PLOS DIGITAL HEALTH

REVIEW

Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible

Tyler J. Loftus[1,2☯], Patrick J. Tighe[3], Tezcan Ozrazgat-Baslanti[2,4], John P. Davis[5], Matthew M. Ruppert[2,4], Yuanfang Ren[2,4], Benjamin Shickel[2,4], Rishikesan Kamaleswaran[6], William R. Hogan[7], J. Randall Moorman[8], Gilbert R. Upchurch, Jr[1], Parisa Rashidi[2,9☯], Azra Bihorac[2,4☯*]

1 Department of Surgery, University of Florida Health, Gainesville, Florida, United States of America, **2** Precision and Intelligent Systems in Medicine (PrismaP), University of Florida, Gainesville, Florida, United States of America, **3** Departments of Anesthesiology, Orthopedics, and Information Systems/Operations Management, University of Florida Health, Gainesville, Florida, United States of America, **4** Department of Medicine, University of Florida Health, Gainesville, Florida, United States of America, **5** Department of Surgery, University of Virginia, Charlottesville, Virginia, United States of America, **6** Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, Georgia, United States of America, **7** Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, United States of America, **8** Department of Medicine, University of Virginia, Charlottesville, Virginia, United States of America, **9** Departments of Biomedical Engineering, Computer and Information Science and Engineering, and Electrical and Computer Engineering, University of Florida, Gainesville, Florida, United States of America

☯ These authors contributed equally to this work.
* abihorac@ufl.edu

Citation: Loftus TJ, Tighe PJ, Ozrazgat-Baslanti T, Davis JP, Ruppert MM, Ren Y, et al. (2022) Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible. PLOS Digit Health 1(1): e0000006. https://doi.org/10.1371/journal.pdig.0000006

Editor: Henry Horng-Shing Lu, National Yang Ming Chiao Tung University, TAIWAN

Published: January 18, 2022

Funding: TJL was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under Award Number K23GM140268. PTJ was supported by R01GM114290 from the NIGMS and R01AG121647 from the National Institute on Aging (NIA). T.O.B. was supported by K01 DK120784, R01 DK123078, and R01 DK121730 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK), R01 GM110240 from the National Institute of General Medical Sciences (NIH/NIGMS), R01 EB029699 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), and R01 NS120924 from the National Institute of Neurological

Abstract

Established guidelines describe minimum requirements for reporting algorithms in healthcare; it is equally important to objectify the characteristics of ideal algorithms that confer maximum potential benefits to patients, clinicians, and investigators. We propose a framework for ideal algorithms, including 6 desiderata: explainable (convey the relative importance of features in determining outputs), dynamic (capture temporal changes in physiologic signals and clinical events), precise (use high-resolution, multimodal data and aptly complex architecture), autonomous (learn with minimal supervision and execute without human input), fair (evaluate and mitigate implicit bias and social inequity), and reproducible (validated externally and prospectively and shared with academic communities). We present an ideal algorithms checklist and apply it to highly cited algorithms. Strategies and tools such as the predictive, descriptive, relevant (PDR) framework, the Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence (SPIRIT-AI) extension, sparse regression methods, and minimizing concept drift can help healthcare algorithms achieve these objectives, toward ideal algorithms in healthcare.

Introduction

The breadth and complexity of human disease confer unique challenges in clinical decision-making. The 10th revision of the International Statistical Classification of Diseases and Related

Disorders and Stroke (NIH/NINDS). PR was supported by National Science Foundation CAREER award 1750192, 1R01EB029699 and 1R21EB027344 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), R01GM-110240 from the National Institute of General Medical Science (NIH/NIGMS), 1R01NS120924 from the National Institute of Neurological Disorders and Stroke (NIH/NINDS), and by R01 DK121730 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK). A.B. was supported R01 GM110240 from the National Institute of General Medical Sciences (NIH/NIGMS), R01 EB029699 and R21 EB027344 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), R01 NS120924 from the National Institute of Neurological Disorders and Stroke (NIH/NINDS), and by R01 DK121730 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Health Problems (ICD) classification system includes approximately 68,000 diagnostic codes. Patients can have nearly any combination of these diagnoses managed with nearly any combination of relevant therapies whose efficacy hinges on underlying behavioral, social, and genetic determinants of health. Patients and clinicians face shared clinical decision-making tasks while under time constraints and high cognitive loads from high volumes of information [1,2]. The average person generates more than 1 million gigabytes of healthcare data during their lifetime or approximately 300 million books; these massive volumes of data far exceed human cognitive capacities, which allow for approximately 5 to 10 facts per decision [3,4]. Abnormal or unexpected data are typically applied to hypothetical-deductive reasoning and heuristic processes that are highly variable and error prone; when collected data are within normal limits, they are often discarded to reduce cognitive load [5,6]. Unsurprisingly, clinical decision-making errors are common and associated with mortality and morbidity [7,8].

By contrast, high-complexity and high-volume data can be parsed by machine learning applications with relative ease. Published algorithms supporting clinical decisions have become ubiquitous. Hundreds of retrospective studies are classified as artificial intelligence (AI) clinical trials, but few are methodologically rigorous [9,10]. Experts have described important components of algorithm-based and AI-enabled decision support and reporting guidelines; the minimum information about clinical artificial intelligence modeling (MI-CLAIM) checklist, Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence (SPIRIT-AI) extension, and Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI) guidelines facilitate consistent reporting, interpretation, and validation of AI applications by establishing minimum requirements [11–15]. We believe that it is equally important to objectify the characteristics of ideal algorithms that confer maximum potential benefits to patients, clinicians, and investigators. To this end, we propose a framework for ideal algorithms consisting of 6 desiderata. This framework is supported by a checklist, which we apply to prominent healthcare algorithms.

The ideal algorithm framework

Ideal algorithms are explainable, dynamic, precise, autonomous, fair, and reproducible, as illustrated in **Fig 1**. These desiderata are independent. Therefore, the degree to which an algorithm achieves maximum potential benefits to patients, clinicians, and investigators can be conceptualized as a continuum ranging from 0 desiderata (least ideal) to all 6 (most ideal). Desiderata may have unique applications for different algorithms types; our framework is designed to apply broadly to any algorithm type using objective criteria. A checklist of criteria for ideal algorithms is provided in **Table 1**. Each desideratum is evaluated as being met, partially met, not met, or not applicable by 1 or more objective criteria, which are each evaluated as being met, not met, or not applicable.

Ideal algorithms are explainable

Explainable algorithms convey the relative importance of features in determining outputs. Informed patients, diligent clinicians, and scrupulous investigators want to know how algorithm predictions are made. We recommend the predictive, descriptive, relevant (PDR) framework for achieving optimal explainability. PDR standardizes discussions regarding machine learning explanations according to predictive accuracy, descriptive accuracy (i.e., the ability of explainability mechanisms to describe objectively what the model has learned), and relevancy as judged by the algorithm's target human audience for its ability to provide insight into a chosen problem [16].

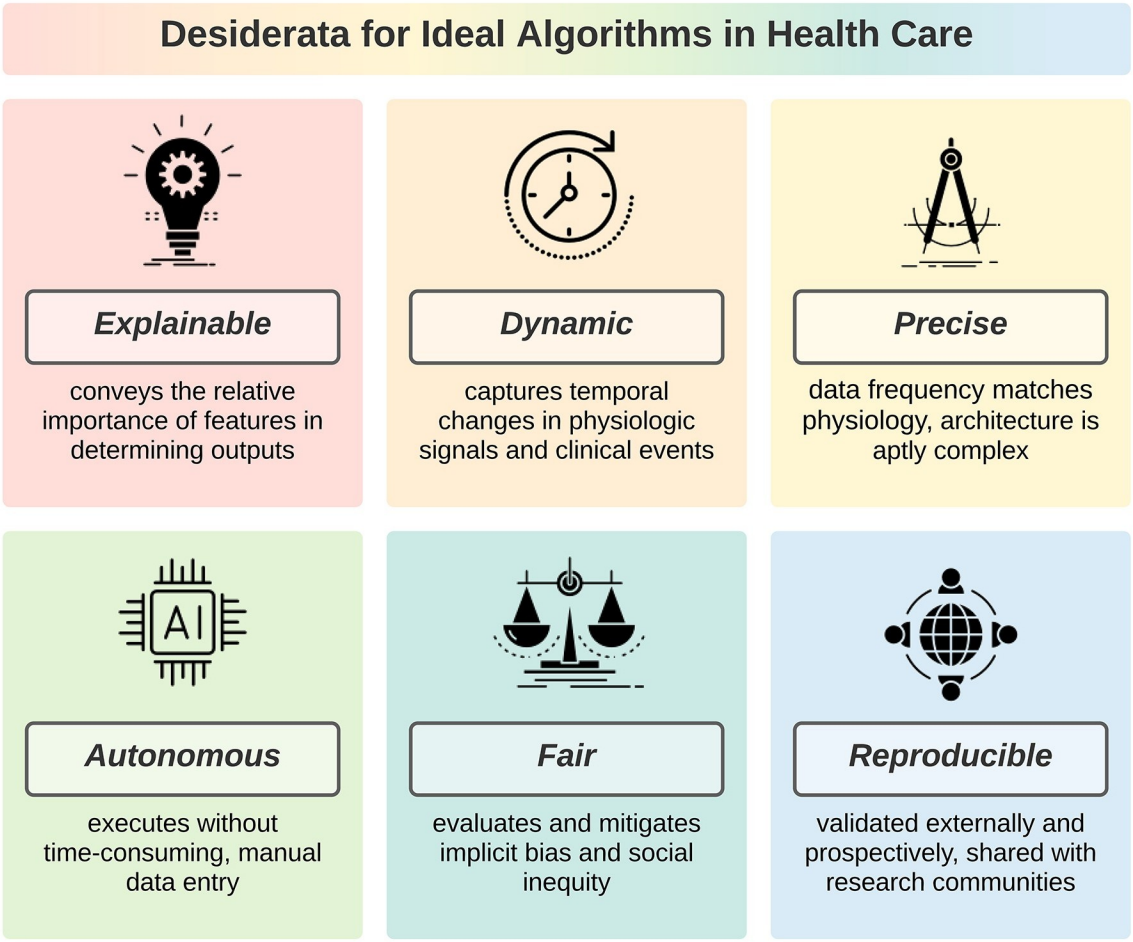

Fig 1. Ideal algorithms in healthcare have 6 desirable characteristics: explainable, dynamic, precise, autonomous, fair, and reproducible.

https://doi.org/10.1371/journal.pdig.0000006.g001

Algorithm predictive accuracy is commonly described and easily interpreted by most clinicians and scientists. Yet, one underappreciated aspect of predictive accuracy can affect the explainability of model outputs: In some cases, prediction error varies substantially by class. When applying an algorithm to a patient in a class with disproportionately high prediction error, one should have less confidence that model outputs are accurate and deemphasize algorithm outputs in the decision-making process. Descriptive accuracy, or objective indicators of what the model learned (e.g., coefficients in a regression model or weights in a neural network), is less commonly described and is difficult to achieve with complex, "black box" models such as deep neural networks. By contrast, the odds ratios produced by simple logistic regression are relatively easy to interpret, allowing clinicians to understand and mentally simulate the model's process for generating predictions. Despite the greater descriptive accuracy of simple models, complex models are often needed to solve complex, nonlinear problems for which simple models suffer from poor predictive accuracy. Therefore, algorithm explainability methods have focused on complex machine and deep learning models. We note, though, recent studies showing no great superiority of deep learning over regression in this field of classifying illness severity of individual patients using readily available clinical data [15,17]. Descriptive accuracy can be improved by choosing a simple, highly explainable model or performing post

Table 1. Checklist for ideal algorithms in healthcare.

Desiderata	Criteria	Yes	Location	No	N/A
Explainable Yes[a] Partially[b] No[c] N/A[d]	*Feature importance*: conveys the relative importance of features in determining algorithm outputs				
	Descriptive accuracy: describes what the algorithm has learned (e.g., weights in a neural network)				
	Simulatability: clinicians can understand and mentally simulate the model's process for generating predictions				
	Relevance: describes relevancy as judged by the algorithm's target human audience				
Dynamic Yes[a] Partially[b] No[c] N/A[d]	*Temporality*: captures temporal changes in physiologic signals and clinical events				
	Continuous monitoring: performance is reassessed at several time points, including the point at which performance is expected to plateau				
Precise Yes[a] Partially[b] No[c] N/A[d]	*Data frequency*: rate of data collection matches the rate of physiologic changes				
	Complexity: algorithm complexity matches the complexity of the prediction or classification task				
Autonomous Yes[a] Partially[b] No[c] N/A[d]	*Efficiency*: the algorithm executes without the need for time-consuming, manual data entry by the end user (i.e., patient, provider, or investigator)				
Fair Yes[a] Partially[b] No[c] N/A[d]	*Generalizability*: algorithm is developed and validated across diverse patient demographics and practice settings				
	Selectivity: excludes features that lack pathophysiologic or linguistic association with outcomes, but may introduce bias				
	Objectivity: includes variables that are minimally influenced by clinician judgments (e.g., vital signs)				
Reproducible Yes[a] Partially[b] No[c] N/A[d]	*Generalizability*: validated externally, prospectively				
	Collaboration: algorithm is shared with the research community				
	Compliance: fulfills SPIRIT-AI extension guidelines (if trial) and fulfills CONSORT-AI guidelines				

[a]Overall adjudication is "Yes" when all criteria are either met or not applicable.

[b]Overall adjudication is "Partially" when some but not all criteria are either met or not applicable.

[c]Overall adjudication is "No" when no criteria are met.

[d]Overall adjudication is "N/A" when all criteria are not applicable.

CONSORT-AI, Consolidated Standards of Reporting Trials-Artificial Intelligence; N/A, not applicable; SPIRIT-AI, Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence.

https://doi.org/10.1371/journal.pdig.0000006.t001

hoc analyses on a trained, complex model to understand what relationships the model has learned [18]. Finally, the PDR framework holds that relevancy is context specific, i.e., the usefulness of model explainability mechanisms depends on criteria that are unique to different people groups. Therefore, relevancy should be graded by the intended human audience and their intended use of predictions generated by the model.

Examination of relevancy can resolve trade-offs between predictive and descriptive accuracy [16]. Consider an algorithm that is predicting the risk for complications after surgery. To target researchers who seek the greatest predictive accuracy, explainability mechanisms could be used to optimize feature engineering. To target patients who are planning to undergo elective surgery, explainability mechanisms could be used to identify the most important modifiable risk factors for complications (i.e., modifiable predictors of wound infection could include poor blood glucose control and ongoing tobacco use). Notably, the PDR framework (intentionally) does not address causal inference or methods for determining the degree to which altering one variable changes another. In its purest form, explainability describes general relationships and does not distinguish between causal and noncausal effects. Therefore,

PDR is a simple and effective framework for evaluating and discussing the full range of user-specific machine learning interpretations without confusing explainability with causality.

Ideal algorithms are dynamic

Dynamic algorithms capture temporal changes in physiologic signals and clinical events via time series or sequence modeling. When algorithms are intended to improve clinical trial design, statistical adjustment, or patient enrollment strategies, static predictions at a single time point are adequate. When algorithms are intended to augment real-time, clinical decision-making as conditions evolve, the algorithm should make dynamic predictions using new data as it become available. Dynamic algorithm predictions are useful because continuous manual recalculations are burdensome for individual patients, caregivers, and clinicians, and the cognitive load imposed by serial reassessments of continuously accumulating data is substantial. Potentially valuable information is easily missed and underutilized for risk stratification and clinical decision-making, as it often requires computational capacity beyond human ability [1,2]. Instead, humans tend to rely on heuristics, or cognitive shortcuts, which can lead to bias, error, and preventable harm [6,19]. By contrast, large volume electronic health record (EHR) data are well suited to dynamic predictive analytics that capture trends over time; physiologic time series data have been used to predict mortality and specific conditions such as acute kidney injury [20–23].

Algorithm dynamicity is especially important when modeling conditions that change rapidly. For instance, intracranial and cerebral perfusion pressure can vacillate substantially after traumatic brain injury. Delayed recognition of rapid changes in intracranial and cerebral perfusion pressure can worsen outcomes because brain ischemia is exquisitely time sensitive. Classical traumatic brain injury prediction models only used static variables present on admission [24–26]. These models may be useful for research purposes, early prognostication, and early resource use decisions, but they do not perform the critically important function of updating predictions as new data become available. For example, an algorithm using 5-minute median values of intracranial pressure, mean arterial pressure, cerebral perfusion pressure, and Glasgow coma scale scores predicts 30-day mortality with approximately 84% discrimination 48 hours after admission [25]. Using 5-minute median values rather than continuous data streams may be favorable for implementation in clinical settings, where data collection is frequently interrupted.

Dynamic algorithms face challenges in evaluating performance over time and explainability. In some cases, algorithms learn to predict which action a clinician will take next, rather than physiologic events [27,28]. In addition, there are no standards for evaluating model performance when predictions are made in a continuous or nearly continuous fashion. We suggest evaluating standard model performance metrics at several predetermined, discrete time points, including the point at which enough information has become available that calibration is expected to plateau, achieving continuous monitoring of predictive performance. To optimize explainability for dynamic algorithms, attention mechanisms can reveal periods during which certain features make significant contributions to algorithm outputs [20,29]. For example, the DeepSOFA algorithm uses time series measurements of the same input variables as the sequential organ failure assessment (SOFA) score, passing those values through a recurrent neural network with gated return and self-attention units. In 2 independent datasets of intensive care unit (ICU) patients, DeepSOFA predicted in-hospital mortality with accuracy greater than that of the traditional SOFA score [20]. Model explainability was promoted by generating heatmaps that illustrate each variable's relative contributions at each time step to the model's ultimate mortality prediction. Using time series measurements in dynamic algorithms relates to the next desideratum of ideal algorithms: precision.

Ideal algorithms are precise

Precise algorithms use data collection rates that are proportional to rates of physiologic changes and machine learning techniques whose complexity matches the target outcome. Precision is important because human diseases are complex and nonlinear [30,31]. Simple, additive models often demonstrate poor predictive performance [32–34]. Three days after colorectal surgery, a serum C-reactive protein level less than 172 mg/L has a 97% negative predictive value for the occurrence of anastomotic leak [34]. This finding may facilitate early discharge home after major surgery. However, high C-reactive protein levels are nonspecific: As a general marker of systemic inflammation, one would expect that C-reactive protein has a poor positive predictive value, and it does (21%). To perform a complex task, such as differentiating between an anastomotic leak and other pro-inflammatory postoperative complications, it is potentially advantageous to incorporate high-resolution, multimodal patient data and machine learning modeling [35–39].

For a given algorithm, the ideal rate of data collection should exceed by several fold the rate of salient physiologic changes, similar to the manner in which Harry Nyquist noted that to represent a signal with fidelity, sampling should occur at twice the highest frequency of the signal [40]. In many disease processes, this will require high-resolution data that are sampled at a frequency that allows for early diagnosis, prevention, or treatment by capturing subtle but clinically significant physiologic changes. Generally, longer intervals are more likely to miss critical physiologic changes that occur between measurements [41–44]. For hospitalized patients, high-frequency assessments are associated with greater accuracy in predicting decompensation. Subtle signs of physiologic instability often occur hours before organ failure and cardiac arrest, representing opportunities for prevention [45,46]. This is discordant with standard practices on hospital wards, where vital signs are typically measured every 4 hours. Unsurprisingly, continuous vital sign monitoring is associated with fewer rescue events, respiratory decompensation events, unplanned ICU transfers, and ICU days, as well as shorter hospital length of stay [47,48]. Yet, continuous monitoring can be expensive, can generate distracting false-positive alarms, might impair patient comfort and mobility, and is not supported by a great deal of level 1 evidence, apart from heart rate characteristics monitoring for neonatal sepsis [49–54]. Therefore, continuous monitoring is often reserved for high-risk patients that are most likely to manifest time-sensitive clinically significant physiologic changes, for whom continuous data have a proven ability to stand alone [55–60] and to add information to EHR data elements [36,59,61–64] in predictive analytics. In designing algorithms, we suggest resampling data at intervals that align with the expected velocity with which changes in physiologic signals lead to clinically significant events, with sampling frequency equal to or greater than the Nyquist rate [40].

In many healthcare settings, highly granular data are routinely recorded from multiple sources for clinical purposes. For example, clinical surveillance of critically ill patients often includes not only vital sign and laboratory measurements but also assessments of mental status, pain, respiratory mechanics, and mobility. Historically, these assessments are performed and recorded by hospital staff in a subjective fashion. With improvements in sensor technologies and machine learning applications in healthcare, it has become feasible to automatically capture and analyze data from ICU patients and environments tracked by accelerometers, light sensors, sound sensors, and high-resolution cameras [65]. Wearable sensors can also capture meaningful, multimodal physiologic data from community-dwelling participants. Notably, high-resolution, multimodal data often suffer from high dimensionality, rendering simple algorithms inaccurate.

Conversely, when algorithms have too many inputs relative to their application, generalizability is compromised due to overfitting. The optimal approach balances predictive accuracy

and input complexity by using the fewest variables necessary to maintain high performance. This can be accomplished with sparse regression methods [66]. Generating parsimonious models, although harboring the potential to compromise predictive performance, has the additional advantage of improving the descriptive accuracy for input features, as described above in the "Ideal algorithms are explainable" section.

Ideal algorithms are autonomous

Autonomous algorithms execute with minimal human input. Beyond the training and testing autonomy shared by all unsupervised machine learning algorithms, autonomous algorithms in healthcare can be implemented with minimal input by users. Manual data entry by the end user imposes time constraints that hinder the clinical application of nonautonomous decision support algorithms [67]. For dynamic models that capture temporal changes by frequently resampling high-resolution data, the cost of manual data entry is even greater. Fortunately, the widespread availability of high-volume EHR data and open-source machine learning code promotes algorithm autonomy [68,69].

Autonomous algorithms have substantial potential to augment decision-making for clinical scenarios in which many input features have complex associations with outcomes. Predicting risk for complications after surgery is one such instance. Accurate predictions of postoperative complications can influence patients' decisions whether to undergo surgery, identify risk factors that are amenable to risk reduction strategies, and inform decisions regarding appropriate postoperative triage destination and resource use. Regrettably, clinicians demonstrate variable performance in predicting risk for postoperative complications, and surgeons frequently commit judgment errors that confer preventable harm [70–72]. Several accurate predictive analytic decision support algorithms have been developed and validated to augment clinical risk predictions, but most are hindered by time-consuming manual data entry requirements and lack of integration with clinical workflow [67,73–77]. Yet, autonomous prediction of postoperative complications is possible. One machine learning platform autonomously imports EHR input data to predict 8 postoperative complications with area under the receiver operating characteristic curve (AUC) 0.82 to 0.94, exhibiting accuracy greater than that of physicians [69,70].

Potential advantages of autonomy also apply to algorithm training. Supervised machine learning algorithms use training data that are labeled by humans and then classifies or makes predictions on new, unseen data; in unsupervised learning, algorithms generate their own labels according to the structure and distribution of input data, discovering patterns and associations. Deep learning models avoid time-intensive, handcrafted feature engineering by autonomously learning feature representations from raw data. In addition to efficiency and pragmatism, autonomous learning offers performance advantages, as has been demonstrated in the gaming industry. "Go" has 32,490 possible first moves, precluding an exhaustive search of all possible moves for each board configuration. Instead, a combination of deep and reinforcement learning can predict outcomes following sequences of actions and efficiently identify optimal moves. This approach was initially applied in learning 30 million positions and instructions from a human Go expert, allowing the algorithm to build a decision policy network. The program then played against itself, attempting to maximize the chance of beating previous versions of its own decision-making policy. Next, a value network predicted the final outcome of a game based on any board configuration. Finally, the policy and value networks were combined, and an optimized search algorithm was used to select the next move for any board configuration. This approach defeated the European Go champion 5 games to 0 [78]. Subsequently, a completely autonomous model was trained exclusively on self-play. This model defeated the human input model 100 games to 0 [79]. For healthcare applications, it

remains plausible that performance is greatest for completely autonomous learning approaches for instances in which high-quality training data exist. Unfortunately, most healthcare data sources are compromised by a lack of granularity, generalizability, volume, or a combination thereof.

Ideal algorithms are fair

Fair algorithms evaluate and mitigate implicit bias and social inequity. In theory, algorithms use mathematical formulas and functions to produce objective outputs, offering a bulwark against subjectivity with resultant bias and inequity. In practice, many algorithms are trained on biased source data and produce biased outputs [80]. In healthcare, single-center source data may disproportionately represent certain demographics. When these data are used for algorithm training, that algorithm may perform poorly when applied to a patient that is sparsely represented in the source data. Poor performance may be especially harmful when it has directionality, i.e., the algorithm consistently overestimates or underestimates risk in a manner that affects decision-making. For example, if a decision support tool incorporates the observation that Black patients have increased risk for mortality after coronary artery bypass, then model outputs could decrease the likelihood that Black patients will garner the benefits of an indicated procedure [81,82]. To determine whether a demographic or socioeconomic factor should be included in a prediction model, it is necessary to assess whether that factor has a plausible or proven pathophysiologic association with the outcome of interest. To do so, we recommend machine learning explainability mechanisms, causal inference, and clinical interpretation of biologic plausibility. If this analysis reveals no evidence of a pathophysiologic association, then it is likely that the demographic or socioeconomic factor is an indicator of suboptimal access to care, referral patterns, or systemic bias and should be excluded from the algorithm.

Algorithm bias can be evaluated by assessing calibration across demographic and socioeconomic variables. If an observed outcome matches algorithm-predicted probabilities for men but not women, then the algorithm exhibits bias against women. This method was used to evaluate racial bias in an algorithm that predicts healthcare needs [83]. The authors compared observed versus predicted healthcare needs for primary care patients who self-identified as Black versus White. When comparing Black and White patients with similar predicted risk, Black patients had greater illness severity. The algorithm was designed to identify patients at or above the 97th percentile of risk and allocate them to receive extra care. At the 97th percentile, Black patients had 4.8 chronic illnesses, and White patients had 3.8 chronic illnesses ($p < 0.001$). The likely mechanism for this discrepancy was the use of healthcare expenditures as a proxy for health needs. If less money is spent on Black patients than on White patients who have the same illness severity, then the algorithm will errantly learn that Black patients have lesser health needs than White patients who have the same illness severity. Racial discrepancies were eliminated by modifying the algorithm so that expenditures were not a proxy for health needs.

Several other methods for promoting algorithm fairness have been described [4]. Models should be reevaluated over time to determine whether temporal changes in study populations, healthcare systems, and medical practices have affected relationships between features and outcomes. This phenomenon, concept drift, undermines algorithm performance by several mechanisms, including algorithm bias. During preprocessing, individual patient data can be mapped to probability distributions that obfuscates information about membership in a protected subgroup (e.g., race, ethnicity, sex, gender, etc.) while retaining as much other information about the patient as possible [84]. During postprocessing, the open-source What-If Tool

allows interactive model testing under user-controlled hypothetical circumstances, which can quantify the effects of different demographic and socioeconomic factors on model outputs [85]. In addition, the What-If Tool can demonstrate whether model performance varies across subgroups, which may be useful in determining whether the model should be applied for a patient that is poorly represented in model training data.

Ideal algorithms are reproducible

Reproducible algorithms are validated both externally and prospectively and are shared with academic communities. In a survey distributed by *Nature*, greater than 70% of all researchers had attempted and failed to reproduce another scientist's experiments, and 90% reported that science is facing a reproducibility crisis [86]. Reproducibility, a critically important element of any scientific inquiry, is especially important for machine learning algorithms because it establishes trustworthiness and credibility. Prior to successful clinical implementation, "black box" algorithms must earn the trust of patients, clinicians, and investigators. Even when explainability is suboptimal, people may be willing to use an algorithm that is well validated and freely available to academicians. In addition, a reproducible algorithm can be tuned and optimized over time, offering a performance advantage.

There are several major barriers to algorithm reproducibility. Prominent EHR platforms are not designed to accommodate algorithm scalability across institutions and platforms. This produces an "analytic bottleneck" in which investigators must process, harmonize, and validate massive amounts of data within institutional silos. Many researchers do not possess the necessary resources to work at such a large computational scale, much less keep track of which data were used for different studies and evaluate the impact of data reuse on the statistical bias. In addition, there are limited cloud resources for sharing multiple, large, healthcare data repositories among research groups that have their own algorithm pipelines and tools. Given these obstacles, many algorithms are never shared and validated externally. To ensure that algorithms are suitable for external validation, results from interventions using AI algorithms should be reported in a standardized fashion, as proposed by the SPIRIT-AI extension, which was developed in parallel with CONSORT-AI guidelines [11,12]. Compliance with these protocols will promote the reproducibility of findings. Yet, most reports involving algorithms in healthcare do not involve implementation in a clinical trial. Noninterventional studies that involve prediction models should comply with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement [87,88]. Finally, generalizability can be enhanced by using input features that are collected routinely in clinical care, excluding features whose collection requires specialized measurement tools that are unavailable in most settings.

Federated learning offers opportunities to ensure the external validity, generalizability, and reproducibility of algorithms via collaborative machine learning without data sharing [89–92]. When sensitive patient data are shared between institutions, there is risk for unintended data disclosures and piracy by adversarial third parties. In federated learning, local models are trained separately and consolidated into a global model [89–92]. As local models train, they send local updates in the form of gradients or coefficients for incorporation in the global model. Even when these relatively secure methods are applied, privacy leakage can occur when adversaries infer whether a given attribute belongs to the model's training data or infer class representatives from collaborative models [93–97]. To mitigate privacy leakage in federated learning, the risks for privacy-sensitive information and privacy leakage can be quantified for each data record, with subsequent obfuscation of high-risk records.

Table 2. Highly cited AI algorithms graded by their interpretability, dynamicity, precision, autonomy, fairness, and reproducibility.

Primary author	Algorithm application	Explainable	Dynamic	Precise	Autonomous	Fair	Reproducible
Gulshan	Detecting diabetic retinopathy	No	N/A	Yes	Yes	Yes	No
Iorio	Predicting tumor sensitivity to pharmacotherapies	Yes	N/A	Yes	No	Yes	No
Kamnitsas	Brain lesion segmentation	Yes	N/A	Yes	Yes	Yes	Yes
Ott	Predicting human lymphocyte antigen binding	No	N/A	Yes	Yes	Yes	Yes
Savova	Extracting information from clinical free text in EHRs	No	N/A	Yes	Yes	No	Yes
Tajbakhsh	Medical image classification, detection, and segmentation	No	N/A	Yes	Yes	Yes	No
Wolfe	Identifying and assessing severity of fibromyalgia	Yes	N/A	Yes	No	No	No
Xiong	Predicting splicing regulation for mRNA sequences	Yes	N/A	Yes	Yes	No	No

AI, artificial intelligence; EHR, electronic health record; N/A, not applicable (i.e., temporal changes and continuous monitoring were not applicable to these algorithms and their intended use).

https://doi.org/10.1371/journal.pdig.0000006.t002

Application of the ideal algorithms framework to prominent algorithms in healthcare

To identify prominent examples of algorithms in healthcare, we reviewed the 20 most highly cited articles in medical AI, as identified in a bibliometric analysis by Nadri and colleagues [98]. Among these 20 articles, 8 described an algorithm. **Table 2** applies the ideal algorithm framework to these 8 algorithms by the majority vote of 3 independent raters. Fleiss kappa statistic was 0.708, suggesting substantial interrater agreement [99,100]. All 8 algorithms met criteria for precision, 6 of the algorithms were autonomous, 5 were fair, 4 were explainable, and 3 were reproducible. Dynamicity was not applicable to any of the algorithms. These findings suggest opportunities to enhance the autonomy, fairness, explainability, and reproducibility of algorithms in healthcare.

Conclusions

While the breadth and complexity of human disease compromise the efficacy of hypothetical-deductive reasoning and heuristic decision-making, high-complexity and high-volume data can be parsed by machine learning applications with relative ease. Established guidelines describe minimum requirements for reporting algorithm healthcare applications; it is equally important to describe the maximum potential of ideal algorithms. We propose that ideal algorithms have 6 desiderata that are represented in a checklist presented herein: explainable (convey the relative importance of features in determining outputs), dynamic (capture temporal changes in physiologic signals and clinical events), precise (use high-resolution, multimodal data and aptly complex architecture), autonomous (learn with minimal supervision and execute without human input), fair (evaluate and mitigate implicit bias and social inequity), and reproducible (are validated externally and prospectively and shared with academic communities). By achieving these objectives, healthcare algorithms confer maximum potential benefits to patients, clinicians, and investigators.

References

1. Jakimowicz JJ, Cuschieri A. Time for evidence-based minimal access surgery training—simulate or sink. Surg Endosc. 2005; 19(12):1521–2. Epub 2005/10/26. https://doi.org/10.1007/s00464-005-0441-x PMID: 16247572.
2. Bitterman N. Technologies and solutions for data display in the operating room. J Clin Monit Comput. 2006; 20(3):165–73. Epub 2006/05/16. https://doi.org/10.1007/s10877-006-9017-0 PMID: 16699740.

3. Ahmed MN, Toor AS, O'Neil K, Friedland D. Cognitive Computing and the Future of Health Care Cognitive Computing and the Future of Healthcare: The Cognitive Power of IBM Watson Has the Potential to Transform Global Personalized Medicine. IEEE Pulse. 2017; 8(3):4–9. Epub 2017/05/24. https://doi.org/10.1109/MPUL.2017.2678098 PMID: 28534755.

4. O'Reilly-Shah VN, Gentry KR, Walters AM, Zivot J, Anderson CT, Tighe PJ. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. Br J Anaesth. 2020. Epub 2020/08/26. https://doi.org/10.1016/j.bja.2020.07.040 PMID: 32838979; PubMed Central PMCID: PMC7442146.

5. Blumenthal-Barby JS, Krieger H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. Med Decis Making. 2015; 35(4):539–57. https://doi.org/10.1177/0272989X14547740 PMID: 25145577.

6. Wolf FM, Gruppen LD, Billi JE. Differential diagnosis and the competing-hypotheses heuristic. A practical approach to judgment under uncertainty and Bayesian probability. JAMA. 1985; 253(19):2858–62. PMID: 3989960

7. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med. 2005; 165 (13):1493–9. https://doi.org/10.1001/archinte.165.13.1493 PMID: 16009864.

8. Kirch W, Schafii C. Misdiagnosis at a university hospital in 4 medical eras. Medicine (Baltimore). 1996; 75(1):29–40. https://doi.org/10.1097/00005792-199601000-00004 PMID: 8569468.

9. Liu XX, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, 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(6):E271–E97. https://doi.org/10.1016/S2589-7500(19)30123-2 WOS:000525871300011. PMID: 33323251

10. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020; 368. https://doi.org/10.1136/bmj.m689 WOS:000523764500002. PMID: 32213531

11. Rivera SC, Liu XX, Chan AW, Denniston AK, Calvert MJ, Grp S-AC-AW. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020; 2(10):E549–E60. https://doi.org/10.1016/S2589-7500(20)30219-3 WOS:000581145100012. PMID: 33328049

12. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Spirit AI, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020; 26(9):1364–74. Epub 2020/09/11. https://doi.org/10.1038/s41591-020-1034-x PMID: 32908283; PubMed Central PMCID: PMC7598943.

13. Tcheng JE. Optimizing strategies for clinical decision support: summary of a meeting series. National Academy of Medicine; 2017.

14. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. JAMA. 2018; 320(21):2199–200. Epub 2018/11/07. https://doi.org/10.1001/jama.2018.17163 PMID: 30398550.

15. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020; 26 (9):1320–4. Epub 2020/09/11. https://doi.org/10.1038/s41591-020-1041-y PMID: 32908275; PubMed Central PMCID: PMC7538196.

16. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A. 2019; 116(44):22071–80. https://doi.org/10.1073/pnas.1900654116 PMID: 31619572; PubMed Central PMCID: PMC6825274.

17. Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. JAMA Cardiol. 2021; 6(6):633–41. Epub 2021/03/11. https://doi.org/10.1001/jamacardio.2021.0122 PMID: 33688915; PubMed Central PMCID: PMC7948114.

18. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019; 1(5):206–15.

19. Bekker HL. Making choices without deliberating. Science. 2006; 312(5779):1472; author reply https://doi.org/10.1126/science.312.5779.1472a PMID: 16763132.

20. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Sci Rep. 2019; 9 (1):1879. https://doi.org/10.1038/s41598-019-38491-0 PMID: 30755689.

21. Kim SY, Kim S, Cho J, Kim YS, Sol IS, Sung Y, et al. A deep learning model for real-time mortality prediction in critically ill children. Crit Care. 2019; 23(1):279. Epub 2019/08/16. https://doi.org/10.1186/s13054-019-2561-z PMID: 31412949; PubMed Central PMCID: PMC6694497.

22. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sundermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med. 2018; 6(12):905–14. Epub 2018/10/03. https://doi.org/10.1016/S2213-2600(18)30300-X PMID: 30274956.

23. Tomasev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019; 572(7767):116–9. Epub 2019/08/02. https://doi.org/10.1038/s41586-019-1390-1 PMID: 31367026; PubMed Central PMCID: PMC6722431.

24. MRC CRASH Trial Collaborators, Perel P, Arango M, Clayton T, Edwards P, Komolafe E, et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. BMJ. 2008; 336(7641):425–9. Epub 2008/02/14. https://doi.org/10.1136/bmj.39461.643438.25 PMID: 18270239; PubMed Central PMCID: PMC2249681.

25. Raj R, Siironen J, Skrifvars MB, Hernesniemi J, Kivisaari R. Predicting outcome in traumatic brain injury: development of a novel computerized tomography classification system (Helsinki computerized tomography score). Neurosurgery. 2014; 75(6):632–46; discussion 46–7. Epub 2014/09/03. https://doi.org/10.1227/NEU.0000000000000533 PMID: 25181434.

26. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med. 2008; 5(8):e165; discussion e. Epub 2008/08/08. https://doi.org/10.1371/journal.pmed.0050165 PMID: 18684008; PubMed Central PMCID: PMC2494563.

27. Beaulieu-Jones BK, Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med. 2021; 4(1):62. Epub 2021/04/01. https://doi.org/10.1038/s41746-021-00426-3 PMID: 33785839; PubMed Central PMCID: PMC8010071 fees from Salutary Inc. A.L.B. received consulting fees and stock options from Generate Biomedicines. The remaining authors declare that there are no competing interests.

28. Davis JP, Wessells DA, Moorman JR. Coronavirus Disease 2019 Calls for Predictive Analytics Monitoring-A New Kind of Illness Scoring System. Crit Care Explor. 2020; 2(12):e0294. Epub 2020/12/29. https://doi.org/10.1097/CCE.0000000000000294 PMID: 33364604; PubMed Central PMCID: PMC7752690.

29. Shamout FE, Zhu T, Sharma P, Watkinson PJ, Clifton DA. Deep Interpretable Early Warning System for the Detection of Clinical Deterioration. IEEE J Biomed Health Inform. 2019. Epub 2019/09/24. https://doi.org/10.1109/JBHI.2019.2937803 PMID: 31545746.

30. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine. Where do we stand? N Engl J Med. 1987; 316(11):685–8. https://doi.org/10.1056/NEJM198703123161109 PMID: 3821801.

31. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. Ann Surg. 2018; 268(1):70–6. https://doi.org/10.1097/SLA.0000000000002693 PMID: 29389679; PubMed Central PMCID: PMC5995666.

32. Dybowski R, Weller P, Chang R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet. 1996; 347(9009):1146–50. https://doi.org/10.1016/s0140-6736(96)90609-1 PMID: 8609749

33. Kim S, Kim W, Park RW. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. Healthc Inform Res. 2011; 17(4):232–43. https://doi.org/10.4258/hir.2011.17.4.232 PMID: 22259725; PubMed Central PMCID: PMC3259558.

34. Singh PP, Zeng IS, Srinivasa S, Lemanu DP, Connolly AB, Hill AG. Systematic review and meta-analysis of use of serum C-reactive protein levels to predict anastomotic leak after colorectal surgery. Br J Surg. 2014; 101(4):339–46. https://doi.org/10.1002/bjs.9354 PMID: 24311257.

35. Bagnall NM, Pring ET, Malietzis G, Athanasiou T, Faiz OD, Kennedy RH, et al. Perioperative risk prediction in the era of enhanced recovery: a comparison of POSSUM, ACPGBI, and E-PASS scoring systems in major surgical procedures of the colorectal surgeon. Int J Colorectal Dis. 2018; 33(11):1627–34. https://doi.org/10.1007/s00384-018-3141-4 PMID: 30078107.

36. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015; 7(299):299ra122. https://doi.org/10.1126/scitranslmed.aab3719 PMID: 26246167.

37. Koyner JL, Carey KA, Edelson DP, Churpek MM. The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model. Crit Care Med. 2018; 46(7):1070–7. https://doi.org/10.1097/CCM.0000000000003123 PMID: 29596073.

38. Delahanty RJ, Kaufman D, Jones SS. Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients. Crit Care Med. 2018; 46(6):e481–e8. https://doi.org/10.1097/CCM.0000000000003011 PMID: 29419557.

39. Adhikari L, Ozrazgat-Baslanti T, Ruppert M, Madushani R, Paliwal S, Hashemighouchani H, et al. Improved predictive models for acute kidney injury with IDEA: Intraoperative Data Embedded Analytics. PLoS ONE. 2019; 14(4):e0214904. https://doi.org/10.1371/journal.pone.0214904 PMID: 30947282; PubMed Central PMCID: PMC6448850.

40. Oppenheim AV, Schafer RW, Buck JR. Discrete-time signal processing. 2nd ed. Upper Saddle River, N.J.: Prentice Hall; 1999.

41. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med. 1981; 9 (8):591–7. https://doi.org/10.1097/00003246-198108000-00008 PMID: 7261642.

42. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. 1996; 22(7):707–10. Epub 1996/07/01. https://doi.org/10.1007/BF01709751 PMID: 8844239.

43. Zimmerman JE, Kramer AA. A history of outcome prediction in the ICU. Curr Opin Crit Care. 2014; 20 (5):550–6. Epub 2014/08/20. https://doi.org/10.1097/MCC.0000000000000138 PMID: 25137400.

44. Stone DJ, Csete M. Actuating critical care therapeutics. J Crit Care. 2016; 35:90–5. Epub 2016/08/03. https://doi.org/10.1016/j.jcrc.2016.05.002 PMID: 27481741.

45. Franklin C, Mathew J. Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. Crit Care Med. 1994; 22(2):244–7. PMID: 8306682.

46. Berlot G, Pangher A, Petrucci L, Bussani R, Lucangelo U. Anticipating events of in-hospital cardiac arrest. Eur J Emerg Med. 2004; 11(1):24–8. https://doi.org/10.1097/00063110-200402000-00005 PMID: 15167189.

47. Taenzer AH, Pyke JB, McGrath SP, Blike GT. Impact of pulse oximetry surveillance on rescue events and intensive care unit transfers: a before-and-after concurrence study. Anesthesiology. 2010; 112 (2):282–7. https://doi.org/10.1097/ALN.0b013e3181ca7a9b PMID: 20098128.

48. Brown H, Terrence J, Vasquez P, Bates DW, Zimlichman E. Continuous monitoring in an inpatient medical-surgical unit: a controlled clinical trial. Am J Med. 2014; 127(3):226–32. https://doi.org/10.1016/j.amjmed.2013.12.004 PMID: 24342543.

49. Slight SP, Franz C, Olugbile M, Brown HV, Bates DW, Zimlichman E. The return on investment of implementing a continuous monitoring system in general medical-surgical units. Crit Care Med. 2014; 42(8):1862–8. https://doi.org/10.1097/CCM.0000000000000340 PMID: 24717454.

50. Prgomet M, Cardona-Morrell M, Nicholson M, Lake R, Long J, Westbrook J, et al. Vital signs monitoring on general wards: clinical staff perceptions of current practices and the planned introduction of continuous monitoring technology. Int J Qual Health Care. 2016; 28(4):515–21. https://doi.org/10.1093/intqhc/mzw062 PMID: 27317251.

51. Watkinson PJ, Barber VS, Price JD, Hann A, Tarassenko L, Young JD. A randomised controlled trial of the effect of continuous electronic physiological monitoring on the adverse event rate in high risk medical and surgical patients. Anaesthesia. 2006; 61(11):1031–9. Epub 2006/10/18. https://doi.org/10.1111/j.1365-2044.2006.04818.x PMID: 17042839.

52. Moorman JR, Delos JB, Flower AA, Cao H, Kovatchev BP, Richman JS, et al. Cardiovascular oscillations at the bedside: early diagnosis of neonatal sepsis using heart rate characteristics monitoring. Physiol Meas. 2011; 32(11):1821–32. Epub 2011/10/27. https://doi.org/10.1088/0967-3334/32/11/S08 PMID: 22026974; PubMed Central PMCID: PMC4898648.

53. Stone ML, Tatum PM, Weitkamp JH, Mukherjee AB, Attridge J, McGahren ED, et al. Abnormal heart rate characteristics before clinical diagnosis of necrotizing enterocolitis. J Perinatol. 2013; 33(11):847–50. Epub 2013/06/01. https://doi.org/10.1038/jp.2013.63 PMID: 23722974; PubMed Central PMCID: PMC4026091.

54. Schelonka RL, Carlo WA, Bauer CR, Peralta-Carcelen M, Phillips V, Helderman J, et al. Mortality and Neurodevelopmental Outcomes in the Heart Rate Characteristics Monitoring Randomized Controlled Trial. J Pediatr. 2020; 219:48–53. Epub 2020/02/09. https://doi.org/10.1016/j.jpeds.2019.12.066 PMID: 32033793; PubMed Central PMCID: PMC7096280.

55. Griffin MP, Moorman JR. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. Pediatrics. 2001; 107(1):97–104. Epub 2001/01/03. https://doi.org/10.1542/peds.107.1.97 PMID: 11134441.

56. Saria S, Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. Sci Transl Med. 2010; 2(48):48ra65. Epub 2010/09/10. https://doi.org/10.1126/scitranslmed.3001304 PMID: 20826840; PubMed Central PMCID: PMC3564961.

57. Tarassenko L, Hann A, Young D. Integrated monitoring and analysis for early warning of patient deterioration. Br J Anaesth. 2006; 97(1):64–8. Epub 2006/05/19. https://doi.org/10.1093/bja/ael113 PMID: 16707529.

58. Politano AD, Riccio LM, Lake DE, Rusin CG, Guin LE, Josef CS, et al. Predicting the need for urgent intubation in a surgical/trauma intensive care unit. Surgery. 2013; 154(5):1110–6. Epub 2013/10/01. https://doi.org/10.1016/j.surg.2013.05.025 PMID: 24075272; PubMed Central PMCID: PMC3805718.

59. Moss TJ, Lake DE, Calland JF, Enfield KB, Delos JB, Fairchild KD, et al. Signatures of Subacute Potentially Catastrophic Illness in the ICU: Model Development and Validation. Crit Care Med. 2016; 44(9):1639–48. Epub 2016/07/28. https://doi.org/10.1097/CCM.0000000000001738 PMID: 27452809; PubMed Central PMCID: PMC4987175.

60. Ruminski CM, Clark MT, Lake DE, Kitzmiller RR, Keim-Malpass J, Robertson MP, et al. Impact of predictive analytics based on continuous cardiorespiratory monitoring in a surgical and trauma intensive care unit. J Clin Monit Comput. 2019; 33(4):703–11. Epub 2018/08/20. https://doi.org/10.1007/s10877-018-0194-4 PMID: 30121744.

61. Griffin MP, Lake DE, Moorman JR. Heart rate characteristics and laboratory tests in neonatal sepsis. Pediatrics. 2005; 115(4):937–41. Epub 2005/04/05. https://doi.org/10.1542/peds.2004-1393 PMID: 15805367.

62. Griffin MP, Lake DE, O'Shea TM, Moorman JR. Heart rate characteristics and clinical signs in neonatal sepsis. Pediatr Res. 2007; 61(2):222–7. Epub 2007/01/24. https://doi.org/10.1203/01.pdr.0000252438.65759.af PMID: 17237726.

63. De Pasquale M, Moss TJ, Cerutti S, Calland JF, Lake DE, Moorman JR, et al. Hemorrhage Prediction Models in Surgical Intensive Care: Bedside Monitoring Data Adds Information to Lab Values. IEEE J Biomed Health Inform. 2017; 21(6):1703–10. Epub 2017/04/20. https://doi.org/10.1109/JBHI.2017.2653849 PMID: 28422699.

64. Moss TJ, Clark MT, Calland JF, Enfield KB, Voss JD, Lake DE, et al. Cardiorespiratory dynamics measured from continuous ECG monitoring improves detection of deterioration in acute care patients: A retrospective cohort study. PLoS ONE. 2017; 12(8):e0181448. Epub 2017/08/05. https://doi.org/10.1371/journal.pone.0181448 PMID: 28771487; PubMed Central PMCID: PMC5542430.

65. Davoudi A, Malhotra KR, Shickel B, Siegel S, Williams S, Ruppert M, et al. Intelligent ICU for Autonomous Patient Monitoring Using Pervasive Sensing and Deep Learning. Sci Rep. 2019; 9. https://doi.org/10.1038/s41598-019-44004-w WOS:000469318500008. PMID: 31142754

66. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc Natl Acad Sci U S A. 2016; 113(15):3932–7. Epub 2016/04/02. https://doi.org/10.1073/pnas.1517384113 PMID: 27035946; PubMed Central PMCID: PMC4839439.

67. Leeds IL, Rosenblum AJ, Wise PE, Watkins AC, Goldblatt MI, Haut ER, et al. Eye of the beholder: Risk calculators and barriers to adoption in surgical trainees. Surgery. 2018; 164(5):1117–23. https://doi.org/10.1016/j.surg.2018.07.002 PMID: 30149939.

68. Stanford Medicine 2017 Health Trends Report: Harnessing the Power of Data in Health. Accessed 23 Feb 2019. Available at: http://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf.

69. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, et al. MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery. Ann Surg. 2018; 269(4):652–62. https://doi.org/10.1097/SLA.0000000000002706 PMID: 29489489.

70. Brennan M, Puri S, Ozrazgat-Baslanti T, Feng Z, Ruppert M, Hashemighouchani H, et al. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study. Surgery. 2019; 165(5):1035–45. https://doi.org/10.1016/j.surg.2019.01.002 PMID: 30792011.

71. Healey MA, Shackford SR, Osler TM, Rogers FB, Burns E. Complications in surgical patients. Arch Surg. 2002; 137(5):611–7; discussion 7–8. https://doi.org/10.1001/archsurg.137.5.611 PMID: 11982478.

72. Shanafelt TD, Balch CM, Bechamps G, Russell T, Dyrbye L, Satele D, et al. Burnout and medical errors among American surgeons. Ann Surg. 2010; 251(6):995–1000. https://doi.org/10.1097/SLA.0b013e3181bfdab3 PMID: 19934755.

73. Raymond BL, Wanderer JP, Hawkins AT, Geiger TM, Ehrenfeld JM, Stokes JW, et al. Use of the American College of Surgeons National Surgical Quality Improvement Program Surgical Risk Calculator During Preoperative Risk Discussion: The Patient Perspective. Anesth Analg. 2019; 128(4):643–50. Epub 2018/09/01. https://doi.org/10.1213/ANE.0000000000003718 PMID: 30169413.

74. Clark DE, Fitzgerald TL, Dibbins AW. Procedure-based postoperative risk prediction using NSQIP data. J Surg Res. 2018; 221:322–7. https://doi.org/10.1016/j.jss.2017.09.003 PMID: 29229146.

75. Lubitz AL, Chan E, Zarif D, Ross H, Philp M, Goldberg AJ, et al. American College of Surgeons NSQIP Risk Calculator Accuracy for Emergent and Elective Colorectal Operations. J Am Coll Surg. 2017; 225 (5):601–11. https://doi.org/10.1016/j.jamcollsurg.2017.07.1069 PMID: 28826803.

76. Cohen ME, Liu Y, Ko CY, Hall BL. An Examination of American College of Surgeons NSQIP Surgical Risk Calculator Accuracy. J Am Coll Surg. 2017; 224(5):787–95 e1. https://doi.org/10.1016/j.jamcollsurg.2016.12.057 PMID: 28389191.

77. Hyde LZ, Valizadeh N, Al-Mazrou AM, Kiran RP. ACS-NSQIP risk calculator predicts cohort but not individual risk of complication following colorectal resection. Am J Surg. 2019; 218(1):131–5. https://doi.org/10.1016/j.amjsurg.2018.11.017 PMID: 30522696.

78. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016; 529(7587):484–9. https://doi.org/10.1038/nature16961 PMID: 26819042.

79. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. 2017; 550(7676):354–9. Epub 2017/10/21. https://doi.org/10.1038/nature24270 PMID: 29052630.

80. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. ProPublica. 2016 May 23:Accessed 24 Jan 2019 [https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing].

81. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC Jr., et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. Ann Thorac Surg. 2018; 105(5):1411–8. Epub 2018/03/27. https://doi.org/10.1016/j.athoracsur.2018.03.002 PMID: 29577925.

82. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms. New Engl J Med. 2020; 383(9):873–81. https://doi.org/10.1056/NEJMms2004740 WOS:000563821400017. PMID: 32853499

83. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019; 366(6464):447–+. https://doi.org/10.1126/science.aax2342 WOS:000493177900040. PMID: 31649194

84. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning Fair Representations. In: Sanjoy D, David M, editors. Proceedings of the 30th International Conference on Machine Learning; Proceedings of Machine Learning Research: PMLR; 2013. p. 325–33.

85. Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viegas F, Wilson J. The What-If Tool: Interactive Probing of Machine Learning Models. IEEE Trans Vis Comput Graph. 2020; 26(1):56–65. https://doi.org/10.1109/TVCG.2019.2934619 WOS:000506166100006. PMID: 31442996

86. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016; 533(7604):452–4. Epub 2016/05/27. https://doi.org/10.1038/533452a PMID: 27225100.

87. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015; 162(1):55–63. Epub 2015/01/07. https://doi.org/10.7326/M14-0697 PMID: 25560714.

88. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015; 162(1):W1–73. Epub 2015/01/07. https://doi.org/10.7326/M14-0698 PMID: 25560730.

89. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020; 3(1):119. https://doi.org/10.1038/s41746-020-00323-1 PMID: 34518641

90. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards Federated Learning at Scale: System Design. arXiv preprint arXiv:190201046 [Internet]. 2020.

91. Yang Q, Liu Y, Chen T, Tong Y. Federated Machine Learning: Concept and Applications. ACM Trans Intell Syst Technol. 2019; 10(2):Article 12. https://doi.org/10.1145/3298981

92. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAy. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Aarti S, Jerry Z, editors. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics; Proceedings of Machine Learning Research: PMLR; 2017. p. 1273–82.

93. Melis L, Song C, Cristofaro ED, Shmatikov V, editors. Exploiting Unintended Feature Leakage in Collaborative Learning. 2019 IEEE Symposium on Security and Privacy (SP); 2019 19–23 May 2019.

94. Nasr M, Shokri R, Houmansadr A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. 2019. p. 739–53.

95. Wei W, Liu L, Loper M, Chow K, Gursoy M, Truex S, et al. A Framework for Evaluating Gradient Leakage Attacks in Federated Learning. arXiv preprint arXiv:200410397 [Internet]. 2020.

96. Hitaj B, Ateniese G, Perez-Cruz F. Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security; Dallas, Texas, USA: Association for Computing Machinery; 2017. p. 603–18.

97. Wang Z, Song M, Zhang Z, Song Y, Wang Q, Qi H, editors. Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning. IEEE Conference on Computer Communications INFOCOM; 2019 29 April-2 May 2019.

98. Nadri H, Rahimi B, Timpka T, Sedghi S. The Top 100 Articles in the Medical Informatics: a Bibliometric Analysis. J Med Syst. 2017; 41(10):150. Epub 2017/08/22. https://doi.org/10.1007/s10916-017-0794-4 PMID: 28825158.

99. Fleiss JL. Measuring Nominal Scale Agreement among Many Raters. Psychol Bull. 1971; 76(5):378. https://doi.org/10.1037/h0031619 WOS:A1971K852700006.

100. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–74. Epub 1977/03/01. PMID: 843571.