Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness

Sebastian Vollmer,[1,2] Bilal A Mateen,[1,3,4] Gergo Bohner,[1,2] Franz J Király,[1,5] Rayid Ghani,[6] Pall Jonsson,[7] Sarah Cumbers,[8] Adrian Jonas,[9] Katherine S L McAllister,[9] Puja Myles,[10] David Grainger,[11] Mark Birse,[11] Richard Branson,[11] Karel G M Moons,[12] Gary S Collins,[13] John P A Ioannidis,[14] Chris Holmes,[1,15] Harry Hemingway[16,17,18]

For numbered affiliations see end of the article.

Correspondence to: C Holmes cholmes@stats.ox.ac.uk (ORCID 0000-0002-6667-4943)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;368:l6927
http://dx.doi.org/10.1136/bmj.l6927

Accepted: 22 October 2019

Machine learning, artificial intelligence, and other modern statistical methods are providing new opportunities to operationalise previously untapped and rapidly growing sources of data for patient benefit. Despite much promising research currently being undertaken, particularly in imaging, the literature as a whole lacks transparency, clear reporting to facilitate replicability, exploration for potential ethical concerns, and clear demonstrations of effectiveness. Among the many reasons why these problems exist, one of the most important (for which we provide a preliminary solution here) is the current lack of best practice guidance specific to machine learning and artificial intelligence. However, we believe that interdisciplinary groups pursuing research and impact projects involving machine learning and artificial intelligence for health would benefit from explicitly addressing a series of questions concerning transparency, reproducibility, ethics, and effectiveness (TREE). The 20 critical questions proposed here provide a framework for research groups to inform the design, conduct, and reporting; for editors and peer reviewers to evaluate contributions to the literature; and for patients, clinicians and policy makers to critically appraise where new findings may deliver patient benefit.

SUMMARY POINTS

Patients and healthcare professionals require clinical prediction models to accurately guide healthcare decisions

Larger sample sizes lead to the development of more robust models

Data should be of sufficient quality and representative of the target population and settings of application

It is better to use all available data for model development (ie, avoid data splitting), with resampling methods (such as bootstrapping) used for internal validation

When developing prediction models for binary or time-to-event outcomes, a well known rule of thumb for the required sample size is to ensure at least 10 events for each predictor parameter

The actual required sample size is, however, context specific and depends not only on the number of events relative to the number of candidate predictor parameters but also on the total number of participants, the outcome proportion (incidence) in the study population, and the expected predictive performance of the model

We propose to use such information to tailor sample size requirements to the specific setting of interest, with the aim of minimising the potential for model overfitting while targeting precise estimates of key parameters

Our proposal can be implemented in a four step procedure and is applicable for continuous, binary, or time-to-event outcomes

The pmsampsize package in Stata or R allows researchers to implement the procedure

Machine learning (ML), artificial intelligence (AI), and other modern statistical methods are providing new opportunities to operationalise previously untapped and rapidly growing sources of data for patient benefit. The potential uses include improving diagnostic accuracy,[1] more reliably predicting prognosis,[2] targeting treatments,[3] and increasing the operational efficiency of health systems.[4] Examples of potentially disruptive technology with early promise include image based diagnostic applications of ML/AI, which have shown the most early clinical promise (eg, deep learning based algorithms improving accuracy in diagnosing retinal pathology compared with that of specialist physicians[5]), or natural language processing used as a tool to extract information from structured and unstructured (that is, free) text embedded in electronic health records.[2] Although we are only just beginning to understand the wealth of opportunities afforded by these methods, there is growing concern in the academic community that because the products

of these methods are not perceived in the same way as other medical (eg, pharmacological) interventions,[6] they do not have well defined guidelines for development and use and rarely undergo the same degree of scrutiny.

Need for guidance

Several high profile publications have shown a lack of transparency,[7 8] replicability,[9] ethics,[10] and effectiveness[11] in the reporting and assessment of ML/AI based prediction models. This growing body of evidence suggests that while many best practice recommendations for design, conduct, analysis, reporting, impact assessment, and clinical implementation can be borrowed from the traditional biostatistics and medical statistics literature,[12] they are not sufficient to guide the use of ML/AI in research. Producing such guidance is a major undertaking due to the ever-growing battery of ML/AI algorithms and the multifaceted nature of assessing performance and clinical impact. Not taking action is unacceptable, and if we wait for a more definitive solution, we risk wasting valuable work,[13-17] while allowing futile research to continue unchecked, or worse, translation of ineffective (or even harmful) algorithms from the computer bench to the bedside.

Summary points

- Clinically relevant research using modern statistical methods (such as machine learning and artificial intelligence) is too often limited by one or more of TREE concerns (transparency, reproducibility, ethics, and effectiveness); addressing these concerns can facilitate appropriate translation from computer bench to patient benefit
- Here we propose 20 critical questions that offer a framework for users and generators of ML/AI research
- For research generators, the 20 questions can inform the way research groups design, conduct, and report their research
- For editors and peer reviewers, the checklist provides a starting point for evaluating the quality and clinical relevance of articles
- For the users of such research findings—including healthcare professionals, patients, and the public—the 20 questions highlight important issues for critical appraisal

Initial framework

We propose a series of 20 critical questions (box 1) to help identify common pitfalls that can undermine ML/AI based applications in health. The questions span issues of transparency, reproducibility, ethics, and effectiveness (TREE). Appendix 1 includes a brief description of how these questions were generated. The questions are not only relevant for those who use the findings (that is, patients and policy makers), but also for those who generate ML/AI health research. We envision this checklist of questions as providing a framework for journal editors, peer reviewers, and those who critically evaluate contributions to the literature; for researchers as a reference to inform the way that research groups design and conduct ML/AI research; for regulators judging algorithm approval; and for educators of clinicians and academic disciplines involved. Current practice in research publication is heterogeneous with relevant questions not clearly dealt with. Clearly further work is needed to build consensus on what constitutes acceptable practice and reporting, but we believe that adoption of this framework as a starting point, and of other related publications,[18] will help to build trust in the underlying processes and results of health related ML/AI research.

Critical questions

Inception (questions 1-2)

What is the health question relating to patient benefit?

The vast majority of published clinical prediction models are never used in clinical practice.[19] One reason for this is the lack of a specific clinical decision making process that the model could meaningfully inform or optimise; simply predicting future events on their own might not help a clinician do anything differently[20] (in other words: just because we can, it does not mean we should). This is an important departure from the lone wolf attitude, which has helped foster innovation over the past few decades in ML/AI for health. However, it is being increasingly recognised that such research needs to be seen in a wider organisational context to be made most useful. Therefore, we strongly urge researchers embarking on a new project, at the outset, to clarify and state the relevance of their work to healthcare system and patients. In essence, researchers should be cognisant of the path from development to implementation, and be able to describe which parts of the healthcare data science cycle their proposed research engages with. Note that this does not preclude theoretical, proof of concept, or operational research, which either only occupies a small angle of the healthcare data science cycle or only tangentially affects patients (eg, efficiency related gains in an administrative task). What is important, much like the principles on which registration of research is built, is that this expectation is stated up front.

What evidence is there that the development of the algorithm was informed by best practices in clinical research and epidemiological study design?

Similar themes to that of historical issues with clinical research are beginning to present in ML/AI based research, such as using outcome variables are predictors, paying little attention to causal pathways, insufficiently detailed descriptions of the conceptualisation of an inception cohort, and documenting exactly what sort of patients made their way into the analysis.[21] The PECO principles of

Box 1: Critical questions for health related technology involving machine learning and artificial intelligence

Inception
1. What is the health question relating to patient benefit?
2. What evidence is there that the development of the algorithm was informed by best practices in clinical research and epidemiological study design?

Study
1. When and how should patients be involved in data collection, analysis, deployment, and use?
2. Are the data suitable to answer the clinical question—that is, do they capture the relevant real world heterogeneity, and are they of sufficient detail and quality?
3. Does the validation methodology reflect the real world constraints and operational procedures associated with data collection and storage?
4. What computational and software resources are required for the task, and are the available resources sufficient to tackle this problem?

Statistical methods
1. Are the reported performance metrics relevant for the clinical context in which the model will be used?
2. Is the ML/AI algorithm compared to the current best technology, and against other appropriate baselines?
3. Is the reported gain in statistical performance with the ML/AI algorithm justified in the context of any trade-offs?

Reproducibility
1. On what basis are data accessible to other researchers?
2. Are the code, software, and all other relevant parts of the prediction modelling pipeline available to others to facilitate replicability?
3. Is there organisational transparency about the flow of data and results?

Impact evaluation
1. Are the results generalisable to settings beyond where the system was developed (that is, results reproducibility/external validity)?
2. Does the model create or exacerbate inequities in healthcare by age, sex, ethnicity, or other protected characteristics?
3. What evidence is there that clinicians and patients find the model and its output (reasonably) interpretable?
4. How will evidence of real world model effectiveness in the proposed clinical setting be generated, and how will unintended consequences be prevented?

Implementation
1. How is the model being regularly reassessed, and updated as data quality and clinical practice changes (that is, post-deployment monitoring)?
2. Is the ML/AI model cost effective to build, implement, and maintain?
3. How will the potential financial benefits be distributed if the ML/AI model is commercialised?
4. How have the regulatory requirements for accreditation/approval been addressed?

epidemiological study design (that is, defining a study population, the exposures used, the key comparators, and the clinical outcomes) have an important role in some these issues when they originally arose in health research, and have now become a useful guide for assessments of the quality and relevance of research evidence.[22] Although developed in the clinical domain, these principles are still highly relevant to ML/AI research, especially in providing a framework on which to ground large scale projects involving electronic health records. This is just one example of how researchers can use clinical frameworks that exist to inform best practice research in the development of ML/AI based projects.

Study (questions 3-6)

When and how should patients be involved in data collection, analysis, deployment, and use?

With the growing use of routinely collected individual participant data (in addition to researcher collected data), often with an alternative legal basis (that is, legitimate interests) to individual consent, it is more important now than ever that patient and public involvement is seen as an adjunct to all research in healthcare, including work related to machine learning. The exemption from seeking individual consent does not mean that the researchers are exempt from engaging patients and public altogether. Thus (where appropriate), healthcare ML/AI projects should include a clear mechanism to evaluate the acceptability of the proposed model and outcomes to those individuals from whom the data was collected, the users (that is, clinicians), and the affected individuals (that is, those for whom the model will be used to inform clinical management).

Several established frameworks[23] illustrate how patients and the public might be involved in a research project. We would highly encourage researchers to determine which stages of their project, if any, are amenable to patient and public involvement (at inception), for example, identifying the need for a predictive modelling solution, supporting the development of the algorithm (that is, selection of relevant targets, framing of how outcomes are presented), and determining the acceptability of the algorithm in practice. Arguments suggesting that policies pertaining to patient and public involvement should be decided at the political or institutional level does not recognise the agency of individual researchers, and it is for that reason we have included this question, in an effort to reassign the responsibility to those undertaking the work.

Are the data suitable to answer the clinical question—that is, do they capture the relevant real world heterogeneity, and are they of sufficient detail and quality?

The key issue here is whether the clinical question can be answered with the data available. For example, a dataset not containing the (known) relevant or important predictors of an outcome is unlikely to satisfactorily answer questions about it. No ML/AI algorithm can produce something from nothing. To help illustrate some of the potential issues involved in determining whether data are of sufficient quality and detail to inform the clinical question of interest, we have briefly described two core areas where researchers frequently have difficulties when attempting to apply ML methods to healthcare related data:

- Intrinsic sample characteristics. If data are available, but are of poor quality or are not relevant, development of a good ML/AI application is unlikely.[24] The accuracy of data collection methods, sampling of participants, eligibility criteria, and missing data all need to be considered when assessing the potential of developing useful and generalisable ML/AI algorithms.
- Relevance to task. Models are often unable to attain the levels of accuracy seen in training, owing to the likelihood of failure when operating outside the training data range. For example, the decision making system for an image recognition/self-driving car could fail when encountering a cyclist at night for the first time. Hence, the data—including timescale, heterogeneity (differences in data collection such as measuring devices or compliance), population, and situation—should accord with and represent the envisioned clinical application scenario.

Does the validation methodology reflect the real world constraints and operational procedures associated with data collection and storage?

Increasingly, ML/AI research is making use of routinely collected data, including healthcare data (eg, electronic health records, clinical imaging, and genomic information), civil administrative data (eg, death records, and educational achievement), and data[25] from mobile and wearable devices. Information from these sources can arrive in batches, or via a continuous stream, and is often stored in different locations requiring reconciliation, which in and of itself introduces a delay in when specific pieces of data are available for use. In contrast to these real world constraints, ML/AI algorithms are often validated on historical data, yielding performance guaranteed only under the assumption that the data generating process does not change (eg, over time, or across hospitals). In practice, these assumptions are often violated and result in ML/AI models underperforming when deployed in comparison to performance reported during development.[26]

Researchers could consider this problem as two different but related difficulties. The first is the issue of ensuring that a robust validation scheme is developed. For example, methods that take time into account and create temporally disjointed training and test sets[27 28] might be needed to account for how the data are collected and stored. The second issue is to prevent a useful solution from becoming redundant owing to drift in institutional data collection, or storage methods. However, little can be done by developers and researchers to future proof their work, other than using best practices for reproducibility (that is, clear descriptions of dependencies and modular development of the data ingress pathway, cleaning, pre-processing, and modelling), in order to reduce the amount of work necessary to redeploy a relevant version of the solution.

What computational and software resources are required for the task, and are the available resources sufficient to tackle this problem?

Working with millions of parameters is common in many areas of health related prediction modelling, such as image based deep learning[29] and statistical genetics.[30] Therefore, it is common practice to ascertain not only the complexity of the data, but also the computational resources available, because these resources can be the limiting factor (much more often than with traditional statistical models) in determining what analyses can be undertaken.[31] In some situations, more computational resources could in fact allow better models to be trained. For example, without sufficient computer resources, use of models based on complex neural networks could be prohibitively difficult, especially if these large scale models require additional complex operations (eg, regularisation) to prevent overfitting.[32 33] Ideally, analysis would not be limited by the availability of computational resources, but researchers should understand the constraints within which they are working so that any analysis can be tailored to requirements. Similar problems can arise when using secure computer environments, such as data enclaves or data safe havens, where the relevant software frameworks might not be available and thus would warrant implementation from scratch. Therefore, it is also important to understand the implications of using specific software, because the underlying licence can have far reaching consequences on the commercial potential and other aspects of the algorithm's future. A brief overview of software licensing for scientist programmers has been published elsewhere.[34]

Statistical methods (questions 7-9)

Are the reported performance metrics relevant for the clinical context in which the model will be used?

The choice of performance metric matters in order to translate good performance in the (training data) evaluation setup to good performance in the eventual clinical setting with patient benefit. This discrepancy in model performance can arise for multiple reasons; the most common of which is that the evaluation

metrics are not good proxies for demonstrating improved outcomes for patients (eg, misclassification error for a screening application with imbalanced classes). Another common mistake is choosing a performance metric that is vaguely related to, but not indicative or demonstrative of, improved clinical outcomes for patients. For example, IBM's Watson For Oncology (WFO)[35] is an expert system used in several hospitals worldwide to support decision making. However, published works describing WFO do not report relevant statistical (eg, discrimination, calibration) and clinically oriented (eg, net benefit type) performance metrics. Instead, they focus on concordance (true positive rate where the ground truth is provided by physician—that is, the proportion of instances where WFO's recommendation agrees with that of the treating physician[36-38]). We recommend the following guidance for researchers to avoid such pitfalls:

- Consult all relevant parties (eg, patients, data scientists/statisticians, clinicians) to determine the most appropriate formulation of the statistical goal, such as predicting the absolute risk of an event, or establishing a rank ordering or a pattern detection or classification (see question 3).
- Select the appropriate performance metrics. Each goal has its own unique requirements, and making explicit the statistical goal will help researchers ascertain what the relevant measures of predictive performance are for each specific situation. For example, if prediction (not classification) is the goal, then calibration and discrimination are the minimum requirements for reporting. Furthermore, for comparing two models, proper scoring rules should be used (or at least side-by-side histograms). The TRIPOD explanation and elaboration paper provides a reasonable starting point for researcher seeking more information on this issue.[12]
- Report all results. Although training results are unlikely to be sufficient to evidence the usefulness of the model, they provide important insights in the context of the sample characteristics and any out-of-sample results that are also provided. However, unbiased estimates (that is, those that have been adjusted appropriately for overfitting) are the most important to report.

Is the ML/AI algorithm compared to the current best technology, and against other appropriate baselines?

ML/AI algorithms should be viewed as health technologies, and at the design stage consideration should be given to identifying the approach that the algorithm might replace. One common way to exaggerate the benefit of ML/AI approaches is to avoid any comparison of ML/AI with null models or the currently used approach and instead compare to sub-par competitors (including inappropriately or weakly developed statistical models), or to avoid a comparison altogether. This "weak comparator" bias has been generally seen in reports of new versus existing prognostic models.[33] One such example comes from a systematic review of proposed modifications to the Framingham risk score for predicting the risk of a heart attack within 10 years; the review found that most proposed alternatives had flaws in their design, analyses, and reporting that cast doubt on the reliability of the claims for improved prediction.[39] To simplify this process, we have summarised the three baselines that together form the basis of a robust comparison:

- Model proxies for uninformed guessing, such as predicting the majority class in a classification task. This is the simplest form of sanity check that researchers can use to demonstrate that their ML/AI model is actually learning something. In some instances, probabilistic guessing could be a more appropriate baseline, but the decision of which one to use should be task specific.
- For almost all clinical questions, there will be a standard statistical approach that is well accepted from decades of biostatistics research, for example, proportional hazards models for survival modelling. The impetus is on developers and researchers to show some demonstrable value in using machine learning instead of the standard approach. Recent evidence has shown that these comparisons are often not fair, and favour one set of methods (commonly ML) over classical statistical methods.[40] We would urge researchers to keep this in mind when carrying out such comparisons.
- The current preferred method standard, whether it is a clinical diagnosis, biochemical test, or pre-existing model. Researchers should show how the model compares to a relevant gold standard. The ML/AI tool does not need to be better than the gold standard, but it is informative to know how the model compares to it. There might be use cases outside of improved accuracy (eg, prediction can be made on a larger class of patients because less data are required). It is the responsibility of the researcher to articulate this in their specific circumstances.

Is the reported gain in statistical performance with the ML/AI algorithm justified in the context of any trade-offs?

For a new diagnostic or prognostic tool to be justified for routine use, it must offer a (clinically) meaningful advantage over existing approaches in addressing a specific need,[41] which requires the use of an appropriate performance metric as discussed previously. Although necessary, the presence of a (clinically) meaningful advantage alone is not sufficient justification, because any improvement must be weighed against the cost of any changes it necessitates (eg, the resource requirement to collect additional data). In a recent paper published by Google, researchers investigated

the accuracy of deep learning methods in combination with electronic health records for predicting mortality, readmission, and length of stay.[2] In the appendix, the paper's authors compared their deep learning model against a logistic regression model. The area-under-the-curve improvement reported for each of the three tasks ranged from 0.01 to 0.02. If we assume that all caveats pertaining to statistical significance, and the sufficiency of the reported metric for making this next decision are met, is the marginal gain of implementing a complex ML/AI solution worth it, and is the need any more effectively addressed by the deep learning model? Although the answer to that question will certainly be situation specific, it will (at minimum) need to justify the following:

- The cost of developing, deploying, using, and maintaining a deep learning model such as the one described relative to the improvement observed; and
- The need for additional subsidiary models to increase the explainability lost in the transition away from a model with a human interpretable model (eg, with simple coefficients or consisting of a decision tree)

Reproducibility (questions 10-12)

On what basis are data accessible to other researchers?

Data sharing is not an endpoint in itself but rather a means to enhance, verify, and distribute the knowledge generated by the ML/AI algorithm.[42] Most major funding sources now require applicants to outline a data management and data sharing plan; this can entail (among other things) storing the data in a convenient format along with a data dictionary, a long term archiving plan, and providing an independent access mechanism (eg, a university ethics committee, or a research and development department). Where data used to develop the ML/AI algorithm have been accessed via national data custodians (eg, Clinical Practice Research Datalink,[43] NHS Digital,[44] Healthcare Quality Improvement Partnership[45]), clear data access processes have been put in place for independent validation by other researchers. Additionally, data sharing can be undertaken by a wide range of mechanisms, including:

- Making the data available in open repositories such as datadryad.org[46] (after being anonymised using tools such as Amnesia[47]), or restricted access repositories such as the UK Data Archive[48];
- Signing data sharing agreements;
- Providing remote access to local computing facilities where the data are stored, as is possible with specific restricted access data enclaves such as NORC at the University of Chicago,[49] and the electronic Data Research and Innovation Service;[50]
- Open sharing of data modified by privacy-preserving methods.[51]

We acknowledge that free and open sharing of all data are a distant goal, however, our expectation of the near future is that all descriptions of ML/AI algorithm development will be accompanied by clear statements of what tools and mechanism will be used to support access to the data used, for the purposes of replication of reported results. The advent of the facilities described above means that there are fewer reasons to be unable to share data from publicly funded research with other researchers, and as such, we would strongly recommend that investigators establish early on what mechanisms they think are most appropriate and ensure their relevant partners are in agreement.

Are the code, software, and all other relevant parts of the prediction modelling pipeline available to others to facilitate replicability[52]?

Reproducibility of research has become a growing concern across many scientific domains,[53 54] and in the ML/AI field, access to the underlying code and raw data are central to preventing and mitigating reproducibility concerns. A recent example of how concerns regarding reproducibility in medical modelling research have manifested comes from a review of studies published using the Massachusetts Institute of Technology critical care database (MIMIC), which illustrates the degree to which inadequate reporting can affect replication in the prediction modelling.[9] Specifically, the reproducibility issues that have been identified in the literature occur not only in attempts to recreate reported findings, but also in how authors report data characteristics, such as the inclusion and exclusion criteria used to arrive at the final population of interest. In the review,[28] studies based on the same core dataset (MIMIC) predicting mortality were investigated, and two important results were identified. For more than half of the studies examined, the reproduced sample size differed by more than 25% from the reported sample size because of insufficiently clear descriptions of the inclusion or exclusion criteria. The result of inadequate reporting was that in the replication, the use of off-the-shelf logistic regression and boosted trees on the reproduced samples produced better results in 64% and 82% of the 28 studies, respectively, than the ML/AI model reported in the original study.

These problems could have been easily avoided by providing the project code, specifically the code relating to data cleaning and pre-processing. The RECORD reporting guidelines for studies using routinely collected health data already recommend providing detailed information to this effect,[55] and several potential solutions can facilitate this process, including code sharing and project curation platforms such GitHub. However, we acknowledge that the ideal level of sharing is not always achievable for many different reasons.[56] We would highly recommend that, where possible, researchers archive annotated code and include adequate information about software version control to support attempts to reproduce their results.[57]

Is there organisational transparency about the flow of data and results?

Patients have strong views about transparency in the flow of data, and how their data are secured.[58] For patients and their clinicians to trust ML/AI models, they need to understand the interactions that led to the development of the model, whether they are between organisations in the public, not-for-profit and industrial sectors, or within them (eg, transfer from one hospital department to another). Complying with the aforementioned legislative frameworks (eg, the European Union's General Data Protection Regulation) is necessary, but is not sufficient to show the transparency required to produce trustworthy ML/AI research. The degree of detail needed will differ depending on the institutions involved and the nature of the work being undertaken. Therefore, the responsibility lies with ML/AI algorithm developers and those involved in accessing, transferring, or storing the data, to engage key stakeholders to understand what is required in each particular case. One aspect of the reporting procedure that can help ensure transparency regarding the aforementioned interactions is the inclusion of clear declarations of interest by all involved parties.

Impact evaluation (questions 13-16)

Are the results generalisable to settings beyond where the system was developed (that is, results reproducibility/external validity[59])?

Even before ML/AI had become established, few validation studies had been done on diagnostic and prognostic tools.[60] In external validation studies, reductions in the predictive accuracy of models (relative to their original performance in development studies) is expected.[61 62] Systematic reviews have repeatedly observed this reduced accuracy in the applications of classical statistical models to various healthcare related prediction tasks, from mortality risk prediction in patients with acute kidney injury[63] to risk prediction of falls in older people.[64] How this phenomenon is associated with results reproducibility (that is, the production of corroborating results in a new study, having followed the same experimental methods[65]), whether it is a consequence of the inadequate reporting observed in the modelling literature,[59] or other related issues is unclear. Given the additional complexities introduced by ML/AI algorithms, developers should be proactive in ensuring that sufficient information is provided to allow their models to undergo rigorous but fair[66] external validation (ideally by independent investigators). This work could include identifying potential datasets for validation experiments at the planning stage, parallel data collection of a validation dataset, or using simulated data to illustrate that the model performs as expected.

Does the model create or exacerbate inequities in healthcare by age, sex, ethnicity, or other protected characteristics?

Systematic testing for bias and fairness is the first decision making step in informed model selection, to minimise inequities that could be caused by ML/AI algorithms use.[67] Although many of the ML/AI algorithms developed will often have bias, it should be compared with the bias in the existing systems being used. One way in which ML/AI algorithms result in bias is by making disproportionate errors in different populations. Depending on how the ML/AI algorithm has been developed (including whether key populations (defined by sex, age, and ethnicity) are sufficiently represented in the data, and included in the training of the algorithm) can influence the predictive accuracy of the algorithm in different subgroups. Thus, when these predictions are used to take actions on individuals, they can create or exacerbate inequities.[68] The issue of data that are not truly representative of the entire target population is particularly important,[69] because it highlights the importance of fairness considerations at every point in the project cycle. Other examples of how these issues can manifest in the real world can be found in ProPublica's analysis of a recidivism prediction tool (the Correctional Offender Management Profiling for Alternative Sanctions software)[10] and the United States' diabetes screening criterion,[70] which both illustrate variation in performance of an algorithm based on race.

The types of performance variation to be investigated depend on the consequent actions (or interventions) that the algorithm is helping to decide between. If the interventions are expensive or have unwanted side-effects, then we would want to minimise disparities in the number of false positive predictions from different subgroups, to prevent unnecessary harm. If the interventions are predominantly assistive, we should be more concerned with disparities in false negatives, to prevent individuals missing out on a potentially beneficial input. The explanation above presupposes that a decision threshold has been set, which might sometimes be outside of a developer's remit. However, developers still need to demonstrate that when using sensible thresholds, the algorithm does not create or exacerbate inequalities. In fact, several methodological developments in the area of fairness evaluation support this type of analysis,[71-73] and ML/AI developers and health practitioners should engage with these tools. One way in which researchers might demonstrate bias in key subgroups (eg, in minority ethnic groups, or by age) would be to explicitly present these findings so that users of the algorithm know where it has good or poor predictive accuracy.

What evidence is there that clinicians and patients find the model and its output (reasonably) interpretable?

Clinical adoption of an algorithm depends on two main factors: its clinical usefulness and its trustworthiness. When the outputs of a prediction model do not directly answer a specific clinical question, its usefulness is limited (as discussed in earlier questions), whereas models whose processing pipeline is difficult to explain and justify to a general audience will invariably limit

the trust placed in its outputs,[74] despite robust and demonstrated statistical gains. However, trust is not the only reason that interpretability in important.[75] Recent changes in legislation (eg, the EU General Data Protection Regulation) have introduced additional protections for individuals (including a right to an explanation for how a decision was made and where it pertains them[76]), thereby creating a legal requirement to provide insight into the underlying decision making process an algorithm learns. Several partial solutions, including model specific and model agnostic methods (eg, LIME[77]), can be used to claw back interpretability when using ML/AI methods. Legal and moral burdens of explanation to establish trust will vary with the nature of the decision—that is, ML/AI applications in health that influence the allocation of potentially life-prolonging treatments will necessitate a much higher explanatory burden to satisfy those individuals who are affected. Therefore, the sufficiency of any explanations and adequacy of any insight producing method can only be determined through consultation and collaboration with the end users (clinicians), and target audience (patients).

How will evidence of real world model effectiveness in the proposed clinical setting be generated, and how will unintended consequences be prevented?

ML/AI tools often carry the misleading aura of self-evident advanced technology, which falsely limits the perceived need for careful validation and verification of their performance, clinical use, and overall use once they begin being used in the routine clinical practice. A recent systematic review showed that only a couple of hundred randomised clinical trials (of a million trials in total) examined how the use of diagnostic tests affected clinical outcomes (and therefore clinical utility).[78] With regards to the ML/AI domain, Babylon Health's symptom checker for triage was piloted at a small number of general practices. During early testing, patient focus groups had concerns that there might be "gaming [of] the symptom checker to achieve a GP appointment."[4] This example demonstrates how algorithms are not always used as intended in the real world, and that these factors need to be assessed using pragmatic clinical trials.[79] Early consideration of what the potential pitfalls are of the proposed ML/AL based solution and how it could be manipulated (among other issues) would help researchers develop a better informed framework with which to decide how their tool should be built.

Implementation (questions 17-20)

How is the model being regularly reassessed, and updated as data quality and clinical practice changes (that is, post-deployment monitoring)?

Even if evidence of efficacy and real world effectiveness of a model is sufficient to endorse its widespread use in clinical practice, the effectiveness requires constant review given the dynamic landscape of the healthcare environment. For example, computer aided diagnosis programs have become an integral part of breast cancer screening programmes worldwide since the US Food and Drug Administration (FDA) first approved one for use in 1998,[80] but are they still as useful as they were 20 years ago? Most of the commercially available tools are based on neural networks which identify regions of interest, and diagnose the identified abnormality (eg, calcification or mass). Early studies showed modest increases in detection rates of breast cancer using computer aided diagnosis or detection (CAD), compared with clinicians working without the aid of a CAD system.[1 81] However, almost 20 years after the FDA's first license for a mammography based CAD system, national registry based studies have shown no significant improvement in diagnostic accuracy associated with CAD use in mammography interpretation.[82] Moreover, researchers have recently demonstrated that incorrect prompts from mammography based CAD systems can actually decrease the sensitivity of more discerning users by up to 0.145 (95% confidence interval 0.034 to 0.257) for difficult diagnoses.[83] Although the work discussed is not a comprehensive review of CAD in breast cancer, the results suggest the importance for constant re-evaluation of technologies, as their usefulness can change over time. Researchers should aim to plan and develop model performance with the intention to reassess; thus, they need to discuss early on what the necessary mechanisms are to facilitate this process, and how these mechanisms can be integrated at the start of implementation (instead of unplanned additional years later).

Is the ML/AI model cost effective to build, implement, and maintain?

Although ML/AI algorithms might offer transformational benefits to healthcare systems and patients, substantial costs can be associated with the development of software, generation and use of data, implementation of a new system in practice, and acting on the additional information provided. Understanding the potential clinical benefit of new models (over and above current practice) alongside the cost or savings introduced by using these models should form part of any healthcare decision maker's appraisal of ML/AI based technologies. Effective appraisal will require the development of assessment frameworks that take into account both the evidence for effectiveness and the evidence for economic impact. In this area, healthcare decision makers (such as the National Institute for Health and Care Excellence and the FDA) are crucial. They can help developers of ML/AI models by providing clear guidance on the appropriate evidence that should be generated to demonstrate both effectiveness and economic impact, including: credible evidence relating to technical accuracy of the models; the relevant outcomes that show clinical effectiveness in general practice; and, as appropriate, evidence to inform decision makers on the budget impact or the cost effectiveness. Researchers should plan projects with an understanding of how their tool or algorithm will eventually be operationalised.

How will the potential financial benefits be distributed if the ML/AI model is commercialised?

Like all technologies, ML/AI algorithms could have a market value. In situations where commercialisation is a goal, health systems and governmental research funding can make a substantial contribution to the creation of an algorithm via the associated unrecoverable costs, such as data acquisition (clinicians' time, scanners), data annotation (training clinicians who eventually interpret the data generated), and developers' time (that is, when they are publicly funded researchers). This issue is even more important in a publicly funded health system, because the symbiotic relationship between data-generating institutions and those with the capabilities to create ML/AI algorithms is only possible because of expectation that benefits arising from the data use will be retained (to some degree) by the health system, thereby satisfying the social contract with the public. Therefore, the investment and contributions of a health system or institution to an algorithm should be recognised, and a mechanism to compensate them for having done so should be put in place. Answering this question after the development of an algorithm or ML/AI based tool is often fraught with complexities that can take years to untangle, and thus, we would strongly advise researchers and developers near the end of the planning stages of any project to clarify their institution's innovation pathway, including the routes to commercialisation and the framework through which this could be achieved.

How have the regulatory requirements for accreditation/approval been addressed?

Software products including ML/AI algorithms can be subject to many regulatory requirements, depending on the setting in which the product will be used, from research and development to placing the product on the market (box 2 provides a high level overview of the UK's regulatory framework). In our experience, while most clinicians are aware of "CE" marking of physical devices (the regulatory framework in the EU and the UK), its application to software products can often be a surprise, which is also true of software developers. Given that the regulatory landscape for health related ML/AI based software has changed substantially over the past decade, and will continue to respond dynamically to innovation for the foreseeable future, discussions regarding the regulatory requirements for products in development should be made early in the planning process of a research project. However, having this conversation once is clearly not sufficient. For example, devices that are developed and used in-house (in the UK) are not currently subject to device regulations, but this will change in 2020 when new updated regulations apply,[87 88] and as such, regular review of regulatory compliance is necessary.

Conclusion: from critical questions to a consensus TREE framework

Similar to how clinicians have been aided by frameworks to evaluate the strength of evidence over previous decades, the ML/AI field should usefully build on what has been learned in traditional statistical approaches for clinical evidence and the quality assurance pipeline.[6 19 89-93] However, as shown here, some of the challenges are new and different. Encouraging patients, clinicians, academics, and all manner of healthcare decision makers to ask the challenging questions raised here will hopefully contribute to the development of safe and effective ML/

Box 2: Overview of the UK's regulatory framework for health related algorithms involving machine learning (ML) and artificial intelligence (AI)

Developers should determine whether their ML/AI algorithm falls under the Medical Device Regulations' remit,[84] which until 2010 did not regulate independent software products. These regulations cover products that make claims with a medical nature such as: providing diagnostic information, making recommendations for treatment, or providing risk predictions of disease. The Medicines and Healthcare products Regulatory Agency has published guidance for developers that covers this in greater detail.[85] If an algorithm does fall within the remit of the aforementioned regulation, the developer must then seek regulatory approval or accreditation in the form of a "CE" mark before marketing it. To CE mark an algorithm, the developer must follow one of the applicable conformity assessment routes that, for medium and high risk products, will require the involvement of a notified body to assure the process. The developer must ensure that the device meets the relevant essential requirements before applying the CE mark. These requirements include:

- Benefits to the patient shall outweigh any risks
- Manufacture and design shall take account of the generally acknowledged gold standard
- Devices shall achieve the performance intended by the manufacturer
- Software must be validated according to the gold standard, taking into account the principles of development lifecycle, risk management, validation, and verification
- Confirmation of conformity must be based on clinical data; evaluation of these data must follow a defined and methodologically sound procedure.

In addition to the above, manufacturers are required to have post market surveillance provision to review experience gained from device use and to apply any necessary corrective actions.

Moreover, the use of ML/AI algorithms might be regulated indirectly by other legislation or regulatory agencies. The highest profile additional legislative framework to be aware of might be the European Union's General Data Protection Regulations, the relevance of which has been discussed elsewhere (questions 3 and 16). In terms of other regulatory agencies who have an important role in the regulation of ML/AI software in health, the United Kingdom's Care Quality Commission is one group to be aware of, as they are tasked with monitoring compliance with NHS Digital's Clinical risk management standards[86]; a contractual requirement placed on developers engaging in service provision to the UK's health service.

AI based tools in healthcare. Developing a definitive framework for how to undertake effective and ethical research in ML/AI will involve many challenges. These challenges include finding common terminology (where key terms partly or fully overlap in meaning), balancing the need for robust empirical evidence of effectiveness without stifling innovation, identifying how best to manage the many open questions regarding best practices of development and communication of results, the role of different venues of communication and reporting, simultaneously providing sufficiently detailed advice to produce actionable guidance for non-experts, and balancing the need for transparency against the risk of undermining intellectual property rights. Addressing these challenges of transparency, reproducibility, ethics, and effectiveness are important in delivering health benefits from ML/AI.

Author affiliations

[1]Alan Turing Institute, Kings Cross, London, UK

[2]Departments of Mathematics and Statistics, University of Warwick, Coventry, UK

[3]Warwick Medical School, University of Warwick, Coventry, UK

[4]Kings College Hospital, Denmark Hill, London, UK

[5]Department of Statistical Science, University College London, London, UK

[6]University of Chicago, Chicago, IL, USA

[7]Science Policy and Research, National Institute for Health and Care Excellence, Manchester, UK

[8]Health and Social Care Directorate, National Institute for Health and Care Excellence, London, UK

[9]Data and Analytics Group, National Institute for Health and Care Excellence, London, UK

[10]Clinical Practice Research Datalink, Medicines and Healthcare products Regulatory Agency, London, UK

[11]Medicines and Healthcare products Regulatory Agency, London, UK

[12]Julius Centre for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, Utrecht, Netherlands

[13]UK EQUATOR Centre, Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK

[14]Meta-Research Innovation Centre at Stanford, Stanford University, Stanford, CA, USA

[15]Department of Statistics, University of Oxford, Oxford OX1 3LB, UK

[16]Health Data Research UK London, University College London, London, UK

[17]Institute of Health Informatics, University College London, London, UK

[18]National Institute for Health Research, University College London Hospitals Biomedical Research Centre, University College London, London, UK

We thank all those at the Alan Turing Institute, HDR UK, National Institute for Clinical and Care Excellence (NICE), Medicines and Healthcare products Regulatory Agency (MHRA), Clinical Practice Research Datalink (CPRD), Enhancing the Quality and Transparency of Health Research (EQUATOR) Network, Meta-Research Innovation Centre at Stanford (METRICS), and Data Science for Social Good (DSSG) programme at the University of Chicago who supported this project.

Contributors: SV and BAM contributed equally to the manuscript. SV, BAM, and HH conceived the study. BAM, GB, FJK, and SV wrote the first version of the manuscript. The second version of the manuscript, which formed the basis of the submission to *The BMJ*, was written and edited by all the stated authors. All authors read and approved the final and accepted version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: The work presented here did not receive any particular funding. SJV, BAM, GB, FJK, and CH are employees of the Alan Turing Institute (support from Engineering and Physical Sciences Research Council grant EP/N510129/1). GSC was supported by the NIHR Biomedical Research Centre, Oxford. HH is a National Institute for Health Research (NIHR) senior investigator. HH's work is supported by (1) Health Data Research UK (grant No LOND1), which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; (2) BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement number 116074 (this joint undertaking receives support from the European Union's Horizon 2020 research and innovation programme and European Federation of Pharmaceutical Industries and European Society of Cardiology and is chaired by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and European Society of Cardiology); (3) NIHR University College London Hospitals Biomedical Research Centre. JPAI is supported by an unrestricted gift from Sue and Bob O'Donnell to the Stanford Prevention Research Center. METRICS is supported by a grant from the Laura and John Arnold Foundation. KGMM is supported by the Netherlands Organisation for Health Research and Development. GB and SV are supported by the University of Warwick's Impact Acceleration, funded by the EPSRC. PJ, SC, KSLM, and AJ are employees of NICE. PM, DG, MB, and RB are employees of the MHRA. SJV is supported by the data study group funding as its director (/TU/B/000012). The authors confirm that the funders had no role in the writing or editing of the manuscript.

Competing interests: We have read and understood BMJ policy on declaration of interests and declare the following interests: GSC and KGMM are part of the TRIPOD steering group. GSC is director of the UK EQUATOR Centre. The remaining authors have no additional declarations.

The lead author affirms that the manuscript is an honest, accurate, and transparent account of the work undertaken and being reported; that no important aspects of the work have been purposefully omitted without explanation; and that any discrepancies from the original manuscript as planned have been explained.

Patient and public involvement: No patients were directly involved in the inception of the manuscript, development of the questions, or review of the text before publication.

1 Gilbert FJ, Astley SM, Gillan MG, et al, CADET II Group. Single reading with computer-aided detection for screening mammography. *N Engl J Med* 2008;359:1675-84. doi:10.1056/NEJMoa0803545

2 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *Digital Med* 2018;1:18.

3 Tsigelny IF. Artificial intelligence in drug combination therapy. *Brief Bioinform* 2019;20:1434-48.

4 Carney S. Report of the NW London CCGs' collaboration board. November 2017. https://www.centrallondonccg.nhs.uk/media/70538/150-updated-with-clarification-november-collaboration-board-update.pdf

5 De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342-50. doi:10.1038/s41591-018-0107-6

6 Dhindsa K, Bhandari M, Sonnadara RR. What's holding up the big data revolution in healthcare?*BMJ* 2018;363:k5357.

7 Office of the President, Executive. 2014. Big data: seizing opportunities, preserving values. Big data: an exploration of opportunities, values, and privacy issues. 1-85. https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

8 Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Arxiv 1711.00399 [Preprint]. 2016. https://arxiv.org/abs/1711.00399

9 Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. Proceedings of the 2nd Machine Learning for Healthcare Conference, in Proceedings of Machine Learning Research 2017;68:361-76.

10 Larson J, Mattu S, Kirchner L, Angwin J. 2016. Compas Analysis. [Github Repository] https://github.com/propublica/compas-analysis.

11 Kiraly FJ, Mateen BA, Sonabend R. NIPS - not even wrong? - a systematic review of empirically complete demonstrations of algorithmic effectiveness in the machine learning and artificial intelligence literature. arXiv [Preprint] 2018. https://arxiv.org/abs/1812.07519
12 Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73. doi:10.7326/M14-0698
13 Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *Lancet* 2014;383:156-65. doi:10.1016/S0140-6736(13)62229-1
14 Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014;383:166-75. doi:10.1016/S0140-6736(13)62227-8
15 Al-Shahi Salman R, Beller E, Kagan J, et al. Increasing value and reducing waste in biomedical research regulation and management. *Lancet* 2014;383:176-85. doi:10.1016/S0140-6736(13)62297-7
16 Chan AW, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014;383:257-66. doi:10.1016/S0140-6736(13)62296-5
17 Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383:267-76. doi:10.1016/S0140-6736(13)62228-X
18 Department of Health and Social Care Guidance. initial code of conduct for data-driven health and care technology. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology 2018 Sept.
19 Steyerberg EW, Moons KG, van der Windt DA, et al, PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381. doi:10.1371/journal.pmed.1001381
20 Snooks H, Bailey-Jones K, Burge-Jones D, et al. Effects and costs of implementing predictive risk stratification in primary care: a randomised stepped wedge trial. *BMJ Qual Saf* 2019;28:697-705. doi:10.1136/bmjqs-2018-007976
21 Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. *BMC Med Inform Decis Mak* 2018;18(Suppl 4):122. doi:10.1186/s12911-018-0677-8
22 Morgan RL, Whaley P, Thayer KA, Schünemann HJ. Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environ Int* 2018;121:1027-31. doi:10.1016/j.envint.2018.07.015
23 UK Standards for Public Involvement in Research. Homepage. 2018. https://sites.google.com/nihr.ac.uk/pi-standards/home
24 Cortes C, Jackel LD, Chiang WP. *Limits on learning machine accuracy imposed by data quality.* In: Advances in neural information processing systems, 1995: 239-46.
25 Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Sci Rep* 2018;8:7961. doi:10.1038/s41598-018-26174-1
26 Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;68:25-34. doi:10.1016/j.jclinepi.2014.09.007
27 Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 8 May 2018. https://otexts.com/fpp2/
28 Lyddon S, Walker S, Holmes CC. *Nonparametric learning from Bayesian models with randomized objective functions.* In: Advances in neural information processing systems, 2018:2072-82.
29 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv [Preprint] 2014 Sep 4. https://arxiv.org/abs/1409.1556
30 Inouye M, Abraham G, Nelson CP, et al, UK Biobank CardioMetabolic Consortium CHD Working Group. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J Am Coll Cardiol* 2018;72:1883-93. doi:10.1016/j.jacc.2018.07.079
31 Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. arXiv [Preprint] 2016 May 24. https://arxiv.org/abs/1605.07678
32 Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv [Preprint] 2012 Jul 3. https://arxiv.org/abs/1207.0580
33 Collins GS, Moons KGM. Comparing risk prediction models. *BMJ* 2012;344:e3186.
34 Morin A, Urban J, Sliz P. A quick guide to software licensing for the scientist-programmer. *PLoS Comput Biol* 2012;8:e1002598. doi:10.1371/journal.pcbi.1002598
35 Epstein AS, Zauderer MG, Gucalp A, et al. Next steps for IBM Watson Oncology: scalability to additional malignancies. *J Clin Oncol* 2014;32(suppl):6618. doi:10.1200/jco.2014.32.15_suppl.6618
36 Suwanvecho S, Suwanrusme H, Sangtian M, Norden A, Urman A, Hicks A, et al. Concordance assessment of a cognitive computing system in Thailand. *J Clin Oncol* 2017;35(15_suppl):6589.
37 Somashekhar S, Kumarc R, Rauthan A, Arun K, Patil P, Ramya Y.Double blinded validation study to assess performance of IBM artificial intelligence platform, Watson for oncology in comparison with Manipal multidisciplinary tumour board – First study of 638 breast cancer cases. *Cancer Res* 2017;77(4 suppl):S6-07.
38 Baek J, Ahn S, Urman A, et al. Use of a cognitive computing system for treatment of colon and gastric cancer in South Korea. *J Clin Oncol* 2017;35(15_suppl):e18204.
39 Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;302:2345-52. doi:10.1001/jama.2009.1757
40 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
41 Ioannidis JPA, Bossuyt PMM. Waste, leaks, and failures in the biomarker pipeline. *Clin Chem* 2017;63:963-72. doi:10.1373/clinchem.2016.254649
42 Pasquetto IV, Randles BM, Borgman CL. On the reuse of scientific data. *Data Sci J* 2017;16:8.
43 Clinical Practice Research Datalink. Homepage. 2018. https://www.cprd.com
44 NHS Digital. Homepage. 2018. https://digital.nhs.uk
45 Health Quality Improvement Partnership. Homepage. 2018. https://www.hqip.org.uk/
46 Dryad Digital Repository. Homepage. 2018. https://datadryad.org
47 Amnesia. What is Amnesia? 2015. https://amnesia.openaire.eu/amnesiaInfo.html
48 UK Data Archive. Homepage. 2018. https://data-archive.ac.uk
49 NORC. Data enclave. 2018. http://www.norc.org/Research/Capabilities/Pages/data-enclave.aspx
50 ISD Services | Electronic Data Research and Innovation Service (eDRIS) | ISD Scotland. Isdscotland.org. 2018. https://www.isdscotland.org/Products-and-Services/eDRIS/
51 Rao PR, Krishna SM, Kumar AS. Privacy preservation techniques in big data analytics: a survey. *J Big Data* 2018;5:33. doi:10.1186/s40537-018-0141-8
52 Claerbout JF, Karrenbach M. Electronic documents give reproducible research a new meaning. *SEG Expanded Abstracts* 1992;11:601-4. doi:10.1190/1.1822162
53 Ioannidis JPA. How to make more published research true. *PLoS Med* 2014:e1001747.
54 Pashler H, Wagenmakers EJ. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect Psychol Sci* 2012;7:528-30. doi:10.1177/1745691612465253
55 Benchimol EI, Smeeth L, Guttmann A, et al, RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;12:e1001885. doi:10.1371/journal.pmed.1001885
56 Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works?/ *Am Med Inform Assoc* 2019;26:1651-4. doi:10.1093/jamia/ocz130
57 GitHub. The Turing Way. 2019.https://github.com/alan-turing-institute/the-turing-way.
58 Stockdale J, Cassell J, Ford E. "Giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. *Wellcome Open Res* 2019;3:6. doi:10.12688/wellcomeopenres.13531.2
59 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515-24. doi:10.7326/0003-4819-130-6-199903160-00016
60 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. doi:10.1186/1471-2288-14-40
61 Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment of validation practices for molecular classifiers. *Brief Bioinform* 2011;12:189-202. doi:10.1093/bib/bbq073
62 Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;68:25-34. doi:10.1016/j.jclinepi.2014.09.007
63 Ohnuma T, Uchino S. Prediction models and their external validation studies for mortality of patients with acute kidney injury: a systematic review. *PLoS One* 2017;12:e0169341. doi:10.1371/journal.pone.0169341

64 Billington J, Fahey T, Galvin R. Diagnostic accuracy of the STRATIFY clinical prediction rule for falls: a systematic review and meta-analysis. *BMC Fam Pract* 2012;13:76. doi:10.1186/1471-2296-13-76
65 Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean?*Sci Transl Med* 2016;8:341ps12. doi:10.1126/scitranslmed.aaf5027
66 Dagan N, Cohen-Stavi C, Leventer-Roberts M, Balicer RD. External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study. *BMJ* 2017;356:i6755.
67 Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544-7. doi:10.1001/jamainternmed.2018.3763
68 Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr Genet* 2018;28:85-9. doi:10.1097/YPG.0000000000000206
69 Morales J, Welter D, Bowler EH, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol* 2018;19:21. doi:10.1186/s13059-018-1396-2
70 O'Brien MJ, Lee JY, Carnethon MR, et al. Detecting dysglycemia using the 2015 United States Preventive Services Task Force screening criteria: a cohort analysis of community health center patients. *PLoS Med* 2016;13:e1002074. doi:10.1371/journal.pmed.1002074
71 Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. In: Advances in neural information processing systems. 2017:4066-76.
72 Saleiro P, Kuester B, Stevens A, Anisfeld A, Hinkson L, London J, Ghani R. Aequitas: a bias and fairness audit toolkit. arXiv [Preprint] 2018. https://arxiv.org/abs/1811.05577
73 Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care?*AMA J Ethics* 2019;21:E167-79. doi:10.1001/amajethics.2019.167
74 Nuffield Council on Bioethics. Artificial intelligence (AI) in healthcare and research. 2018. https://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf
75 Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv [Preprint] 2017. https://arxiv.org/abs/1702.08608.
76 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) Official Journal of the European Union; 2016.
77 Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016 Aug 13:1135-44.
78 Siontis KC, Siontis GC, Contopoulos-Ioannidis DG, Ioannidis JP. Diagnostic tests often fail to lead to changes in patient outcomes. *J Clin Epidemiol* 2014;67:612-21. doi:10.1016/j.jclinepi.2013.12.008
79 Ford I, Norrie J. Pragmatic trials. *N Engl J Med* 2016;375:454-63. doi:10.1056/NEJMra1510059
80 Summary of safety and effectiveness data. *R2 technologies.* US Food and Drug Administration, 1998: 970058.
81 Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 2001;220:781-6. doi:10.1148/radiol.2203001282
82 Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL, Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828-37. doi:10.1001/jamainternmed.2015.5231
83 Povyakalo AA, Alberdi E, Strigini L, Ayton P. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Med Decis Making* 2013;33:98-107. doi:10.1177/0272989X12465490
84 Medical Devices Regulations. SI 2002 No 618, as amended. London, 2002. http://www.legislation.gov.uk/uksi/2002/618/contents/made.
85 Medicines and Healthcare products Regulatory Agency (MHRA). Medical devices: software applications (apps). 2014. https://www.gov.uk/government/publications/medical-devices-software-applications-apps
86 NHS Digital. Clinical risk management standards. 2012. https://digital.nhs.uk/services/solution-assurance/the-clinical-safety-team/clinical-risk-management-standards
87 Regulation (EU) 2017/745 of The European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. Official Journal of the European Union; 2017.
88 Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU. Official Journal of the European Union; 2018.
89 Hemingway H, Croft P, Perel P, et al, PROGRESS Group. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ* 2013;346:e5595. doi:10.1136/bmj.e5595
90 Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605. doi:10.1136/bmj.b605
91 Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009;338:b604. doi:10.1136/bmj.b604
92 Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606. doi:10.1136/bmj.b606
93 HM Treasury. The Aqua Book: guidance on producing quality analysis for government. 2015. https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government.

Web appendix: Supplementary material