

GUIDELINE

Open Access



Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology

Nikos Sourlos^{1,2}, Rozemarijn Vliegenthart^{1,2}, Joao Santinha³, Michail E. Klontzas^{4,5}, Renato Cuocolo⁶, Merel Huisman⁷ and Peter van Ooijen^{2,8*} 

Abstract

Various healthcare domains have witnessed successful preliminary implementation of artificial intelligence (AI) solutions, including radiology, though limited generalizability hinders their widespread adoption. Currently, most research groups and industry have limited access to the data needed for external validation studies. The creation and accessibility of benchmark datasets to validate such solutions represents a critical step towards generalizability, for which an array of aspects ranging from preprocessing to regulatory issues and biostatistical principles come into play. In this article, the authors provide recommendations for the creation of benchmark datasets in radiology, explain current limitations in this realm, and explore potential new approaches.

Clinical relevance statement Benchmark datasets, facilitating validation of AI software performance can contribute to the adoption of AI in clinical practice.

Key Points

- Benchmark datasets are essential for the validation of AI software performance.
- Factors like image quality and representativeness of cases should be considered.
- Benchmark datasets can help adoption by increasing the trustworthiness and robustness of AI.

Keywords Benchmark dataset, Validation, Bias, Artificial intelligence (AI) software, Radiology

*Correspondence:

Peter van Ooijen

p.m.a.van.ooijen@umcg.nl

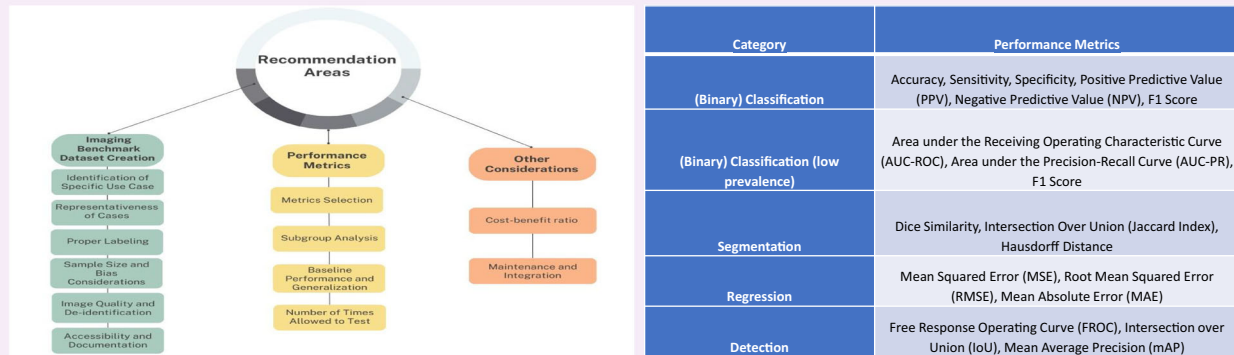
Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Graphical Abstract

Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology



Category	Performance Metrics
(Binary) Classification	Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1 Score
(Binary) Classification (low prevalence)	Area under the Receiving Operating Characteristic Curve (AUC-ROC), Area under the Precision-Recall Curve (AUC-PR), F1 Score
Segmentation	Dice Similarity, Intersection Over Union (Jaccard Index), Hausdorff Distance
Regression	Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)
Detection	Free Response Operating Curve (FROC), Intersection over Union (IoU), Mean Average Precision (mAP)

Benchmark datasets, facilitating validation of AI software performance can contribute to the adoption of AI in clinical practice.



Insights Imaging (2024) Sourlos N, Villegenthart R, Santinha J et al; DOI: 10.1186/s13244-024-01833-2

Introduction

The development of artificial intelligence (AI) algorithms in healthcare has gained significant momentum in recent years, including in radiology. In early 2023, there were more than 200 commercially available AI software solutions for radiology alone [1]. One important aspect of external validation of AI models is the creation of benchmark datasets. A benchmark dataset is a well-curated collection of expert-labeled data that represents the entire spectrum of diseases of interest and reflects the diversity of the targeted population and variation in data collection systems and methods. Such datasets are vital for validating, in the sense of establishing the reliability and accuracy of, AI models, increasing trustworthiness, and the chance of robust performance in real-world applications [2–4].

If the dataset used to develop and validate an AI algorithm is not representative of the target population, biases could arise that could have severe consequences for a large group of patients [3]. For instance, if a dataset is derived from a relatively homogenous source population from within a well-established healthcare system, the developed algorithms may not generalize effectively to, for example, a limited-resource setting with different demographic and pathophysiological features of the population. This may further amplify health inequities, potentially

leading to worse healthcare outcomes for those marginalized populations [5]. Also, algorithms developed on over-used public datasets derived from a hospital population may exhibit subpar performance if applied in a screening setting on individuals with similar demographics but different disease prevalence [6, 7]. This could lead to missed diagnoses on a large scale, especially in the light of automation bias [8]. Logullo et al [9] reviewed studies in which AI was trained to diagnose lung nodules (detect, segment, or classify them) using public datasets. They showed that 49% of their included studies used LIDC-IDRI [10] or LUNA [11] or a dataset derived from them during model development and/or validation. The characteristics of such public datasets might differ from those of the intended use case of an AI algorithm that utilized them for training/validation. For example, the volume quantification of nodules might have been derived from manual diameter measurements, which will give different results compared to fully automated measurements. In addition, these public datasets might have been preprocessed and their quality might differ from those used in clinical practice. It is therefore essential to perform further analysis to ensure the clinical utility of the dataset prior to deciding if it should be used for the particular task of interest.

It is imperative to create and enable access to benchmark datasets encompassing diverse populations and disease characteristics to validate the performance of an AI algorithm and test its generalizability. Moreover, the benchmark creation process must be transparent and rigorously documented. Furthermore, the dataset should be representative of the clinical context it is designed to address (e.g., screening and clinical diagnosis). Consequently, creating a benchmark dataset is not a straightforward task, as biases could arise in various steps in its formation process [3]. Factors to limit bias include the data sources used, anonymization steps, data format, and annotation methods.

There are initiatives to standardize infrastructure for validating AI software in imaging, enhancing transparency [1]. Furthermore, recommendations for a benchmark dataset for medical imaging in screening settings exist, but no standardized approach for clinical applications [12]. In pathology, proposals for creating test datasets to validate AI performance are already in place [13]. For more general AI solutions, it might be argued that local fine-tuning of a model and strict post-market surveillance is most efficient since data are scarce. However, before model deployment, the models' weaknesses need to be established before introduction in the clinic, especially in rare diseases.

This paper explores the key considerations in creating imaging benchmark datasets (Fig. 1) to validate the performance of AI software, addressing challenges like data quality and data heterogeneity, and emphasizing domain experts' input. Finally, it discusses metrics for evaluating model performance and provides recommendations for creating benchmark datasets in clinical practice. The primary objective of this paper is to guide the development of these datasets for AI software assessment in hospitals.

Imaging benchmark dataset creation

When developing a benchmark dataset, there are several steps to be taken [4, 12]. The following section highlights and examines the most crucial of these steps.

Identification of specific use case

It is essential to identify the specific use case(s) prior to creating a benchmark dataset. This involves considering various tasks such as object detection, binary or multiclass classification, segmentation, and regression, and their requirements (e.g., correct bounding box for detection, correct contour for segmentations, etc.). The clinical context, including the disease(s) of interest, modality, target population, and healthcare setting, should be clear, such as detecting chest X-ray abnormalities vs a normal chest X-ray in patients presenting to the emergency department of a secondary or tertiary referral

center. Furthermore, it is important to identify the most accurate ground truth labels. In many cases the expert user is regarded as the ground truth, but more on practical grounds than based on actual proof. Follow-up of patients or more extensive diagnostics are often lacking resulting in the absence of a definitive ground truth. For example, biopsy results should be preferred to clinical observations to decide if a lung nodule is malignant, but they are either not available at all (yet) or just not included in the data collection. Furthermore, in this case, the required 2-year follow-up data that could be used to confirm the benign nature of nodules is also often lacking.

Representativeness of cases

A crucial aspect to consider is the representativeness of cases encountered in clinical practice. The dataset must reflect real-world scenarios, including the disease severity spectrum, and ensure diversity in terms of demographics, vendors, and other factors.

One challenge that is difficult to address is the inclusion of rare diseases. Given their low prevalence, a very large sample size would be needed for these cases to be properly represented. Since it is commonly unfeasible to acquire a sufficiently large dataset, one proposed method is augmenting the dataset by generating synthetic data including variants of the underrepresented subsets [14]. For segmentation tasks, the inclusion of synthetic cases has been shown to lead to an improvement of the intersection over union (IoU) of up to 30% [15]. For detection tasks like that of the chromophobe subtype, synthetic histology images improved accuracy in clinical settings [16]. However, potential biases introduced by synthetic dataset heterogeneity in clinical practice are still under research [17].

Considering all the above-mentioned factors (spectrum of disease, diverse demographics, etc.) will help guarantee that the dataset is representative of the patient population and the intended clinical setting (e.g., primary care, public hospital, academic centers, or population screening).

For instance, a dataset derived from a population-based screening cohort is unsuitable for validating algorithms intended for routine computed tomography (CT) scans of the hospital population due to differences in scan protocols and disease prevalence. Validating algorithms is challenging due to clinical indication heterogeneity and incidental findings leading to new diagnoses, especially in broader clinical settings like abdominal CT scans. In these cases, there may be patients with varying indications ranging from analysis of an incidental finding to periodic oncologic follow-up [18, 19]. This is why it may be more straightforward to implement or evaluate AI techniques in highly specialized environments characterized by well-defined indications and a limited spectrum of findings,

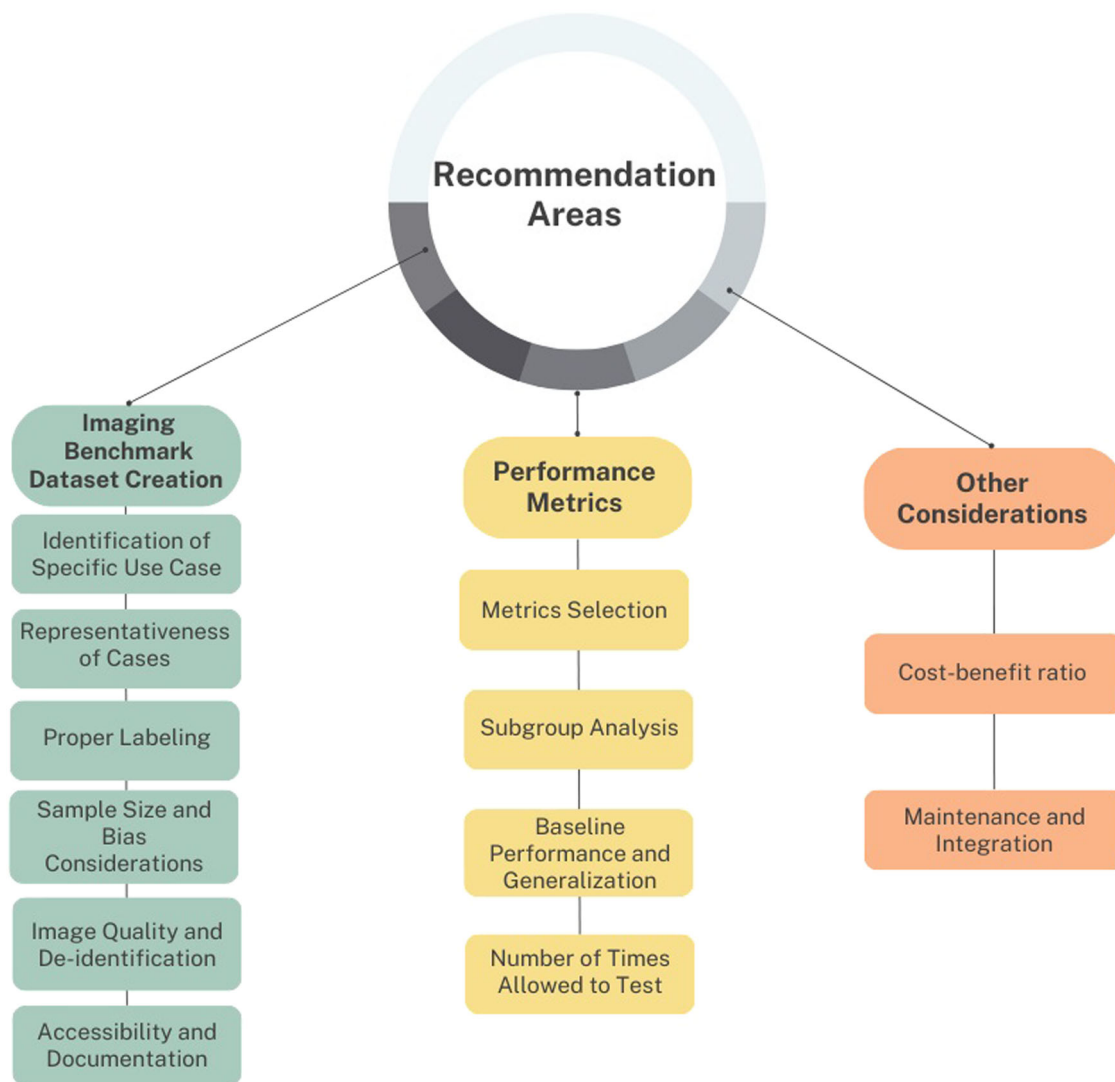


Fig. 1 Considerations for the creation of a benchmark dataset

such as mammography screening [20, 21] and prostate cancer detection on MRI [22].

An example of a non-representative dataset in terms of population characteristics is the MIMIC-CXR dataset [23, 24], which consists mainly of data from a single hospital's emergency department [6]. MIMIC-CXR is a large-scale dataset of chest X-ray images with associated radiology reports. For chest CT for lung nodule detection tasks well-known datasets are the LIDC-IDRI [10] and its derived LUNA16 [11]. Their popularity among researchers is due to being the only publicly available datasets providing lung nodule coordinates. However, AI solutions based on these datasets may have limited generalizability. A study by Li et al [25] showed that algorithms trained on independent datasets and LUNA16

maintained high performance when tested on a non-LIDC-IDRI dataset. In contrast to that study, Ahluwalia et al [6] showed that when chest radiograph classifiers are validated in a geographically and temporally different real-world dataset their diagnostic performance may drop in certain subgroups. Thus, caution must be exercised when applying a solution developed based only on, for example, the public LUNA16 dataset to real-world scenarios.

Proper labeling

The main characteristic of a well-curated benchmark dataset is that it should be properly labeled to be used as a reference standard for validation studies, ideally by having sufficiently long follow-up, or pathological proof

(biopsy and/or histology). Often, reader consensus or majority voting is taken as a proxy, since histology or cross-sectional imaging of all participants is usually not available in a retrospective study design, nor ethically feasible in a prospective setting. This (inherently imperfect) approach requires the involvement of domain experts, including radiologists. The years of experience of these experts should be considered and reported, and cases with poor interobserver agreement should be identified and analyzed for any (systematic) errors. Another consideration related to the labeling process is the types of labels that should be accompanied by proper instructions, especially when these labels are collected from different hospitals, to ensure homogeneous results. It is also crucial to decide on the annotation format like DICOM (DICOM-SEG, RTSTRUCT), NIFTI, or BIDS [26]. For ultrasound images, any image annotation format that either preserves the grayscale image or the RGB colors is sufficient [27]. Diaz et al [26] provided a comprehensive guide of (open access) data curation tools and Willeminck et al [28] presented a list of steps for pre-processing medical imaging data and explained the difficulties in data curation and data availability.

Another important consideration is the types of metadata that should be included along with the annotations. Metadata can include information such as de-identified patient demographics, relevant clinical history, etc., which can help contextualize the labeled data and provide useful information for downstream analysis. The inclusion and analysis of metadata should be done with caution since there might be correlations between metadata of different formats [29]. In addition to the above, metadata should also reflect the information available to an AI model in clinical practice, if it is to be used directly for inference in clinical cases [30]. At last, it is possible to include in metadata (like in DICOM-SEG), information on whether the labels were obtained manually, semi-automated, or fully automated using an AI algorithm, to ensure anonymity, as well as to allow the evaluation of inter and intra-observer variability. For cases with multiple binary segmentations (e.g., one from each radiologist) some approaches that can be used to select the input mask to an AI algorithm are taking the intersection of the masks, their mean, their union, or randomly selecting one of them. It is also possible to perform a majority vote on a pixel basis [31]. The above methods are two-stage approaches in which curated labels are created based on the available ones [32]. There is also the need to provide specific recommendations on how to deal with regions where radiologists are uncertain if they belong to a tumor or not [33].

Of equal importance to the type and format of metadata is the issue of data harmonization. Data collected from multiple centers is needed to enhance stability and robustness but exhibit variations in clinical and/or

imaging characteristics obtained from diverse scanners and protocols [34]. Common harmonization techniques for tabular data include standard scaling and ComBat [35], whereas histogram equalization, adaptive histogram equalization, and contrast-limited adaptive histogram equalization are commonly used to harmonize medical images [36]. There are still open research questions regarding the limitations of reproducibility of harmonization methods, especially when the variations are related to radiomic features [37]. For example, the ComBat harmonization is a statistical method developed to remove the batch effects in microarray expression. However, unlike gene expression arrays for which ComBat was designed, radiomic features have different complexity levels, which are expected to be non-uniformly affected by variations in imaging parameters [38]. Furthermore, ComBat harmonization aims only to remove the variance attributed to the batch effects while maintaining the biological information, but using ComBat to correct these effects directly on patient data without providing the correct biological covariates that actually do have an effect on radiomic feature values will lead to a loss of biological signals. This is because ComBat will assume that the variations in radiomic feature values are only attributed to the defined batch, and thus would not perform uniformly [39]. For the above reasons, ComBat corrections cannot just be applied during inference, and it rather requires both the training and test data to be processed together by a model, changing the feature values as well. Even in the case of a single participant, the entire harmonization process should be repeated from scratch, and the model would have to be retrained as well. Therefore, the ComBat method cannot deal with prospective data (impractical to be used in clinical settings), since its performance depends on variations between batches, making its use not optimal and not applicable to clinical practice [39–42]. Currently, European Horizon 2020 projects work on data harmonization methods [43]. One of them is the ChAImeleon project [44] which recently announced a challenge in which harmonized multimodality imaging and clinical data will be provided for many types of cancer, allowing development and comparison of algorithms.

Sample size and bias considerations

A benchmark dataset should be appropriately sized for the task at hand, and should consider the clinically relevant difference in effect size, and the desired level of statistical significance to be achieved. Preferably, sample size calculations are performed, although no standardized method is available for modern AI tools to date due to their complexity [45, 46]. Generic sample size calculations can be performed in cases where areas under the curve (AUCs) are calculated, with a minimum sample size for a

given AUC, confidence interval, and confidence level provided [47]. A review performed by Balki et al [46] showed that evidence of sample size calculations is scarce in AI applications in medical research. Only a few studies performed any kind of sample size analysis. Rajput et al [48] showed that to consider the sample size adequate, the classification accuracy of a model should be above 80% and the effect size should be bigger than 0.5 according to Cohen's scale. For sample size calculation of the validation dataset, Goldenholz et al [45] developed a model-agnostic open-source tool that can be used for any clinical validation study. Balki et al showed that both pre- and post-hoc methods have their strengths and weaknesses and advise that researchers should try to perform both to estimate sample sizes or consult a biostatistical when conceptualizing a study [46]. It should also be noted that the choice of sample size also depends on the algorithm that will be used. More complex models (based on deep learning (DL)) usually require more data compared to machine learning algorithms (e.g., decision trees). As traditional sample size estimations cannot derive a conclusion regarding the clinical value of a machine learning algorithm due to its complexity; tools like sample size analysis for machine learning can be useful [45]. Using this tool, by specifying the performance metrics to calculate, and some other parameters such as the required precision, accuracy, and the 'coverage probability', an estimation of the minimum sample size required to achieve metric values above a cut-off value can be provided. For machine learning solutions other than predictive models there is still no consensus on the sample size, but the more variables and the rarer the outcome, the larger the sample size needed.

The availability of resources to collect the data, and the rarity of diseases of interest, may limit the number of cases unless the dataset is augmented. The dataset's balance—whether maintaining natural disease prevalence or having equal normal and disease cases—depends on its intended use. If the dataset will be used to validate the real-world applicability of an AI algorithm, then the natural disease prevalence as present in the target population should be maintained. If the dataset's purpose is to be used to train machine learning algorithms, then a balanced dataset is preferable since otherwise a very large sample size is needed to obtain optimal performance. Furthermore, it is not guaranteed that increasing the sample size will lead to a more accurate AI algorithm, as demonstrated in the case of distinguishing various clinical conditions that could indicate the presence of prodromal Alzheimer's disease [49]. Efforts should be made to ensure that the risk of bias is low by considering possible factors of bias during dataset creation [30].

One dataset that is frequently used in the literature [50, 51] as an external validation for AI tools is the MIMIC-CXR

dataset [22, 23]. Caution should be given to the fact that it consists of single-center data and might not be representative of geographically different populations. A study by Ahluwalia et al [6] showed that if a subgroup analysis is performed, the performance of chest radiograph classifiers is dependent on patient characteristics, clinical setting, and pathology. Still, the creation of such large databases can facilitate progress in creating AI solutions that could potentially be implemented in clinical practice and should be promoted, especially given the fact that they are still largely lacking for other imaging modalities like CT, MRI, and PET/CT. Many other forms of bias can arise during the data collection and annotation phases. A detailed overview is provided in a recent review [3].

Image quality and de-identification

When creating a benchmark dataset in radiology, image quality is crucial. Images must be free of artifacts that render them undiagnostic and should be correctly pre-processed [28]. Furthermore, to ensure reproducibility, any preprocessing of the images (e.g., noise reduction, intensity normalization, or augmentation) should be thoroughly described and the software (code) used should be made available to the researchers who will perform the validation. Images in a benchmark dataset should be acquired using appropriate acquisition settings and parameters, similar to those of the intended use. Be aware that images from older scanners in open datasets might differ from current clinical practice, making them unsuitable for benchmarks. Detecting the performance drift of an algorithm that was trained with such images, can be done with different methods such as just using the scan date to exclude them or unsupervised prediction alignment [52] to correct for that drift. Other methods include checking the metadata for parameters that indicate the year of the scanner, or the image quality of the scans and confirming that it is not of low resolution, that there are no signs of degradation, and that there are normal levels of noise present. Apart from the above, data drift can also be caused by changes in clinical population (demographic or disease prevalence changes), and/or changes in clinical guidelines, diagnostic criteria, and treatment protocols used in clinical practice. Therefore, these factors should always be assessed to evaluate if a data drift occurred.'

Data privacy and security are legally required in healthcare. Protection of personal data can be achieved through different techniques like randomization (deletion of identifiers), cryptographic techniques, restricted access, etc. [53], which also must comply with relevant regulations. In the case of a restricted dataset, hosting it using privacy-preserving techniques (e.g., encryption) can ensure the protection of sensitive information.

In the European Union (EU), privacy and security laws, especially Europe's General Data Protection Regulation

(GDPR), do not allow unrestricted data sharing with other institutions to improve models. Even with de-identified metadata, it has been shown that it is for example still possible to reconstruct the face of the individual who underwent an MRI scan of the head [54]. One promising solution to the privacy preservation issue is federated learning (FL) strategies, where the model is brought to data from different institutions (and therefore heterogeneous patient data) to train and test without compromising privacy and security as the data do not leave the center's server [55, 56]. In the case of FL, as image and label quality verifications cannot be done in a centralized approach, data quality becomes the sole responsibility of the data-providing institution. At last, special caution should be taken for cases in which patient data are burnt in the DICOM images and/or secondary captures. Some methods to automate the process of removing burnt patient data exist [57], but manual intervention might still be needed to confirm the correctness of these methods.

Accessibility and documentation

An important concern when developing benchmark datasets, taking the findable, accessible, interoperable, and reusable principles into account [58], is easy accessibility for researchers. The dataset should come with a metadata file containing the information needed to access and handle data. Moreover, the manuscript describing the dataset and possible use cases should follow specific reporting guidelines appropriate to the type of application [59]. Relevant clinical and demographic information should also be made available to allow subgraph analysis [60].

Xie et al [61] used the MIMIC-IV-ED database [62] to create a publicly available benchmark dataset of electronic health records of more than 400,000 adults admitted to the emergency department of a hospital. By making such a large dataset publicly available, they stimulate other researchers and companies to use that database to develop and test their solutions. Another example of a large dataset available to researchers is the NLST dataset [63]. The dataset consists of either low-dose CT scans or chest radiographs, along with accompanying clinical data. It is

maintained and can be accessed through the cancer imaging archive [64]. Both these datasets are easily accessible and are accompanied by participants' clinical and demographic characteristics.

Performance metrics

Apart from dataset creation considerations, different aspects of the performance metrics chosen to evaluate the model should also be taken into account. Performance metrics help identify the inherent weaknesses of the model that could cause bias.

Performance metrics selection

The selection of performance metrics is crucial in assessing an algorithm's performance on a benchmark dataset in radiology. Performance metric selection depends on the model's objectives or desired outcomes, and different metrics may be more appropriate for different tasks [65, 66]. For example, metrics such as sensitivity and negative predictive value (NPV) are relevant for a dataset designed for screening purposes given the low prevalence of disease. Providing recommendations on metrics' relevance in clinical scenarios can improve dataset usage and awareness of pitfalls [65].

For most clinical tasks, multiple performance metrics should be reported to give an overall impression of the model performance, including its inherent errors given a specific clinical setting (e.g., low prevalence) [67]. It should be ensured that they provide clinically relevant information that is easily interpretable by the end-user. Table 1 shows some commonly used performance metrics and their categories.

Importantly, some of those metrics like the AUC-ROC and accuracy derived from a balanced dataset do not directly translate to low prevalence settings due to the naturally large proportion of false negatives, even in a poor classifier. It should be noted that the same metric may be referred to in different ways based on the domain it is applied to. For example, the Dice coefficient could be the same as the F1 score for segmentation, or recall could be the same as sensitivity depending on the profession of the end-user. It is also recommended to report confidence intervals since they are of high importance for

Table 1 Commonly used performance metrics and their categories [67, 96]

Category	Performance metrics
(Binary) Classification	Accuracy, sensitivity, specificity, positive predictive value, NPV, and F1 score
(Binary) Classification (low prevalence)	The area under the receiving operating characteristic curve (AUC-ROC), the area under the precision-recall curve, and the F1 score
Segmentation	Dice similarity, IoU (Jaccard Index), and Hausdorff distance
Regression	Mean squared error, root mean squared error, and mean absolute error
Detection	Free response operating curve, IoU, and mean average precision (mAP)

performance metrics in biomedical research due to the extra information they provide for the samples used [68].

Subgroup analysis

Reporting performance metrics for subgroups, such as by age, sex, or race [5, 69, 70], can help to assess bias and identify specific subgroups in which the model might underperform. It should be noted that GDPR does not allow requests from participants to declare their race unless this is the study's primary goal, limiting the possibilities for this subgroup analysis. A workaround can be using summary demographics at a group level, which has disadvantages. Tripathi et al [71] reviewed publicly accessible imaging datasets and found that there are many issues related to, among others, demographics, race, and geographic diversity of different populations.

Tools like Aequitas [72], and FUTURE-AI [73], can help to analyze the fairness and bias of models and provide guidance on how to address any issues that arise, and PROBAST-AI [30, 74] will provide guidelines on assessing the risk of bias. However, the final version of PROBAST-AI has not yet been published [59]. Regulations and recommendations on how to avoid biases can be found in the European Parliament's document for AI in healthcare [75]. Furthermore, information about specific subgroups and the data used to develop an algorithm can be provided through Model Cards [76], helping to enhance transparency and accountability in model deployment.

It is beneficial to utilize a benchmark dataset to evaluate the presence of bias within specific subgroups of the populations mentioned above. However, in addition to this approach, various techniques can be employed during the development and post-processing of the model to mitigate these biases [77] such as generative AI techniques to augment the training data. For instance, Burlina et al [78] demonstrated that by generating synthetic fundus images of the eye, the discrepancies between individuals with dark and light skin tones were minimized. Another approach is the application of adversarial methods, which not only enhance a model's performance on a specific variable of interest but also minimize the ability of a second model to correctly identify protected attributes from the features learned by the first model [79]. Li et al [80] successfully demonstrated this approach for skin lesion classification. Finally, model predictions can be calibrated across different subgroups as part of the post-processing stage. Ultimately, the effectiveness of these methods can be assessed by comparing the model's performance to the benchmark dataset, which can also include examples generated with these techniques (e.g., synthetic images).

Baseline performance and generalization

Establishing and reporting a reference (baseline) performance based on criteria set by clinicians on how well a

model should perform on a particular task of interest, can provide context for the lower bound of required performance. Comparing AI software's performance on benchmark datasets with that of radiologists or other expert clinicians reveals areas where AI or clinicians are superior, indicating the potential added value of the software. For example, for medical images, a carefully designed study, e.g., according to the multi-reader multi-case design, is recommended to establish if AI could be beneficial, although this could be very resource-intensive given the number of human readers required [81].

The comparison of the AI performance vs clinical experts is challenging due to the fact that the clinically preferred settings of the algorithm depend on the context. Efforts to create open-source datasets include the WILDS benchmark dataset [82], aiming to address naturally occurring distribution shifts (changes in imaging characteristics) in a diverse set of problems (e.g., in tumor identification tasks across different acquisition sites), BenchMD for variations across hospitals [83], and the DomainBed suite [84], consisting of multi-domain datasets, and focusing on assessing the generalizability of AI algorithms in real-world settings. Another great resource of publicly available datasets, along with their performance on a dataset of interest can be found in the papers with code website [85], and datasets focusing on medical imaging tasks in the GitHub repository of Adalca [86]. After establishing a baseline performance on an open-source dataset, a restricted-access benchmark dataset that has not been used for model development can then be utilized to get an estimate of the true performance of the developed AI algorithm in new, unseen cases.

Another way to assess the limitations of an algorithm developed on a different source population is to conduct a failure analysis [4] using a benchmark dataset. Oakden-Rayner et al [87] evaluated the performance of a DL model designed to detect fractures on X-rays. Even though the model maintained a very good performance during external validation, an algorithmic audit revealed an elevated error rate in unexpected edge cases, such as Paget's disease, along with a significant alteration in the model's operating point.

Number of times allowed to test

Finally, it is essential to consider how many times an algorithm is allowed to run on the same benchmark dataset used for external validation only [12]. Providing the dataset and allowing many evaluations of the algorithm in the benchmark dataset can increase the risk of overfitting, resulting in misleading performance results. Establishing a limit to the number of runs or providing a different fraction of the dataset in each test run can help mitigate that risk. Ideally, the benchmark dataset should not be

directly accessible to the users and the specific cases used during validation should be selected randomly each time (given that the dataset size allows that). A study by Roelofs et al [88] demonstrated, that contrary to popular belief, when a separate test set is used only once to obtain the final ranking in Kaggle competitions (although a holdout set with similar characteristics could have been used multiple times for the public ranking), there were limited indications of significant overfitting, showing that the test set could potentially be used multiple times. At last, an agreement should be achieved prior to performing the validation on where the results would be available (peer-reviewed journal, website, etc.) and ensure that they are reported correctly using the designated reporting guidelines [59].

Other considerations before creating and using a benchmark dataset

Apart from the dataset and the performance metrics used to evaluate a software's performance, other factors can affect the creation and use of a benchmark dataset. These are listed below.

Regulatory compliance

Creating a benchmark dataset requires adherence to regulations like GDPR or HIPAA [89], ensuring data privacy and security, and addressing ethical considerations such as transparency and fairness. These regulations evolve constantly (upcoming AI Act in EU [73]) necessitating regular dataset updates and maintenance. This involves allocating resources and expertise in regulatory compliance throughout the workflow, from data acquisition to reporting validation results. It is also important to thoroughly vet the privileges and access granted to the software provider when validating their software to ensure no compromise of patient privacy and security. This can also be achieved by installing the software locally and granting it access to the data offline or by using encryption. Moreover, compliance with local Institutional Review Board regulations must be achieved prior to using patient data for model development/validation.

Maintenance and integration

Providing the dataset together with technical support is essential. This includes assistance with software installation and evaluating algorithms against the benchmark. Furthermore, the dataset's interoperability with various picture archiving and communication systems, tailored for either clinical or research purposes [26], is important. A user-friendly interface with clear instructions for various actions is needed for that. Alternatively, the dataset can be distributed and securely accessed through platforms like the cancer imaging archive [90, 91].

Discussion

In recent years, numerous vendors have entered the medical imaging market with AI products to assist clinicians, and even though external validation might have been performed in a limited form in some cases [92], generalizability issues persist with CE-marked or FDA-cleared models, depending on the end-users clinical context. While recommendations on reducing biases exist [3, 59, 74, 77, 93], they do not provide a foolproof guarantee against it. Besides this, AI companies most often do not disclose what data were used exactly to train their models making it hard to compare the training data to the data used in the local clinical setting.

To deal with the absence of benchmark datasets, this publication provides valuable insights for creating such datasets, selecting relevant performance assessment metrics, and considerations on how AI software can be integrated into the clinical workflow. By addressing these aspects, the implementation of AI in radiology has the potential to become more reliable, effective, and ethically sound, ultimately leading to improved patient outcomes. Moreover, recent initiatives like the European Cancer Imaging Initiative (EUCAIM), a federated European digital infrastructure, will result in a large-scale, high-quality dataset ideal for benchmarking [43].

From a stakeholder perspective, choosing an AI software is a non-trivial task since it requires considering parameters like its diagnostic or prognostic performance, interpretability, usability, error rate, integrated workflow, turnaround time, etc., as well as providing services concerning maintenance, post-market surveillance, etc. [94]. Even though there exist publicly available imaging datasets, these cannot be used for validation of AI software since vendors might have used part of that dataset to develop their algorithm and therefore, if this dataset is used for validation it will result in overestimation of the true performance of the algorithm. Caution should also be taken during inference to apply the same preprocessing steps as those used during the training of the developed algorithm. Moreover, caution should be given to the fact that the equipment and the acquisition methods constantly improve (e.g., photon counting CT [95]) and benchmark datasets might end up being outdated at the point of release or some time afterward. Other limitations include the need for an expert opinion to establish the reference standard and possibly the fine-tuning of the parameters of the algorithm that might be required to fit those of the benchmark dataset. At last, the recent AI act in the EU [73] poses new challenges in the adaptation and use of AI solutions in clinical practice that should be considered.

Prior to creating a benchmark, it is important to consider the task in which this dataset would be used. Efforts should be made towards creating more benchmark datasets since they are essential for the validation of AI

software before it can be used in clinical practice. Furthermore, a direct comparison of the performance of different vendors on those datasets would allow clinicians to decide which software performs better on a given task.

Conclusion

In this paper, we provided detailed recommendations regarding benchmark dataset creation, aiming to assist researchers, clinicians, and data scientists in creating high-quality benchmark datasets that are reliable, diverse, and representative of real-world medical data. Ultimately, we believe that the creation of benchmark datasets will facilitate the development of more effective AI models by increasing trust in them, and potentially lead to improved patient outcomes and better healthcare delivery.

Abbreviations

AUC	Area under the curve
AUC-ROC	Area under the receiving operating characteristic curve
AI	Artificial intelligence
DL	Deep learning
EU	European Union
FL	Federated learning
GDPR	General data protection regulation
IoU	Intersection over Union
NPV	Negative predictive value

Acknowledgements

This manuscript was a collaboration of experts from the European Society of Medical Informatics (EuSoMI).

Author contributions

N.S. was responsible for creating the first version of the manuscript. M.H., J.S., M.E.K., R.C., P.v.O., and R.V. all contributed by providing useful and detailed feedback with suggestions for improvement. The final version would not be as concise and informative without each author's contribution from its field of expertise. N.S.: conceptualization, methodology, investigation, and writing—original draft. R.V.: supervision and writing—review and editing. J.S.: writing—review and editing. M.E.K.: writing—review and editing. R.C.: writing—review and editing. M.H.: validation, supervision, and writing—review and editing. P.v.O.: supervision and writing—review and editing. All authors read and approved the final manuscript.

Funding

The position of the first author is supported by a research grant from the Netherlands Organisation for Scientific Research/Innovative Medical Device Initiative.

Data availability

No dataset or any other information can be shared.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Rozemarijn Vliegenthart declares institutional research grants from Siemens Healthineers. The other authors of this manuscript declare no relationships with any companies.

Author details

¹Department of Radiology, University Medical Center of Groningen, Groningen, The Netherlands. ²DataScience Center in Health, University Medical Center Groningen, Groningen, The Netherlands. ³Digital Surgery LAB, Champalimaud Foundation, Champalimaud Clinical Centre, Lisbon, Portugal. ⁴Department of Medical Imaging, University Hospital of Heraklion, Heraklion, Greece. ⁵Department of Radiology, School of Medicine, University of Crete, Heraklion, Greece. ⁶Department of Medicine, Surgery, and Dentistry, University of Salerno, Baronissi, Italy. ⁷Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands. ⁸Department of Radiation Oncology, University Medical Center Groningen, Groningen, The Netherlands.

Received: 22 February 2024 Accepted: 20 September 2024

Published online: 14 October 2024

References

1. Radiology Health AI Register (2023) "AI for Radiology" (Online). www.aiforradiology.com. Accessed 19 Sep 2023
2. Dhar S, Shamir L (2021) Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks. *Vis Inform* 5:92–101. <https://doi.org/10.1016/j.visinf.2021.10.001>
3. Sourlos N, Wang J, Nagaraj Y, Van Ooijen P, Vliegenthart R (2022) Possible bias in supervised deep learning algorithms for CT lung nodule detection and classification. *Cancers* 14:3867. <https://doi.org/10.3390/cancers14163867>
4. Mincu D, Roy S (2022) Developing robust benchmarks for driving forward AI innovation in healthcare. *Nat Mach Intell* 4:1–6. <https://doi.org/10.1038/s42256-022-00559-4>
5. Celi LA, Cellini J, Charpignon ML et al (2022) Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digit Health* 1:e0000022. <https://doi.org/10.1371/journal.pdig.0000022>
6. Ahluwalia M, Abdalla M, Sanayei J et al (2023) The subgroup imperative: chest radiograph classifier generalization gaps in patient, setting, and pathology subgroups. *Radiol Artif Intell* 5:e220270. <https://doi.org/10.1148/ryai.220270>
7. Huisman M, Hannink G (2023) The AI generalization gap: one size does not fit all. *Radiol Artif Intell* 5:e230246. <https://doi.org/10.1148/ryai.230246>
8. Dratsch T, Chen X, Rezazade Mehrizi M et al (2023) Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* 307:e222176. <https://doi.org/10.1148/radiol.222176>
9. Logullo P, MacCarthy A, Dhiman P et al (2023) Artificial intelligence in lung cancer diagnostic imaging: a review of the reporting and conduct of research published 2018–2019. *BJR Open* 5:20220033. <https://doi.org/10.1259/bjro.20220033>
10. Armato SG, McLennan G, Bidaut L et al (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans: the LIDC/IDRI thoracic CT database of lung nodules. *Med Phys* 38:915–931. <https://doi.org/10.1118/1.3528204>
11. Setio AAA, Traverso A, de Bel T et al (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Med Image Anal* 42:1–13. <https://doi.org/10.1016/j.media.2017.06.015>
12. Chalkidou A, Shokraneh F, Kijauskaite G et al (2022) Recommendations for the development and use of imaging test sets to investigate the test performance of artificial intelligence in health screening. *Lancet Digit Health* 4:e899–e905. [https://doi.org/10.1016/S2589-7500\(22\)00186-8](https://doi.org/10.1016/S2589-7500(22)00186-8)
13. Homeyer A, Geißler C, Schwen LO et al (2022) Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology. *Mod Pathol*. <https://doi.org/10.1038/s41379-022-01147-y>
14. Shorten C, Khoshgoftaar TM (2019) A survey on Image data augmentation for deep learning. *J Big Data* 6:60. <https://doi.org/10.1186/s40537-019-0197-0>
15. Thambawita V, Salehi P, Sheshkal SA et al (2022) SinGAN-Seg: synthetic training data generation for medical image segmentation. *PLoS One* 17:e0267976. <https://doi.org/10.1371/journal.pone.0267976>
16. Alhussaini AJ, Steele JD, Nabi G (2022) Comparative analysis for the distinction of chromophobe renal cell carcinoma from renal oncocytoma in

- computed tomography imaging using machine learning radiomics analysis. *Cancers* 14:3609. <https://doi.org/10.3390/cancers14153609>
17. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F (2021) Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* 5:493–497. <https://doi.org/10.1038/s41551-021-00751-8>
 18. Kelly ME, Heeney A, Redmond CE et al (2015) Incidental findings detected on emergency abdominal CT scans: a 1-year review. *Abdom Imaging* 40:1853–1857. <https://doi.org/10.1007/s00261-015-0349-4>
 19. Berland LL, Silverman SG, Gore RM et al (2010) Managing incidental findings on abdominal CT: white paper of the ACR incidental findings committee. *J Am Coll Radiol* 7:754–773. <https://doi.org/10.1016/j.jacr.2010.06.013>
 20. Spak DA, Plaxco JS, Santiago L, Dryden RM, Dogan BE (2017) BI-RADS® fifth edition: a summary of changes. *Diagn Interv Imaging* 98:179–190. <https://doi.org/10.1016/j.diii.2017.01.001>
 21. Lång K, Josefsson V, Larsson AM et al (2023) Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol* 24:936–944. [https://doi.org/10.1016/S1470-2045\(23\)00298-X](https://doi.org/10.1016/S1470-2045(23)00298-X)
 22. Beyer T, Schlemmer H-P, Weber M-A, Thierfelder KM (2021) PI-RADS 2.1—image interpretation: the most important updates and their clinical implications. *Rofo* 193:787–796. <https://doi.org/10.1055/a-1324-4010>
 23. Johnson AEW, Pollard TJ, Greenbaum NR et al (2019) MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. Preprint at <https://doi.org/10.48550/arXiv.1901.07042>
 24. Johnson AEW, Pollard TJ, Berkowitz SJ et al (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. <https://doi.org/10.1038/s41597-019-0322-0>
 25. Li D, Mikela Vilmun B, Frederik Carlsen J et al (2019) The performance of deep learning algorithms on automatic pulmonary nodule detection and classification tested on different datasets that are not derived from LIDC-IDRI: a systematic review. *Diagnostics* 9:207. <https://doi.org/10.3390/diagnostics9040207>
 26. Diaz O, Kushibar K, Osuala R et al (2021) Data preparation for artificial intelligence in medical imaging: a comprehensive guide to open-access platforms and tools. *Phys Med* 83:25–37. <https://doi.org/10.1016/j.ejpm.2021.02.007>
 27. Larobina M, Murino L (2014) Medical image file formats. *J Digit Imaging* 27:200–206. <https://doi.org/10.1007/s10278-013-9657-9>
 28. Willemink MJ, Koszek WA, Hardell C et al (2020) Preparing medical imaging data for machine learning. *Radiology* 295:4–15. <https://doi.org/10.1148/radiol.2020192224>
 29. Kondylakis H, Ciarrocchi E, Cerda-Alberich L et al (2022) Position of the AI for health imaging (AI4HI) network on metadata models for imaging biobanks. *Eur Radiol Exp* 6:29. <https://doi.org/10.1186/s41747-022-00281-1>
 30. Wolff RF, Moons K, Riley RD et al (2019) PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 170:51–58. <https://doi.org/10.7326/M18-1376>
 31. Wang L, Ye X, Ju L et al (2022) Medical matting: a new perspective on medical segmentation with uncertainty. Preprint at <https://doi.org/10.48550/arXiv.2106.09887>
 32. Le KH, Tran TV, Pham HU, Nguyen HT, Le TT, Nguyen HQ (2022) Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. *IEEE Xplore*. <https://ieeexplore.ieee.org/document/10041153>. Accessed 12 Apr 2024
 33. De Biase A, Sijtsema NM, van Dijk LV, Langendijk JA, van Ooijen PMA (2023) Deep learning aided oropharyngeal cancer segmentation with adaptive thresholding for predicted tumor probability in FDG PET and CT images. *Phys Med Biol*. <https://doi.org/10.1088/1361-6560/acb9cf>
 34. Nan Y, Ser JD, Walsh S et al (2022) Data harmonisation for information fusion in digital healthcare: a state-of-the-art systematic review, meta-analysis and future research directions. *Int J Inf Fusion* 82:99–122. <https://doi.org/10.1016/j.inffus.2022.01.001>
 35. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118–127. <https://doi.org/10.1093/biostatistics/kxj037>
 36. Mustafa WA, Kader MMM (2018) A review of histogram equalization techniques in image enhancement application. *J Phys Conf Ser* 1019:012026. <https://doi.org/10.1088/1742-6596/1019/1/012026>
 37. Mali SA, Ibrahim A, Woodruff HC et al (2021) Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *J Pers Med* 11:842. <https://doi.org/10.3390/jpm11090842>
 38. Ibrahim A, Primakov S, Beauque M et al (2021) Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new framework. *Methods* 188:20–29. <https://doi.org/10.1016/j.jymeth.2020.05.022>
 39. Ibrahim A, Refaee T, Leijenaar R et al (2021) The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS One* 16:e0251147. <https://doi.org/10.1371/journal.pone.0251147>
 40. Ibrahim A, Widaatalla Y, Refaee T et al (2021) Reproducibility of CT-based hepatocellular carcinoma radiomic features across different contrast imaging phases: a proof of concept on SORAMIC trial data. *Cancers*. <https://doi.org/10.3390/cancers13184638>
 41. Ibrahim A, Refaee T, Primakov S et al (2021) The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization. *Cancers*. <https://doi.org/10.3390/cancers13081848>
 42. Orhac F, Eertink JJ, Cottreau AS et al (2022) A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J Nucl Med* 63:172–179. <https://doi.org/10.2967/jnumed.121.262464>
 43. EU (2024) European cancer imaging initiative. Shaping Europe's digital future. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/cancer-imaging>. Accessed 28 Nov 2023
 44. Bonmatí LM, Miguel A, Suárez A et al (2022) CHAIMELEON project: creation of a pan-European repository of health imaging data for the development of AI-powered cancer management tools. *Front Oncol* 12:742701. <https://www.frontiersin.org/articles/10.3389/fonc.2022.742701>
 45. Goldenholz DM, Sun H, Ganglberger W, Westover MB (2023) Sample size analysis for machine learning clinical validation studies. *Biomedicines* 11:685. <https://doi.org/10.3390/biomedicines11030685>
 46. Balki I, Amirabadi A, Levman J et al (2019) Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J* 70:344–353. <https://doi.org/10.1016/j.carj.2019.06.002>
 47. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143:29–36. <https://doi.org/10.1148/radiology.143.1.7063747>
 48. Rajput D, Wang W-J, Chen C-C (2023) Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics* 24:48. <https://doi.org/10.1186/s12859-023-05156-9>
 49. Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med* 5:48. <https://doi.org/10.1038/s41746-022-00592-y>
 50. Arora M, Davis CM, Gowda NR et al (2023) Uncertainty-aware convolutional neural network for identifying bilateral opacities on chest X-rays: a tool to aid diagnosis of acute respiratory distress syndrome. *Bioengineering*. <https://doi.org/10.3390/bioengineering10080946>
 51. Oliver M, Renou A, Allou N, Moscatelli L, Ferdynus C, Allyn J (2023) Image augmentation and automated measurement of endotracheal-tube-to-carina distance on chest radiographs in intensive care unit using a deep learning model with external validation. *Crit Care* 27:40. <https://doi.org/10.1186/s13054-023-04320-0>
 52. Roschewitz M, Khara G, Yearsley J et al (2023) Automatic correction of performance drift under acquisition shift in medical image classification. *Nat Commun* 14:6608. <https://doi.org/10.1038/s41467-023-42396-y>
 53. Ram Mohan Rao P, Murali Krishna S, Siva Kumar AP (2018) Privacy preservation techniques in big data analytics: a survey. *J Big Data* 5:33. <https://doi.org/10.1186/s40537-018-0141-8>
 54. Schwarz CG, Kremers WK, Therneau TM et al (2019) Identification of anonymous MRI research participants with face-recognition software. *N. Engl J Med* 381:1684–1686. <https://doi.org/10.1056/NEJMc1908881>
 55. Darzidehkalani E, Ghasemi-rad M, van Ooijen PMA (2022) Federated learning in medical imaging: part I: toward multicenter health care ecosystems. *J Am Coll Radiol* 19:969–974. <https://doi.org/10.1016/j.jacr.2022.03.015>
 56. Darzidehkalani E, Ghasemi-Rad M, van Ooijen PMA (2022) Federated learning in medical imaging: part II: methods, challenges, and considerations. *J Am Coll Radiol* 19:975–982. <https://doi.org/10.1016/j.jacr.2022.03.016>

57. Tsui GK, Chan T (2012) Automatic selective removal of embedded patient information from image content of DICOM files. *AJR Am J Roentgenol* 198:769–772. <https://doi.org/10.2214/AJR.10.6352>
58. Wilkinson MD, Dumontier M, Aalbersberg IJ et al (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>
59. Klontzas ME, Gatti AA, Tejani AS, Kahn CE (2023) AI reporting guidelines: how to select the best one for your research. *Radiol Artif Intell* 5:e230055. <https://doi.org/10.1148/ryai.230055>
60. Gebru T, Morgenstern J, Vecchione B et al (2021) Datasheets for datasets. Preprint at <https://doi.org/10.48550/arXiv.1803.09010>
61. Xie F, Zhou J, Lee JW et al (2022) Benchmarking emergency department prediction models with machine learning and public electronic health records. *Sci Data* 9:658. <https://doi.org/10.1038/s41597-022-01782-9>
62. Johnson A, Bulgarelli L, Pollard T, Celi LA, Mark R, Horng S (2023) MIMIC-IV-ED. *PhysioNet*. <https://doi.org/10.13026/5N7K-KM72>
63. T National Lung Screening Trial Research, Aberle DR, Adams AM et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 365:395–409. <https://doi.org/10.1056/NEJMoa1102873>
64. National Lung Screening Trial Research Team (2013) Data from the National Lung Screening Trial (NLST). The Cancer Imaging Archive. <https://doi.org/10.7937/TCIA.HMQ8-J677>
65. Maier-Hein, L., Reinke, A., Godau, P. et al. (2024) Metrics reloaded: recommendations for image analysis validation. *Nat Methods* 195–212. <https://doi.org/10.1038/s41592-023-02151-z>
66. Park SH, Han K, Jang HY et al (2023) Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* 306:20–31. <https://doi.org/10.1148/radiol.220182>
67. Hicks SA, Strümke I, Thambawita V et al (2022) On evaluation metrics for medical applications of artificial intelligence. *Sci Rep* 12:5979. <https://doi.org/10.1038/s41598-022-09954-8>
68. Simundic A-M (2008) Confidence interval. *Biochem Med* 18:154–161. <https://doi.org/10.11613/BM.2008.015>
69. Gichoya JW, Banerjee I, Bhimireddy AR et al (2022) AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* 4:e406–e414. [https://doi.org/10.1016/S2589-7500\(22\)00063-2](https://doi.org/10.1016/S2589-7500(22)00063-2)
70. Gichoya JW, Banerjee I, Bhimireddy AR et al (2022) Reading race: AI recognises patient's racial identity in medical images. *Lancet Digit Health* 4:e406–e414. [https://doi.org/10.1016/S2589-7500\(22\)00063-2](https://doi.org/10.1016/S2589-7500(22)00063-2)
71. Tripathi S, Gabriel K, Dheer S et al (2023) Dataset development review. *J Am Coll Radiol*. <https://doi.org/10.1016/j.jacr.2023.06.015>
72. Saleiro P, Kuester B, Hinkson L et al (2019) Aequitas: a bias and fairness audit toolkit. Preprint at <https://doi.org/10.48550/arXiv.1811.05577>
73. EU Artificial Intelligence Act (2023) The act texts. The Artificial Intelligence Act. <https://artificialintelligenceact.eu/the-act/>. Accessed 19 Sep 2023
74. Collins GS, Dhiman P, Andaur Navarro CL et al (2021) Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 11:e048008. <https://doi.org/10.1136/bmjopen-2020-048008>
75. Artificial intelligence in healthcare: applications, risks, and ethical and societal impacts (2022) Think Tank European Parliament. [https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU\(2022\)729512](https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729512). Accessed 19 Sep 2023
76. Mitchell M, Wu S, Zaldivar A et al (2019) Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. <https://doi.org/10.1145/3287560.3287596>
77. Ricci Lara MA, Echeveste R, Ferrante E (2022) Addressing fairness in artificial intelligence for medical imaging. *Nat Commun*. <https://doi.org/10.1038/s41467-022-32186-3>
78. Burlina P, Joshi N, Paul W, Pacheco KD, Bressler NM (2021) Addressing artificial intelligence bias in retinal diagnostics. *Transl Vis Sci Technol* 10:13. <https://doi.org/10.1167/tvst.10.2.13>
79. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. *ACM Comput Surv* 54:115:1–115:35. <https://doi.org/10.1145/3457607>
80. Li X, Cui Z, Wu Y, Gu L, Harada T (2021) Estimating and improving fairness with adversarial learning. Preprint at <https://doi.org/10.48550/arXiv.2103.04243>
81. Gennaro G (2018) The 'perfect' reader study. *Eur J Radiol* 103:139–146. <https://doi.org/10.1016/j.ejrad.2018.03.014>
82. Koh PW, Sagawa S, Marklund H et al (2023) WILDS: a benchmark of in-the-wild distribution shifts. In: Proceedings of the 38th international conference on machine learning. PMLR, 18–24 Jul 2021, pp 5637–5664. <https://proceedings.mlr.press/v139/koh21a.html>
83. Wantlin K, Wu C, Huang S-C et al (2023) BenchMD: a benchmark for unified learning on medical images and sensors. Preprint at <https://doi.org/10.48550/arXiv.2304.08486>
84. Gulrajani I, Lopez-Paz D (2020) In search of lost domain generalization. Preprint at <https://doi.org/10.48550/arXiv.2007.01434>
85. Papers with code (2023) "Papers with code" (Online). <https://paperswithcode.com/>. Accessed 19 Sep 2023
86. Dalca A (2023) List of medical (imaging) datasets. <https://github.com/adalca/medical-datasets>. Accessed 19 Sep 2023
87. Oakden-Rayner L, Gale W, Bonham TA et al (2022) Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet Digit Health* 4:e351–e358. [https://doi.org/10.1016/S2589-7500\(22\)00004-8](https://doi.org/10.1016/S2589-7500(22)00004-8)
88. Roelofs R, Shankar V, Recht B et al (2019) A meta-analysis of overfitting in machine learning. Presented at the neural information processing systems, 2019. <https://www.semanticscholar.org/paper/A-Meta-Analysis-of-Overfitting-in-Machine-Learning-Roelofs-Shankar/c0904fba1814c37e957e401219d0fdab4be5eb7>. Accessed 19 Sep 2023
89. Sion R, Winslett M (2004) Regulatory-compliant data management. In: Proceedings of the 33rd international conference on very large data-bases, in VLDB '07. VLDB Endowment, Vienna pp 1433–1434
90. Fedorov A, Longabaugh W, Pot D et al (2021) NCI imaging data commons. *Cancer Res* 81:4188–4193. <https://doi.org/10.1158/0008-5472.CAN-21-0950>
91. The Cancer Imaging Archive (TCIA) (2023) "The Cancer Imaging Archive" (Online) <https://www.cancerimagingarchive.net/>. Accessed 19 Sep 2023
92. Faghani S, Khosravi B, Zhang K et al (2022) Mitigating bias in radiology machine learning: 3. Performance metrics. *Radiol Artif Intell* 4:e220061. <https://doi.org/10.1148/ryai.220061>
93. Roselli D, Matthews J, Talagala N (2019) Managing bias in AI. In: Companion proceedings of the 2019 world wide web conference, in WWW '19. Association for Computing Machinery, New York, pp 539–544. <https://doi.org/10.1145/3308560.3317590>
94. Omoumi P, Ducarouge A, Tournier A et al (2021) To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol* 31:3786–3796. <https://doi.org/10.1007/s00330-020-07684-x>
95. Hsieh SS, Leng S, Rajendran K, Tao S, McCollough CH (2021) Photon counting CT: clinical applications and future developments. *IEEE Trans Radiat Plasma Med Sci* 5:441–452. <https://doi.org/10.1109/trpms.2020.3020212>
96. Reinke A, Tizabi MD, Sudre CH et al (2021) Common limitations of image processing metrics: a picture story. Preprint at <https://doi.org/10.48550/arXiv.2104.05642>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.