this study.

Contributors ZAM conceived the study. KAG, MCW and BT guided the design and conduct of the study. All the authors were involved in data analysis and manuscript write-up. All authors read and approved the final manuscript.

Funding This work was supported by the University of Gondar with grant number 57428660. The University of Gondar has no role in the design, data collection, analysis and interpretation of the study findings.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study obtained ethical approval from the University of Gondar Institutional Ethical Review Board (ref no: O/V/P/RCS/05/060/2018). Also, study permission was acquired at all levels, and informed written consent was obtained from study participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data will be available from the corresponding author upon reasonable request.

ORCID iD

Zeleke Abebaw Mekonnen http://orcid.org/0000-0003-2923-468X

REFERENCES

1 Machingaidze S, Wiysonge CS, Hussey GD. Strengthening the expanded programme on immunization in Africa: looking beyond 2015. *PLoS Med* 2013;10:e1001405.
2 Patel TA, Pandit NB. Why infants miss vaccination during routine immunization sessions? study in a rural area of Anand district, Gujarat. *Indian J Public Health* 2011;55:321–3.
3 FMOH. *Ethiopia national expanded programme on immunization comprehensive multi- multi - year plan 2016 - 2020 Federal Ministry of Health, Addis Ababa*, 2016.
4 Oladepo O, Dipeolu IO, Oladunni O. Nigerian rural mothers' knowledge of routine childhood immunizations and attitudes about use of reminder text messages for promoting timely completion. *J Public Health Policy* 2019;40:459–77.
5 Odinaka K, Edelu B, Achigbu K. Acceptance of mobile phone short message service for childhood immunisation reminders by Nigerian mothers. *Port Harcourt Med J* 2018;12:127.
6 Abahussin AA, Albarrak AI. Vaccination adherence: review and proposed model. *J Infect Public Health* 2016;9:781–9.
7 Akinrinade OT, Ajayi IO, Fatiregun AA. Ownership of mobile phones and willingness to receive childhood immunisation reminder messages among caregivers of infants in Ondo state, south-western Nigeria. *South African J Child Heal* 2018;12.
8 Kazi AM, Carmichael J-L, Hapanna GW, *et al*. Assessing mobile phone access and perceptions for Texting-Based mHealth interventions among expectant mothers and child caregivers in remote regions of northern Kenya: a survey-based descriptive study. *JMIR Public Health Surveill* 2017;3:e5.
9 GSMA. *The mobile economy*, 2019.
10 Bank TW. *Ethiopia digital foundations project*, 2019.
11 Endehabtu B, Weldeab A, Were M, *et al*. Mobile phone access and willingness among mothers to receive a Text-Based mHealth intervention to improve prenatal care in Northwest Ethiopia: cross-sectional study. *JMIR Pediatr Parent* 2018;1:e9.
12 John R, Giudicessi BA. Text messaging as a tool for behavior change in disease prevention and management. *Bone* 2013;23:1–7.
13 WHO. *mHealth new horizons for health through mobile technologies. based on the findings of the second global survey on eHealth*, 2011.
14 Lin C-L, Mistry N, Boneh J, *et al*. Text message reminders increase appointment adherence in a pediatric clinic: a randomized controlled trial. *Int J Pediatr* 2016;2016:1–6.
15 Georgette N, Siedner MJ, Zanoni B, *et al*. The acceptability and perceived usefulness of a Weekly clinical SMS program to promote HIV antiretroviral medication adherence in KwaZulu-Natal, South Africa. *AIDS Behav* 2016;20:2629–38.
16 Jemere AT, Yeneneh YE, Tilahun B, *et al*. Access to mobile phone and willingness to receive mHealth services among patients with diabetes in Northwest Ethiopia: a cross-sectional study. *BMJ Open* 2019;9:1–11.
17 Kebede M, Zeleke A, Asemahagn M, *et al*. Willingness to receive text message medication reminders among patients on antiretroviral treatment in North West Ethiopia: A cross-sectional study. *BMC Med Inform Decis Mak* 2015;15:1–10.
18 Domek GJ, Contreras-Roldan IL, O'Leary ST, *et al*. Sms text message reminders to improve infant vaccination coverage in Guatemala: a pilot randomized controlled trial. *Vaccine* 2016;34:2437–43.
19 Wakadha H, Chandir S, Were EV, *et al*. The feasibility of using mobile-phone based SMS reminders and conditional cash transfers to improve timely immunization in rural Kenya. *Vaccine* 2013;31:987–93.

20 Higgs ES, Goldberg AB, Labrique AB, *et al*. Understanding the role of mHealth and other media interventions for behavior change to enhance child survival and development in low- and middle-income countries: an evidence review. *J Health Commun* 2014;19 Suppl 1:164–89.
21 Cormick G, Kim NA, Rodgers A, *et al*. Interest of pregnant women in the use of SMS (short message service) text messages for the improvement of perinatal and postnatal care. *Reprod Health* 2012;9:1–7.
22 Albino S, Tabb KM, Requena D, *et al*. Perceptions and acceptability of short message services technology to improve treatment adherence amongst tuberculosis patients in Peru: a focus group study. *PLoS One* 2014;9:1–6.
23 Oluwatosin A, Ogundeji MO, Ogundeji MO. Experiences, perceptions and preferences of mothers towards childhood immunization reminder/recall in Ibadan, Nigeria: a cross-sectional study. *Pan Afr Med J* 2015;20:243.
24 Kalan R, Wiysonge CS, Ramafuthole T, *et al*. Mobile phone text messaging for improving the uptake of vaccinations: a systematic review protocol. *BMJ Open* 2014;4:1–5.
25 Manakongtreecheep K. SMS-reminder for vaccination in Africa: research from published, unpublished and grey literature. *Pan Afr Med J* 2017;27:23.
26 Datta SS, Ranganathan P, Sivakumar KS. A study to assess the feasibility of text messaging service in delivering maternal and child healthcare messages in a rural area of Tamil Nadu, India. *Australas Med J* 2014;7:175–80.
27 Greenhalgh T, Wherton J, Papoutsi C, *et al*. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J Med Internet Res* 2017;19:e367.
28 Aranda-Jan CB, Mohutsiwa-Dibe N, Loukanova S. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. *BMC Public Health* 2014;14.
29 VitalWaveconsulting, Vital Wave Consulting, VitalWaveconsulting, Vital Wave Consulting.,Vital Wave Consulting. *mHealth in Ethiopia: strategies for a new framework* 2011:1–65.
30 World Health Organization. *Global diffusion of eHealth: making universal health coverage achievable. Report of the third global survey on eHealth*, 2016.
31 Gao S, Krogstie J, Siau K. Developing an instrument to measure the adoption of mobile services. *Mob Inf Syst* 2011;7:45–67.
32 El-Wajeeh M, H. Galal-Edeen PG, Mokhtar DH. Technology acceptance model for mobile health systems. *IOSRJMCA* 2014;1:21–33.
33 Bastawrous A, Armstrong MJ. Mobile health use in low- and high-income countries: an overview of the peer-reviewed literature. *J R Soc Med* 2013;106:130–42.
34 Hofstetter AM, Vargas CY, Kennedy A, *et al*. Parental and provider preferences and concerns regarding text message reminder/recall for early childhood vaccinations. *Prev Med* 2013;57:75–80.
35 Campbell JI, Aturinda I, Mwesigwa E, *et al*. The technology acceptance model for resource-limited settings (TAM-RLS): a novel framework for mobile health interventions targeted to Low-Literacy End-Users in resource-limited settings. *AIDS Behav* 2017;21:3129–40.
36 InAhlers-Schmidt CR, Chesser AK, Paschal AM, *et al*. Parent opinions about use of text messaging for immunization reminders. *J Med Internet Res* 2012;14:e83.
37 NÖ G, ÖS B. Technology acceptance in health care: an integrative review of predictive factors and intervention programs. *Procedia - Soc Behav Sci* 2015;195.
38 Gurupur VP, Wan TTH. Challenges in implementing mHealth interventions: a technical perspective. *Mhealth* 2017;3:32.
39 Buabeng-Andoh C. Predicting students' intention to adopt mobile learning: A combination of theory of reasoned action and technology acceptance model. *J Res Innov Teach Learn* 2018;11.
40 Cho Y-M, Lee S, Islam SMS, *et al*. Theories applied to m-health interventions for behavior change in low- and middle-income countries: a systematic review. *Telemed J E Health* 2018;24:727–41.
41 Article O. *A comparative analysis of mothers ' preference for specific type of phone-derived reminders for routine immunization appointments in Ilorin, Nigeria*, 2018.
42 Vital Wave Consulting. *Mhealth in Ethiopia. strategies for a new framework*, 2011.
43 Karkonasasi K, Yu-n C, Mousavi SA. *Intention to Use SMS Vaccination Reminder and Management System among Health Centers in Malaysia : The Mediating Effect of Attitude*, 2011.
44 Ibraheem RM, Akintola MA. Acceptability of reminders for immunization appointments via mobile devices by mothers in Ilorin, Nigeria: a cross-sectional study. *Oman Med J* 2017;32:471–6.
45 Chao C-M. Factors determining the behavioral intention to use mobile learning: an application and extension of the UTAUT model. *Front Psychol* 2019;10:1652.
46 Lepère P, Touré Y, Bitty-Anderson AM, *et al*. Exploring the patterns of use and acceptability of mobile phones among people living with HIV to improve care and treatment: cross-sectional study in three francophone West African countries. *JMIR Mhealth Uhealth* 2019;7:e13741.
47 Zaunbrecher BS, Kowalewski S, Ziefle M. *The willingness to adopt technologies: A cross-sectional study on the influence of technical self-efficacy on acceptance. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2014;8512 LNCS(PART 3)*: 764–75.
48 Department GTH. *Gondar town administration health office report*, 2018.
49 Lee Y, Kozar KA, Larsen KRT. The technology acceptance model: past, present, and future. *CAIS* 2003;12.
50 Sek Y-W, Lau S-H, Teoh K-K, *et al*. Prediction of user acceptance and adoption of smart phone for learning with technology acceptance model. *J. of Applied Sciences* 2010;10:2395–402.
51 Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 1989;13:319.
52 Amoroso DL, Hunsinger S. Measuring the acceptance of Internet technology by consumers. *IJEA* 2009;1:48–81.
53 Venkatesh V, Davis FD. A model of the antecedents of perceived ease of use: development and test. *Decis Sci* 1996;27:451–81.
54 Park DY, Goering EM, Head KJ, *et al*. Implications for training on smartphone medication reminder APP use by adults with chronic conditions: pilot study applying the technology acceptance model. *JMIR Form Res* 2017;1:e5.
55 Venkatesh V, Morris MG, Davis GB, *et al*. User acceptance of information technology: toward a unified view. *MIS Quarterly* 2003;27:425.
56 Gagnon MP, Orruño E, Asua J, *et al*. Using a modified technology acceptance model to evaluate healthcare professionals' adoption of a new telemonitoring system. *Telemedicine and e-Health* 2012;18:54–9.
57 Balogun MR, Sekoni AO, Okafor IP, Odukoya OO, *et al*. Access to information technology and willingness to receive text message reminders for childhood immunisation among mothers attending a tertiary facility in Lagos, Nigeria. *S Afr J CH* 2012;6.
58 de Veer AJE, Peeters JM, Brabers AEM, *et al*. Determinants of the intention to use e-health by community dwelling older people. *BMC Health Serv Res* 2015;15:1–9.
59 Deng Z, Hong Z, Ren C, *et al*. What predicts patients' adoption intention toward mhealth services in China: Empirical study. *JMIR Mhealth Uhealth* 2018;6:e172–14.
60 Tao D. Intention to use and actual use of electronic information resources: further exploring technology acceptance model (TAM). *AMIA Annu Symp Proc* 2009;2009:629–33.
61 Liébana-Cabanillas F, Ramos de Luna I, Montoro-Ríos F. Intention to use new mobile payment systems: a comparative analysis of SMS and NFC payments. *Econ Res Istraz* 2017;30:892–910.
62 Puhan MA, Chandra D, Mosenifar Z, *et al*. Trust, perceived risk, perceived ease of use and perceived usefulness as factors related to mHealth technology use. *Study Heal Technol Inf* 2017;37:784–90.

BMJ Health & Care Informatics

Clinician checklist for assessing suitability of machine learning applications in healthcare

Ian Scott,[1,2] Stacey Carter,[3] Enrico Coiera[4]

To cite: Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform* 2021;**28**:e100251. doi:10.1136/bmjhci-2020-100251

Received 03 October 2020
Accepted 12 January 2021

[1]Internal Medicine and Clinical Epidemiology, Princess Alexandra Hospital, Brisbane, Queensland, Australia
[2]School of Clinical Medicine, Univeristy of Queensland, Brisbane, Queensland, Australia
[3]Australian Centre for Health Engagement Evidence and Values, University of Woolloongong, Woollongong, New South Wales, Australia
[4]Centre for Health Informatics, Macquarie University, Sydney, New South Wales, Australia

Correspondence to
Professor Ian Scott;
ian.scott@health.qld.gov.au

ABSTRACT

Machine learning algorithms are being used to screen and diagnose disease, prognosticate and predict therapeutic responses. Hundreds of new algorithms are being developed, but whether they improve clinical decision making and patient outcomes remains uncertain. If clinicians are to use algorithms, they need to be reassured that key issues relating to their validity, utility, feasibility, safety and ethical use have been addressed. We propose a checklist of 10 questions that clinicians can ask of those advocating for the use of a particular algorithm, but which do not expect clinicians, as non-experts, to demonstrate mastery over what can be highly complex statistical and computational concepts. The questions are: (1) What is the purpose and context of the algorithm? (2) How good were the data used to train the algorithm? (3) Were there sufficient data to train the algorithm? (4) How well does the algorithm perform? (5) Is the algorithm transferable to new clinical settings? (6) Are the outputs of the algorithm clinically intelligible? (7) How will this algorithm fit into and complement current workflows? (8) Has use of the algorithm been shown to improve patient care and outcomes? (9) Could the algorithm cause patient harm? and (10) Does use of the algorithm raise ethical, legal or social concerns? We provide examples where an algorithm may raise concerns and apply the checklist to a recent review of diagnostic imaging applications. This checklist aims to assist clinicians in assessing algorithm readiness for routine care and identify situations where further refinement and evaluation is required prior to large-scale use.

As a subset of artificial intelligence, machine learning (ML) is being used to create algorithms to screen and diagnose disease, prognosticate, and predict response to clinical interventions (box 1). Deep learning (DL), which uses massive artificial neural networks, has been responsible for much recent progress in ML. More than 150 clinical DL algorithms have now passed proof-of-concept phase,[1] and over 50 have been approved for routine use by the US Food and Drug Administration.[2]

However, before adopting algorithms into routine care, practising clinicians will seek reassurance from their professional bodies and healthcare institutions about their validity, utility, feasibility, safety and ethical use. Amidst the hype and opaque nature of many ML applications, and contestable claims of superior performance of some algorithms compared with clinical experts,[3] clinicians need to have some understanding of how algorithms are developed and how to assess their clinical worth.

Recent commentaries have identified several important challenges relating to ML applications in healthcare which end-users need to be aware of when deciding whether to adopt them into routine care.[4–8] We developed a checklist that reflect these challenges in a manner suitable to the needs and training of practising clinicians. It contains questions clinicians should ask of algorithm developers, vendors and implementers. In so doing, we recognise that, as non-experts in ML, clinicians cannot be expected to demonstrate mastery over what can be highly complex statistical and computational concepts. In seeking answers to certain questions, they may need to depend on the expertise of data scientists or health informaticians. In formulating the checklist, we made reference to recent narrative reviews,[1 9–12] a report from the US National Academy of Medicine,[13] and recent studies (from 2000) published in PubMed using search terms 'ML,' 'DL' and related synonyms.

Q1. WHAT IS THE PURPOSE AND CONTEXT OF THE ALGORITHM?

Algorithm development should be driven by a clinical need or 'pain point', not what is simply technically feasible by virtue of available data. Clinicians should ask if, at the design phase, developers collaborated with end-users in agreeing: (1) the specific clinical task or function of the algorithm (diagnosis, prognostication, treatment response); (2) the target population and clinical setting and (3) the intended method of algorithm implementation.[4]

Q2. HOW GOOD WERE THE DATA USED TO TRAIN THE ALGORITHM?

Algorithms can only be as good as the data they were trained on, and that data need to be easily accessible where the algorithm is to be used, easily migrated into different computer programmes (interoperable), and able to be stored and reused.

Q2a. To what extent were the data accurate and free of bias?

In assuring algorithm accuracy, clinicians should confirm that datasets used to train an algorithm were of high quality, representative of the population of interest, derived from reliable sources and had minimal missing data.[14] Many algorithms use transactional data from electronic medical records (EMRs) or administrative datasets—typically of poorer quality than clinical registry and trial datasets. However, given their extensive coverage of clinical care and their availability, such data will continue to be used. However, clinicians should note that incomplete, inaccurate, poorly described or incorrectly labelled data are more likely to introduce error.

Even more important are systematic biases in what data were collected, how and on whom. Some variables highly relevant to clinical outcomes (ancestry, language, socioeconomic status, laboratory tests, health-related circumstances, such as substance abuse, physical activity and homelessness) may not be routinely captured.[6] For example, a cardiovascular risk prediction algorithm was inaccurate in marginalised populations because training data were never obtained from them (selection bias).[15] An algorithm predicting survival of post-menopausal women using electrocardiographic markers, clinical characteristics and demographic variables performed worse than conventional Framingham scores, partly because it lacked important blood test results (measurement bias).[16] Recent research detected racial bias in an algorithm that could potentially affect millions of patients.[17]

Clinicians need to ask: what were the criteria for selecting patients for the training dataset, how many were screened and included, were all relevant baseline characteristics measured in all individuals, and what was done

Box 1 Machine learning (ML)—background concepts and examples

ML is the process whereby advanced computer programs (machines), often with minimal human instruction, process often huge datasets (big data), potentially from many sources, to discern patterns and associations which are then used to iteratively encode (or learn) a process or system model (algorithm). This algorithm, when applied to new data, aims to produce a prediction or outcome more quickly and accurately than clinical experts, devoid of errors due to human cognitive bias and fatigue.

Algorithms are developed (or trained) using training datasets derived from medical imaging devices, electronic medical records, administrative datasets or wearable biosensors. The trained algorithms may be tuned and then tested on samples of the training datasets to gauge accuracy and reproducibility, and then validated on new unseen datasets in assessing their generalisability to new populations and settings.

Types of ML

- Supervised learning maps input data from a training set of labelled (or known) examples to generate a model which can be applied to new data in making predictions. As the examples are already known, the model learns 'under supervision'. Supervised learning is used for classification (eg, discriminating between different items, categories or subgroups in making a diagnosis) and regression (prediction) (eg, estimating the likelihood of a future clinical event).
- Unsupervised learning uses input data from unlabelled examples and groups them according to some attribute (or pattern) of shared commonality. Unsupervised learning is used for: clustering, that is, identifying and characterising clusters of variables that appear to share latent similarities; and anomaly detection, that is, identifying unusual patterns of outlier or dissimilar values for different variables. An example is where clinical and genetic data from thousands of patients with a certain diagnosis, and who have been managed in different ways, are processed in identifying genotypic or phenotypic features associated with favourable or unfavourable response to certain treatments.
- Reinforcement learning processes dynamic data that is constantly changing and where the algorithm adapts to change and learns an optimised set of rules for achieving a goal or maximising an expected return (or reward) by a process of trial and error. Model behaviour is 'reinforced' by the level of reward achieved. Examples may include controlling an artificial pancreas system to fine-tune the measurement and delivery of insulin to patients with diabetes, or adjusting ventilator and vasopressor infusion rates in seriously ill patients in intensive care units.

Classes of ML algorithms

There are more than 20 different classes of ML algorithms; the following are the most commonly encountered.

- Artificial neural networks are non-linear algorithms loosely inspired by human brain synapses, with the most common being convoluted neural networks (or deep learning). These networks comprise input nodes, output nodes and intervening or hidden layers of nodes, which may number up to 100. Each node within a layer involves two or more inputs and applies an activation and weighting function to produce an output which serves as the input data for the next layer of nodes. In deep learning, data from imaging devices is passed through successive layers of nodes which convolute (transform) and pool the data and extract high order features such as contrast, colour, shapes, edges and patterns. These feature maps are successively pooled to produce the final outputs.
- Support vector machines (SVMs) transform input data into two classes or categories by choosing the boundary or widest plane (or support vector) that separates them to the maximal degree. SVMs can map examples to other dimensions which have non-linear relationships, and by transforming low dimensional input data into high-dimensional space using mathematical tools (kernel functions), they can separate such examples linearly by determining a hyperplane as the decision surface.
- Decision trees choose a series of sequential branching decisions on features in the training data which map the features to a known outcome with the most accuracy. They may use naïve Bayesian methods which assign pretest probabilities or prevalence to certain features and assume all features are independent of one another, or use random forests which adopt a completely random order of branching steps in a subset of training examples. Similar to SVMs, the goal is to optimally separate the classes in training examples.

to account for missing data or time varying confounders, such as downstream clinical management decisions? Because algorithms can learn, automate and accentuate existing biases in training datasets, thereby worsening healthcare inequities,[18] strategies for mitigating these biases during the training process[19] should be stated.

Q2b. Were data labelled correctly?

Supervised learning, currently the most common type of ML, may require training data to be labelled with the category or class of interest. For example, a retinal image might be labelled as showing diabetic retinopathy, where diabetes can be confirmed by a glycosylated haemoglobin test, but diagnosing retinopathy relies on subjective judgement of ophthalmologists. In avoiding algorithms developed using unreliable labels, clinicians should ask what reference standards (or 'ground truths') were used in deciding whether, in this case, diabetic retinopathy was the correct diagnosis. The ideal standard is often consensus adjudication by panels of expert clinicians, blind to algorithm predictions and given sufficient time and clinical information—reflecting normal clinical practice—to make well-considered predictions of whether a particular abnormality is present, absent or indeterminate.[20]

Q2c. Were the data standardised and interoperable?

Most algorithms are initially programmed to have data presented to them in a format (or 'common data model') that accords with a specific data standard. Imaging data are typically well standardised and interoperable using the Digital Imaging and Communications in Medicine and Picture Archiving and Communication System standards. However, for structured data within clinical

Box 2 Performance measures for machine learning algorithms

Area under receiver operating characteristic curve (AUROC)

For binary outcomes involving numerical samples (such as disease or event present or absent), the receiver operating characteristic (ROC) curve plots the true positive (TP) rate (sensitivity) against the false positive rate (1 minus specificity). An AUROC of 1.0 represents perfect prediction; an AUROC equal to or above 0.8 is preferred.

For binary outcomes involving imaging data, a modification of the ROC is the free-response ROC, or FROC* where a FROC curve comprising a 45º diagonal line indicates the algorithm is useless, while the steeper and more convex the slope of the curve, the greater the accuracy.

In situations where outcomes are not binary and multidimensional, or where data are highly skewed with disproportionately large numbers of true negatives, other methods such as the volume under the surface of the ROC curve and false discovery rate-controlled area under the ROC curve have been suggested; values equal to or above 0.8 are again preferred.**

Confusion matrix

A confusion matrix is a contingency table which yields several metrics, with optimal performance represented by values approaching 100% or 1.0.

- Positive predictive value (PPV) or precision: the proportion of positive cases that are TP rather than false positives (FP): $PPV=TP/TP+FP$.
- Negative predictive value (NPV): the proportion of negative cases that are true negatives (TN) rather than false negatives (FN): $NPV=TN/TN+FN$.
- Sensitivity (Sn) or recall: the proportion of TP cases that are correctly identified: $Sn=TP/TP+FN$.
- Specificity (Sp): the proportion of true negative (TN) cases which are correctly identified: $Sp=TN/TN+FP$.
- Accuracy: the proportion of the total number of predictions that are correct: $TP+TN/TP+FP+TN+FN$.
- F1 score: this measure represents the harmonic mean of precision (or PPV) and recall (sensitivity) in which both are maximised to the largest extent possible, given that one comes at the expense of the other. It is reported as a single score from 0 to 1 using the formula: $2\times TP/(2\times TP+FP+FN)$. The higher the score, the better the performance.
- Matthew's correlation coefficient**:** This coefficient takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes: $TP\times TN-FP\times FN/\sqrt{}\,(TP+FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)$. A coefficient of +1 represents a perfect prediction, 0 no better than random, and −1 total disagreement between prediction and actual outcome.

Precision-recall (PR) curve

The PR curve is a graphical plot of PPV (or precision) against sensitivity (or recall) to show the trade-off between the two measures for different feature (or parameter) settings. The area under the PR curve is a better measure of accuracy for classification tasks involving highly imbalanced datasets (ie, very few positive cases and large numbers of negative cases). An area under the PR curve (AUPRC) of 0.5 is preferred. Ideally, algorithm developers should report both AUROC and AUPRC, along with figures of the actual curves.

Regression metrics

Various metrics can be used to measure performance of algorithms performing regression functions (ie, predicting a continuous outcome). They include mean absolute error (mean of the absolute differences between actual and predicted values), mean squared error (calculated by summing the differences between actual and predicted values, squaring the results, and dividing by the total number of instances) and root mean squared error (standard deviation of all errors). In all cases, values closer to 0 indicate better performance.

Another commonly used metric is the coefficient of determination (R^2), which represents how much of the variation in the output variable (or Y—dependent variable) of the algorithm is explained by variation in its input variables (X—independent variables). An R^2 of 0 means prediction is impossible based on input variables and R^2 of 1 means completely accurate prediction with no variability. Generally R^2 should be above 0.6 for the algorithm to be useful.

*See Moskowitz CS. Using free-response receiver operating characteristic curves to assess the accuracy of machine diagnosis of cancer. *JAMA* 2017;318:2250–2251.

**See Yu T. ROCS: Receiver operating characteristic surface for class-skewed high-throughput data. *PLoS One* 2012;7:e40598.

records, different standards exist, for example, Systematised Nomenclature of Medicine-Clinical Terms[21] or the Observational Medical Outcomes Partnership standard.[22] In mapping data from one standard to another, the more mapping required, the greater the cost and risk of inducing errors.[23] Fortunately, the HL7-Fast Healthcare Interoperability Resources is emerging as a robust, standard-agnostic messaging system which facilitates data migration with minimal need for mapping.[24] Mapping unstructured, free-text clinical data is more challenging, although natural language processing algorithms can map words to clinical concepts.[25] Clinicians should ask if significant mapping work is required to meet local data standards before implementing an algorithm, and inquire into the costs and risks of doing so.

Q3. WERE THERE SUFFICIENT DATA TO TRAIN THE ALGORITHM?

In general, the more complex the algorithm, in having to make more distinctions between a larger number of different things, the more data required. Convolutional neural networks used to process medical images or text or huge numerical datasets may require many thousands of training examples.[26] However, methods for determining a priori just how many examples are required are yet to be agreed.[27] If more data continues to improve algorithm performance, more data should be supplied. Clinicians should be informed of how much data were used, how that sample size decision was reached, and what techniques (such as feature engineering and regularisation procedures) were used to deal with data of high dimensionality (ie, possessing many different attributes, as in imaging data) or of limited availability, as these all bear on algorithm performance.[28]

Box 3 Ethical, legal and social issues of using algorithms[61–66]

- How were consent issues handled in collecting data used for algorithm training and validation?
- Who owns, or has stewardship of, the data and determines how it is to be used in training and testing of algorithms?
- How are data confidentiality and patient privacy ensured when data is stored (in the cloud) and used and shared across different platforms?
- How much responsibility for care should clinicians be expected to assume when using algorithms they cannot control or explain?
- Who carries liability if patients are injured by a faulty or misapplied algorithm (developers who trained and tested the algorithm, vendors who integrated the algorithm into electronic medical records or imaging software, or clinicians using the algorithm to make decisions)?
- Who takes responsibility for postimplementation monitoring of the safety and efficacy of an algorithm throughout its life cycle, and determine when an algorithm needs updating, retraining or even withdrawal because of emerging inaccuracies?
- Will the majority of clinicians (and patients) be literate enough to understand how, when and in whom machine learning algorithms are safe and effective to use?
- How equitable and inclusive are the algorithms? Is there risk of a digital divide between healthcare institutions (and their catchment populations) who can or cannot deploy or access algorithm systems (for various reasons)?
- Who might have conflicts of interest in developing, disseminating, using or advocating a particular algorithm?
- Who owns the intellectual property pertaining to an algorithm; who owns the patent rights; who and what factors determine whether an algorithm is able to be commercialised for profit?

Q4. HOW WELL DOES THE ALGORITHM PERFORM?

Just as with a diagnostic test or a prediction rule, clinicians should be told the accuracy and reproducibility of algorithm outputs. A process of internal (or in-sample) validation should have tested and refined the algorithm on datasets resampled from the original training datasets,[29] either by bootstrapping (multiple sampling in random order) or cross-validation (datasets segmented into different testing sets multiple times [or 'folds'], hence the term k-fold cross-validation where k=number of folds, usually 5 or 10).

This is followed by a process of external (out-of-sample) validation on previously unseen data, preferably taken from a temporally or geographically different population. This step, which is often omitted, is crucial as it often reveals overfitting, where the algorithm has learnt features of the training dataset too perfectly, including minor random fluctuations, and consequently, may not perform well on new datasets. For classification tasks which are most common, metrics of discrimination should be reported (box 2), and chosen sensitivity/specificity thresholds justified in maximising clinical utility.[30] For regression-based prediction tasks, clinicians should ask if an algorithm performs better than existing regression models, in case it may not,[31] and ask if replication studies of the same algorithm by independent investigators have yielded the same performance results.[32]

Q5. IS THE ALGORITHM TRANSFERABLE TO NEW CLINICAL SETTINGS?

A crucial question for clinicians is whether the algorithm performs equally well across a range of new clinical settings and, if not, can the algorithm be retuned or recalibrated using local data to account for differences in population characteristics, type or reporting formats of imaging devices, or care protocols.[33 34] For example, a DL system for interpreting thyroid ultrasound images in detecting cancers saw sensitivity drop from 92% (human equivalent) to 84% (below human), with no change in specificity, when applied to different hospitals.[35] An algorithm used to diagnose pneumonia on chest X-rays in one hospital system failed to generalise to radiographs from another hospital system, due to differences in prevalence of pneumonia between populations[36] (class imbalance). Differences in illness severity can also degrade performance of algorithms trained on more severely diseased populations when applied to those with mild or moderate

Table 1 Application of the checklist

Liu *et al*[67] analysed 82 studies published between January 2012 and June 2019 which compared diagnostic performance of deep learning algorithms and healthcare professionals based on medical imaging for 17 different clinical conditions. The authors extracted diagnostic accuracy data and constructed contingency tables to derive the measures of interest. In generating responses to each item on the checklist, we used information stated in the review or, if certain information was missing, retrieved from the individual full-text articles.

Item	Response
1. What is the purpose of the algorithm?	Objective and context of the algorithms were adequately stated in included studies.
2a. How good were the data used to train the algorithm? 2b. To what extent were the data accurate and free of bias? 2c. Were the data standardised and interoperable?	26 studies (32%) did not report patient inclusion criteria; 33 studies (40%) did not report exclusion criteria; 30 studies (37%) did not report age and 43 studies (52%) did not report sex. 72 studies (88%) used retrospectively collected data from historical routine care (48 studies) or open source (24 studies) registries which are rarely quality controlled for images or accompanying labels, and in which population characteristics are either not collected or inaccessible; only 10 studies (12%) used prospectively collected data specific to a research setting. 26 studies (32%) excluded low-quality images; 18 (22%) retained low-quality images; 38 (46%) did not report this. The extent of missing data, and how this was handled, was poorly reported in all studies. All data used in 36 studies (44%) were obtained at a single hospital or medical centre. The extent to which data were standardised and rendered interoperable across sites in multisite studies was not reported in any study.
3. Were there sufficient data to train the algorithm?	57 studies (69%) did not report the number of participants represented by the training data; in remaining studies, the numbers ranged from 40 to 200 000. No study pre-specified a sample size.
4. How well does the algorithm perform?	For internal validation, 22 studies (27%) used resampling methods, 29 studies (35%) used random split sampling, 1 study (1%) used stratified random sampling, and 30 studies (37%) did not report any form of internal validation. 69 studies (84%) provided adequate data to construct contingency tables. In these studies sensitivity ranged from 9.7% to 100.0% (mean±SD 79.1%±0.2%); specificity ranged from 38.9% to 100.0% (mean±SD 88.3%±0.1%). Only 12 studies (14.6%) reported cut-points for determining sensitivity and specificity for which no justification was provided. The same reference standard was used across internal validation datasets in 61 studies (74%). Reference standards varied widely according to target condition and imaging modality. More rigorous expert group consensus standards were used in 66 studies (80%); remaining studies relied on single expert consensus (n=1), existing clinical care notes or imaging reports or existing labels (n=11), clinical follow-up (n=9), surgical confirmation (n=2), another imaging modality (n=1) and laboratory testing (n=3). No comments were made about outlier studies although AUROC curves depicted within the review clearly indicated there were such studies. Only 25 of 82 studies (36%) performed external validation. In these studies, the pooled sensitivity was 88.6% (95% CI 85.7 to 90.9) and pooled specificity was 93.9% (95% CI 92.2 to 95.3). Studies were inconsistent in their use of the term ‘validation’ as it applied to testing datasets; there was often lack of transparency as to whether testing sets were truly independent of training sets.
5. Is the algorithm transferable to new clinical settings?	Only 9 studies (11%) assessed algorithm performance in real-world contexts where clinicians received additional clinical information alongside the image, rather than just view the image in isolation.
6. Are the outputs of the algorithm clinically intelligible?	81 studies (99%) used artificial or convoluted neural networks; 1 study did not report algorithm architecture. Only 32 studies (39%) provided a heat map of salient features.
7. How will this algorithm fit into and complement current workflows?	No studies reported how their algorithms impacted real-world clinical workflows. In one study which compared algorithm performance among pathologists simulating normal workflows (ie, imposed time constraints) with that of a single pathologist with no time constraint, the AUROC were the same (0.96).*
8. Has use of the algorithm been shown to improve patient care and outcomes?	None of the algorithms in these studies have been subjected to clinical trials aimed at demonstrating improved care or patient outcomes.
9. Could the algorithm cause patient harm?	No comments were made about potential harms.
10. Does use of the algorithm raise ethical, legal or social concerns?	No comments were made about any such concerns.

*Bhteshami Bejnordi BE, Veta M, van Diest PJ, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318(22):2199–2210.
AUROC, area under receiving operator characteristic curve.

disease (spectrum bias). Variations in data quality, clinical actions included in the algorithm (causality leakage) or classification of outcomes (label leakage) can also affect local performance. While methods are emerging to minimise these problems,[37 38] clinicians should ask if the algorithm is applicable to their local setting, and whether it may need recalibration using local data.

Q6. ARE THE OUTPUTS OF THE ALGORITHM CLINICALLY INTELLIGIBLE?

Clinicians may not trust 'black box' algorithms which produce diagnoses or predictions in difficult-to-interpret formats, or provide little explanation of how these outputs were generated, especially those that appear counterintuitive. For the former, output formats may need to be customised to those that facilitate rapid clinical interpretation.[39] For the latter, decision trees and Bayesian networks are readily explainable in how they model causality, but data-driven methods, such as DL do so only implicitly, and may confuse association with causation, leading in some cases to clinically incorrect inferences. For example, an algorithm predicting low-risk patients with pneumonia who could be safely discharged from hospital was found to have incorrectly classified high risk asthmatic patients as low risk,[40] unaware that, by being routinely admitted to intensive care units, such patients had better survival. Another algorithm for detecting pneumothoraces on chest X-rays was trained on films taken after chest tube insertion, thus learning to identify chest tubes rather than pneumothoraces.[41]

In affording clinicians a better understanding of how algorithms generate their conclusions, various software tools can identify the features an algorithm chose as being critical in forming its predictions (eg, Local Interpretable Algorithm-Agnostic Explanations and Shapley Values in Machine Learning (SHAP)). These programmes can produce saliency or heat maps, pinpointing the exact areas and features in an image the algorithm has decided are abnormal,[42] and deconvolution graphs, highlighting the variables the algorithm regards as being most informative in predicting risk.[43]

Q7. HOW WILL THIS ALGORITHM FIT INTO AND COMPLEMENT CURRENT WORKFLOWS?

The utility of any algorithm in routine practice depends greatly on its 'fit' into clinical work and its impact on clinician time, efficiency and cognitive load. For example, in detecting metastatic breast tumours in sentinel lymph node biopsies, highlighting only the most suspicious regions expedited image review by pathologists, while showing raw algorithm predictions of each region of the image slowed them down.[44] Research into the ergonomics of using algorithms in routine clinical care is currently very limited, especially as the effort required for successful implementation can vary widely across even similar healthcare organisations because of subtle variations in workflows, tasks and patient needs.

Automating entry of imaging or EMR data into algorithms which self-activate in response to specific orders or requests can potentially help generate timely, actionable outputs.[45 46] The absence of such automation may simply increase burden of work on users, causing them to devise workarounds to avoid using an algorithm or abandoning it altogether.[47] Clinicians should therefore consider: (1) the exact point in the clinical trajectory where the algorithm will be applied; (2) the way the algorithm would actually be implemented in a specific clinical setting, and the technical and staff training effort required; (3) the resulting workflow changes and (4) the level of use the algorithm would likely receive from its intended users.

Q8. HAS USE OF THE ALGORITHM BEEN SHOWN TO IMPROVE PATIENT CARE AND OUTCOMES?

An algorithm will likely be ignored if clinicians do not perceive it as improving patient care and outcomes, either because the current human system is already optimal, or the algorithm is too far removed from critical decision points. Screening applications in otherwise healthy populations,[48] in whom inaccurate algorithms may cause significant harm, warrant careful attention. Rigorous clinical impact studies of DL algorithms are, to date, infrequent,[3 49] most are uncontrolled pre-post or cohort studies, and clinical effects are sometimes very marginal.[50] Ideally, the algorithm should be implemented and tested for utility in pilot studies in 'silent' mode (real-time predictions exposed to clinical experts but not acted on, so errors can be identified), then tested for efficacy in prospective clinical trials, and finally assessed for effectiveness and cost-effectiveness in large-scale studies.[51 52] Importantly, more rigorous testing should apply as algorithms move from narrow diagnostic imaging applications to more complex therapeutic scenarios, and from assistive applications informing decisions to fully automated applications determining patient management independently of clinicians.

Q9. COULD THE ALGORITHM CAUSE PATIENT HARM?

Poorly calibrated algorithms applied to insurance risk, employability and other forms of social profiling have generated false and detrimental predictions.[53] ML algorithms have generated unsafe drug recommendations in oncology.[54] Algorithms can quickly become inaccurate or out of date, and need retraining due to changes in background characteristics, exposures or outcomes of patient populations (distributional shifts), unanticipated changes in clinical practices or patient behaviour (calibration drift), and persistence of outmoded clinical technologies.[55 56] Even changes in clinical care due to algorithm implementation can, in itself, cause data shifts.[57] Adversarial cyber attacks can corrupt either the datasets or the computer programmes underpinning the algorithm, with effects potentially indiscernible to humans.[58] Automation bias may see clinicians become deskilled over time by over-reliance on algorithms,[59] leading to misdiagnoses and inappropriate therapeutics. Algorithms may encourage overdiagnosis by detecting subclinical anomalies that prompt unwarranted intervention.[60] Algorithms are unlikely to recognise when their outputs are false or affected by bias, and hence clinician must continue to question counter-intuitive or potentially harmful predictions.

Q10. DOES THE ALGORITHM RAISE ETHICAL, LEGAL OR SOCIAL CONCERNS?

Several contestable and intertwined ethical, legal and social issues are raised in using algorithms (box 3)[61–63] that clinicians need to consider, particularly personal liability for algorithm-induced harm[64] and blatant misuse of patient data that breaches privacy rules[65] enshrined in the US Health Insurance Portability and Insurance Act, the UK Data Protection Bill and the European General Data Protection Regulation. Numerous reports[66] provide guidance around clinician and patient autonomy, data privacy and governance processes, potential commercial conflicts of interest, openness (open data sets, methods and source code) and transparency, non-discrimination and fairness.

Application of the checklist

As a test of its potential utility, we applied our checklist to a recent systematic review of studies comparing accuracy of diagnostic imaging algorithms with that of clinical experts[67] (table 1). While this exercise did not target a single algorithm, which may be a limitation, our impression was that many studies demonstrated shortcomings for virtually every question—a problem which recently issued reporting guidelines for ML studies[68 69] will hopefully improve. In the meantime, our checklist may serve to protect clinicians from premature adoption of algorithms of uncertain worth.

CONCLUSION

Most clinicians will likely see ML algorithms increasingly used to augment their decision making. Image-intensive disciplines will likely see major reconfiguration of roles as algorithms are adopted to improve diagnostic accuracy. Algorithms will not replace clinicians, but clinicians who use well-designed and validated algorithms appropriately may replace those who do not. Clinicians need to be able to judge algorithm readiness for use and identify situations where further refinement and evaluation are needed prior to large-scale use.

Acknowledgements The authors wish to acknowledge Professor Jenny Doust (Bond University, Gold Coast, Australia), Nazanin hahreman-Falconer (Pharmacy PhD student, Princess Alexandra Hospital, Brisbane, Australia) and Ahmad Abdel-Hafez (Principal Data Analyst, Clinical Informatics, Metro South Health Service, Brisbane, Australia) for their comments on previous drafts.

Contributors IAS conceived the idea, undertook the primary research and drafted the first manuscript. EC and SC critically reviewed the manuscript, added new themes and references and assisted in redrafting the manuscript. IAS is the guarantor who accepts full responsibility for the finished article and had access to any data from literature searches.

Competing interests None declared.

Patient and public involvement statement We did not involve patients or the public in this work which comprised secondary research of published literature. There are no individual patient results to disseminate to these groups.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–58.
2 US Food and Drug Administration. Fda cleared AI algorithms. data science Institute. Available: https://www.acrdsi.org/DSI-Services/FDA-cleared-ai-algorithms [Accessed 9 Sep 2020].
3 Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
4 Gilvary C, Madhukar N, Elkhader J, *et al*. The missing pieces of artificial intelligence in medicine. *Trends Pharmacol Sci* 2019;40:555–64.
5 Lindsell CJ, Stead WW, Johnson KB. Action-Informed artificial Intelligence-Matching the algorithm to the problem. *JAMA* 2020;323:2141.
6 Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–7.
7 Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: Humanism and artificial intelligence. *JAMA* 2018;319:19–20.
8 Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA* 2018;320:2199–200.
9 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
10 Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. *Nat Med* 2019;25:24–9.
11 He J, Baxter SL, Xu J, *et al*. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30–6.
12 Liu Y, Chen P-HC, Krause J, *et al*. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322:1806–16.
13 Matheny MS, Israni T, Ahmed M, *et al*, eds. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. NAM Special Publication*. Washington, DC: National Academy of Medicine, 2019.
14 Goldstein BA, Navar AM, Pencina MJ, *et al*. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24:198–208.
15 Gijsberts CM, Groenewegen KA, Hoefer IE, *et al*. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One* 2015;10:e0132321.
16 Gorodeski EZ, Ishwaran H, Kogalur UB, *et al*. Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the women's health Initiative. *Circ Cardiovasc Qual Outcomes* 2011;4:521–32.
17 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
18 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
19 Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019;322:2377–8.

20 Krause J, Gulshan V, Rahimy E, *et al*. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* 2018;125:1264–72.

21 Benson T. *Principles of health Interoperability HL7 and SNOMED*. London, England: Springer, 2012. ISBN: 978-1-4471-2800-7.

22 FitzHenry F, Resnic FS, Robbins SL, *et al*. Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. *Appl Clin Inform* 2015;6:536–47.

23 Rosenbloom ST, Carroll RJ, Warner JL, *et al*. Representing knowledge consistently across health systems. *Yearb Med Inform* 2017;26:139–47.

24 Lehne M, Luijten S, Vom Felde Genannt Imbusch P, *et al*. The use of FHIR in digital health - A review of the scientific literature. *Stud Health Technol Inform* 2019;267:52–8.

25 Bruland P, McGilchrist M, Zapletal E, *et al*. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. *BMC Med Res Methodol* 2016;16:1.

26 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.

27 Balki I, Amirabadi A, Levman J, *et al*. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J* 2019;70:344–53.

28 Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;3:1157–82.

29 Moons KGM, de Groot JAH, Bouwmeester W, *et al*. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist. *PLoS Med* 2014;11:e1001744.

30 Shah NH, Milstein A, Bagley PhD SC. Making machine learning models clinically useful. *JAMA* 2019;322:1351–2.

31 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.

32 Coiera E, Ammenwerth E, Georgiou A, *et al*. Does health informatics have a replication crisis? *J Am Med Inform Assoc* 2018;25:963–8.

33 Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A* 2016;113:7345–52.

34 Saria S, Subbaswamy A. Tutorial: safe and reliable machine learning. arXiv.org, 2019. Available: https:// arxiv.org/abs/1904.07204

35 Li X, Zhang S, Zhang Q, *et al*. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol* 2019;20:193–201.

36 Zech JR, Badgeley MA, Liu M, *et al*. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018;15:e1002683.

37 Soleimani H, Hensman J, Saria S. Scalable joint models for reliable Uncertainty-Aware event prediction. *IEEE Trans Pattern Anal Mach Intell* 2018;40:1948–63.

38 Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data* 2016;3:9.

39 Tschandl P, Rinner C, Apalla Z, *et al*. Human-computer collaboration for skin cancer recognition. *Nat Med* 2020;26:1229–34.

40 et alCaruana R, Lou Y, Gehrke J. Intelligible algorithms for healthcare: predicting pneumonia risk and hospital 30-day readmission. *Paper presented at: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*2015.

41 Oakden-Rayner L. *Exploring the ChestXray14 dataset: problems*. Wordpress: Luke Oakden Rayner, 2017.

42 Zhang Z, Beck MW, Winkler DA, *et al*. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann Transl Med* 2018;6:216.

43 Nielsen AB, Thorsen-Meyer H-C, Belling K, *et al*. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish national patient registry and electronic patient records. *Lancet Digit Health* 2019;1:e78–89.

44 Steiner DF, MacDonald R, Liu Y, *et al*. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol* 2018;42:1636–46.

45 Kuzniewicz MW, Puopolo KM, Fischer A, *et al*. A quantitative, risk-based approach to the management of neonatal early-onset sepsis. *JAMA Pediatr* 2017;171:365–71.

46 Cronin PR, Greenwald JL, Crevensten GC, *et al*. Development and implementation of a real-time 30-day readmission predictive model. *AMIA Annu Symp Proc* 2014;2014:424–31.

47 Miller A, Koola JD, Matheny ME, *et al*. Application of contextual design methods to inform targeted clinical decision support interventions in sub-specialty care environments. *Int J Med Inform* 2018;117:55–65.

48 Houssami N, Lee CI, Buist DSM, *et al*. Artificial intelligence for breast cancer screening: opportunity or hype? *Breast* 2017;36:31–3.

49 Clifton DA, Niehaus KE, Charlton P, *et al*. Health informatics via machine learning for the clinical management of patients. *Yearb Med Inform* 2015;10:38–43.

50 Giannini HM, Ginestra JC, Chivers C, *et al*. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med* 2019;47:1485–92.

51 Khalifa M, Magrabi F, Gallego B. Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support. *BMC Med Inform Decis Mak* 2019;19:207.

52 Xie Y, Gunasekeran DV, Balaskas K, *et al*. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. *Transl Vis Sci Technol* 2020;9:22.

53 O'Neil C. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. London: Allen Lane, 2016.

54 Palmer A. *IBM's Watson AI suggested "often inaccurate" and "unsafe" treatment recommendations for cancer patients, internal documents show*. DailyMail.com, 2018. https://www.dailymail.co.uk/sciencetech/article-6001141/IBMs-Watson-suggested-inaccurate-unsafe-treatment-recommendations-cancer-patients.html?ito=email_share_article-top

55 Challen R, Denny J, Pitt M. Artificial intelligence. *bias and clinical safety BMJ Qual Saf* 2019;28:231–7.

56 Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle regulation of artificial intelligence– and machine learning–based software devices in medicine. *JAMA* 2019;322:2285–6.

57 Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. *J Am Med Inform Assoc* 2019;26:1645–50.

58 Finlayson SG, Bowers JD, Ito J, *et al*. Adversarial attacks on medical machine learning. *Science* 2019;363:1287–9.

59 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017;24:423–31.

60 Komorowski M, Celi LA. Will artificial intelligence contribute to overuse in healthcare? *Crit Care Med* 2017;45:912–3.

61 Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med* 2018;378:981–3.

62 Carter SM, Rogers W, Win KT, *et al*. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast* 2020;49:25–32.

63 Abràmoff MD, Tobey D, Char DS. Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. *Am J Ophthalmol* 2020;214:134–42.

64 Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA* 2019;322:1765–6.

65 Jiang JX, Bai G. Types of information compromised in breaches of protected health information. *Ann Intern Med* 2020;172:159–60.

66 AI ethics guidelines global inventory. Available: https://inventory.algorithmwatch.org/;

67 Liu X, Faes L, Kale AU, *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1:e271–97.

68 Cruz Rivera S, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020;26:1351–63.

69 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364–74.