CellPress

Med

Review

Machine learning in clinical decision making

Lorenz Adlung,[1,3] Yotam Cohen,[1,3] Uria Mor,[1,3] and Eran Elinav[1,2,*]

SUMMARY

Machine learning is increasingly integrated into clinical practice, with applications ranging from pre-clinical data processing, bedside diagnosis assistance, patient stratification, treatment decision making, and early warning as part of primary and secondary prevention. However, a multitude of technological, medical, and ethical considerations are critical in machine-learning utilization, including the necessity for careful validation of machine-learning-based technologies in real-life contexts, unbiased evaluation of benefits and risks, and avoidance of technological over-dependence and associated loss of clinical, ethical, and social-related decision-making capacities. Other challenges include the need for careful benchmarking and external validations, dissemination of end-user knowledge from computational experts to field users, and responsible code and data sharing, enabling transparent assessment of pipelines. In this review, we highlight key promises and achievements in integration of machine-learning platforms into clinical medicine while highlighting limitations, pitfalls, and challenges toward enhanced integration of learning systems into the medical realm.

INTRODUCTION

Computational support has been gradually integrated into clinical decision making since the 1970s and 1980s, with computer-aided technologies developed to assist in medical diagnosis of conditions such as acute abdominal pain[1,2] or in modeling of mortality in intensive-care units.[3] While integration of such approaches was technologically difficult and clinically limited at the time, rapid improvements were gradually noted in the following decades, facilitated by adaptation of new models including conditional probabilities.[4] The implementation of such knowledge-based systems in hospitals and primary care centers has further expanded in the past few years[4–6] through the use of data-driven machine learning (ML)-based support systems and medical devices approved by the Food and Drugs Administration (FDA).[7] Numerous studies are increasingly incorporating ML algorithms across a large variety of domains.[8–10]

As ML decision-support systems are gradually entering the clinics, it is important to carefully evaluate their utility and performance while realizing their limitations. Importantly, the know-how of these platforms should be expanded from computational biologists to the clinical end users, who should be informed and made aware of the trade-offs between these factors[11] in determining the performance and accuracy of ML pipelines in various clinical contexts. ML can improve clinical decision making in multiple ways by providing early warning, facilitating diagnosis, performing wide screenings, individualizing treatment, and assessing patients' responsiveness to therapy.[9,10] To evaluate the contribution of a given ML pipeline, it must be tested with respect to the clinical features of the question faced and current

[1]Immunology Department, Weizmann Institute of Science, Rehovot 7610001, Israel

[2]Cancer-Microbiome Division Deutsches Krebsforschungszentrum (DKFZ), Neuenheimer Feld 280, Heidelberg 69120, Germany

[3]These authors contributed equally

*Correspondence: eran.elinav@weizmann.ac.il

https://doi.org/10.1016/j.medj.2021.04.006

gold-standard capacities and in particular clinician knowledge, intuition, and experience.[7,12–14] In some cases, the ML model has been shown to perform at least as well as clinicians, yet many studies provide poor accuracy reports and neither perform "fair" comparisons nor external validations.[7,14] Recently suggested guidelines for ML-associated clinical trials[15,16] attempt to address these challenges by specifying frameworks aimed at achieving better uniformity in ML assessment.

In this review, we provide examples of recent studies involving cutting-edge ML technology aimed at facilitating medical care while evaluating their performance. We highlight the characteristics of widely used types of ML systems, elaborate on general evaluation of ML models and their wide spectrum of uses, and critically discuss the challenges and pitfalls in their application into clinical decision making. We aim to present an overview of ML clinical use for non-computational biologists and clinicians. More in-depth descriptions of theoretical, statistical, and technical considerations of ML development and implementations are presented elsewhere.[17,18]

AN OVERVIEW OF ML SPECIFICATIONS

ML constitutes a field of artificial intelligence (AI) concerned with the question of how to construct computer programs that automatically improve the accuracy of their output with experience. ML algorithms use sample data (also known as a training dataset) for training computational models to generate predictions that will fit the sample data.[19] These predictions can guide clinical decisions in various contexts that may be divided into the following topics: pre-clinical solution (PCS); computer aiding detection system (CADe); computer-aided diagnosis system (CADx); recommendation system (RecSys); and early warning system (EWS) (Figure 1). PCSs are ML systems that assist in creating the infrastructure that is subsequently utilized by downstream analytical tools, such as marker-gene identification or drug discovery. For example, PCS systems may be used for validation of questionnaires prior to their assessment in clinical diagnosis. CADe and CADx refer to software-based analyses used in automated or assisted interpretation of medical imaging, such as X-ray radiography, computed tomography, magnetic resonance imaging, or ultrasonography. CADe is aimed at assisting physicians to interpret medical imaging by identifying suspicious features and bringing them to the clinician's attention, while CADx provides a more-structured diagnosis of a given image. For instance, during a colonoscopy, CADe would highlight suspicious polyps, while a CADx would predict the pathological entity of such polyp (adenoma, carcinoma, or another benign lesion). As indicated by its name, RecSys utilizes one of several user interfaces (e.g., Netflix, Spotify, Youtube, Facebook, among others) to provide health-related and at times individualized, dietary, medical, or pharmacological recommendations for a patient. EWS produces warning information, which enables individuals, communities, and organizations to prepare and proactively respond in advance to a specific risk, thereby providing additional information while saving precious anticipatory time. Examples for EWS include electronic health records (EHRs) screening, intraoperative warnings, and systems utilized in patient triage.

To enable the above applications, ML approaches can be canonically classified into the following major types:

(1) Reinforcement learning. Models are trained to fulfill tasks by means of being rewarded for good operations (i.e., decisions) of the model and being punished for bad actions. An example for reinforcement learning is longitudinal

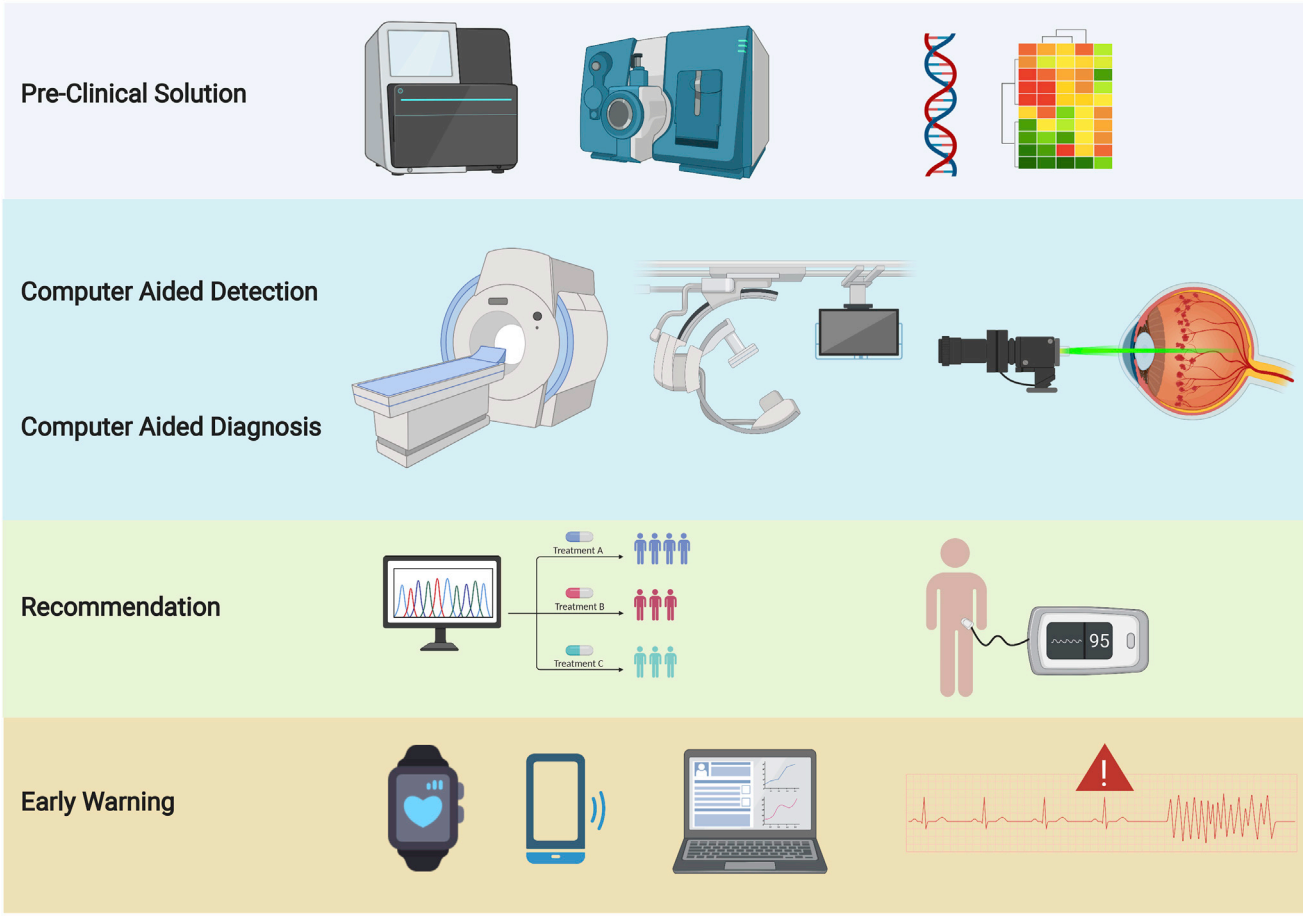

Figure 1. Graphical representation of ML specifications in clinical decision making

From top to bottom: DNA and metabolome sequencing devices for biomarker discovery for pre-clinical solutions; image acquisition techniques for CAD, such as magnet resonance imaging, X-ray, and retinal photography; precision medicine drug recommendation systems and a smart insulin recommendation pump; smart wearables and electronic health record surveillance providing early warning.

evaluations of treatment options for patients, in which the model fulfills the task of suggesting which drug to take, with the reward of a good clinical outcome.

(2) Unsupervised learning. Methods aim to unravel underlying complex structures in the data and enable to reason about these patterns. Models such as principal-component analysis, multidimensional scaling, and autoencoders attempt to find lower dimensional representations of the data that provide insights into associations between the different features in the data, while mixture modeling and cluster analysis aim to identify sub-populations in the data or they group a set of samples enabling an exploratory observation on subsets of samples.

(3) Supervised learning. Labeled data are utilized by the model to learn a mapping between the input data and the outcome variables. This modality is commonly utilized in clinical decision making, in which ML models produce an inferred function that maps labeled input data to outcome variables. Once trained, the inferred function can be utilized to make predictions of the outcome variables given new input data. Supervised learning can be further classified into (1) regression, in which the outcome variable is continuous, and (2) classification, in which the outcome variable takes discrete

categorical values. Examples for commonly used supervised learning algorithms are linear regression (LR), support vector machines (SVMs), decision trees, and artificial neural networks (aNNs). LR and SVMs are models seeking linear relationships between the input data and outcome variables, which are simple and easy to interpret. A decision tree is a model that, during training, recursively splits the feature space into subdomains and associates each subdomain with an outcome value that is based on the outcome values of training data points that remain within it. At evaluation, given an unlabeled sample, the decision tree locates the test cases' relevant subdomain according to the learned partitioning and outputs the value associated with this subdomain. When unrestricted (via hyper-parametrization), decision trees may split the feature space so that each subdomain contains only a single data point and will most often suffer poor performance when presented with examples outside of the training set due to the high sensitivity implied by such overfitting. Conversely, "shallow" decision trees, which allow only for a small number of splits, tend to miss complex structures and subtle relationships in the data. Ensemble approaches (random forests, AdaBoost, and XGBoost) seek to overcome such dilemmas by combining a number of "weak" models to form a diverse and accurate model that generalizes well. Models such as LR, SVMs, and decision trees seek to exploit direct relation between the features of the input data and the outcome variables. For them to perform well, the data must exhibit a "straightforward geometric structure."

In contrast to the aforementioned models, which make predictions based on the original features in the data, aNNs transform the data into a latent feature space, such that only the latent representation of the data features a straightforward relationship with the outcome variable. Then, a simple regression can be successfully applied to the latent representation and yield accurate results. Learning this latent representation is performed by applying to the input data a sequence of linear mappings, where each linear transformation is followed by a non-linear activation function to allow for composite transformations. Each pair of linear and activation functions results in an intermediate representation of the data, usually referred to as a layer. The input layer contains the original features of the data, and layers in between the input and output layers are termed hidden layers. aNNs with a large number of hidden layers are called deep neural networks (DNNs) (hence "deep learning") and are tasked with fitting a very large set of parameters (in some cases millions), and the number of parameters increases with the depth of the network, as well as the dimension of each hidden layer. Fitting parameters in quantities traditionally associated with aNNs requires a training dataset that is large enough to avoid "memorization" of the data, i.e., learning a degenerate "dictionary," which maps every sample from the training dataset to its label instead of learning a meaningful mapping, which makes use of true relationships between features in the data.

To obtain good predictive results, supervised ML tasks require large amounts of qualitatively labeled datasets, which might be impractical due to the high volume required, or they require expansive domain-experts time to create such annotations. Therefore, a data acquisition bottleneck exists in this context. Weakly supervised models tackle this challenge by harnessing noisy, limited, and/or unprecise data annotation to provide an appropriate signal for supervised models. For example, in a task to identify cancerous lesions, in which a given set of histopathology images is solely labeled as negative or positive for cancer, weak supervision methods are applied to learn these specific cancerous lesions using multiple-instance learning.[20]

Table 1. Overview of supervised learning

Supervised learning: fitting a model to approximate known values using the input data	regression: continuous outcome predictions	root-mean squared error ([R]MSE), R^2, maximal error
	binary/multiclass classification: estimating the probability of sample membership in a class	sensitivity, specificity, auROC, f1

Another example for weak supervision is the recent work by Xue et al.,[21] in which a model that was trained for the self-supervised task of future values forecasting of an unlabeled EHR longitudinal dataset was then fine-tuned for the supervised task of estimating time to event for scenarios such as mortality and kidney failure by utilizing a smaller, labeled dataset. This pre-train fine-tune technique is a special instance of the transfer learning methodology, which is an additional solution for the lack of large-scale, highly curated data. Transfer learning utilizes the knowledge of pre-trained, well-validated models as basis for training a new model to either fulfill a different task or extend the current models' capabilities to a new dataset.[22] The underlying assumption is that solutions for tasks such as contour or edge detection of any given domain share fundamental components that are helpful in fulfilling additional tasks, as histopathology image segmentation for cancer detection, and also in assessing severity levels of diabetic retinopathy from fundoscopy images.

Performance evaluation

Quantitative outcome predictions, such as viral load or intracranial pressure, often involve the use of regression models, which can be evaluated using the mean-squared error (MSE), the maximal error of the prediction, or the coefficient of determination R^2 measuring proportion of the variance in the outcome variable that is predictable from the input, etc. (Table 1). In clinical practice, health-related outcome variables are often discretized into class labels to form more clinically convenient classifications. For example, instead of actual viral load estimation, only infected and non-infected individuals are discerned.

Sensitivity and specificity are common measures to evaluate the performance of a given ML system prediction. Sensitivity measures the ability of the classifier to identify samples from the positive class (disease, type of polyp, and treatment) out of all positive samples, simply stating how sensitive the model is in detecting this positive class. Specificity measures the ability of the model to identify samples from the negative class out of all negative samples. The trade-off between these measures is inherent. If an extremely sensitive classifier is required, then its sensitivity will come at the expense of specificity and vice versa. Of note, the terms sensitivity, recall, and true positive rate (TPR) are synonyms and so are specificity and true negative rate (TNR). Precision, or interchangeably positive predictive value (PPV), refers to a measure evaluating how well a model identifies the positive class from all the cases as positive, and negative predictive value (NPV) measures the same for the negative class. In contrast to sensitivity and specificity referring to the identification of true positive and negative cases among the positive and negative cases, PPV and NPV quantify the positive and negative cases among all cases, which depends on the prevalence of cases (or "pre-test probability"), independent of the model performance.

In the case of binary classifiers, it is convenient to assess the models' performance by analysis of receiver operating characteristic (ROC) curve, in which the sensitivity of a model is plotted against 1 − specificity. It is common to evaluate the performance of

a model by the area under the ROC curve (auROC). Each point on the ROC curve represents the sensitivity and specificity of the classifier at a certain decision threshold, resulting in a detailed description of the trade-off between the two measures. To exemplify this concept, one can imagine an ML system predicting, based on some input data, whether an individual would be positive or negative for coronavirus disease 2019 (COVID-19). If we adjust the ML system to exhibit high sensitivity, nearing perfection, many individuals being truly positive will be predicted as such; however, such perfect sensitivity will affect the specificity of the model leading to COVID-19-negative individuals being predicted as positive, resulting in unnecessary quarantining, hospital admission, treatment, and its potential adverse effects and patient anxiety contributing to an enhanced healthcare system burden. Random chance results in a 50% probability of predicting a positive class as positive and a negative class as negative (corresponding to an auROC of 0.5), and thus a diagonal ROC, while a ML model featuring benefits to clinical decision making is expected to perform at a higher sensitivity and specificity. Examples for performance measures of ML systems at different clinical contexts are discussed in the following sections within the particular presented contexts.

PCSs

ML-based PCSs are based on ML models trained on data to identify associations and make predictions. These, in turn, could support disease detection, diagnosis, or patient stratification, or alternatively be integrated into RecSys or EWSs in generating data-driven recommendations based on PCS findings. A plethora of PCSs based on molecular data are being developed in both academia and industry. Since the completion of the Human Genome Project in 2003, technology advances increasingly enable the assessment of molecular entities from human samples on a global scale at steadily decreasing costs.[23] This allows for significant multiplexing, in which multiple genes, sequences, or molecules associated with a disease are assessed in parallel, using ML pipelines, in search for associating patterns and correlating or predicting disease manifestations. In contrast to traditional bottom-up approaches focusing on a single gene or molecule of interest, such global assessment of features with high-throughput technologies is feasible at different layers. Related datasets and approaches are usually referred to with the suffix -omics; e.g., genomics, transcriptomics, metabolomics. These enable ML-based assessment even if the nature and mechanism of individual feature impact on disease course, and their respective cross-interactions remain elusive. From a biological standpoint, integrating multiple features may enable to capture more signals within a signaling network associated with a disease trait, thus optimizing the chances to achieve predictive accuracy. With these exciting prospects notwithstanding, there are data-type-specific constraints to be considered when utilizing omics data as input for ML algorithms, including the high dimensionality of omics data with thousands of features and the high costs to generate such comprehensive datasets. Additionally, omics data are prone to spurious correlations between features or with metadata. To prevent ML models from picking up such random associations, the training dataset must be sufficiently large to reduce the probability of irrelevant correlations between the many features in the dataset and the outcome variable. Even with the reduction in the costs of high-throughput data acquisition methods,[23] generating a sufficiently large training dataset of dozens of thousands of omics samples remains a formidable and often prohibitive challenge.

Utilizing genomic data for ML-based predictions

The human genome has been utilized as a feature for clinical decision making ever since the establishment of karyotyping via Giemsa staining. The detection of chromosomal abnormalities developed into a quantitative method to systematically

detect genome-wide DNA copy number changes.[24] In the past decade, deep sequencing of exomes or even whole genomes is rapidly integrating into clinical practice. Well-curated databases exist with annotations of clinically relevant metadata, such as The Cancer Genome Atlas.[25] These datasets are suitable as inputs training AI algorithms and may even allow the detection of microbial genetic signatures from human samples of the so-called "metagenome" for cancer diagnosis and patient stratification.[26] Such approaches may yield ML models that inform personalized, rational interventions in human pathologies even beyond cancer.[27] In fact, the sheer number of sequences publicly available allows for the classification of homogeneous matching groups of disease versus control cases,[28] even though the diversity with respect to the ethnicity of the people is largely underrepresented in existing datasets.

The transcriptome can be measured from human biopsies at single-cell resolution. PCSs in the field of single-cell RNA sequencing (scRNA-seq) involve the robust identification of individual cell types and cellular states from transcriptome data. Deep learning allows to train and predict the cell-type composition from scRNA-seq datasets.[29,30] In addition, methods exist to infer cellular states, e.g., aging, or responding versus non-responding cells upon treatment.[31,32] Human peripheral blood holds great promise as a potential "liquid biopsy." The serum metabolome determines various diseases and was recently predicted with an ML model based on host genetics, gut microbiome, clinical parameters, diet, lifestyle, and anthropometric measurements.[33] Besides the blood metabolome being a non-genetic, individualized reference map, recent advances in the collection and analysis of blood-cell-free DNA harbor huge potential to enter the clinics in the context of cancer,[34] e.g., for the postoperative evaluation of gastric cancer.[35] Such approaches are suitable to train predictive ML models for disease,[36] which might guide clinical decision making in the future. Genetic information was recently utilized as input for a deep-learning algorithm to predict the transcriptome based on DNA sequence information.[37]

In addition to DNA sequences, epigenetic data can also serve as input for ML models. DNA methylation was utilized by an ML approach to identify differentially methylated loci as biomarkers of human leukocyte types.[38] Single-cell data on the complex chromatin landscape of the human brain have been integrated into an ML framework to predict single-nucleotide polymorphisms associated with Alzheimer's and Parkinson's diseases.[39] The integration of several omics layers (e.g., genomics, transcriptomics, DNA methylome, and histone marks) may provide future guidance to clinical decision making. The microbiome, increasingly considered to constitute a "second genome," may provide another level of functional complexity to be harnessed by ML approaches for prevention, diagnosis, and therapy in a personalized manner. Successful examples, in which microbiome data were utilized by ML models, include cancer,[26] microbiome-based predictions of the human blood metabolome,[33] and microbiome utilization in personalization of nutrition.[40] The latter will be further elaborated as an example of the RecSys.

In the following, we will showcase two examples of clinical contexts, in which ML models have been trained as PCSs.

Screening for drug discovery

In the quest to identify new precision drug interventions, large-scale screening aimed at drug discovery and testing pipelines may be greatly facilitated by utilization of ML.[41] Predictive ML algorithms have been used in identifying promising candidate molecules in drug development or in decoding potential mechanisms

of action of promising compounds, such as binding of the drug to its molecular target (reviewed in Santos et al.[42]). ML can be utilized to discern features indicative of drug-target binding.[43]

One recent ML breakthrough, which may greatly facilitate drug discovery, is an AI pipeline developed by Google, predicting 3D protein structures from amino acid sequences.[44] This impressive development may enable to generate predictive knowledge on active sites within the tertiary structures of proteins, thereby facilitating the design of small-molecule inhibitors. As such, ML algorithms may be projected to increasingly help in reaching a rational drug design, thereby accelerating identification of compounds, while reducing the costs associated with unbiased high-throughput experimental screening.[45] One example for such ML-assisted drug design is a small-molecule inhibitor against the discoidin domain receptor 1 (DDR1), which plays a role in pulmonary and renal fibrosis. With a ML approach utilized for *de novo* drug design, potent DDR1 inhibitors were discovered and validated in only 46 days.[46] First, the mapping of the chemical space was learned by the algorithm and then this space was explored for the discovery of new compounds. A large ensemble of pre-existing molecules as well as known DDR1 inhibitors and similar structures served as input data for the model, which in turn generated 30,000 structures as output, of which 40 representative structures across the chemical space were randomly selected for further analysis. Six of those molecules, which were also not related to any patented structure, were selected for synthesis and experimental validation. Two compounds featured potent inhibition of DDR1, including in a human bone osteosarcoma epithelial cell line. *In vitro* studies of microsomal stability and *in vivo* assessment of the pharmacokinetic properties of these candidate compounds (half-life ~3.5 h) are currently underway in animal models.

SARS-CoV-2 patient stratification

Another recent example for the utilization of ML in PCSs is the establishment of a prediction model to prioritize individuals for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) testing based on a simple questionnaire.[47] During the COVID-19 pandemic, demands for systematic testing of people and hospitalization of patients are often exceeding available resources. ML can help to predict demands for testing and hospitalization based on large public datasets, thereby identifying priorities so that resources are made available for the people who need them the most. This challenge was addressed by generating a model that estimates the probability of an individual to test positive for SARS-CoV-2 by reverse-transcriptase polymerase chain reaction (RT-PCR). As input data, Shoer et al.[47] utilized the answers of 43,752 adults from Israel to nine straight-forward questions on, e.g., age, gender, and self-reported symptoms. With these data, a logistic regression model was trained to predict the outcome of a COVID-19 test, which the authors defined as a self-reported diagnosis confirmed by a laboratory. The model achieved an auROC of 0.737 (confidence interval [CI]: 0.712–0.759). To evaluate the contribution of the different features (i.e., answers to the questionnaire) to the model output (i.e., predicted probability of being diagnosed with COVID-19), a gradient-boosting decision trees algorithm was applied. With this approach, the loss of taste or smell was identified to be the most contributing feature. The model was validated with independent datasets from the US, the UK, and Sweden. With these data, the trained model achieved an auROC of 0.727 (CI: 0.711–0.739). Major limitations of this work include the reliance on the presence of symptoms and thus asymptomatic patients were largely neglected, the exclusion of children, and an inherent selection bias created by the willingness to self-report. Further work is required to overcome existing challenges in the pre-clinical settings and translate these approaches to

accurate bedside predictive and stratification tools. Of note in the context of COVID-19 diagnosis is the application of ML models to computed tomography data.[48,49] Such approaches are CADes, which will be presented in the following section.

CADes

Starting as early as in the 1970s, CADe has become one of the most promising ML-based applications in medical image analysis,[50,51] as well as in sound processing, signal processing, and interpretation of EHRs data.[52] ML CADe could help clinicians by providing feedback in the form of alerts, segmentation results, or interpretations of complex signals.[52] Innovative advancements and immense industry efforts in image analysis using neural networks, along with an exponential increase in data availability, enabled substantial improvement of CADe platforms.[53,54] The existing architecture and platforms enabled neural networks to discover imaging structures that are obscure to the naked human eye, thus detecting hidden layers of clinically important information in medical images. Collectively, numerous applications of image classification arose in multiple domains, with two use cases described below.[8,12,55–58]

CADe in endoscopic imaging

Colorectal cancer is the second leading cause of cancer-related deaths in the US and the fifth leading overall cause of death in China.[59,60] Screening colonoscopy remains the most efficient modality recommended for early diagnosis and is recommended to all individuals older than 50 years.[61,62] However, adenoma miss rate has been a rising concern in endoscopy-based diagnosis of colorectal cancer, despite the high accuracy of detection of colorectal cancer.[63–65] This is of major clinical importance, as adenomas may progress into life-risking colorectal carcinomas. Efforts to further increase the detection rate of adenomas, e.g., by chromo-endoscopy or extra lenses, have been limited in success.[66] ML has been recently offered to constitute an attractive potential solution to this challenge,[67] with numerous randomized control trials attempting to address this challenge, mostly harnessing the ability of neural networks to provide high-quality image analysis.[16,66,68–74] In these studies, CADe patients were randomized into control groups, receiving diagnosis without use of an AI system, or to a treatment group receiving care from an ML computer-aided physician. Adenoma detection rates were reportedly increased in all of these studies, compared to the non-ML group, and performed particularly better at detecting small polyps (<5 mm).

A systematic review[68] encompassing five randomized controlled trials has quantified adenoma detection rate while considering endoscopic withdrawal time in assessing both the advantages and drawbacks of CADe. The reviewed studies collectively included a total of 4,354 colonoscopies with 2,163 cases in the CADe group and 2,191 cases in the control group. An improved adenoma detection rate of 36.6% was observed in the CADe group compared to 25.2% in the control group (relative risk [RR], 1.44; 95% CI, 1.27–1.62; $p < 0.01$), which was mainly attributed to superior ML-associated detection of smaller adenomas (<5 mm). Despite no statistically significant difference observed between the CADe and control groups in terms of mean colonoscopy withdrawal time (representing the duration of the procedure), it was higher on average in the CADe group of all of the reviewed studies compared to the controls. As such, CADe may enable better identifying early onset of colorectal cancer but is associated with an extended procedure duration, potentially resulting in a heavier hospital burden and associated adverse effects. Additionally, the majority of trials were only performed to date at a single country, China, and merit prospective duplication in other localities and ethnic groups.

CADe in radiology imaging

Breast cancer constitutes the second leading cause of death of women.[75,76] Large screening attempts are underway to enable early detection of breast cancer as a primary preventive measure.[77,78] Interpretation of these medical images remains an enormous challenge. In a large-scale survey spanning 205 radiologists, the median scored a sensitivity of 83.8% across >1,000,000 mammograms,[79] while in another,[80,81] 271 radiologists scored a median sensitivity of 87.3%. Given the increased prevalence and associated screening, the workload of radiologists and clinicians needed for this assignment has dramatically increased.[82,83] CADe, without ML algorithms, have been widely used since the first FDA-approved device in 1998, with mixed results noted regarding their effectiveness.[80,81,84]

ML algorithms, and more particularly neural networks, have shown promising results in improving the accuracy of image analysis.[8,14,56,85,86] However, the accelerated availability of ML CADe is often challenged with a need for external validation and sufficient controls.[58,87,88] A recent comparison[89] of three CADe-based mammography interpretation pipelines, applied to a database of ~9,000 women (739 of whom ultimately were diagnosed with breast cancer), suggested that only one of the tested pipelines achieved a sensitivity higher than human manual interpretation. Combining this computerized pipeline with human interpretation led to a marked improvement in diagnostic accuracy over either manual or computed interpretation, reaching 88.6% sensitivity and 93.5% specificity for accurate detection, suggesting integrating computer-assisted diagnostics in human interpretation rather than replacing it. One concern, to be assessed by "real-life" trials, is the risk of creating spurious overdiagnosis, which may result in patient stress and anxiety,[90,91] and an excess of health system workload and associated costs.

CADx

CADe and CADx may be regarded as overlapping to some extent, as both are designed to provide physician support in the diagnosis process. In contrast to CADe, which mainly involves ML-mediated assistance in improving diagnostic accuracy (such as in accurately evaluating medical images), CADx adds a ML-based facet focused on stratifying a suggested follow-up evaluation or treatment modality, based on CADe inputs. In all, ML is aimed at both reducing miss rates and increasing accuracy or treatment success while avoiding unnecessary procedures and their associated adverse effects. Two examples of CADx-based uses are depicted below.

CADx-based pipelines in cancer diagnosis

One example involves a mammography screen for breast cancer, in which a CADe-based approach may help the radiologists to more accurately pinpoint lesions that are suspected as being potentially malignant, while a CADx-helped approach may further help the radiologists to utilize the data in deciding which lesion merits a biopsy versus other, less-invasive follow-up diagnostic approaches.[92] In breast cancer, the second leading cause of death for women,[75,76] CADx based on deep learning of multiparametric magnetic resonance imaging improved the radiologists' performance by reducing the false-positive rate of diagnosis.[93] Besides breast cancer, such evaluation-facilitating ML pipelines have been proposed in the context of lung cancer[94] and gastrointestinal disorders.[95]

CADx-based pipelines in diabetic retinopathy

Diabetic retinopathy (DR) is a common micro-vascular long-term complication of uncontrolled diabetes mellitus and a leading cause of consequent blindness. Early diagnosis of DR can lead to an important reduction in sight loss.[96,97] However, diabetic patients

often feature poor adherence to the recommended annual eye examinations aimed at diagnosing early signs of DR.[96–99] A few attempts have been made to increase screening compliance, including training primary care clinicians to identify DR, or through use of telemedicine, which involves acquiring retinal images that are later interpreted by an expert.[100,101] Lately, ML-based solutions were also integrated into this effort, as means of enabling an automated diagnosis of DR at the primary care setting.

Following several experimental attempts at achieving high accuracy in DR diagnosis,[102–105] the first automated diagnosis system, IDx-DR, was approved in 2018. The ML model consists of many input features associated with DR, such as microaneurysms, hemorrhages, and lipoprotein exudates, that are, in turn, implemented in a convolutional neural network to provide a disease level output. In a prospective study,[106] 900 participants with type 1 diabetes mellitus (T1D) and T2D, who had not been diagnosed with DR, were enrolled in ten different primary care sites, and DR was detected using an automated AI system, capturing the retinal image and utilizing ML in stratifying between retinas that are healthy or featuring mild DR and those featuring more-advanced DR. Results were compared to the gold standard Wisconsin Fundus Photograph Reading Center (FPRC) widefield stereoscopic photography and macular optical coherence tomography taken by official FPRC photographers. Out of 819 patients featuring proper readouts by both methods, FPRC detected 198 cases having more than mild DR, while the ML-based system detected 173 cases, achieving a sensitivity of 87.3% and surpassing the FDA's designated 85% superiority endpoint. Specificity using the ML approach reached 90.7%, with 556 patients out of 621 mild DR patients accurately diagnosed by the system, surpassing the 82.5% FDA's superiority endpoint. The IDx-DR has been adapted and continuously validated in pilot autonomous AI-based DR screenings in Poland[107,108] and Spain.[109] Currently, automated AI-based DR diagnosis is being further adapted to smartphone technology, which could additionally enhance the adherence to annual eye examinations.[109,110] Further evaluations are needed to enable this potentially exciting technology to be integrated into widespread clinical use.

RecSys

RecSys applies knowledge-discovery techniques to make personalized recommendations for information, products, or services during a live interaction.[111] RecSys is commonly used for content suggestions. For example, the Netflix content-recommender algorithm is considered a canonical example for a system that gathers ratings from all of its users about featured shows and makes predictions about ratings of users who did not see these shows. These predictions are then used to direct users to movies for which the predicted rating is high, based on their previous preferences and similarities to other users.

In the clinical context, RecSys aims to suggest the most appropriate course of action for a given clinical condition in diverse situations, for example, insulin administration dosage for T1D patients or in optimizing dietary regimes for pre-diabetic individuals.[40,112–114] Of note, direct interaction of the patient with a RecSys, without a professional medical mediation of the recommendations and the required actions, may lead to misinterpretation that would be detrimental to the patient. Thus, gradual acclimation and proper training sessions, coupled with medical caregiver support, should be considered in mitigating these risks.

Automated insulin dose optimization

T1D is a metabolic disorder characterized by increased blood glucose concentrations, caused by the loss of insulin-producing beta cells in the pancreas. It is

estimated that T1D prevalence rate is increasing by about 3% annually, with millions impacted by this life-risking disease worldwide.[115,116] T1D is associated with acute life-risking complications, such as diabetic ketoacidosis and hypoglycemia, and long-term complications, such as cardiovascular disease, diabetic neuropathy, nephropathy, and retinopathy. Hyperglycemia is considered a major player in the formation of all of these complications.[117,118] The current standard of care for individuals with T1D includes an intensive insulin therapy (IIT), which is key to controlling hyperglycemia and thus reduce the risk of diabetes-related complications. Nathan et al.[119] defined IIT as comprising at least three daily injections of insulin, or alternatively, treatment with external insulin pumps. In both modalities, insulin dosage is adjusted based on at least four self-monitored glucose measurements per day. The aim of IIT is to achieve glucose control within the nondiabetic range as continuously as possible while minimizing both hypoglycemic and hyperglycemic episodes.[119]

Despite increased adoption of insulin pumps and continuous-glucose-monitoring (CGM) devices in this setting, most people with T1D do not achieve their glycemic goals.[120,121] Reasons for this may be related to the lack of expertise of clinicians in analyzing complex sensor-augmented pump data[112] or the inherent inaccuracy of the modification system, which is often based on a rough and inaccurate assessment of previous meal carbohydrate content. Nimri and colleagues[112] compared the outcomes of dose adjustments delivered by an automated AI-based decision support system (AI-DSS) against those given by expert physicians in a 6-month, multicenter, multinational parallel, randomized control study in 108 T1D patients aged 10–21. Importantly, the AI-based solution was shown to be non-inferior to the expert guidance in terms of maintaining the primary efficacy measure, which is the percentage of time spent within the target blood-glucose levels range (70–180 mL/dL). Moreover, three severe adverse events were reported in the physicians' arm compared to zero in the AI-DSS arm. Of note, the specific solution discussed by Nimri et al.[112] did not disclose open-source implementation enabling the scientific community to test and validate the specificities of the algorithm, thus substantially limiting its critical evaluation and implementation. On the positive side, the findings encourage further research aimed at ML-based detection and prediction of CGM patterns associated with impending loss of glycemic control.

Personalized nutrition for blood sugar control

T2D and related metabolic disorders represent an even greater major global health and economic burden worldwide. This disease constitutes a central component of a spectrum of cardio-metabolic disorders, including obesity, non-alcoholic fatty liver disease, hyperlipidemia, and their common cardiovascular complications. Overt T2D is often preceded by a milder form of glucose intolerance termed pre-diabetes, which is reversible upon lifestyle modification but often remains undiagnosed. A common first-line prevention and treatment strategy in both T2D and pre-diabetes involves reduction of caloric intake and implementation of a carbohydrate-depleted diet, which, however, remains limited in its long-term success on an individual and population-wide level.[122] A potential reason for the failure of these nutritional interventions is, in part, related to poor long-term compliance given strict specifications.

This paradigm was recently revisited in a study demonstrating individualized glycemic responses of individuals consuming identical amounts of non-nutritive sweeteners,[123] followed by a larger scale study, the personalized nutrition study, in which people's glycemic responses to identical foods were surprisingly found to vary between individuals.[40] These studies raised the possibility that dietary approaches

aimed at normalizing blood glucose levels may have to be tailored to the individual. Toward this aim, the authors recruited a cohort of 800 people and continuously monitored their blood glucose levels over a week in response to a total of 46,898 meals. In addition to postprandial glycemic responses (PPGRs), data were collected on anthropometrics and the gut microbiome, as well as a questionnaire on medical history, lifestyle, and eating habits, together with a panel of blood tests. With this input data, Zeevi et al.[40] devised a model based on stochastic gradient boosting regression. The performance of the model was evaluated by a leave-one-out cross validation scheme, which resulted in a correlation between predicted and measured PPGRs of held-out individuals of R = 0.68. The trained model was further validated on an independent cohort of 100 individuals, in which a similar performance was obtained: R = 0.70, featuring a high degree of generalizability.

The clinical relevance of the model-based RecSys was demonstrated in a two-arm blinded, randomized, controlled trial with 26 independently recruited participants, in which the rational (i.e., ML-based), personalized dietary intervention improved the PPGRs of individuals. These conceptual findings were repeated with a cohort consuming whole-wheat versus white bread[124] and in additional cohorts in distinct territories in which the same[125] or alternative[126,127] ML-based predictive pipelines were utilized. Potential limitations of these approaches could include a loss of predictive power following prolonged ML-based interventions inducing changes in the input features (such as microbiome configuration) as well as emergence of counter-regulatory mechanisms limiting the long-term patient responsiveness to this approach. Long-term clinical utility of such personalized dietary approaches in diabetes and potentially in other nutrition-dependent multi-factorial diseases merits further studies.

EWSs

EWSs and early warning scores are tools used by clinicians and hospital care teams to prematurely identify the deterioration of the clinical state and initiate early intervention and treatment.[128,129] The context in which EWSs are used might vary between intensive care or intraoperative alerts to detect an onset of a new clinical condition among the surveillance population. Although early warning scores for monitoring and reducing cardiac arrest and overall hospital deaths have been widely and effectively used,[130] attempts to facilitate additional EWSs for a variety of clinical conditions are constantly being pursued. Precision-medicine approaches and improved available measurement devices, along with the rise of electronic health records, have provided a solid ground for the introduction of AI to the field of EWSs.[131–133] Examples of clinically utilized EWSs include predictors of acute illness at the intensive care setting,[134,135] predictors of kidney failure,[136,137] sepsis,[138,139] glaucoma, age-related macular degeneration, diabetic retinopathy,[140] detection and monitoring of Parkinson's disease,[141,142] fall risk estimation,[143] and heart monitoring.[144] We want to highlight the potential power of natural language processing, a deep-learning approach particularly suitable to process large amounts of textual structured and unstructured data. In some cases, such approaches meet the expectations by predicting future acute kidney injury[137] or pediatric disease[145] from EHRs.

With these advances notwithstanding, most such EWS studies to date provide a proof of concept, while faithful and safe translation into the clinical setting await the completion of rigorous clinical trials. Below, we provide examples of two unique implementations of EWSs in both an intensive care setting and as a screening approach in the community.

Intraoperative support with EWSs

Intraoperative hypotension (IOH) in non-cardiac surgeries constitutes a major complication resulting from general anesthesia, intra-operative bleeding, and other surgery-related events.[146] With no clear definition of IOH, the prevalence of this complication ranges between 12% and 94% of operations[147,148] and is associated with postoperative kidney failure, myocardial infarction, and various long-term patient outcomes.[149,150] Monitoring arterial blood pressure is considered important during standard general anesthesia management, either by non-invasive oscillometry in 2- to 5-min intervals or continuously using an invasive arterial catheter.[151] In a randomized controlled study, twice as many hypotensive minutes were detected on average using an arterial catheter compared to oscillometry.[152] Another randomized control study demonstrated that individualized intraoperative systolic blood pressure monitoring was more effective than standard management monitoring in the prevention of systemic inflammatory response syndrome or organ dysfunction.[153]

The first attempted implementation of ML in aiming to control IOH included the Hypotension Prediction Index (HPI), a metrics based on the analysis of features in high-fidelity arterial pressure waveform recordings.[153,154] Use of intense signal processing enabled feature extraction, followed by training of an IOH predictor utilizing logistic regression of 1,334 patients' records. The predictor was prospectively validated using a dataset of 204 patients by predicting hypotension incidents (mean arterial pressure <65 mmHg for at least 1 min) 5, 10, or 15 min before the event. An impressive 92% specificity and sensitivity, 89% and 90% specificity and sensitivity, and 88% specificity and 87% sensitivity were noted in the prediction of IOH 5, 10, and 15 min prior to the actual event, respectively. A follow-up assessment[155] retrospectively validated these findings, albeit with a slightly decreased performance for 5-min predictions of 85.8% sensitivity and specificity. In two small-scale intervention studies, this system was used as an EWS in high-risk surgical cases and managed to significantly decrease IOH time.[156,157] Limitations of these trials include the relatively small groups of participants enrolled, specific surgical procedures tested using the ML system, and their validation mainly in severely ill populations that are invasively monitored. In addition, generalization of the IOH shortening to clinically meaningful complications was not measured and merits future studies.

Cardiovascular disease detection with smartwatches and smartphones

Cardiovascular diseases (CVDs) are considered the leading cause of mortality globally, with an estimated 17.3 million people succumbing to cardiovascular-related disease in 2008.[158] People with cardiovascular conditions may feature a variety of clinical presentations, ranging, among others, from cardiovascular ischemia, heart failure, peripheral vascular disease, stroke, or heart-rhythm disturbances. An immense need exists for early diagnosis of CVD to provide primary care for improved quality of life and longevity while reducing the burden of CVD imposed on healthcare systems.[159,160] In the past 3 decades, the prospect of remote cardiac monitoring via telemonitoring has attracted much attention as a possible early-detection point-of-care modality for CVD patients and has managed to increase the quality of life while reducing mortality, hospitalizations, and health care costs in several clinical studies.[161–165]

Attempts to integrate AI for improvement of CVD remote monitoring have been proven most useful in improving the New York Heart Association functional class, preventing heart failure and subsequent heart failure hospitalization, and reducing mortality.[166–168] Yet these modalities involved cumbersome, complex, and often invasive monitoring equipment and have not been translated well into clinical practice,[169,170] which precludes them from routine, widespread use.[171] Even though

advancements were achieved in heart failure prevention with invasive monitoring for at-risk populations,[161–165] detection and treatment for early CVD onset has remained under-explored.

Increased availability of wearable technology and smartphones has provided a new opportunity for collection of a wide, comprehensive, and continuous array of features related to heart function.[133,144,172] Smart wearables, coupled with compatible software enabling the support for advanced analytics and the use of ML, are increasingly used as EWSs for CVD. Such a framework has been shown to provide beneficial predictive values in a few proof-of-concept studies.[173,174] One study enrolled 9,750 participants, 347 of which with atrial fibrillation (AF). A neural network was established through heuristic pre-training, in which the network approximated the time between heartbeats with manual annotation of the training data. The authors validated the performance of their algorithm in two independent cohorts. In the first cohort, encompassing 51 patients undergoing cardioversion, the neural network exhibited high sensitivity and specificity of 90.2% and 98%, respectively. In the second validation cohort, encompassing 1,617 ambulatory patients, the pre-trained neural network exhibited worse performance with only 67.7% and 67.6% sensitivity and specificity, respectively.[174] In another prospective trial, a smartwatch signal-processing algorithm was tested against an electrocardiogram of 508 hospitalized patients, 237 of which presented AF. The EWS achieved a sensitivity of 93.7% and specificity of 98.2%.[173,174]

Powered with the Apple watch as an EWS, an extensively large-scale study, including 419,297 participants, aimed at screening detection of new-onset AF.[175] In case of an alert, a telemedicine meeting was scheduled and an electrocardiography (ECG) patch was sent via email and worn for 7 days to decide whether the participant was indeed diagnosed with AF by the gold-standard heart monitoring. In total, 2,161 participants received a notice of irregular pulse, and of them, only 450 returned ECG patches containing data that could be analyzed, of which 34% of cases have been correctly diagnosed with AF. Elderly individuals (>65 years old) featured a higher diagnostic accuracy compared to young adults (<40 years old). Among 86 patients that featured an AF episode during continuous ECG monitoring, 84% also noted an AF by their smartwatch.

Despite the relatively low 34% accurate detection rate of AF by smartwatches, the price of mistake in this clinical context was negligible compared to a late diagnosis of AF. Moreover, the irregular pulse monitored by a smartwatch could stratify a risk group for other heart conditions besides AF.[176] However, a high miss rate in the general population could lead to product uncertainty, anxiety, ignored notifications, or, at the other extreme, to overdiagnosis and overtreatment. Additionally, hardware and software differences of various brands of smartphones and smartwatches may possibly impact the performances of ML monitoring systems. Addressing these challenges may potentially enable to utilize these exciting wearable technologies for a variety of medical readouts as well of other potential smartwatch readouts and merit future studies.

CHALLENGES AND OBSTACLES

Despite the enthusiasm, technological developments, encouraging results, and a hype featured by integration of ML into the clinical realm, many challenges and pitfalls still limit the widespread implementation to clinical decision making.

Explainability

ML algorithms for regression and classification are based on discovering associations in the data, which are strong and reliable enough to make predictions from previously

unseen data. An explanation of the underlying data-driven ML model only refers to the statistical sense, in which it is possible to measure the amount of variance in the output labels that is explained by the trained predictor. Therefore, it is important to stress that explainability of ML algorithms must not be confused with causal inference or "mechanistic observations." A simple example is the case of a linear regression model, in which the absolute value and the sign of a coefficient assigned to a predictor describe how strong its impact on the predicted outcome and the directionality of the association. Another example is a decision tree, which can be considered as a decision path formed by answering a sequence of yes and no questions.

Some models aim for the discovery of associations that are obvious or easy to explain. An example of this important concept is the work by O'Shea[177] that utilized a Bayesian network to discover risk factors for post-stroke mortality. The high extent of agreement between the data-driven network model and common expert knowledge demonstrates the potential of such causal models not only to reconfirm previously known associations but also to discover new relationships between factors and better understand hypothesized relations between such factors. As the complexity of the model increases, the ability to provide explainable results decreases, which leads to complex models, such as DNNs, frequently referred to as "black box" algorithms. These are further discussed elsewhere.[10,178,179]

Causability

Beyond explainability, the term "causability"[180] is referring to measures for the quality of explanations provided by ML algorithms.[181] This is of particular importance in clinical decision making, because medical professionals require a measure to judge and understand why an ML algorithm generated a certain output. These proceedings can be facilitated by interactive ML,[182] in which the human domain expert provides implicit and contextual knowledge to the learning process. A solid understanding of the causability is the foundation for a rational attribution of ML-based discoveries in medicine in accordance with legal constraints and regulatory hurdles,[183] which will be discussed further below.

Availability of data

Despite the increase in public availability of clinical data,[184,185] high-quality data representative of the global population remains limited,[186] especially when considering rare medical cases which, according to their prevalence, are hardly documented. A key example is the challenge of EHR data collection, where the task of integrating data from multiple sources is challenging due to usage of different formats and the rich and complex nature of the data. When using inadequate datasets for training of an ML model, there is a risk of picking up incidental correlations that appear due to sample size or bias therein, which have nothing to do with the true underlying relationships between the variables of interest.[187] In such cases, models are prone to suffer from poor generalizability and provide misguided predictions. A related challenge relates to public and transparent sharing of data and codes enabling practical implementation of models, which remains disappointingly limited. One must stress the importance of such community sharing to enable validation, corroboration, and comparisons of results and their faithful representation between researchers and study populations. These must be balanced with preservation of privacy, data protection, and avoidance of unauthorized commercial exploitation of clinical data.

Generalizability

It is not uncommon for an ML system to perform better on data collected in the same study or from similar sources than of newly collected data.[188] Generalizability is the

measure of how useful an existing model will be on real-life data. For instance, if an ML system is trained and tested on data from one hospital and performs badly on data from another hospital, the ML system may exhibit a lower degree of generalizability. ML systems with higher generalizability are often preferable for clinical use. There are multiple ways to reason about the expected generalizability of a trained model, but in the absence of a rigorous definition for this concept, some scientific journals provided new guidelines demanding external cohort validation.[189–191] Panch et al.[192] discussed the necessity of benchmark datasets for such external validation and provided detailed guidelines for constructing such databases. Another perspective by Futoma et al.[188] raised concerns that the pursuit of generalizable models would become counterproductive. The authors explain that, although universal generalization is important, it is not always most relevant; thus, the concept of generalizability should be adapted to the specific setting in which the model operates and be measured accordingly.[188] The authors discuss the relevance of geographic generalizability in a successful nationwide DR screening in India[193] yet warn that unrealistic expectation that every ML system will generalize may risk compromising excellent and often important local performance.[194–196]

Incorporating ML of multi-omics

Several studies combine in their ML input features obtained from different omics layers.[33] For such approaches to enter the clinics, several existing hurdles have to be overcome. Despite the gradual decrease noted in the cost of obtaining diverse omics data, such as whole-genome sequences, scRNA-seq data, and metabolomics, the acquisition of such datasets from entire patient cohorts longitudinally during a study is still expensive and laborious and may become prohibitive, especially in less-affluent population settings.

Additionally, such omics often require characterization using a standard reference, which could exhibit different biases. For example, common pitfalls in the field of ML with scRNA-seq data are the confounders related to reference transcriptomes for the supervised *a priori* assignment of cell types. Inter-cellular and inter-personal variability may not allow unique assignments of cell types for model training. Traditionally, cell types were defined by the presence or absence of surface markers acquired by fluorescence-activated cell sorting (FACS). However, these data are not always available in addition to the transcriptomic readout. While ML has been utilized to tackle that problem,[197] dynamic transitions of cellular states were recently modeled within a deterministic framework, more suitable than ML algorithms in this case.[198]

Randomized control trials versus real-life data evaluation

Estimating whether a ML system is appropriately suited for clinical usage currently constitutes one of the most challenging aspects in this emerging field. Mostly, ML systems are validated in retrospective or prospective trials and, in some occasions, with simulated data.[15,199] As these trials could be useful in a proof-of-concept manner, more randomized control trials are needed to allow for a better evaluation of the ML system, as it was lately discussed in guidelines for AI-related publication.[15] Although randomized control trials may provide solid evidence for the performance of an ML system, they may face interpretive problems and challenges due to lack of fine-grained controls.[89,200] Recently, a team from Google Research, following a successful trial assessing a ML pipeline in diagnosing diabetic retinopathy (DR),[201] deployed their ML system in a nationwide screening of DR in Thailand.[200] While closely monitoring the deployment of this ML system, the team noticed many unexpected issues that were never encountered in their clinical

trials,[200] including low internet speed and poor lighting slowing down the screening process, preventing diagnosis in some cases, and adding unexpected costs. This emphasizes the importance of follow-up research for the ML system to face real-life problems, which may require adjustments to the model itself or to the way it is deployed in the clinics.

Regulatory hurdles

Regulatry hurdles are often encountered in the process of certification of a ML-based pipeline or software for use in prospective clinical trials, or during adaptation into the real-life setting. While the legal process differs markedly between the US, EU, and other global regions, in most regulatory settings the concept of a learning and thus an inherently shifting algorithm constitutes a formidable regulatory challenge and merits adaptation of some of the basic concepts of clinical testing specifications. Ethical consideration also constitutes an important unresolved topic with use of predictive ML pipelines. For example, data-sharing requests need to be balanced with data protection considerations. The implications of unsupervised personalized health predictions on the individual may bear dramatic economic, medical, and personal ramifications that merit a broader discussion and careful consideration in balancing population and personal benefits.

CONCLUSIONS

ML systems are gradually being adopted from well-established pre-clinical scenarios to the bedside in multiple domains and various use cases. The extensive and unique contributions of ML systems to clinical decision making may generate a substantial impact but necessitate continued rigorous research and tackling of challenges and biases. Regulatory authorization of ML-based medical pipelines constitutes another challenge, given their inherent dynamic and constantly improving analytical nature. Much-needed uniformity is sought in terms of guidelines for protocols and publications. These require more randomized control trials evaluating, in a standardized manner, the performance of new ML systems in real-life settings. As these challenges are gradually met and existing hurdles are overcome, ML may eventually meet its expectations in integrating into clinical decision making and transforming the data-driven evolution of precision medicine.

ACKNOWLEDGMENTS

No funding sources were used in this manuscript. We thank the members of the Elinav lab, Weizmann Institute of Science, and the Cancer-Microbiome Division, DKFZ, for discussions and apologize for authors whose work was not cited because of space constraints. L.A. received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 842203. E.E. is the incumbent of the Sir Marc and Lady Tania Feldmann Professorial Chair, a senior fellow at the Canadian Institute of Advanced Research, and an international scholar at the Bill & Melinda Gates Foundation and the Howard Hughes Medical Institute (HHMI). Figure 1 was created with BioRender.com.

AUTHOR CONTRIBUTIONS

All authors performed an extensive literature research, contributed substantially to discussion of the content, and wrote and edited the manuscript.

DECLARATION OF INTERESTS

E.E. is a salaried scientific consultant for DayTwo and BiomX and an editorial board member in Med.

REFERENCES

1. Adams, I.D., Chan, M., Clifford, P.C., Cooke, W.M., Dallos, V., de Dombal, F.T., Edwards, M.H., Hancock, D.M., Hewett, D.J., McIntyre, N., et al. (1986). Computer aided diagnosis of acute abdominal pain: a multicentre study. Br. Med. J. (Clin. Res. Ed.) *293*, 800–804.
2. de Dombal, F.T., Leaper, D.J., Staniland, J.R., McCann, A.P., and Horrocks, J.C. (1972). Computer-aided diagnosis of acute abdominal pain. BMJ *2*, 9–13.
3. Doig, G.S., Inman, K.J., Sibbald, W.J., Martin, C.M., and Robertson, J.M. (1993). Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. Proc. Annu. Symp. Comput. Appl. Med. Care, 361–365.
4. Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., and Kroeker, K.I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit. Med. *3*, 17.
5. Moja, L., Kwag, K.H., Lytras, T., Bertizzolo, L., Brandt, L., Pecoraro, V., Rigon, G., Vaona, A., Ruggiero, F., Mangia, M., et al. (2014). Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. Am. J. Public Health *104*, e12–e22.
6. Varghese, J., Kleine, M., Gessner, S.I., Sandmann, S., and Dugas, M. (2018). Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. J. Am. Med. Inform. Assoc. *25*, 593–602.
7. Benjamens, S., Dhunnoo, P., and Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit. Med. *3*, 118.
8. Dong, J., Geng, Y., Lu, D., Li, B., Tian, L., Lin, D., and Zhang, Y. (2020). Clinical trials for artificial intelligence in cancer diagnosis: a cross-sectional study of registered trials in ClinicalTrials.gov. Front. Oncol. *10*, 1629.
9. Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. N. Engl. J. Med. *380*, 1347–1358.
10. Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. *25*, 44–56.
11. Al'Aref, S.J., Anchouche, K., Singh, G., Slomka, P.J., Kolli, K.K., Kumar, A., Pandey, M., Maliakal, G., van Rosendael, A.R., Beecy, A.N., et al. (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur. Heart J. *40*, 1975–1986.
12. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.
13. Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., et al.; Reader study level-I and level-II Groups (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann. Oncol. *29*, 1836–1842.
14. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al. (2019). 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 *1*, e271–e297.
15. Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A.K., and Calvert, M.J.; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat. Med. *26*, 1351–1363.
16. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M.J., and Denniston, A.K.; SPIRIT-AI and CONSORT-AI Working Group (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat. Med. *26*, 1364–1374.
17. Alpaydin, E. (2020). Introduction to Machine Learning (MIT).
18. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning (Springer).
19. Mitchell, T.M. (1997). Machine Learning (McGraw-Hill).
20. Xu, Y., Zhu, J.-Y., Chang, E.I.C., Lai, M., and Tu, Z. (2014). Weakly supervised histopathology cancer image segmentation and classification. Med. Image Anal. *18*, 591–604.
21. Xue, Y., Du, N., Mottram, A., Seneviratne, M., and Dai, A.M. (2020). Learning to select the best forecasting tasks for clinical outcome prediction. Advances in Neural Information Processing Systems*33* (Vancouver, Canada: NeurIPS).
22. Chen, D., Liu, S., Kingsbury, P., Sohn, S., Storlie, C.B., Habermann, E.B., Naessens, J.M., Larson, D.W., and Liu, H. (2019). Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit. Med. *2*, 43.
23. National Human Genome Research Institute (2020). The cost of sequencing a human genome. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
24. Wang, T.-L., Maierhofer, C., Speicher, M.R., Lengauer, C., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E. (2002). Digital karyotyping. Proc. Natl. Acad. Sci. USA *99*, 16156–16161.
25. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M.; Cancer Genome Atlas Research Network (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.
26. Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S.J., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature *579*, 567–574.
27. Adlung, L., Elinav, E., Greten, T.F., and Korangy, F. (2020). Microbiome genomics for cancer prediction. Nat. Can. *1*, 379–381.
28. Vujkovic-Cvijin, I., Sklar, J., Jiang, L., Natarajan, L., Knight, R., and Belkaid, Y. (2020). Host variables confound gut microbiota studies of human disease. Nature *587*, 448–454.
29. Deng, Y., Bao, F., Dai, Q., Wu, L.F., and Altschuler, S.J. (2019). Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. Nat. Methods *16*, 311–314.
30. Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D.S., Kloiber, K., Heutink, P., and Bonn, S. (2020). Deep learning-based cell composition analysis from tissue expression profiles. Sci. Adv. *6*, eaba2619.
31. Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721.
32. Singh, S.P., Janjuha, S., Chaudhuri, S., Reinhardt, S., Kränkel, A., Dietz, S., Eugster, A., Bilgin, H., Korkmaz, S., Zararsız, G., et al. (2018). Machine learning based classification of cells into chronological stages using single-cell transcriptomics. Sci. Rep. *8*, 17156.
33. Bar, N., Korem, T., Weissbrod, O., Zeevi, D., Rothschild, D., Leviatan, S., Kosower, N., Lotan-Pompan, M., Weinberger, A., Le Roy, C.I., et al.; IMI DIRECT consortium (2020). A reference map of potential determinants for the human serum metabolome. Nature *588*, 135–140.
34. Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D.C., Jensen, S.Ø., Medina, J.E., Hruban, C., White, J.R., et al. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature *570*, 385–389.
35. Leal, A., van Grieken, N.C.T., Palsgrove, D.N., Phallen, J., Medina, J.E., Hruban, C., Broeckaert, M.A.M., Anagnostou, V., Adleff, V., Bruhm, D.C., et al. (2020). White blood cell and cell-free DNA analyses for detection of residual disease in gastric cancer. Nat. Commun. *11*, 525.
36. Wan, N., Weinberg, D., Liu, T.Y., Niehaus, K., Ariazi, E.A., Delubac, D., Kannan, A., White, B., Bailey, M., Bertin, M., et al. (2019). Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. BMC Cancer *19*, 832.
37. Zrimec, J., Börlin, C.S., Buric, F., Muhammad, A.S., Chen, R., Siewers, V., Verendel, V., Nielsen, J., Töpel, M., and Zelezniak, A. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-

evolving interacting gene regulatory structure. Nat. Commun. *11*, 6141.

38. Macartney-Coxson, D., Cameron, A.M., Clapham, J., and Benton, M.C. (2020). DNA methylation in blood-Potential to provide new insights into cell biology. PLoS ONE *15*, e0241367.

39. Corces, M.R., Shcherbina, A., Kundu, S., Gloudemans, M.J., Frésard, L., Granja, J.M., Louie, B.H., Eulalio, T., Shams, S., Bagdatli, S.T., et al. (2020). Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. Nat. Genet. *52*, 1158–1168.

40. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized nutrition by prediction of glycemic responses. Cell *163*, 1079–1094.

41. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477.

42. Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I., and Overington, J.P. (2017). A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. *16*, 19–34.

43. Piazza, I., Beaton, N., Bruderer, R., Knobloch, T., Barbisan, C., Chandat, L., Sudau, A., Siepe, I., Rinner, O., de Souza, N., et al. (2020). A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes. Nat. Commun. *11*, 4200.

44. Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature, November 30, 2020. https://www.nature.com/articles/d41586-020-03348-4.

45. Häse, F., Roch, L.M., and Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. Trends Chem. *1*, 282–291.

46. Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. *37*, 1038–1040.

47. Shoer, S., Karady, T., Keshet, A., Shilo, S., Rossman, H., Gavrieli, A., Meir, T., Lavon, A., Kolobkov, D., Kalka, I., et al. (2020). A prediction model to prioritize individuals for SARS-CoV-2 test built from national symptom surveys. Med (NY) *2*, 196–208.e4.

48. Mei, X., Lee, H.C., Diao, K.Y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al. (2020). Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat. Med. *26*, 1224–1228.

49. Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., et al. (2020). Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell *181*, 1423–1433.e11.

50. Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput. Med. Imaging Graph. *31*, 198–211.

51. Giger, M.L. (2018). Machine learning in medical imaging. J. Am. Coll. Radiol. *15* (3 Pt B), 512–520.

52. Yanase, J., and Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. Expert Syst. Appl. *138*, 112821.

53. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 770–778.

54. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM *60*, 84–90.

55. Dalmiş, M.U., Gubern-Mérida, A., Vreemann, S., Bult, P., Karssemeijer, N., Mann, R., and Teuwen, J. (2019). Artificial intelligence-based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, T2, and DWI. Invest. Radiol. *54*, 325–332.

56. Firmino, M., Angelo, G., Morais, H., Dantas, M.R., and Valentim, R. (2016). Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. Biomed. Eng. Online *15*, 2.

57. Hahn, S., Perry, M., Morris, C.S., Wshah, S., and Bertges, D.J. (2020). Machine deep learning accurately detects endoleak after endovascular abdominal aortic aneurysm repair. JVS Vasc. Sci. *1*, 5–12.

58. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature *577*, 89–94.

59. Chen, W., Zheng, R., Zhang, S., Zeng, H., Zuo, T., Xia, C., Yang, Z., and He, J. (2017). Cancer incidence and mortality in China in 2013: an analysis based on urbanization level. Chin. J. Cancer Res. *29*, 1–10.

60. Edwards, B.K., Ward, E., Kohler, B.A., Eheman, C., Zauber, A.G., Anderson, R.N., Jemal, A., Schymura, M.J., Lansdorp-Vogelaar, I., Seeff, L.C., et al. (2010). Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer *116*, 544–573.

61. Lieberman, D.A., Rex, D.K., Winawer, S.J., Giardiello, F.M., Johnson, D.A., and Levin, T.R. (2012). Guidelines for colonoscopy surveillance after screening and polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology *143*, 844–857.

62. Seeff, L.C., Richards, T.B., Shapiro, J.A., Nadel, M.R., Manninen, D.L., Given, L.S., Dong, F.B., Winges, L.D., and McKenna, M.T. (2004). How many endoscopies are performed for colorectal cancer screening? Results from CDC's survey of endoscopic capacity. Gastroenterology *127*, 1670–1677.

63. Dawwas, M.F. (2014). Adenoma detection rate and risk of colorectal cancer and death. N. Engl. J. Med. *370*, 2539–2540.

64. Robertson, D.J., Greenberg, E.R., Beach, M., Sandler, R.S., Ahnen, D., Haile, R.W., Burke, C.A., Snover, D.C., Bresalier, R.S., McKeown-Eyssen, G., et al. (2005). Colorectal cancer in patients under close colonoscopic surveillance. Gastroenterology *129*, 34–41.

65. van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., van Deventer, S.J., and Dekker, E. (2006). Polyp miss rate determined by tandem colonoscopy: a systematic review. Am. J. Gastroenterol. *101*, 343–350.

66. Brand, E.C., and Wallace, M.B. (2017). Strategies to increase adenoma detection rates. Curr. Treat. Options Gastroenterol. *15*, 184–212.

67. Ahmad, O.F., Soares, A.S., Mazomenos, E., Brandao, P., Vega, R., Seward, E., Stoyanov, D., Chand, M., and Lovat, L.B. (2019). Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. Lancet Gastroenterol. Hepatol. *4*, 71–80.

68. Hassan, C., Spadaccini, M., Iannone, A., Maselli, R., Jovani, M., Chandrasekar, V.T., Antonelli, G., Yu, H., Areia, M., Dinis-Ribeiro, M., et al. (2021). Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest. Endosc. *93*, 77–85.e6.

69. Huang, S., Yang, J., Fong, S., and Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Lett. *471*, 61–71.

70. Wang, P., Berzin, T.M., Glissen Brown, J.R., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., et al. (2019). Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut *68*, 1813–1819.

71. Wang, P., Liu, X., Berzin, T.M., Glissen Brown, J.R., Liu, P., Zhou, C., Lei, L., Li, L., Guo, Z., Lei, S., et al. (2020). Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol. Hepatol. *5*, 343–351.

72. Wu, L., Zhang, J., Zhou, W., An, P., Shen, L., Liu, J., Jiang, X., Huang, X., Mu, G., Wan, X., et al. (2019). Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut *68*, 2161–2169.

73. Gong, D., Wu, L., Zhang, J., Mu, G., Shen, L., Liu, J., Wang, Z., Zhou, W., An, P., Huang, X., et al. (2020). Detection of colorectal adenomas with a real-time computer-aided

system (ENDOANGEL): a randomised controlled study. Lancet Gastroenterol. Hepatol. *5*, 352–361.

74. Repici, A., Badalamenti, M., Maselli, R., Correale, L., Radaelli, F., Rondonotti, E., Ferrara, E., Spadaccini, M., Alkandari, A., Fugazza, A., et al. (2020). Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology *159*, 512–520.e7.

75. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.

76. GBD 2017 Causes of Death Collaborators (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet *392*, 1736–1788.

77. Oeffinger, K.C., Fontham, E.T., Etzioni, R., Herzig, A., Michaelson, J.S., Shih, Y.C., Walter, L.C., Church, T.R., Flowers, C.R., LaMonte, S.J., et al.; American Cancer Society (2015). Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. JAMA *314*, 1599–1614.

78. Lee, C.H., Dershaw, D.D., Kopans, D., Evans, P., Monsees, B., Monticciolo, D., Brenner, R.J., Bassett, L., Berg, W., Feig, S., et al. (2010). Breast cancer screening with imaging: recommendations from the Society of Breast Imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. J. Am. Coll. Radiol. *7*, 18–27.

79. Elmore, J.G., Jackson, S.L., Abraham, L., Miglioretti, D.L., Carney, P.A., Geller, B.M., Yankaskas, B.C., Kerlikowske, K., Onega, T., Rosenberg, R.D., et al. (2009). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology *253*, 641–651.

80. Kohli, A., and Jha, S. (2018). Why CAD failed in mammography. J. Am. Coll. Radiol. *15* (3 Pt B), 535–537.

81. Lehman, C.D., Wellman, R.D., Buist, D.S., Kerlikowske, K., Tosteson, A.N., and Miglioretti, D.L.; Breast Cancer Surveillance Consortium (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern. Med. *175*, 1828–1837.

82. Moran, S., and Warren-Forward, H. (2012). The Australian BreastScreen workforce: a snapshot. Radiographer *59*, 26–30.

83. Rimmer, A. (2017). Radiologist shortage leaves patient care at risk, warns royal college. BMJ *359*, j4683.

84. Fenton, J.J., Taplin, S.H., Carney, P.A., Abraham, L., Sickles, E.A., D'Orsi, C., Berns, E.A., Cutter, G., Hendrick, R.E., Barlow, W.E., and Elmore, J.G. (2007). Influence of computer-aided detection on performance of screening mammography. N. Engl. J. Med. *356*, 1399–1409.

85. Hollon, T.C., Pandian, B., Adapa, A.R., Urias, E., Save, A.V., Khalsa, S.S.S., Eichberg, D.G., D'Amico, R.S., Farooq, Z.U., Lewis, S., et al. (2020). Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nat. Med. *26*, 52–58.

86. Steiner, D.F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J.D., Gammage, C., Thng, F., Peng, L., and Stumpe, M.C. (2018). Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am. J. Surg. Pathol. *42*, 1636–1646.

87. Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., et al. (2019). Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology *292*, 331–342.

88. Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T.H., Chevalier, M., Tan, T., Mertelmeier, T., et al. (2019). Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J. Natl. Cancer Inst. *111*, 916–922.

89. Salim, M., Wåhlin, E., Dembrower, K., Azavedo, E., Foukakis, T., Liu, Y., Smith, K., Eklund, M., and Strand, F. (2020). External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol. *6*, 1581–1588.

90. Bond, M., Garside, R., and Hyde, C. (2015). A crisis of visibility: the psychological consequences of false-positive screening mammograms, an interview study. Br. J. Health Psychol. *20*, 792–806.

91. Bond, M., Pavey, T., Welch, K., Cooper, C., Garside, R., Dean, S., and Hyde, C. (2013). Systematic review of the psychological consequences of false-positive screening mammograms. Health Technol. Assess. *17*, 1–170.

92. Nishikawa, R.M. (2010). Computer-aided detection and diagnosis. In Digital Mammography, U. Bick and F. Diekmann, eds. (Springer), pp. 85–106.

93. Hu, Q., Whitney, H.M., and Giger, M.L. (2020). A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. Sci. Rep. *10*, 10536.

94. Zhao, W., Yang, J., Sun, Y., Li, C., Wu, W., Jin, L., Yang, Z., Ni, B., Gao, P., Wang, P., et al. (2018). 3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. Cancer Res. *78*, 6881–6889.

95. Vinsard, D.G., Mori, Y., Misawa, M., Kudo, S.E., Rastogi, A., Bagci, U., Rex, D.K., and Wallace, M.B. (2019). Quality assurance of computer-aided detection and diagnosis in colonoscopy. Gastrointest. Endosc. *90*, 55–63.

96. Chakrabarti, R., Harper, C.A., and Keeffe, J.E. (2012). Diabetic retinopathy management guidelines. Expert Rev. Ophthalmol. *7*, 417–439.

97. Liew, G., Michaelides, M., and Bunce, C. (2014). A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999-2000 with 2009-2010. BMJ Open *4*, e004015.

98. Bragge, P., Gruen, R.L., Chau, M., Forbes, A., and Taylor, H.R. (2011). Screening for presence or absence of diabetic retinopathy: a meta-analysis. Arch. Ophthalmol. *129*, 435–444.

99. Virk, R., Binns, A.M., Chambers, R., and Anderson, J. (2021). How is the risk of being diagnosed with referable diabetic retinopathy affected by failure to attend diabetes eye screening appointments? Eye (Lond.) *35*, 477–483.

100. Abramoff, M.D., and Suttorp-Schulten, M.S.A. (2005). Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. Telemed J. E Health *11*, 668–674.

101. Joshi, G.D., and Sivaswamy, J. (2011). DrishtiCare: a telescreening platform for diabetic retinopathy powered with fundus image analysis. J. Diabetes Sci. Technol. *5*, 23–31.

102. Abràmoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., and Niemeijer, M. (2016). Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest. Ophthalmol. Vis. Sci. *57*, 5200–5206.

103. Gargeya, R., and Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. Ophthalmology *124*, 962–969.

104. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA *316*, 2402–2410.

105. Olson, J.A., Sharp, P.F., Fleming, A., and Philip, S. (2008). Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes: response to Abramoff et al. Diabetes Care *31*, e63–e64.

106. Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., and Folk, J.C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit. Med. *1*, 39.

107. Grzybowski, A., and Brona, P. (2019). A pilot study of autonomous artificial intelligence-based diabetic retinopathy screening in Poland. Acta Ophthalmol. *97*, e1149–e1150.

108. van der Heijden, A.A., Abramoff, M.D., Verbraak, F., van Hecke, M.V., Liem, A., and Nijpels, G. (2018). Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. Acta Ophthalmol. *96*, 63–68.

109. Shah, A., Clarida, W., Amelon, R., Hernaez-Ortega, M.C., Navea, A., Morales-Olivas, J., Dolz-Marco, R., Verbraak, F., Jorda, P.P., van

der Heijden, A.A., and Peris Martinez, C. (2020). Validation of automated screening for referable diabetic retinopathy with an autonomous diagnostic artificial intelligence system in a Spanish population. J. Diabetes Sci. Technol., 1932296820906212.

110. Tan, C.H., Kyaw, B.M., Smith, H., Tan, C.S., and Tudor Car, L. (2020). Use of smartphones to detect diabetic retinopathy: scoping review and meta-analysis of diagnostic test accuracy studies. J. Med. Internet Res. *22*, e16658.

111. Sarwar, B., Karypis, G., Konstan, J., and Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of the Tenth International Conference on World Wide Web - WWW '01 (ACM), pp. 285–295.

112. Nimri, R., Battelino, T., Laffel, L.M., Slover, R.H., Schatz, D., Weinzimer, S.A., Dovc, K., Danne, T., and Phillip, M.; NextDREAM Consortium (2020). Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. Nat. Med. *26*, 1380–1384.

113. Poddar, L., Hsu, W., and Lee, M.L. (2019). Predicting user reported symptoms using a gated neural network. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (IEEE), pp. 368–375.

114. Yang, L., Hsieh, C.K., Yang, H., Pollak, J.P., Dell, N., Belongie, S., Cole, C., and Estrin, D. (2017). Yum-Me: a personalized nutrient-based meal recommender system. ACM Trans. Inf. Syst. *36*, 1–31.

115. Aanstoot, H.-J., Anderson, B.J., Daneman, D., Danne, T., Donaghue, K., Kaufman, F., Réa, R.R., and Uchigata, Y. (2007). Executive summary. Pediatr. Diabetes *8*, 8–9.

116. Daneman, D. (2006). Type 1 diabetes. Lancet *367*, 847–858.

117. Devaraj, S., Glaser, N., Griffen, S., Wang-Polagruto, J., Miguelino, E., and Jialal, I. (2006). Increased monocytic activity and biomarkers of inflammation in patients with type 1 diabetes. Diabetes *55*, 774–779.

118. Nathan, D.M. (1993). Long-term complications of diabetes mellitus. N. Engl. J. Med. *328*, 1676–1685.

119. Nathan, D.M., Cleary, P.A., Backlund, J.Y., Genuth, S.M., Lachin, J.M., Orchard, T.J., Raskin, P., and Zinman, B.; Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) Study Research Group (2005). Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. N. Engl. J. Med. *353*, 2643–2653.

120. Foster, N.C., Miller, K., Dimeglio, L., Maahs, D.M., Tamborlane, W.V., Bergenstal, R.M., Clements, M.A., Rickels, M.R., Smith, E., Olson, B.A., et al. (2018). Marked increases in CGM use has not prevented increases in HbA1c levels in participants in the T1D Exchange (T1DX) Clinic Network. Diabetes *67*, 1689-P.

121. Miller, K.M., Foster, N.C., Beck, R.W., Bergenstal, R.M., DuBose, S.N., DiMeglio, L.A., Maahs, D.M., and Tamborlane, W.V.; T1D Exchange Clinic Network (2015). Current state of type 1 diabetes treatment in the U.S.: updated data from the T1D Exchange clinic registry. Diabetes Care *38*, 971–978.

122. Blüher, M. (2019). Obesity: global epidemiology and pathogenesis. Nat. Rev. Endocrinol. *15*, 288–298.

123. Suez, J., Korem, T., Zeevi, D., Zilberman-Schapira, G., Thaiss, C.A., Maza, O., Israeli, D., Zmora, N., Gilad, S., Weinberger, A., et al. (2014). Artificial sweeteners induce glucose intolerance by altering the gut microbiota. Nature *514*, 181–186.

124. Korem, T., Zeevi, D., Zmora, N., Weissbrod, O., Bar, N., Lotan-Pompan, M., Avnit-Sagi, T., Kosower, N., Malka, G., Rein, M., et al. (2017). Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. Cell Metab. *25*, 1243–1253.e5.

125. Mendes-Soares, H., Raveh-Sadka, T., Azulay, S., Edens, K., Ben-Shlomo, Y., Cohen, Y., Ofek, T., Bachrach, D., Stevens, J., Colibaseanu, D., et al. (2019). Assessment of a personalized approach to predicting postprandial glycemic responses to food among individuals without diabetes. JAMA Netw. Open *2*, e188102.

126. Hall, H., Perelman, D., Breschi, A., Limcaoco, P., Kellogg, R., McLaughlin, T., and Snyder, M. (2018). Glucotypes reveal new patterns of glucose dysregulation. PLoS Biol. *16*, e2005143.

127. Berry, S.E., Valdes, A.M., Drew, D.A., Asnicar, F., Mazidi, M., Wolf, J., Capdevila, J., Hadjigeorgiou, G., Davies, R., Al Khatib, H., et al. (2020). Human postprandial responses to food and potential for precision nutrition. Nat. Med. *26*, 964–973.

128. Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C.K., Suter, P.M., and Thijs, L.G. (1996). 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. *22*, 707–710.

129. Whittington, J., White, R., Haig, K.M., and Slock, M. (2007). Using an automated risk assessment report to identify patients at risk for clinical deterioration. Jt. Comm. J. Qual. Patient Saf. *33*, 569–574.

130. Smith, M.E.B., Chiovaro, J.C., O'Neil, M., Kansagara, D., Quiñones, A.R., Freeman, M., Motu'apuaka, M.L., and Slatore, C.G. (2014). Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. Ann. Am. Thorac. Soc. *11*, 1454–1465.

131. Khennou, F., Khamlichi, Y.I., and Chaoui, N.E.H. (2018). Improving the use of big data analytics within electronic health records: a case study based OpenEHR. Procedia Comput. Sci. *127*, 60–68.

132. Miotto, R., Li, L., Kidd, B.A., and Dudley, J.T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci. Rep. *6*, 26094.

133. Phan, D., Siong, L.Y., Pathirana, P.N., and Seneviratne, A. (2015). Smartwatch: performance evaluation for long-term heart rate monitoring. 2015 International Symposium on Bioelectronics and Bioinformatics (ISBB) (IEEE), pp. 144–147.

134. Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J., and Thiesson, B. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat. Commun. *11*, 3852.

135. Shickel, B., Loftus, T.J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., and Rashidi, P. (2019). DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. Sci. Rep. *9*, 1879.

136. Cheng, P., Waitman, L.R., Hu, Y., and Liu, M. (2018). Predicting inpatient acute kidney injury over different time horizons: how early and accurate? AMIA Annu. Symp. Proc. *2017*, 565–574.

137. Tomašev, N., Glorot, X., Rae, J.W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. Nature *572*, 116–119.

138. Islam, M.M., Nasrin, T., Walther, B.A., Wu, C.C., Yang, H.C., and Li, Y.C. (2019). Prediction of sepsis patients using machine learning approach: a meta-analysis. Comput. Methods Programs Biomed. *170*, 1–9.

139. Lauritsen, S.M., Kalør, M.E., Kongsgaard, E.L., Lauritsen, K.M., Jørgensen, M.J., Lange, J., and Thiesson, B. (2020). Early detection of sepsis utilizing deep learning on electronic health record event sequences. Artif. Intell. Med. *104*, 101820.

140. Lin, W.-C., Chen, J.S., Chiang, M.F., and Hribar, M.R. (2020). Applications of artificial intelligence to electronic health record data in ophthalmology. Transl. Vis. Sci. Technol. *9*, 13.

141. Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K.M., Dorsey, E.R., and Little, M.A. (2015). Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study. Parkinsonism Relat. Disord. *21*, 650–653.

142. Capecci, M., Pepa, L., Verdini, F., and Ceravolo, M.G. (2016). A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease. Gait Posture *50*, 28–33.

143. Majumder, A.J.A. (2014). A real-time smartphone- and smartshoe-based fall prevention system. Proceedings of the 29th Annual ACM Symposium on Applied Computing - SAC '14 (ACM), pp. 470–471.

144. Li, K.H.C., White, F.A., Tipoe, T., Liu, T., Wong, M.C., Jesuthasan, A., Baranchuk, A., Tse, G., and Yan, B.P. (2019). The current state of mobile phone apps for monitoring heart rate, heart rate variability, and atrial fibrillation: Narrative review. JMIR Mhealth Uhealth *7*, e11606.

145. Liang, H., Tsui, B.Y., Ni, H., Valentim, C.C.S., Baxter, S.L., Liu, G., Cai, W., Kermany, D.S., Sun, X., Chen, J., et al. (2019). Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat. Med. *25*, 433–438.

146. Bijker, J.B., van Klei, W.A., Kappen, T.H., van Wolfswinkel, L., Moons, K.G., and Kalkman, C.J. (2007). Incidence of intraoperative hypotension as a function of the chosen definition: literature definitions applied to a retrospective cohort using automated data collection. Anesthesiology *107*, 213–220.

147. Maheshwari, K., Turan, A., Mao, G., Yang, D., Niazi, A.K., Agarwal, D., Sessler, D.I., and Kurz, A. (2018). The association of hypotension during non-cardiac surgery, before and after skin incision, with postoperative acute kidney injury: a retrospective cohort analysis. Anaesthesia *73*, 1223–1228.

148. Vernooij, L.M., van Klei, W.A., Machina, M., Pasma, W., Beattie, W.S., and Peelen, L.M. (2018). Different methods of modelling intraoperative hypotension and their association with postoperative complications in patients undergoing non-cardiac surgery. Br. J. Anaesth. *120*, 1080–1089.

149. Salmasi, V., Maheshwari, K., Yang, D., Mascha, E.J., Singh, A., Sessler, D.I., and Kurz, A. (2017). Relationship between intraoperative hypotension, defined by either reduction from baseline or absolute thresholds, and acute kidney and myocardial injury after noncardiac surgery: a retrospective cohort analysis. Anesthesiology *126*, 47–65.

150. van Waes, J.A.R., van Klei, W.A., Wijeysundera, D.N., van Wolfswinkel, L., Lindsay, T.F., and Beattie, W.S. (2016). Association between intraoperative hypotension and myocardial injury after vascular surgery. Anesthesiology *124*, 35–44.

151. Saugel, B., Dueck, R., and Wagner, J.Y. (2014). Measurement of blood pressure. Best Pract. Res. Clin. Anaesthesiol. *28*, 309–322.

152. Naylor, A.J., Sessler, D.I., Maheshwari, K., Khanna, A.K., Yang, D., Mascha, E.J., Suleiman, I., Reville, E.M., Cote, D., Hutcherson, M.T., et al. (2020). Arterial catheters for early detection and treatment of hypotension during major noncardiac surgery: a randomized trial. Anesth. Analg. *131*, 1540–1550.

153. Futier, E., Lefrant, J.Y., Guinot, P.G., Godet, T., Lorne, E., Cuvillon, P., Bertran, S., Leone, M., Pastene, B., Piriou, V., et al.; INPRESS Study Group (2017). Effect of individualized vs standard blood pressure management strategies on postoperative organ dysfunction among high-risk patients undergoing major surgery: a randomized clinical trial. JAMA *318*, 1346–1357.

154. Hatib, F., Jian, Z., Buddi, S., Lee, C., Settels, J., Sibert, K., Rinehart, J., and Cannesson, M. (2018). Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. Anesthesiology *129*, 663–674.

155. Davies, S.J., Vistisen, S.T., Jian, Z., Hatib, F., and Scheeren, T.W.L. (2020). Ability of an arterial waveform analysis-derived hypotension prediction index to predict future hypotensive events in surgical patients. Anesth. Analg. *130*, 352–359.

156. Schneck, E., Schulte, D., Habig, L., Ruhrmann, S., Edinger, F., Markmann, M., Habicher, M., Rickert, M., Koch, C., and Sander, M. (2020). Hypotension Prediction Index based protocolized haemodynamic management reduces the incidence and duration of intraoperative hypotension in primary total hip arthroplasty: a single centre feasibility randomised blinded prospective interventional trial. J. Clin. Monit. Comput. *34*, 1149–1158.

157. Wijnberge, M., Geerts, B.F., Hol, L., Lemmers, N., Mulder, M.P., Berge, P., Schenk, J., Terwindt, L.E., Hollmann, M.W., Vlaar, A.P., and Veelo, D.P. (2020). Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA *323*, 1052–1060.

158. Mendis, S., Puska, P., and Norrving, B. (2011). Global Atlas on Cardiovascular Disease Prevention and Control (WHO).

159. Webster, R.J., Heeley, E.L., Peiris, D.P., Bayram, C., Cass, A., and Patel, A.A. (2009). Gaps in cardiovascular disease risk management in Australian general practice. Med. J. Aust. *191*, 324–329.

160. Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., and Lisheng, L.; INTERHEART Study Investigators (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet *364*, 937–952.

161. Clark, R.A., Inglis, S.C., McAlister, F.A., Cleland, J.G.F., and Stewart, S. (2007). Telemonitoring or structured telephone support programmes for patients with chronic heart failure: systematic review and meta-analysis. BMJ *334*, 942.

162. Adamson, P.B., Ginn, G., Anker, S.D., Bourge, R.C., and Abraham, W.T. (2017). Remote haemodynamic-guided care for patients with chronic heart failure: a meta-analysis of completed trials. Eur. J. Heart Fail. *19*, 426–433.

163. Gensini, G.F., Alderighi, C., Rasoini, R., Mazzanti, M., and Casolo, G. (2017). Value of telemonitoring and telemedicine in heart failure management. Card. Fail. Rev. *3*, 116–121.

164. Landolina, M., Perego, G.B., Lunati, M., Curnis, A., Guenzati, G., Vicentini, A., Parati, G., Borghi, G., Zanaboni, P., Valsecchi, S., and Marzegalli, M. (2012). Remote monitoring reduces healthcare use and improves quality of care in heart failure patients with implantable defibrillators: the evolution of management strategies of heart failure patients with implantable defibrillators (EVOLVO) study. Circulation *125*, 2985–2992.

165. Purcell, R., McInnes, S., and Halcomb, E.J. (2014). Telemonitoring can assist in managing cardiovascular disease in primary care: a systematic review of systematic reviews. BMC Fam. Pract. *15*, 43.

166. Boehmer, J.P., Hariharan, R., Devecchi, F.G., Smith, A.L., Molon, G., Capucci, A., An, Q., Averina, V., Stolen, C.M., Thakur, P.H., et al. (2017). A multisensor algorithm predicts heart failure events in patients with implanted devices: results from the MultiSENSE study. JACC Heart Fail. *5*, 216–225.

167. Hindricks, G., Taborsky, M., Glikson, M., Heinrich, U., Schumacher, B., Katz, A., Brachmann, J., Lewalter, T., Goette, A., Block, M., et al.; IN-TIME study group* (2014). Implant-based multiparameter telemonitoring of patients with heart failure (IN-TIME): a randomised controlled trial. Lancet *384*, 583–590.

168. Anand, I.S., Tang, W.H., Greenberg, B.H., Chakravarthy, N., Libbus, I., and Katra, R.P.; Music Investigators (2012). Design and performance of a multisensor heart failure monitoring algorithm: results from the multisensor monitoring in congestive heart failure (MUSIC) study. J. Card. Fail. *18*, 289–295.

169. Ponikowski, P., Voors, A.A., Anker, S.D., Bueno, H., Cleland, J.G., Coats, A.J., Falk, V., González-Juanatey, J.R., Harjola, V.P., Jankowska, E.A., et al.; Authors/Task Force Members; Document Reviewers (2016). 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur. J. Heart Fail. *18*, 891–975.

170. Yancy, C.W., Jessup, M., Bozkurt, B., Butler, J., Casey, D.E., Drazner, M.H., Fonarow, G.C., Geraci, S.A., Horwich, T., Januzzi, J.L., et al. (2013). 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation *128*, e240–e327.

171. Dickinson, M.G., Allen, L.A., Albert, N.A., DiSalvo, T., Ewald, G.A., Vest, A.R., Whellan, D.J., Zile, M.R., and Givertz, M.M. (2018). Remote monitoring of patients with heart failure: a white paper from the Heart Failure Society of America Scientific Statements Committee. J. Card. Fail. *24*, 682–694.

172. Nam, Y., Kong, Y., Reyes, B., Reljin, N., and Chon, K.H. (2016). Monitoring of heart and breathing rates using dual cameras on a smartphone. PLoS ONE *11*, e0151013.

173. Dörr, M., Nohturfft, V., Brasier, N., Bosshard, E., Djurdjevic, A., Gross, S., Raichle, C.J., Rhinisperger, M., Stöckli, R., and Eckstein, J. (2019). The WATCH AF Trial: smartwatches for detection of atrial fibrillation. JACC Clin. Electrophysiol. *5*, 199–208.

174. Tison, G.H., Sanchez, J.M., Ballinger, B., Singh, A., Olgin, J.E., Pletcher, M.J., Vittinghoff, E., Lee, E.S., Fan, S.M., Gladstone, R.A., et al. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiol. *3*, 409–416.

175. Perez, M.V., Mahaffey, K.W., Hedlin, H., Rumsfeld, J.S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A.M., Rajmane, A., Cheung, L., et al.; Apple Heart Study Investigators (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. N. Engl. J. Med. *381*, 1909–1917.

176. Stehlik, J., Schmalfuss, C., Bozkurt, B., Nativi-Nicolau, J., Wohlfahrt, P., Wegerich, S., Rose, K., Ray, R., Schofield, R., Deswal, A., et al. (2020). Continuous wearable monitoring analytics predict heart failure hospitalization: The LINK-HF multicenter study. Circ. Hear. Fail. *13*, e006513.

177. O'Shea, R. (2020). Understanding stroke with Bayesian networks. J. Med. Artif. Intell. *3*, 2.

178. Castelvecchi, D. (2016). Can we open the black box of AI? Nature *538*, 20–23.

179. Goodman, B., and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Mag. *38*, 50–57.

180. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs. Data Min. Knowl. Discov. *9*, e1312.

181. Holzinger, A., Carrington, A., and Müller, H. (2020). Measuring the quality of explanations: the system causability scale (SCS). KI - Künstliche Intelligenz *34*, 193–198.

182. Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform. *3*, 119–131.

183. Schneeberger, D., Stöger, K., and Holzinger, A. (2020). The European legal framework for medical AI. In Machine Learning and Knowledge Extraction, A. Holzinger, P. Kieseberg, A.M. Tjoa, and E. Weippl, eds. (Springer International Publishing), pp. 209–226.

184. Garcia-Vidal, C., Sanjuan, G., Puerta-Alcalde, P., Moreno-García, E., and Soriano, A. (2019). Artificial intelligence to support clinical decision-making processes. EBioMedicine *46*, 27–29.

185. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.W., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., and Mark, R.G. (2016). MIMIC-III, a freely accessible critical care database. Sci. Data *3*, 160035.

186. Sendak, M., Gao, M., Nichols, M., Lin, A., and Balu, S. (2019). Machine learning in health care: a critical appraisal of challenges and opportunities. EGEMs *7*, 1.

187. Chen, I.Y., Johansson, F.D., and Sontag, D. (2018). Why is my classifier discriminatory?. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada (NeurIPS).

188. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., and Celi, L.A. (2020). The myth of generalisability in clinical research and machine learning in health care. Lancet Digit. Health *2*, e489–e492.

189. Bluemke, D.A., Moy, L., Bredella, M.A., Ertl-Wagner, B.B., Fowler, K.J., Goh, V.J., Halpern, E.F., Hess, C.P., Schiebler, M.L., and Weiss, C.R. (2020). Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-From the *Radiology* Editorial Board. Radiology *294*, 487–489.

190. Leisman, D.E., Harhay, M.O., Lederer, D.J., Abramson, M., Adjei, A.A., Bakker, J., Ballas, Z.K., Barreiro, E., Bell, S.C., Bellomo, R., et al. (2020). Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. Crit. Care Med. *48*, 623–633.

191. Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T.B., et al. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J. Med. Internet Res. *18*, e323.

192. Panch, T., Pollard, T.J., Mattie, H., Lindemer, E., Keane, P.A., and Celi, L.A. (2020). "Yes, but will it work for *my* patients?" Driving clinically relevant research with benchmark datasets. NPJ Digit. Med. *3*, 87.

193. Gulshan, V., Rajan, R.P., Widner, K., Wu, D., Wubbels, P., Rhodes, T., Whitehouse, K., Coram, M., Corrado, G., Ramasamy, K., et al. (2019). Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol. *137*, 987.

194. Bedoya, A.D., Clement, M.E., Phelan, M., Steorts, R.C., O'Brien, C., and Goldstein, B.A. (2019). Minimal impact of implemented early warning score and best practice alert for patient deterioration. Crit. Care Med. *47*, 49–55.

195. Downey, C.L., Tahir, W., Randell, R., Brown, J.M., and Jayne, D.G. (2017). Strengths and limitations of early warning scores: A systematic review and narrative synthesis. Int. J. Nurs. Stud. *76*, 106–119.

196. Gerry, S., Bonnici, T., Birks, J., Kirtley, S., Virdee, P.S., Watkinson, P.J., and Collins, G.S. (2020). Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. BMJ *369*, m1501.

197. Baron, C.S., Barve, A., Muraro, M.J., van der Linden, R., Dharmadhikari, G., Lyubimova, A., de Koning, E.J.P., and van Oudenaarden, A. (2019). Cell type purification by single-cell transcriptome-trained sorting. Cell *179*, 527–542.e19.

198. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414.

199. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Med. *17*, 195.

200. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., and Vardoulakis, L.M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (ACM), pp. 1–12.

201. Ruamviboonsuk, P., Krause, J., Chotcomwongse, P., Sayres, R., Raman, R., Widner, K., Campana, B.J.L., Phene, S., Hemarat, K., Tadarati, M., et al. (2019). Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. npj. Digit. Med. *2*, 25.