*Balkan Med J 2023;40:3-12*



# Must-have Qualities of Clinical Research on Artificial Intelligence and Machine Learning

Burak Koçak[1], Renato Cuocolo[2], Daniel Pinto dos Santos[3,4], Arnaldo Stanzione[5], Lorenzo Ugga[5]

[1]Clinic of Radiology, University of Health Sciences Turkey, Başakşehir Çam and Sakura City Hospital, İstanbul, Turkey
[2]Department of Medicine, Surgery and Dentistry, University of Salerno, Baronissi, Italy
[3]Department of Radiology, University Hospital of Cologne, Cologne, Germany
[4]Department of Radiology, University Hospital of Frankfurt, Frankfurt, Germany
[5]Department of Advanced Biomedical Sciences, University of Naples "Federico II", Napoli, Italy

In the field of computer science, known as artificial intelligence, algorithms imitate reasoning tasks that are typically performed by humans. The techniques that allow machines to learn and get better at tasks such as recognition and prediction, which form the basis of clinical practice, are referred to as machine learning, which is a subfield of artificial intelligence. The number of artificial intelligence- and machine learnings-related publications in clinical journals has grown exponentially, driven by recent developments in computation and the accessibility of simple tools. However, clinicians are often not included in data science teams, which may limit the clinical relevance, explanability, workflow compatibility, and quality improvement of artificial intelligence solutions. Thus, this results in the language barrier between clinicians and artificial intelligence developers. Healthcare practitioners sometimes lack a basic understanding of artificial intelligence research because the approach is difficult for non-specialists to understand. Furthermore, many editors and reviewers of medical publications might not be familiar with the fundamental ideas behind these technologies, which may prevent journals from publishing high-quality artificial intelligence studies or, worse still, could allow for the publication of low-quality works. In this review, we aim to improve readers' artificial intelligence literacy and critical thinking. As a result, we concentrated on what we consider the 10 most important qualities of artificial intelligence research: valid scientific purpose, high-quality data set, robust reference standard, robust input, no information leakage, optimal bias-variance tradeoff, proper model evaluation, proven clinical utility, transparent reporting, and open science. Before designing a study, one should have defined a sound scientific purpose. Then, it should be backed by a high-quality data set, robust input, and a solid reference standard. The artificial intelligence development pipeline should prevent information leakage. For the models, optimal bias-variance tradeoff should be achieved, and generalizability assessment must be adequately performed. The clinical value of the final models must also be established. After the study, thought should be given to transparency in publishing the process and results as well as open science for sharing data, code, and models. We hope this work may improve the artificial intelligence literacy and mindset of the readers.

Artificial intelligence (AI) is a subfield of computer science that is related to the creation of algorithms to make decisions on tasks that are typically associated with human intelligence.[1] Various machine learning (ML) techniques are under the umbrella term "AI." ML simply refers to the methods that allow computers to learn directly from data and develop models for tasks such as prediction and recognition, which could be valuable in clinical practice. The general purpose of clinical AI is to find relevant information from complex and high-dimensional data to assist decision-making.[2] Clinical AI should be useful to solve several clinical tasks such as diagnosis,[3-5] disease stratification,[6] risk predictions,[7,8] therapeutic decisions,[9] prognostic predictions,[10,11] and drug discovery.[12]

Corresponding author: Burak Koçak, Clinic of Radiology, University of Health Sciences Turkey, Başakşehir Çam and Sakura City Hospital, İstanbul, Turkey
e-mail: drburakkocak@gmail.com
*Received: November 21, 2022 Accepted: December 06, 2022 Available Online Date: Jan 23, 2023 • DOI: 10.4274/balkanmedj.galenos.2022.2022-11-51*

*Available at www.balkanmedicaljournal.org*

ORCID iDs of the authors: B.K. 0000-0002-7307-396X; R.C. 0000-0002-1452-1574; D.P.S. 0000-0003-4785-6394; A.S. 0000-0002-7905-5789; L.U. 0000-0001-7811-4612.

Cite this article as:
Koçak B, Cuocolo R, dos Santos DP, Stanzione A, Ugga L. Must-have Qualities of Clinical Research on Artificial Intelligence and Machine Learning. *Balkan Med J.*; 2023; 40(1):3-12.

The number of AI-related publications in clinical journals has grown exponentially, driven by developments in computation power and accessibility of simple tools. A simple PubMed search for 2010-2021 reveals an annual growth rate of 42% over the last 5 years (Figure 1). Nearly 25% of all diagnostic accuracy studies submitted to a prominent journal are related to AI.[13] However, despite the high expectations and promises of AI, data and convincing proof are lacking.[14] In real-world clinical practice, several AI technologies reported being on par with or better than experts have actually shown large false-positive rates.[14]

Data science teams rarely involve clinicians, potentially limiting the clinical relevance, explanability, workflow compatibility, and quality improvement in AI solutions.[15] This also contributes to a communication gap between clinicians and developers. Therefore, physicians usually are not familiar with the basic concepts of AI research, as the methodology is rather complex for non-specialists.[13,16,17] Furthermore, many editors or reviewers of medical journals may not be aware of the key concepts of AI.[13] As an example of the complexity in interpreting these papers, at a recent ML conference (The Conference and Workshop on Neural Information Processing Systems; NeurIPS), double-blind reviewers were unable to reach an agreement on more than half of the submissions.[18] Such a disagreement among reviewers might prevent journals from correctly identifying high-quality AI works and, even worse, may lead to publishing works with poor quality or critical flaws.[19]

Understanding the fundamental qualities is key to a critical appraisal of clinical AI research. In this study, we aim to increase the AI literacy of the readers. Therefore, we focused on the 10 most important qualities and related considerations of AI research that were carefully selected based on the domain expertise of the authors (Figure 2).

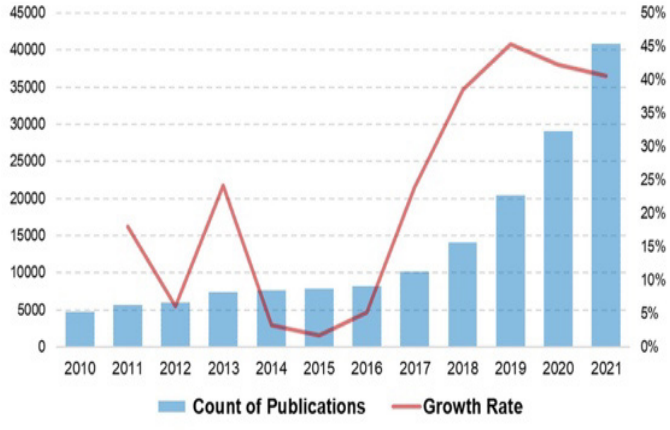


FIG. 1. Recent temporal trend in publications on artificial intelligence or machine learning indexed in PubMed between 2010 and 2021, based on a simple search syntax (artificial intelligence OR machine learning). Average annual growth rates are 23% and 42% for 2010-2021 and 2017-2021, respectively.

## MUST HAVE QUALITIES

### Valid Scientific Purpose

All researchers are familiar with the challenges and complexity behind the conceptualization of a good research question, which is a difficult and recurring task.[20] Indeed, ideas should selected, refined, and finally shaped into valid research questions that must be both interesting and feasible to become the solid foundation for designing a scientific study.[21] A few frameworks can be used to focus on the valid scientific purpose (Figure 3). Frameworks such as the PICOT (population, intervention, comparator, outcome, and time frame) and FINER frameworks can aid in this process,[22] helping researchers to focus on the most promising outcomes or find an unexplored niche with great potential.[20] In this setting, clinical AI research is not an exception. The possibilities for AI in healthcare are apparently limitless, ranging from diagnostics to management and decision-making support. Research efforts should prioritize applications that address a currently unmet

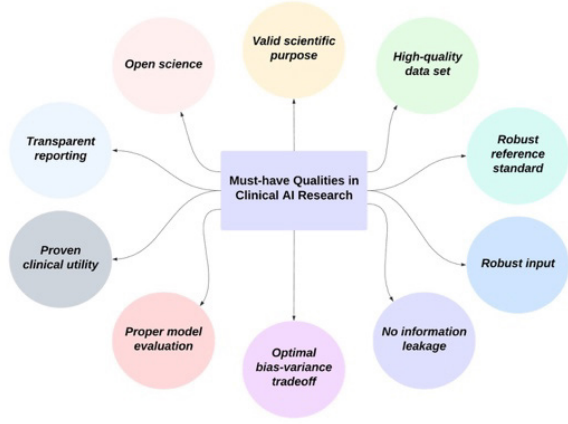


FIG. 2. Top 10 must-have qualities of clinical research on artificial intelligence, being carefully selected by the authors.

*AI, artificial intelligence*

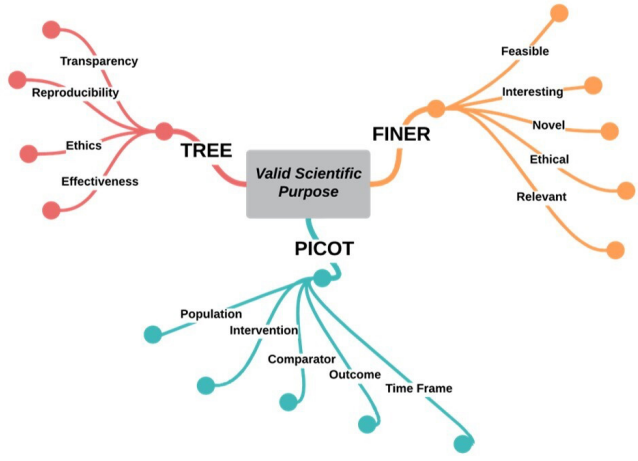


FIG. 3. Frameworks that can be used to establish a valid scientific purpose.

clinical need (e.g., compensating the limitations of current standard-of-care procedures) and exploit the intrinsic advantages of AI (e.g., handling highly dimensional datasets).[23-25] Furthermore, specific frameworks should be considered when formulating a research question for clinical AI studies, such as transparency, reproducibility, ethics, and effectiveness (TREE).[26] Indeed, while keeping in mind what could actually be the ultimate advantage for patients, researchers exploring AI applications in healthcare should promote the paradigm shift toward substantial AI integration in the way healthcare is delivered in clinical practice.[27,28] At present, most patients do not benefit from the steadily increasing research output on AI, which remains in the testing phase and does not move to the bedside.[29] Thus, to be truly valid, the scientific purpose of AI healthcare research should also consider and address the TREE challenges to facilitate its translation into clinical practice. Accordingly, involving experts from different fields (e.g., epidemiologists, physicians with different subspecialties, biostatisticians, engineers, and ethical consultants) is important in the conceptualization phase.

Interestingly, AI does not only represent the object of research but could play an important role in how research is conducted.[30] Regarding clinical trials, AI has been proposed as a solution to optimize protocol design, make patient selection and management more efficient, and, of course, analyze the data collected.[31,32] In the not-so-distant future, AI itself might even find its role alongside researchers to generate valid research questions in the first place.[33,34]

## High-quality Data Set

When designing AI healthcare research, the dataset should be suitable to answer the clinical question.[26] Indeed, good AI applications are highly unlikely obtained when using inadequate data for model training, as the output is heavily dependent on the input, i.e., "garbage in garbage out" (Figure 4).[35] Rather than modifying the model to obtain more reliable performance, working on the two main aspects of dataset appropriateness, namely, quality and quantity might often be more effective. For the latter, a small sample size could lead to unreliable results in AI studies, as confirmed by a recent publication in radiomics.[36] While it might be difficult for each research group to independently obtain a large study population, a possible solution is offered by publicly available datasets.[37,38] However, public datasets might be of heterogeneous quality, and proper controls are advocated to avoid increasing quantity at the expense of quality, which is of course undesirable.[39-41] For some AI tasks, even a relatively small dataset could generate satisfactory results, and the gain from adding new samples for the training dataset tends to decrease, provided that the input data have sufficient quality.[42,43] Furthermore, data-augmentation techniques might be a feasible strategy to compensate for the small sample size in selected AI applications.[44,45] Data quality can be influenced by several factors, such as completeness, accuracy, timeliness, and representativeness. A dataset with some missing values has an obvious completeness issue, which could be either solved by dropping the involved instances (in this scenario, quantity pays the price to ensure the quality) or using imputation to artificially replace missing values.[46] Accuracy refers to how reliable and consistent the dataset is (e.g., free from compilation errors, redundancy, or overlapping categories). Timeliness can be used to define a dataset based on the extent of samples aligned to current practice (e.g., an imaging dataset obtained with a very old magnetice resonance imaging scanner using an obsolete acquisition protocol will likely lead to an AI model that cannot generalize data when applied to more recent acquisition protocols and scanners). Finally, representativeness should be considered to ensure that biomedical AI can be reliably applied to diverse populations (e.g., a dataset highly skewed toward a certain ethnicity might train a model with poor performance on dataset minorities).[47] Similarly, the dataset should be representative of the population suffering from the disorder of interest (e.g., if the prevalence of the disorder in the sample size is significantly different from what is expected, the sample size might not be representative of the target population).

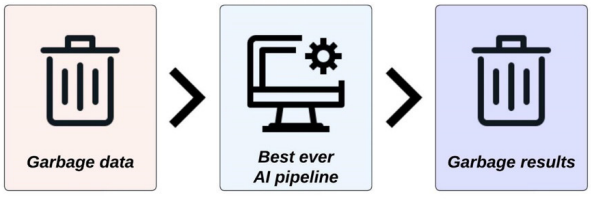


FIG. 4. Importance of high-quality input in artificial intelligence research. The output of models heavily depends on the quality of input data.

*AI, artificial intelligence*

## Robust Reference Standard

Many remember the media attention that ensued when a study from Stanford claimed that deep learning outperformed human radiologists in detecting pneumonia on chest X-ray images (Rajpurkar P. et al. 2017 preprint, https://doi.org/10.48550/arXiv.1711.05225). While many aspects of the study were quite remarkable, one major issue was discussed critically after the study was first published.[48,49] In the initial dataset, over 100,000 frontal-view chest X-ray images were included, and labels were automatically extracted from the radiological reports associated with the images using natural language processing.[50] However, when labels and images were visually inspected by an independent researcher, numerous images were associated with a clearly wrong label.[48] The study was then revised, and the claims were toned down to a more honest statement that the system performs at least on par with human experts in detecting pneumonia-like image features.[49]

Nevertheless, this example highlights a very important issue with research on clinical AI systems. Researchers should ensure that the reference standard the AI system is using during training is of the highest quality that is reasonably achievable. In the example above, it can easily be understood that neither the original report nor the visual inspection of the chest X-ray image alone is ideal in determining whether a patient indeed has pneumonia. Ideally, clinical and laboratory data should be included to establish diagnosis more accurately, especially in cases where visual features alone are ambiguous. The best and most robust reference standard strongly depends on the case selected for the AI system.

For instance, while it is perfectly reasonable to limit the reference standard to visual features that establish the diagnosis (e.g., for obvious intracranial pathologies such as hemorrhages and midline shift),[51] in other cases, histopathological results should be used as the reference standard (e.g., to determine if a breast lesion is benign or malignant).[52] Of course, it is not always feasible to obtain histopathological results from all relevant lesions because patients with suspected benign lesions will often not undergo a biopsy. In such cases, an appropriate follow-up may serve a similar purpose. Other cases may have no final diagnosis (e.g., fractures in pediatric patients). In some instances, a fracture will be clearly visible, whereas in others, some doubts remain about whether a fissure or a bone canal is visible. For such cases multiple expert readings, establishing consensus (Figure 5), or including uncertainty estimation in the model's training or evaluation should be considered.

## Robust Input

The robustness of input refers to the resistance of input data or its derived features to varying conditions. This aspect has been widely studied in medical imaging-related AI. These varying conditions can be acquisition protocols,[53] reconstruction settings,[53] scanners,[54] annotation or segmentation variabilities,[55,56] computational factors,[57] phenotype of interest,[58] and adversarial examples (Figure 6).[59]

Ideally, only inputs and features that are robust to variations should be incorporated into the predictive models to achieve optimal generalizability.[60] Otherwise, these models may fail to predict the outcomes to a large extent.[61] In deep learning models, non-robust features are highly correlated with adversarial examples (Artuso M. et al., 2022 preprint, https://doi.org/10.48550/arXiv.2204.07285). Such a vulnerability not only poses generalizability problems but also leads to security problems.[62] Thus, feature robustness must be assessed to improve the generalizability of AI models. When identified, non-robust ones should be removed from further analysis.

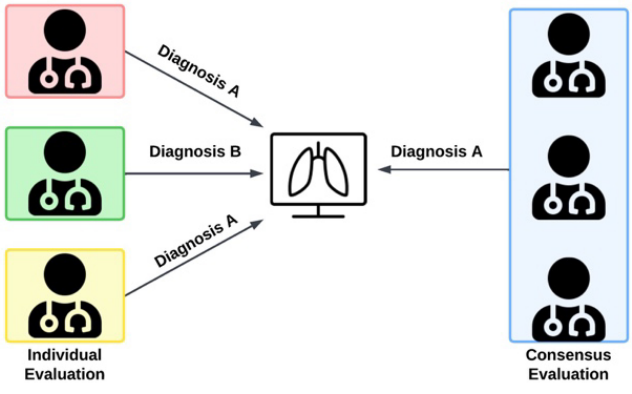


FIG. 5. Robustness of reference standard, highlighting the value of consensus evaluation over individual assessment.

The test-retest analysis is recommended for determining robustness.[56,58,60] However, because it is not a standard part of clinical practice, conducting a test-retest analysis for each research and each susceptibility factor is challenging. Additionally, it could be a partial solution because features' dependence on different factors prevents the transfer of robust information between studies.[58] An alternative to the test–retest method for robustness testing is the use of image perturbations, which enables repeated assessments without the actual acquisition of numerous images.[60]

Different harmonization solutions can be applied to achieve robust input data and features. For medical imaging, these can be evaluated in two main categories: image domain and feature domain.[63] Common methods for the image domain include standardization of image acquisition,[64,65] post-processing of raw sensor-level image data,[66] data augmentation using generative adversarial networks,[67] and style transfer.[68] For the feature domain, identification of reproducible features (e.g., annotation or segmentation reproducibility and computational reproducibility),[55,56,69] normalization techniques,[70] intensity harmonization,[71] ComBat along with its derivatives,[72] and normalization using deep learning[73] are common methods.

Deep learning models are surprisingly susceptible to adversarial attacks, in which tiny input perturbations lead to inaccurate model predictions, notwithstanding their successes in classification and regression tasks. Furthermore, medical image deep learning models are more vulnerable to adversarial attacks than natural image deep neural networks.[74] Universal adversarial perturbations can also cause misdiagnosis with a high success rate.[75] This poses a major security threat to medical deep learning models because an attacker can alter the output of the network.[59] Several defense strategies have been proposed to reduce model sensitivity to adversarial examples, such as detection methods,[76] defensive distillation,[77] adversarial training, and use of simpler models.[59] Adversarial training is considered one of the most effective defense techniques.[59] Recent works have argued that the existence of robust and non-robust features is a primary cause of adversarial examples (Ilyas A. et al., 2019 preprint, https://doi.org/10.48550/arXiv.1905.02175). In this respect, to achieve adversarial robustness, several methods have been described to distill robust and non-robust features. Nevertheless, attaining adversarial robustness of deep neural networks remains an ongoing research effort.

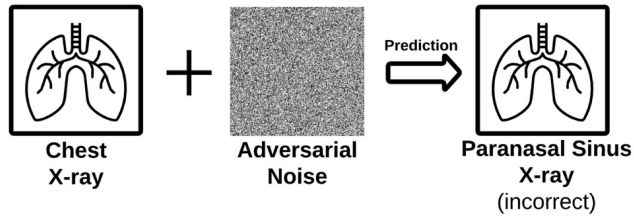


FIG. 6. Robustness of input. Input data should be robust even for external or adversarial attacks. Adding noise that is not discernable by the human eye can lead to wrong prediction by artificial intelligence models.

### No Information Leakage

When training AI models, information leakage (i.e., data leakage or feature leakage) must be avoided. These terms refer to circumstances in which information that would not be available at the time of prediction is made available during the model training process.[78] Researchers must be cautious in preventing leakage of information from data used for testing and validation of an AI model.

Before any other steps are taken, one of the most important steps is to perform dataset splitting into training, validation, and testing (Figure 7). Any preprocessing should be performed solely on the training dataset, and all steps should be recorded to be later applied before testing or validating the model's performance. If dataset splitting is conducted only after preprocessing, the information that should only be contained in the testing dataset leaks to the training of the model through the common preprocessing step. Similarly, if augmentations such as oversampling of the underrepresented class are performed before dataset splitting, researchers risk including oversampled cases in both the training and testing/validation datasets. Lastly, the same applies when a single patient contributes multiple cases to an AI project, and various cases from the same patient are distributed to training and testing/validation during dataset splitting. To prevent this, researchers should carefully distribute cases on a per-patient basis to only one of the dataset splits. Interestingly, in the initial publication of the aforementioned chest X-ray study, only approximately 30,000 patients contributed to over 100,000 cases, but dataset splitting did not consider that distribution should be performed on a per-patient basis. This was later amended in a revision of the study.[79]

Sometimes, information leakage can be very difficult to exclude because subtle information that is not immediately visible to the researchers or is present in the data but unrelated to the used case may be detected by AI models. A typical example of such details may be the subtle differences in image characteristics between different scanners (e.g., different dedicated computed tomography scanners used for outpatients vs. intensive care unit patients - the AI might pick up on the differences in image characteristics and use them as a predictor for more critical conditions as opposed to the images themselves) or variations in radiodense markers included in the image (e.g., in chest X-rays, a "PA" [posterior-anterior] marker may be interpreted as decreasing the probability of pneumonia as opposed to an "AP" marker used ).

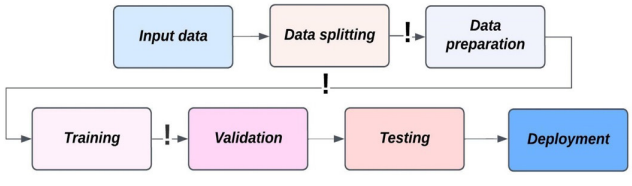


FIG. 7. Information leakage. All data splits should be done before any other step. Exclamation marks indicate potential information leakage zones.

### Optimal Bias-variance Tradeoff

Bias is the difference between the model's prediction and the correct outcome, with a preference for a certain direction. Variance refers to the inconsistency of predictions. Bias can be related to overall model accuracy on historical data, whereas variance to the stability in performance on future data.[19] Bias has an inverse relationship with variance, and vice versa, which is called the bias-variance tradeoff.[80] Briefly, very precise models in training could yield unexpectedly high prediction errors on unseen data, which indicates low bias and high variance. On the contrary, less precise ones in training could perform and generalize well on unseen data, which means high bias and low variance.

To gain more insights regarding bias and variance, researchers should be familiar with the concept of under- and overfitting.[80,81] A high bias leads to underfitting, which means that a model may miss real relationships between the features and the outcome. Underfitting can be detected when the results on the training set are not improving when learning from the present data. By contrast, a high variance leads to overfitting, which means capturing false relationships due to noise or unrelated patterns (e.g., confounders and outliers) between the features and the outcome. Overfitting can be detected when the performance on the training data improves, whereas it deteriorates on previously unseen data.

Although the bias-variance tradeoff is a key concept of the AI field, this classical concept also appears to be at odds with modern ML practice.[82] For instance, in today's practice, very complex models such as deep neural networks are developed to exactly fit the data. These models could be considered overfitted from a classical perspective. However, they usually achieve very high accuracy on unseen test data. In this respect, some authors suggest that classical understanding and modern practice can be reconciled within a unified performance curve (Figure 8).[82]

The ultimate purpose of any ML algorithm is to find the optimal point between bias and variance, which is the key to achieving the most generalizable model. An optimal model should have as low bias and variance as possible. Bias-variance tradeoff is affected by

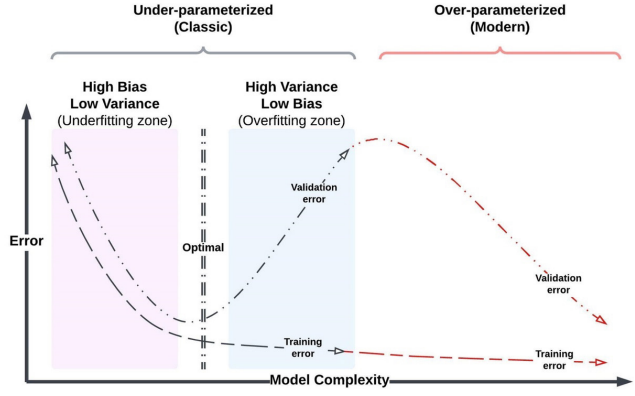


FIG. 8. Classical and modern bias–variance tradeoff.

model complexity that is mainly related to model type, number of instances, number of parameters, and number of features. There is no direct measure of bias and variance. However, to achieve the optimal tradeoff, one can retrain a model several times and measure the performance by partitioning the dataset during model development. To gain a more honest assessment of performance, this assessment should be conducted using development or validation set, but not the test set.[83] To achieve the optimal bias-variance tradeoff, the following strategies can be utilized: early stopping with cross-validation or nested cross-validation, simpler models with fewer parameters (e.g., ResNet18 over ResNet50 or random forest over XGBoost), dimensionality reduction (drop out, feature selection, etc.), data expansion with data-augmentation strategies, appropriate selection of loss functions, regularization techniques, hyperparameter optimization, and use of transfer and ensemble learning.[84]

## Proper Model Evaluation

The evaluation of an ML model presupposes the presence of a test set, distinct from the one on which it was trained, to obtain an unbiased estimate of its generalization performance, i.e., predictive performance on future, unknown data. In this regard, the test set is substantially different from the validation set, although these terms are sometimes used interchangeably. The latter represents the dataset used to select the optimal feature subset or hyperparameters (tuning parameters of an ML algorithm), often through a cross-validation approach. Only when the optimal pipeline of the model has been identified through this process that the model's performance should be evaluated on an external test set (Figure 9) (Raschka S. 2020 preprint, https://doi.org/10.48550/arXiv.1811.12808).[85]

A comprehensive summary of appropriate proper accuracy metrics in relation to a specific model should always be reported in clinical AI research. Several methods can be employed to assess model performance. The *confusion matrix* often represents the basis from which the accuracy metrics of a classification model are obtained. It consists of a matrix in which actual versus predicted outputs are presented. From the *confusion matrix*, several metrics derive *accuracy* (correctly predicted data out of the total), *precision* (percentage of positive instances out of the total predicted positive instances, i.e., positive predictive value), *recall* or *sensitivity* (percentage of positive instances out of the total actual positive instances), *specificity* (percentage of negative instances out of the total actual negative instances), and *F1 score* (harmonic mean of precision and recall). From the output of probabilistic models, the precision-recall and receiver operating characteristic curves can be built, with their respective *area under the curve*, which is another frequently employed metric in this setting.[86] The *logarithmic loss* is a further performance index of a classification model in which the prediction input consists of a probability value between 0 and 1. Besides predicting a class label, obtaining a probability of the respective label can be extremely useful to estimate the confidence level of the prediction. Calibration curves, which plot the true frequency of the positive label against its predicted probability, are available for this purpose. Reporting uncertainty metrics such as confidence intervals and standard deviation is extremely important. As regards regression models, their specific evaluation metrics include the *mean squared error* (the average of squared differences between the predicted and the actual outputs), $R^2$ *coefficient* (the amount of variance in the predictions explained by the dataset) (Botchkarev A. 2018 preprint, https://doi.org/10.48550/arXiv.1809.03006), and *explained variance* [the proportion of the variability of the predictions (i.e., how much variance can be explained by the model)]. Notably, if the error of the predictor is unbiased, the $R^2$ *coefficient* and *explained variance* are the same.

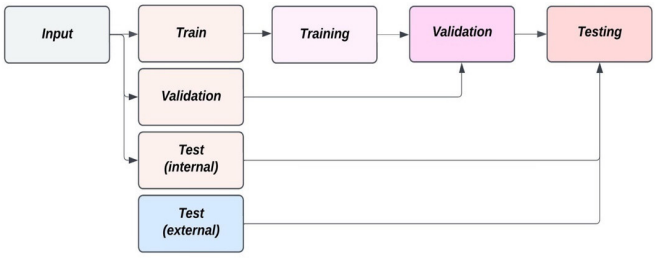


FIG. 9. Optimal data split for proper performance evaluation.

## Proven Clinical Utility

Once model accuracy metrics have been obtained, it is critical to demonstrate the clinical utility of the developed AI application to bridge the development-to-implementation gap to avoid overemphasizing the technical aspects of the proposed algorithms while losing sight of the possible benefits from a clinical perspective. The specific difficulties encountered when deciding to introduce AI-based clinical decision support systems should also be considered, including the frequent lack of the explanability of the model, the so-called black box problem, and the possibility of generating sometimes unexpected results. These elements may contribute to the algorithmic aversion by clinicians, further exacerbated by the ambiguity of who should be responsible for the model's decisions.[87] Thus, bringing these solutions to the patient's bedside can be an extremely complex task.[88] The starting point is definitely to compare standard clinical practices with and without the proposed AI-based decision support system, and this should be addressed in any clinical research relevant to AI to assess its feasibility before simulating real-world conditions in a multi-stage evaluation approach.[89] Embedding the developed model in the clinical environment and not merely providing model outputs are essential. For instance, in a radiomics study, comparing the radiologist's ability to classify different entities with that of the algorithm, but more importantly with that resulting from using a hybrid approach (radiologist with software assistance), may be appropriate.[90-92] Indeed, although the majority of AI clinical studies have focused on a direct comparison of AI with humans, real-life clinical practice is more likely to involve humans actively collaborating with AI systems (Figure 10).[93]

Another key aspect to consider, and only addressed in a minority of clinical research pertaining to AI, is the economic value of

clinical AI. Specifically, the cost-effectiveness ratio, which is the main outcome of the health technology assessment methodology, represents the primary metrics and consists of summed incremental health outcomes divided by the incremental costs associated with using the intervention under consideration.[94] However, acquiring data on AI-associated health outcomes is challenging. As most evidence on clinical AI performance comes from retrospective studies, AI performance is often compared with clinician performance unrealistically, and the effects of AI on clinician productivity are uncertain. These difficulties represent an important opportunity for health economists, who should be prepared to examine AI data collection and methods that may affect AI's future value.

### Transparent Reporting

Presenting experimental details and results with sufficient thoroughness remains an issue in AI research within the medical field.[95] This limitation is relevant as it hinders the build-up of trust in physicians and ultimately patients, limiting clinical adoption of tools based on ML technologies. Accordingly, several entities, including scientific societies, journal editorial offices, and domain experts, have attempted to set common reporting standards for AI studies[61,96-99] These have taken the form of white/position papers or checklists, the second of which may include a quantitative methodological quality assessment, as in the case of the Radiomics Quality Score.[61]

To understand what the current state of the art is, assessing the situation in medical imaging can be useful. This healthcare domain represents one of the fields with more potential applications, such as image quality improvement, automated lesion detection and/or segmentation, pathology characterization, and prediction of clinical outcomes based on imaging data.[100] However, the exponential growth in the number of publications and commercial products has not been matched by an equal increase in the quality or transparency of study methodology.[101,102] This is supported by a recent survey of all systematic review papers using the Radiomics Quality Score to assess methodological quality and transparency in medical imaging. Of the 44 included articles, each evaluating an average of 32 research papers, the median score was 21%, with a stable trend over the years (ranging from 2018 to 2021).

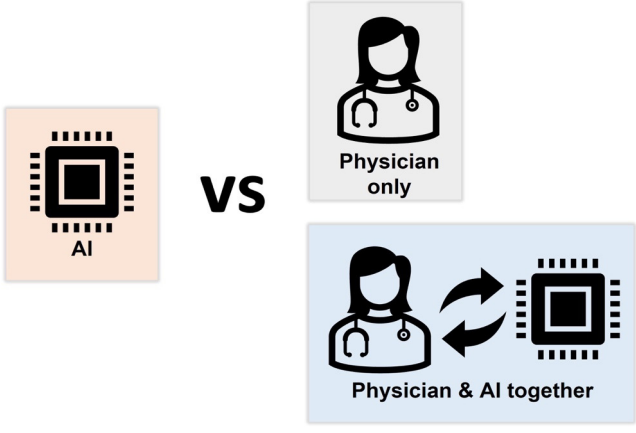


FIG. 10. Clinical utility. Although AI models are usually compared with physicians only, more emphasis should be given to the collaboration of AI and physicians, being closer to real-world clinical practice.

*AI, artificial intelligence*

The lack of transparency is not a novelty in science, and other research fields have gone through reproducibility or replicability crises, with psychology representing one of the most notable recent examples.[103] Even assuming good faith from all actors in the research field, several potential causes for this situation are still possible.[104] In the future, these limitations should be examined to obtain insights on how to avoid repeating the same errors as certainly possible for AI in healthcare.[105] Increasing journal article transparency requirements certainly represents one of the viable solutions to increase study replicability.

However, while detailed methodological reporting (i.e., sufficient detail to exactly reproduce a scientific experiment) should be expected from any single paper, this should not represent the final endpoint in the quest for scientific transparency or robustness. The replicability of the experiment using different data and/or experimental setups (i.e*., inferential replicability*) may be of greater value in developing a more robust theory behind the use of ML in healthcare.[106]

### Open Science

The concept of "open" science is based on the premise of incentivizing public sharing of research data, either raw or processed, experimental methods and results (e.g., trained ML models and/or related code), and freely accessible papers. Intuitively, this should facilitate the development of large datasets that can be the basis for better-performing ML models and easier translation to clinical practice.[107] In healthcare, this is mostly materialized through efforts in building public repositories of data, freely accessible to researchers. Some notable examples are represented by the Genomic Data Commons, National Cancer Institute Imaging Data Commons, Cancer Research Data Commons, and the Cancer Imaging Archive.[108-110] However, several hurdles and issues pertaining to open science practice in healthcare should be acknowledged and accounted for by researchers who are either planning on sharing their datasets or employing already publicly available ones.

Generally, researchers must account for the varied nature of healthcare data, which may be considered more or less challenging to share based on different local legal frameworks. For example, the use of genomic data is extremely restricted under South African legislation,[107] and the European Union's General Data Protection Regulation may have yet unforeseen implications on data-sharing practices.[111] This issue is further compounded by the fact that legislation specifically regarding medical (and ethical) use of AI is not yet well established and can be expected to further evolve over the years as awareness of potential biases and experience on practical implementation grows.[112,113]

From a different perspective, researchers should also consider the potential risks derived from the public sharing of biological data.

Openly accessible information is, by definition, also available to malevolent entities. While this is not an issue affecting all types of patient data equally, it would also be an error not to consider edge cases (e.g., data on infective agents) given the potential risks entailed.[114] Similarly, a lack of quality control or unknown biases in public data may lead to undetected, undesirable issues in models built using these datasets. Problems have often emerged after external auditing, which may not be easily detectable from researchers with less domain-specific knowledge (e.g., ML researchers using public imaging data).[39,40]

Potential misuse is partly tied to misinformation in alternative avenues of article accessibility. In recent years, the use of preprint repositories, either prior to submitting an article to a traditional journal or bypassing the peer review and editorial process entirely, has increased.[115] While this practice has its benefits, as it speeds up the dissemination of novel scientific ideas, it also presents potential limitations in the quality of the presented information. This issue variably affects preprint servers because of different policies employed, but these may not be well known to the general public accessing the papers.[114]

In conclusion, for the critical evaluation of clinical AI research, we believe that knowledge of fundamental characteristics is of utmost relevance. In this context, we discussed a selection of the essential qualities of clinical AI research: valid scientific purpose, high-quality data set, robust reference standard, robust input, no information leakage, optimal bias-variance tradeoff, proper model evaluation, proven clinical utility, transparent reporting, and open science. Although it was not possible to cover all important concepts, we hope that this work may provide a fresh perspective for general readers and thus improve their AI literacy and critical thinking.

**Author Contributions:** Concept – B.K.; Design – B.K.; Literature Review – L.U., A.S., R.C., D.P.S., B.K.; Writing – L.U., A.S., R.C., D.P.S., B.K.

**Conflict of Interest:** D.P.S receives consulting fees from Cook Medical and payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Bayer. Other authors have nothing to declare.

**Funding:** No funding was received for this study.

## REFERENCES

1. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016. at <http://www.deeplearningbook.org>. **[CrossRef]**
2. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309:1351-1352. **[CrossRef]**
3. Li T, Huang H, Zhang S, et al. Predictive models based on machine learning for bone metastasis in patients with diagnosed colorectal cancer. *Front Public Health*. 2022;10:984750. **[CrossRef]**
4. Lee S, Elton DC, Yang AH, et al. Fully Automated and Explainable Liver Segmental Volume Ratio and Spleen Segmentation at CT for Diagnosing Cirrhosis. *Radiol Artif Intell.* 2022;4:e210268. **[CrossRef]**
5. Pickhardt PJ, Nguyen T, Perez AA, et al. Improved CT-based Osteoporosis Assessment with a Fully Automated Deep Learning Tool. *Radiol Artif Intell.* 2022;4:e220042. **[CrossRef]**
6. Johri AM, Singh KV, Mantelsla LE, et al. Deep learning artificial intelligence framework for multiclass coronary artery disease prediction using combination of conventional risk factors, carotid ultrasound, and intraplaque neovascularization. *Comput Biol Med.* 2022;150:106018. **[CrossRef]**
7. Jamthikar AD, Gupta D, Mantella LE, et al. Multiclass machine learning vs. conventional calculators for stroke/CVD risk assessment using carotid plaque predictors with coronary angiography scores as gold standard: a 500 participants study. *Int J Cardiovasc Imaging.* 2021;37:1171-1187. **[CrossRef]**
8. Deng J, He Z. Characterizing Risk of In-Hospital Mortality Following Subarachnoid Hemorrhage Using Machine Learning: A Retrospective Study. *Front Surg.* 2022;9:891984. **[CrossRef]**
9. Zhu F, Pan Z, Tang Y, et al. Machine learning models predict coagulopathy in spontaneous intracerebral hemorrhage patients in ER. *CNS Neurosci Ther.* 2021;27:92-100. **[CrossRef]**
10. Wang H, Liu Y, Xu N, et al. Development and validation of a deep learning model for survival prognosis of transcatheter arterial chemoembolization in patients with intermediate-stage hepatocellular carcinoma. *Eur J Radiol.* 2022;156:110527. **[CrossRef]**
11. Hu G, Hu X, Yang K, et al. Radiomics-Based Machine Learning to Predict Recurrence in Glioma Patients Using Magnetic Resonance Imaging. *J Comput Assist Tomogr.* 2022;doi:10.1097/RCT.0000000000001386. **[CrossRef]**
12. Pandiyan S, Wang L. A comprehensive review on recent approaches for cancer drug discovery associated with artificial intelligence. *Comput Biol Med.* 2022;150:106140. **[CrossRef]**
13. Bluemke DA, Moy L, Bredella MA, et al. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers-From the Radiology Editorial Board. *Radiology.* 2020;294:487-489. **[CrossRef]**
14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44-56. **[CrossRef]**
15. Helman S, Terry MA, Pellathy T, et al. Engaging clinicians early during the development of a graphical user display of an intelligent alerting system at the bedside. *Int J Med Inf.* 2022;159:104643. **[CrossRef]**
16. Nakaura T, Higaki T, Awai K, Ikeda O, Yamashita Y. A primer for understanding radiology articles about machine learning and deep learning. *Diagn Interv Imaging.* 2020;101:765-770. **[CrossRef]**
17. Tan T-E, Xu X, Wang Z, Liu Y, Ting DSW. Interpretation of artificial intelligence studies for the ophthalmologist. *Curr Opin Ophthalmol.* 2020;31:351-356. **[CrossRef]**
18. Pineau J, Vincent-Lamarre P, Sinha K, et al. Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program). *J Mach Learn Res.* 2021;22:1-20. **[CrossRef]**
19. Al-Zaiti SS, Alghwiri AA, Hu X, et al. A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML). *Eur Heart J - Digit Health.* 2022;3:125-140. **[CrossRef]**
20. Kahn CR. Picking a Research Problem - The Critical Decision. *N Engl J Med.* 1994;330:1530-1533. **[CrossRef]**
21. Vandenbroucke JP, Pearce N. From ideas to studies: how to get ideas and sharpen them into research questions. *Clin Epidemiol.* 2018;10:253-264. **[CrossRef]**
22. Fandino W. Formulating a good research question: Pearls and pitfalls. *Indian J Anaesth.* 2019;63:611-616. **[CrossRef]**
23. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am J Med.* 2018;131:129-133. **[CrossRef]**
24. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Decis Mak.* 2021;21:125. **[CrossRef]**
25. Karekar SR, Vazifdar AK. Current status of clinical research using artificial intelligence techniques: A registry-based audit. *Perspect Clin Res.* 2021;12:48-52. **[CrossRef]**
26. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ.* 2020;368:l6927. **[CrossRef]**
27. Pinto Dos Santos D, Dietzel M, Baessler B. A decade of radiomics research: are images really data or just patterns in the noise? *Eur Radiol.* 2021;31:1-4. **[CrossRef]**
28. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. *NPJ Digit Med.* 2019;2:77. **[CrossRef]**

29. van de Sande D, Van Genderen ME, Smit JM, et al. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health Care Inform.* 2022;29:e100495. [CrossRef]
30. Cascini F, Beccia F, Causio FA, Melnyk A, Zaino A, Ricciardi W. Scoping review of the current landscape of AI-based applications in clinical trials. *Front Public Health.* 2022;10:949377. [CrossRef]
31. Harrer S, Shah P, Antony B, Hu J. Artificial Intelligence for Clinical Trial Design. *Trends Pharmacol Sci.* 2019;40:577-591. [CrossRef]
32. Weissler EH, Naumann T, Andersson T, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials.* 2021;22:537. [CrossRef]
33. Hutson M. Could AI help you to write your next paper? *Nature.* 2022;611:192-193. [CrossRef]
34. Extance A. How AI technology can tame the scientific literature. *Nature.* 2018;561:273-274. [CrossRef]
35. Mottaghy FM, Hertel F, Beheshti M. Will we successfully avoid the garbage in garbage out problem in imaging data mining? An overview on current concepts and future directions in molecular imaging. *Methods San Diego Calif.* 2021;188:1-3. [CrossRef]
36. An C, Park YW, Ahn SS, Han K, Kim H, Lee S-K. Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results. *PLOS ONE.* 2021;16:e0256152. [CrossRef]
37. Jin K, Huang X, Zhou J, et al. FIVES: A Fundus Image Dataset for Artificial Intelligence based Vessel Segmentation. *Sci Data.* 2022;9:475. [CrossRef]
38. Sunoqrot MRS, Saha A, Hosseinzadeh M, Elschot M, Huisman H. Artificial intelligence for prostate MRI: open datasets, available applications, and grand challenges. *Eur Radiol Exp.* 2022;6:35. [CrossRef]
39. Oakden-Rayner L. Exploring Large-scale Public Medical Image Datasets. *Acad Radiol*. 2020;27:106-112. [CrossRef]
40. Cuocolo R, Stanzione A, Castaldo A, De Lucia DR, Imbriaco M. Quality control and whole-gland, zonal and lesion annotations for the PROSTATEx challenge public dataset. *Eur J Radiol.* 2021;138:109647. [CrossRef]
41. Elmore JG, Lee CI. Data Quality, Data Sharing, and Moving Artificial Intelligence Forward. *JAMA Netw Open.* 2021;4:e2119345. [CrossRef]
42. Fang Y, Wang J, Ou X, et al. The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. *Phys Med Biol.* 2021;66:185012. [CrossRef]
43. D'souza RN, Huang P-Y, Yeh F-C. Structural Analysis and Optimization of Convolutional Neural Networks with a Small Sample Size. *Sci Rep.* 2020;10:834. [CrossRef]
44. Zaki G, Gudla PR, Lee K, et al. A Deep Learning Pipeline for Nucleus Segmentation. *Cytom Part J Int Soc Anal Cytol.* 2020;97:1248-1264. [CrossRef]
45. Sanford TH, Zhang L, Harmon SA, et al. Data Augmentation and Transfer Learning to Improve Generalizability of an Automated Prostate Segmentation Model. *AJR Am J Roentgenol.* 2020;215:1403-1410. [CrossRef]
46. Stanzione A, Cuocolo R, Ugga L, et al. Oncologic Imaging and Radiomics: A Walkthrough Review of Methodological Challenges. *Cancers*. 2022;14:4871. [CrossRef]
47. Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. *EBioMedicine.* 2021;67:103358. [CrossRef]
48. Oakden-Rayner L. Exploring the ChestXray14 dataset: problems. 2017;at <https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/>. [CrossRef]
49. Oakden-Rayner L. CheXNet: an in-depth review. 2018;at <https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/>. [CrossRef]
50. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *2017 IEEE Conf Comput Vis Pattern Recognit CVPR* 2017. p. 3462-3471.doi:10.1109/CVPR.2017.369. [CrossRef]
51. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet.* 2018;392:2388-2396. [CrossRef]
52. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature.* 2020;577:89-94. [CrossRef]
53. Meyer M, Ronald J, Vernuccio F, et al. Reproducibility of CT Radiomic Features within the Same Patient: Influence of Radiation Dose and CT Reconstruction Settings. *Radiology*. 2019;293:583-591. [CrossRef]
54. Mackin D, Fave X, Zhang L, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. *Invest Radiol*. 2015;50:757-765. [CrossRef]
55. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O. Influence of segmentation margin on machine learning-based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas. *Eur Radiol.* 2019;29:4765-4775. [CrossRef]
56. Kocak B, Durmaz ES, Kaya OK, Ates E, Kilickesmez O. Reliability of Single-Slice–Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility. *Am J Roentgenol.* 2019;213:377-383. [CrossRef]
57. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys.* 2017;44:1050-1062. [CrossRef]
58. van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomogr Ann Arbor Mich.* 2016;2:361-365. [CrossRef]
59. Rodriguez D, Nayak T, Chen Y, Krishnan R, Huang Y. On the role of deep learning model complexity in adversarial robustness for medical images. *BMC Med Inform Decis Mak.* 2022;22:160. [CrossRef]
60. Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. *Sci Rep.* 2019;9:614. [CrossRef]
61. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14:749-762. [CrossRef]
62. Wang X, Li J, Kuang X, Tan Y, Li J. The security of machine learning in an adversarial setting: A survey. *J Parallel Distrib Comput.* 2019;130:12-23. [CrossRef]
63. Mali SA, Ibrahim A, Woodruff HC, et al. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *J Pers Med.* 2021;11:842. [CrossRef]
64. Research C for DE and. Clinical Trial Imaging Endpoint Process Standards Guidance for Industry. *US Food Drug Adm* 2020;at <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-imaging-endpoint-process-standards-guidance-industry>. [CrossRef]
65. Sachs PB, Hunt K, Mansoubi F, Borgstede J. CT and MR Protocol Standardization Across a Large Health System: Providing a Consistent Radiologist, Patient, and Referring Provider Experience. *J Digit Imaging.* 2017;30:11-16. [CrossRef]
66. Lee H, Huang C, Yune S, Tajmir SH, Kim M, Do S. Machine Friendly Machine Learning: Interpretation of Computed Tomography Without Image Reconstruction. *Sci Rep.* 2019;9:15540. [CrossRef]
67. Modanwal G, Vellal A, Buda M, Mazurowski MA. MRI image harmonization using cycle-consistent generative adversarial network. *Med Imaging 2020 Comput-Aided Diagn* SPIE; 2020;259-264. [CrossRef]
68. Fetty L, Bylund M, Kuess P, et al. Latent space manipulation for high-resolution medical image synthesis via the StyleGAN. *Z Für Med Phys.* 2020;30:305-314. [CrossRef]
69. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PloS One.* 2019;14:e0213459. [CrossRef]
70. Haga A, Takahashi W, Aoki S, et al. Standardization of imaging features for radiomics analysis. *J Med Investig JMI.* 2019;66:35-37. [CrossRef]
71. Masson I, Da-ano R, Lucia F, et al. Statistical harmonization can improve the development of a multicenter CT-based radiomic model predictive of nonresponse to induction chemotherapy in laryngeal cancers. *Med Phys.* 2021;48:4099-4109. [CrossRef]
72. Fortin J-P, Parker D, Tunç B, et al. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage.* 2017;161:149-170. [CrossRef]
73. Dinsdale NK, Jenkinson M, Namburete AIL. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage.* 2021;228:117689. [CrossRef]

74. Ma X, Niu Y, Gu L, et al. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognit* 2021;110:107332. **[CrossRef]**
75. Hirano H, Minagi A, Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med Imaging.* 2021;21:9. **[CrossRef]**
76. Xu W, Evans D, Qi Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. *Proc 2018 Netw Distrib Syst Secur Symp* 2018. doi:10.14722/ndss.2018.23198. **[CrossRef]**
77. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. *2016 IEEE Symp Secur Priv SP.* 2016;582-597. **[CrossRef]**
78. Wikipedia. Leakage (machine learning). *Wikipedia* at <https://en.wikipedia.org/wiki/Leakage_(machine_learning)>. **[CrossRef]**
79. Roberts N. Were you concerned that the network could memorize patient anatomy since patients cross train and validation? "ChestX-ray14 dataset contains 112,120 frontal-view X-ray images of 30,805 unique patients. We randomly split the entire dataset into 80% training, and 20% validation." 2017;at <https://twitter.com/nizkroberts/status/931121395748270080>. **[CrossRef]**
80. Geman S, Bienenstock E, Doursat R. Neural Networks and the Bias/Variance Dilemma. *Neural Comput.* 1992;4:1-58. **[CrossRef]**
81. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2009. **[CrossRef]**
82. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc Natl Acad Sci U S A.* 2019;116:15849-15854. **[CrossRef]**
83. Lever J, Krzywinski M, Altman N. Model selection and overfitting. *Nat Methods.* 2016;13:703-704. **[CrossRef]**
84. Zhang K, Khosravi B, Vahdati S, et al. Mitigating Bias in Radiology Machine Learning: 2. Model Development. *Radiol Artif Intell.* 2022;4:e220010. **[CrossRef]**
85. Parmar C, Barry JD, Hosny A, Quackenbush J, Aerts HJWL. Data Analysis Strategies in Medical Imaging. *Clin Cancer Res Off J Am Assoc Cancer Res.* 2018;24:3492-3499. **[CrossRef]**
86. Handelman GS, Kok HK, Chandra RV, et al. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *AJR Am J Roentgenol*. 2019;212:38-43. **[CrossRef]**
87. Aristidou A, Jena R, Topol EJ. Bridging the chasm between AI and clinical implementation. *Lancet Lond Engl*. 2022;399:620. **[CrossRef]**
88. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med.* 2021;27:186-187. **[CrossRef]**
89. Parliament E, Services D-G for PR, Lekadir K, Quaglio G, Tselioudis Garmendia A, Gallin C. Artificial intelligence in healthcare : applications, risks, and ethical and societal impacts. *European Parliament.* 2022. doi:doi/10.2861/568473. **[CrossRef]**
90. Sim Y, Chung MJ, Kotter E, et al. Deep Convolutional Neural Network–based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs. *Radiology.* 2020;294:199-209. **[CrossRef]**
91. Bai HX, Wang R, Xiong Z, et al. Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. *Radiology*. 2020;296:E156-E165. **[CrossRef]**
92. Romeo V, Maurea S, Cuocolo R, et al. Characterization of Adrenal Lesions on Unenhanced MRI Using Texture Analysis: A Machine-Learning Approach. *J Magn Reson Imaging.* 2018;48:198-204. **[CrossRef]**
93. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28:31-38. **[CrossRef]**
94. Hendrix N, Veenstra DL, Cheng M, Anderson NC, Verguet S. Assessing the Economic Value of Clinical Artificial Intelligence: Challenges and Opportunities. *Value Health.* 2022;25:331-339. **[CrossRef]**
95. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. *JAMA Netw Open.* 2022;5:e2233946. **[Crossref]**
96. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform.* 2021;28:e100251. **[Crossref]**
97. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell*. 2020;2:e200029. **[Crossref]**
98. Penzkofer T, Padhani AR, Turkbey B, et al. ESUR/ESUI position paper: developing artificial intelligence for precision diagnosis of prostate cancer using magnetic resonance imaging. *Eur Radiol*. 2021;31:9567-9578. **[Crossref]**
99. Sengupta PP, Shrestha S, Berthon B, et al. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A Checklist: Reviewed by the American College of Cardiology Healthcare Innovation Council. *JACC Cardiovasc Imaging.* 2020;13:2017-2035. **[Crossref]**
100. Choy G, Khalilzadeh O, Michalski M, et al. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology.* 2018;288:318-328. **[Crossref]**
101. Spadarella G, Stanzione A, Akinci D'Antonoli T, Andreychenko A, Fanni SC, Ugga L, Kotter E, Cuocolo R. Systematic review of the radiomics quality score applications: an EuSoMII Radiomics Auditing Group Initiative. *Eur Radiol* 2022;doi:10.1007/s00330-022-09187-3. **[Crossref]**
102. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol*. 2021;31:3797-3804. **[Crossref]**
103. Lewandowsky S, Oberauer K. Low replicability can support robust and efficient science. *Nat Commun.* 2020;11:358. **[Crossref]**
104. Nuzzo R. How scientists fool themselves - and how they can stop. *Nature.* 2015;526:182-185. **[Crossref]**
105. Hullman J, Kapoor S, Nanayakkara P, Gelman A, Narayanan A. The Worst of Both Worlds: A Comparative Analysis of Errors in Learning from Data in Psychology and Machine Learning. *Proc 2022 AAAIACM Conf AI Ethics Soc* New York, NY, USA: Association for Computing Machinery; 2022;335-348. **[Crossref]**
106. Plesser HE. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Front Neuroinformatics.* 2017;11:76. **[Crossref]**
107. Staunton C, Barragán CA, Canali S, et al. Open science, data sharing and solidarity: who benefits? *Hist Philos Life Sci.* 2021;43:115. **[Crossref]**
108. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging.* 2013;26:1045-1057. **[Crossref]**
109. Fedorov A, Longabaugh WJR, Pot D, et al. NCI Imaging Data Commons. *Cancer Res.* 2021;81:4188-4193. **[Crossref]**
110. Grossman RL, Heath AP, Ferretti V, et al. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med.* 2016;375:1109-1112. **[Crossref]**
111. Mascalzoni D, Bentzen HB, Budin-Ljøsne I, et al. Are Requirements to Deposit Data in Research Repositories Compatible With the European Union's General Data Protection Regulation? *Ann Intern Med.* 2019;170:332-334. **[Crossref]**
112. Allen R, Masters D. Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. *ERA Forum.* 2020;20:585-598. **[Crossref]**
113. Regulatory divergences in the draft AI act: Differences in public and private sector obligations | Think Tank | European Parliament. at <https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729507>.**[Crossref]**
114. Smith JA, Sandbrink JB. Biosecurity in an age of open science. *PLOS Biol.* 2022;20:e3001600. **[Crossref]**
115. Rise of the preprints. *Nat Cancer.* 2020;1:1025-1026. **[Crossref]**