Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls

Farzad Maleki, PhD* • Katie Ovens, PhD* • Rajiv Gupta, MD, PhD • Caroline Reinhold, MD, MSc • Alan Spatz, MD • Reza Forghani, MD, PhD

From the Department of Computer Science, University of Calgary, Calgary, Canada (F.M., K.O.); Department of Radiology, Massachusetts General Hospital, Boston, Mass (R.G.); Augmented Intelligence & Precision Health Laboratory (AIPHL); Department of Radiology and the Research Institute of the McGill University Health Centre, McGill University, Montreal, Canada (C.R., R.F.); Montreal Imaging Experts, Montreal, Canada (C.R., R.F.); Division of Pathology, Jewish General Hospital, Montreal, Canada (A.S.); and Radiomics and Augmented Intelligence Laboratory (RAIL), Department of Radiology and the Norman Fixel Institute for Neurologic Diseases, University of Florida College of Medicine, UF Health Shands Hospital, 1600 SW Archer Rd, Gainesville, FL 32610-0374 (R.F.). Received February 13, 2022; revision requested March 24; revision received October 10; accepted October 24. Address correspondence to R.F. (email: r.forghani@ufl.edu).

* F.M. and K.O. contributed equally to this work.

R.F. supported by the Fonds de recherche en santé du Québec (FRQS) and an operating grant jointly funded by the FRQS and the Fondation de l’Association des radiologistes du Québec (FARQ). R.G. supported by the National Institutes of Health (NIH) (grant nos. 5R01CA212382-05, 5R01EB024343-04, 5R01EB024343-04, 1R01EB032038-01). C.R. supported by a grant from Imagia-Medteq.

Conflicts of interest are listed at the end of this article.

Purpose: To investigate the impact of the following three methodological pitfalls on model generalizability: (a) violation of the independence assumption, (b) model evaluation with an inappropriate performance indicator or baseline for comparison, and (c) batch effect.

Materials and Methods: The authors used retrospective CT, histopathologic analysis, and radiography datasets to develop machine learning models with and without the three methodological pitfalls to quantitatively illustrate their effect on model performance and generalizability. F1 score was used to measure performance, and differences in performance between models developed with and without errors were assessed using the Wilcoxon rank sum test when applicable.

Results: Violation of the independence assumption by applying oversampling, feature selection, and data augmentation before splitting data into training, validation, and test sets seemingly improved model F1 scores by 71.2% for predicting local recurrence and 5.0% for predicting 3-year overall survival in head and neck cancer and by 46.0% for distinguishing histopathologic patterns in lung cancer. Randomly distributing data points for a patient across datasets superficially improved the F1 score by 21.8%. High model performance metrics did not indicate high-quality lung segmentation. In the presence of a batch effect, a model built for pneumonia detection had an F1 score of 98.7% but correctly classified only 3.86% of samples from a new dataset of healthy patients.

Conclusion: Machine learning models developed with these methodological pitfalls, which are undetectable during internal evaluation, produce inaccurate predictions; thus, understanding and avoiding these pitfalls is necessary for developing generalizable models.

Supplemental material is available for this article.

Published under a CC BY 4.0 license.
Generalizability of Machine Learning Models: Three Methodological Pitfalls

Abbreviations
AI = artificial intelligence, CLAIM = Checklist for AI in Medical Imaging, DL = deep learning, HNSCC = head and neck squamous cell carcinoma, IoU = intersection over union, ML = machine learning, QUADAS = Quality Assessment of Diagnostic Accuracy Studies, RSNA = Radiological Society of North America, TCIA = The Cancer Imaging Archive, TRIPOD = Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis, 3D = three-dimensional

Summary
Three major methodological pitfalls in developing machine learning and deep learning models remain undetected during internal evaluation, leading to overoptimistic estimation of model performance and consequent lack of generalizability.

Key Points
- The following methodological pitfalls in model development may prevent generalizability: (a) violation of the independence assumption, (b) model evaluation with an inappropriate performance indicator or baseline for performance comparison, and (c) batch effect.
- These pitfalls are often undetected using internal model evaluation and may lead to inaccurate predictions and interpretations.
- Explicit guidelines to avoid these pitfalls are provided.

Keywords
Random Forest, Diagnosis, Prognosis, Convolutional Neural Network (CNN), Medical Image Analysis, Generalizability, Machine Learning, Deep Learning, Model Evaluation

discussion sections of scientific ML and DL articles. Mongan et al (14) published the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) to aid authors and reviewers. Like TRIPOD, CLAIM provides high-level recommendations for preparing scientific manuscripts but focuses on medical imaging, and it is one of the most widely used artificial intelligence (AI) checklists in medical imaging.

These guidelines mainly focus on the reporting and reproducibility aspects of research findings and offer minimal to no guidance regarding good methodological practices in medical applications of ML. Few technical guidelines are available (10), and those that are available are often inaccessible to practitioners in the medical domain. Clear guidelines supported by scientific evidence are essential to promote the development of generalizable ML and DL models that may be clinically deployed.

Here, we identify and investigate the following three major categories of methodological errors in ML and DL model development: (a) violation of the independence assumption, (b) the use of inappropriate performance indicators for model evaluation, and (c) the introduction of batch effect. We also provide guidelines for avoiding these pitfalls.

Materials and Methods

Datasets
This retrospective, institutional review board–exempt study used several imaging modalities to show that the methodological pitfalls are not specific to one type.

Head and Neck Squamous Cell Carcinoma CT Dataset
The CT dataset included pretreatment CT scans in 137 patients with head and neck squamous cell carcinoma (HNSCC) who were treated with radiation therapy (17,18). Hereafter, we refer to this dataset as HNSCC. The dataset is available from The Cancer Imaging Archive (TCIA) (17–19). Table S1 provides a summary of the clinical endpoints.

Lung CT Dataset
This dataset includes 120 CT scan series in 60 patients in TCIA, available from the Lung CT Segmentation Challenge 2017 (19–21). We used this dataset to demonstrate methodological pitfalls related to performance metrics for segmentation.

Digital Histopathologic Analysis Dataset
The histopathologic analysis dataset contains 143 hematoxylin–eosin–stained formalin-fixed paraffin-embedded whole-slide images of lung adenocarcinoma provided by the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (22). The dataset includes five histopathologic patterns: solid (51 slides), lepidic (19 slides), acinar (59 slides), micropapillary (nine slides), and papillary (five slides). We used the 110 slides from patients with solid- and acinar-predominant histopathologic patterns, as they were numerous and relatively balanced.

Due to the high resolution of the histopathologic images, it was computationally impractical to analyze them as individual whole images (23). Therefore, we first downscaled each image by a factor of 4. Then, using color thresholding, we extracted the foreground, that is, the tissue segments on each slide. Next, for each image, we extracted random patches sized 1024 × 1024 pixels. Patches with 75% or more background were excluded during the patch extraction process. The patch extraction process continued until 200 patches were extracted from each image, resulting in 22 000 patches. Figure 1 illustrates an example whole-slide image, as well as a selection of random patches.

Chest Radiograph Datasets
To demonstrate the impact of batch effects, we used two radiography datasets: 8851 normal chest radiographs with no findings from the Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset, available on Kaggle (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge), and a chest radiograph dataset from Kermany et al (24) that included 1349 normal radiographs with no findings and 3883 radiographs showing pneumonia in pediatric patients. Figure 2 illustrates samples from each dataset.

Experiments

Breaking the assumption of independence.— ML and DL approaches assume that data used for model training and evaluation are independent and identically distributed. To develop ML models, it is a common practice to split the available data into training, validation, and test sets (Fig 3). The training set...
is used to learn model parameters, the validation set to select model hyperparameters, and the test set to provide an unbiased estimate of model generalization error. However, the validity of this design is contingent on the assumption of independence.

Data used for model training should be independent of the data used for model evaluation. This assumption can be violated when some data not expected to be available in the prediction and evaluation phase are used for model training, a phenomenon referred to as data leakage. We investigated the impact of four common practices associated with data leakage and violation of the independence assumption on model generalizability: (a) oversampling, (b) data augmentation, (c) using several data points from a patient, and (d) feature selection.

Oversampling: Class imbalance represents a substantial difference between the number of samples between classes. Often, models developed using imbalanced datasets tend to undermine minority classes. Oversampling can alleviate this challenge by sampling with replacements from the original minority classes to artificially increase the number of samples. Class imbalance is common in medical imaging datasets because of disease rarity and difficulties in imaging certain conditions (25).

To quantitatively show how applying oversampling before splitting data into training, validation, and test sets could affect model generalizability, we evaluated two binary classifiers—as described in Appendix S1—for predicting local recurrence in head and neck cancer by employing a radiomics approach using the HNSCC dataset. The models differed only in the order that oversampling was applied. The first model (model A) was developed by conducting oversampling before data splitting and the second model (model B) by conducting oversampling after data splitting.

Data augmentation: Data augmentation refers to computational methods used to generate new data points from existing ones. This technique is commonly used when developing ML and DL models for image analysis, as developing large-scale datasets is often impractical (26), and it can improve the performance and generalizability of resulting models.

To quantitatively show how applying data augmentation before splitting data into training, validation, and test sets
could affect model generalizability, we developed two DL-based binary classifiers—as described in Appendix S2—for distinguishing solid- and acinar-predominant histopathologic patterns in patients with lung adenocarcinoma. For these models, every component of the model building and evaluation pipeline was kept the same; however, the first model (model C) was developed by conducting data augmentation before data splitting and the second model (model D) by conducting data augmentation after data splitting.

Using several data points from a patient: To investigate how distributing data points for a patient across training, validation, and test sets could impact model generalizability, we used the pathologic analysis dataset to build two DL-based binary classifiers—as described in Appendix S2—for distinguishing solid- and acinar-predominant histopathologic patterns for patients with lung adenocarcinoma. Model E was developed by randomly distributing the image patches across the training, validation, and test sets. Model F was developed by assigning image patches for each patient to either the training, validation, or test sets. All other model components were identical.

Feature selection: To demonstrate the impact on model generalizability of applying feature selection before splitting data into training, validation, and test sets, we used the HNSCC dataset to develop two binary classifiers—as described in Appendix S1—to predict overall survival using a radiomics approach. The first model (model G) was developed by conducting feature selection before data splitting. The second model (model H) was developed by conducting feature selection after data splitting.

To achieve statistically reliable results for the conventional radiomics analysis, we repeated the model-building process 100 times and calculated the F1 score as the performance measure. We then used the Wilcoxon rank sum test to assess if there was a statistically significant difference between the performance measures derived from model A versus model B, as well as model G versus model H. We used the stats.mannwhitneyu function from the SciPy (version 1.7.3) Python package for statistical analysis, where a $P$ value less than .05 was considered as significant.

**Evaluating models with an inappropriate performance indicator or baseline for comparison.**— Selecting the appropriate quantitative measures for model evaluation is essential for developing predictive models. Setting the baseline for acceptable performance is also required to determine if a model could be used and deployed in real-world settings.

Accuracy, which represents the proportion of samples correctly classified by a model, is commonly used to measure the performance of classification models. We used a model for distant metastasis prediction in the HNSCC dataset to show how accuracy as a performance indicator for imbalanced datasets inappropriately reflects model performance.

We also investigated how selecting an inappropriate baseline for performance comparison impacts interpretations of model outcome. The performance of segmentation models is commonly measured using intersection over union (IoU) and Dice scores.

Dice score and IoU result in values between 0 and 1, where a value of 1 represents a perfect overlap between $X$ and $Y$, and 0 represents no overlap.

To demonstrate the role of setting a baseline expectation when evaluating segmentation model performance, we developed a threshold-based approach to segment air inside the body for samples in the lung CT dataset as a proxy for lung segmentation to serve as a baseline model. Any
We trained a model (model I) on the Batch x-ray dataset. Batch x-ray dataset by Kermany et al (24) and the normal samples (ie, images with no findings) from the RSNA dataset. This data contains a batch effect, because the pneumonia samples come from children and the normal samples come from adults. Hereafter, we refer to this dataset as Batch x-ray. We trained a model (model I) on the Batch x-ray dataset.

Batch effects.— A batch effect occurs when data from several sources are aggregated to develop a larger dataset and the class distribution samples from these sources substantially vary. For example, suppose all malignant tumors in a cohort were imaged with MRI machine X and all benign tumors with MRI machine Y. In this case, the model may learn to differentiate malignant from benign tumors on the basis of MRI machine—attributed differences rather than intrinsic tumor characteristics. To investigate how batch effects impact model generalizability, we simulated a dataset in which data was imaged with MRI machine X and all benign tumors were imaged with MRI machine Y. In this case, the model may initially vary. For example, suppose all malignant tumors in one trained model; in contrast, for radiomics models A, B, G, and H, we repeated model building 100 times and reported the first and third quartile in parentheses (Q1, Q3) for each performance measure to show the spread of the middle half of its distribution.

Results

Violation of Independence Assumption

Table 1 shows model performance when incorporating the described methodological pitfalls versus when these pitfalls are avoided. Oversampling was conducted before splitting data for model A and after splitting data for model B. The results showed a statistically significant gap between the performance of these two models. While model B—the correct approach—showed poor performance, model A seemed to offer promising results. Incorrect application of oversampling led to a statistically significant but superficial improvement in model performance (Wilcoxon rank sum test = −12.22, P < .001) for predicting local recurrence in head and neck cancer.

Models C, D, E, and F were built for distinguishing solid- and acinar-predominant histopathologic patterns in patients with lung adenocarcinoma. Feature selection was conducted before splitting data into training, validation, and test sets for model D. Model E was developed by randomly distributing all data points across training, validation, and test sets for model D. Both models were built to predict 3-year overall survival using the HNSCC dataset. Note that for models A, B, G, and H, performance measures were acquired as average of performance measures for 100 runs of the model development pipeline. For these models, we also report the first and third quartile in parentheses (Q1, Q3) for each performance measure to show the spread of the middle half of its distribution.

We then tested this model on an external dataset of normal chest radiographs from Kermany et al (24).

Table 1: Performance Measures of Models Built When Incorporating Methodological Pitfalls versus Models Built When Avoiding Pitfalls

| Model | Algorithm | Approach | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) |
|-------|-----------|----------|--------------|---------------|------------|--------------|
| A     | Random forest | Incorrect oversampling | 85 (82, 88) | 82 (78, 86) | 89 (85, 94) | 85 (82, 88) |
| B     | Random forest | Correct oversampling | 72 (67, 76) | 16 (0, 22) | 15 (0, 22) | 14 (0, 21) |
| C     | ResNet50 | Incorrect data augmentation | 99.9 | 99.9 | 99.9 | 99.9 |
| D     | ResNet50 | Correct data augmentation | 65.9 | 53.0 | 54.8 | 53.9 |
| E     | ResNet50 | Incorrect distribution of data points from the same patient | 78.8 | 79.9 | 71.9 | 75.7 |
| F     | ResNet50 | Correct distribution of data points from the same patient | 65.9 | 53.0 | 54.8 | 53.9 |
| G     | Random forest | Incorrect feature selection | 72 (67, 76) | 80 (76, 84) | 81 (77, 87) | 80 (77, 85) |
| H     | Random forest | Correct feature selection | 64 (61, 70) | 76 (72, 80) | 76 (70, 82) | 75 (72, 79) |

Note.—Model A was built with oversampling performed before splitting data into training, validation, and test sets, and model B was built with oversampling performed after splitting data. Both models were trained to predict local recurrence in head and neck cancer using the head and neck squamous cell carcinoma (HNSCC) dataset. For model C, augmentation was performed before splitting data into training, validation, and test sets and after splitting data for model D. Model E was developed by randomly distributing all data points across training, validation, and test sets. For model F, all data points for a patient were assigned to each set. Models C, D, E, and F were trained to distinguish solid- and acinar-predominant histopathologic patterns in patients with lung adenocarcinoma. Feature selection was conducted before splitting data into training, validation, and test sets in model G and after splitting data in model H. Both models were built to predict 3-year overall survival using the HNSCC dataset. Note that for models A, B, G, and H, performance measures were acquired as the average of performance measures for 100 runs of the model development pipeline. For these models, we also report the first and third quartile in parentheses (Q1, Q3) for each performance measure to show the spread of the middle half of its distribution.

We used any voxel with a Hounsfield unit value less than −400 as air. Then, we removed the segment of air outside the bodies of patients. This results in segmenting air within the body, which is used as an erroneous proxy for lung segmentation. We compared the performance of this simple model with the ground truth (lung contours). Any model with a performance less than such a baseline model should be considered irrelevant.

Table 1 shows model performance when incorporating the described methodological pitfalls versus when these pitfalls are avoided. Oversampling was conducted before splitting data for model A and after splitting data for model B. The results showed a statistically significant gap between the performance of these two models. While model B—the correct approach—showed poor performance, model A seemed to offer promising results. Incorrect application of oversampling led to a statistically significant but superficial improvement in model performance (Wilcoxon rank sum test = −12.22, P < .001) for predicting local recurrence in head and neck cancer.

Models C, D, E, and F were built for distinguishing solid- and acinar-predominant histopathologic patterns in patients with lung adenocarcinoma. Note that for models A, B, G, and H, performance measures were acquired as the average of performance measures for 100 runs of the model development pipeline. For these models, we also report the first and third quartile in parentheses (Q1, Q3) for each performance measure to show the spread of the middle half of its distribution.

A Random forest | Incorrect oversampling | 85 (82, 88) | 82 (78, 86) | 89 (85, 94) | 85 (82, 88) |
B Random forest | Correct oversampling | 72 (67, 76) | 16 (0, 22) | 15 (0, 22) | 14 (0, 21) |
C ResNet50 | Incorrect data augmentation | 99.9 | 99.9 | 99.9 | 99.9 |
D ResNet50 | Correct data augmentation | 65.9 | 53.0 | 54.8 | 53.9 |
E ResNet50 | Incorrect distribution of data points from the same patient | 78.8 | 79.9 | 71.9 | 75.7 |
F ResNet50 | Correct distribution of data points from the same patient | 65.9 | 53.0 | 54.8 | 53.9 |
G Random forest | Incorrect feature selection | 72 (67, 76) | 80 (76, 84) | 81 (77, 87) | 80 (77, 85) |
H Random forest | Correct feature selection | 64 (61, 70) | 76 (72, 80) | 76 (70, 82) | 75 (72, 79) |
The effect of breaking the independence assumption by distributing data points for a patient across training and test sets is demonstrated with models E and F. For model E, data points were randomly distributed across training, validation, and test sets. Therefore, data points for the same patient could appear in both training and test sets. For model F, the independence assumption was preserved by assigning data points for each patient to either the training, validation, or test set. Note that model F is the same as model D—that is, the correct approach for distinguishing solid- and acinar-predominant histopathologic patterns of the same patients with lung adenocarcinoma. Distributing data points of patients across training and test sets led to a seemingly higher performance of model E compared with model F.

Table 1 also shows how applying feature selection prior to splitting data into training, validation, and test sets could violate the independence assumption. Feature selection was conducted before data splitting for model G and after data splitting for model H. Applying feature selection before splitting data into training, validation, and test sets led to a significant superficial boost in model performance for predicting 3-year overall survival in head and neck cancer (Wilcoxon rank sum test = −5.87, \( P < .001 \)).

Model Evaluation with an Inappropriate Performance Indicator or Baseline for Comparison

A naive model that predicts all samples as nondistant metastasis would achieve an accuracy of 94% \( (\frac{4k+81}{4k+1681} - 1) \) for our HNSCC dataset. However, such a model has no medical use, as it achieves a recall of zero and fails to diagnose any distant metastasis. This example shows that using accuracy for highly imbalanced datasets might lead to developing models that cannot be deployed in a clinical setting.

Figure 4 illustrates the ground truth segmentation, as well as the predicted segmentation of a randomly chosen image from the lung CT dataset, where the prediction was made by a simple baseline model that detects air inside the body. While the predicted lung segmentation was not medically acceptable, for this example, the baseline model achieved a Dice score of 0.94 and an IoU of 0.88. Table 2 shows the minimum, mean, maximum, and SD of the baseline model performance measures when applied to samples in the lung CT dataset. The simple baseline model achieved a mean Dice score of 0.92 and a mean IoU of 0.86. Therefore, models with average performance measures lower than this baseline model should not be used.

Batch Effect

Our results also showed that batch effects could be a substantial barrier to model generalizability. We observed that model I, which was trained and tested on a dataset with batch effect (Batch x-ray dataset), achieved accuracy, precision, recall, and an F1 score of 99.7%, 97.9%, 99.5%, and 98.7%, respectively. However, when this model was applied to normal pediatric chest radiograph samples from the dataset by Kermany et al (24), only 3.86% of samples were correctly classified.

The attribution of each image pixel to the model prediction for that image can be calculated using the integrated gradient method (27). Figure 5 overlays the attribution values for each pixel of a normal pediatric radiograph. The figure shows that the pneumonia prediction model trained using the Batch x-ray dataset focuses on anatomic structures and body position rather than image characteristics in the lung.

Figure 6 presents a guideline to avoid the methodological errors discussed in this article.

Discussion

With many recent studies highlighting the potential of ML and DL approaches for medical image analysis, the natural expectation is widespread use of these approaches in clinical settings. However, when applied prospectively, the lack of model generalizability is the main challenge. We investigated and highlighted some of the key methodological errors that lead to models that are not generalizable despite achieving deceptively promising results during internal evaluation. Such errors may be difficult to capture by readers, reviewers, or authors, especially if certain methodological details are not presented.

ML model performance is typically assessed using internal evaluation, where the available data are partitioned into training, validation, and test sets. To achieve an unbiased estimate of performance measures and generalization error for a model, data from the training and test sets must be independent. Any violation of the independence assumption should be avoided. When oversampling is conducted before randomly splitting data into training, validation, and test sets, the same copy of a data point could appear in training and test sets. Therefore, the training and test sets are no longer independent. Our results demonstrated that when oversampling was incorrectly applied to the HNSCC dataset, the model achieved superficially high performance; however, a correct approach led to poor results, as expected, due to the very small number of local recurrences in the dataset.

Similar superficial boosts were observed for incorrect application of data augmentation. When data augmentation is applied to an image, some of its characteristics change, but many will still be shared by the original and augmented images. If data augmentation is applied before data splitting, just as with oversampling, these highly correlated samples can be spread across the training, validation, and test sets, potentially leading to high performance on the internal test set and poor performance on external data.

Often, there are several data points associated with each patient. Extracting image patches is a common practice when analyzing histopathologic images or three-dimensional (3D) images (28). These patches might share characteristics irrelevant to the study goal. Distributing the different patches derived from a single patient between training, validation, and test sets could artificially boost model performance, but the resulting model will not be generalizable when applied to external data. Multiple data points for a single patient should be assigned to only the training, validation, or test set. For example, one should not assign a T2-weighted sequence of a patient to the training set and the corresponding T1-weighted sequence to the test set.

In radiomics, there are often several features representing the statistical characteristics, shape, and texture of a region of interest. Feature selection is an important step in developing ML models with high-dimensional features. If applied before...
splitting data, the information from all samples in the dataset is used to select a subset of features that work best for all samples. This partially exposes the test set, which will be selected in the next step, to the model and breaks the independence assumption. In this work, we observed that exposing the test samples to feature selection methods led to a superficial boost in performance measures. The selected features are discriminative for the given test set, leading to degraded performance on external data and a lack of generalizability.

Another key consideration in developing predictive models is selecting the appropriate performance indicators. For example, using accuracy for a diagnostic model for a rare condition is often misleading, as the model may disregard all cases of the rare condition in a dataset with high class imbalance and still achieve high accuracy. The cost associated with misclassification should also be considered. Consider a diagnostic test for a life-threatening condition. A method with high accuracy but low recall (sensitivity) cannot be used clinically. Similarly, high false-positive rates potentially leading to highly invasive procedures prohibit

Table 2: Summary of Scores for a Simple Model Detecting Air Inside the Body as an Estimate for Lung Segmentation in Chest CT Scans of the Lung Segmentation Challenge 2017

| Scoring Method | Mean | Min | Max | SD |
|----------------|------|-----|-----|----|
| Dice           | 0.92 | 0.84| 0.97| 0.02|
| IoU            | 0.86 | 0.72| 0.94| 0.04|

Note.—The summary highlights the mean, minimum (min), and maximum (max) Dice scores and intersection over union (IoU) between the prediction made by the thresholding approach and the ground truth resulting from manual segmentation. Considering the thresholding approach as a computationally efficient but medically unreliable baseline for lung segmentation, any sound lung segmentation model should at least outperform this baseline. For example, a deep learning model with a Dice score of 0.92 should be considered a low-performing model.
the deployment of predictive models in clinical settings. In such scenarios, precision should also be considered as a performance indicator to guide model development and evaluation.

Dice score and IoU are commonly used as performance measures for evaluating segmentation models. From a mathematical perspective, the Dice score is always larger than or equal to IoU (see Appendix S3 for a mathematical proof), which encourages reporting the Dice score. In our example in Figure 4, we observed that both metrics achieved a high value (IoU: 0.88 and Dice: 0.94), despite obvious flaws in lung segmentation. Therefore, we encourage visual inspection of segmentation model outcomes as a qualitative analysis. Additionally, pixel-level accuracy measurement should be avoided when evaluating small regions or volumes of interest.

During model development, it is desirable to collect and analyze imaging data from different sites. This is beneficial for increasing sample size and model generalizability because of the use of samples more representative of the target population. However, in addition to imaging techniques (29,30), the class distribution of samples substantially varies across sites. Thus, the aggregated dataset and resulting models could suffer from a batch effect. Of note, batch effect is sometimes disregarded. In a study of ML models used for COVID-19 diagnosis, Roberts et al (31) reported that some studies had used normal images from healthy pediatric patients, while the COVID-19 samples came from adults. Even for cases where training, validation, and test sets have the same

Figure 5: The attribution of each pixel of a radiograph in a healthy pediatric patient to model prediction is represented using integrated gradients. The radiograph pixels most attributed to model prediction are shown in red. The model, trained in the presence of a batch effect, incorrectly classifies this sample as pneumonia on the basis of anatomic structures and body position, which are substantially different between pediatric and adult patients.

Avoiding Data Leakage
- Oversampling of data should be done after splitting data into training, validation, and test sets
- Data augmentation should be applied after splitting data into training, validation, and test sets
- Feature selection should be conducted after splitting data into training, validation, and test sets
- Data points for a patient should not be distributed across training, validation, and test sets

Reporting Performance Measures
- Accuracy should not be used as a sole performance indicator for imbalanced datasets
- Other performance indicators, such as precision and recall, should be reported alongside accuracy
- A baseline for acceptable model performance should be provided

Sample Origin
- In multicenter studies, aim to keep similar data distributions across centers to avoid batch effect
- If possible, use an external dataset to provide an unbiased estimate of generalization error

Figure 6: Methodological guidelines for developing generalizable machine learning and deep learning models.
Radiology: Artificial Intelligence Volume 5: Number 1—2023 • radiology-ai.rsna.org

distribution, the performance of a model should be regularly and systematically monitored. Due to the dynamic nature of data in real-world settings, the data distribution might change as time passes. This phenomenon—referred to as distribution shift, domain shift, or domain drift—might happen because of factors such as changes in imaging hardware, software, or protocols. In such cases, the performance of a model might degrade as time passes, demonstrating that training and evaluating ML and DL models is a nonstationary and iterative process.

Other ML and DL challenges beyond the described methodological errors, including data quality and availability, bias, and explainability, are outside the scope of this article. Recent literature covers potential AI misuse and provides suggestions and guidelines for how AI research can be used responsibly (31,32). Current guidelines also provide the framework for presenting ML and DL research to ensure the reproducibility of results (14). We recommend reviewing this literature for further knowledge, as the recommendations in this article are complementary to these works.

Developing ML and DL models for medical image analysis is challenging. Compared with natural images, medical images are often high dimensional due to their high resolutions (for example, in histopathologic analysis, whole-slide images) or their 3D nature (for example, in MR and CT images). Consequently, analyzing such data is challenging, specifically for domains where large-scale annotated datasets are unavailable. Among the challenges posed by high data dimensionality in medical imaging are the impracticality of whole-image analysis, the demanding nature of data annotation for medical images, and the increased possibility of overfitting. Whole-image analysis is impractical due to the limited amount of video random-access memory offered by graphics processing units. Therefore, patch-based approaches are used to analyze 3D and microscopic images, which in turn pose the challenge of combining the predictions made for these patches to achieve an image-level or patient-level outcome or prediction. This also complicates the data analysis pipeline and exposes the developed models to methodological pitfalls that might affect model generalizability. Pixel- or voxel-level annotation for medical images is also more tedious and time-consuming compared with that for natural images, as annotating a single 3D image often requires manual processing of hundreds of two-dimensional sections. Microscopic images also often have high resolutions, for example, 100000 × 100000 pixels, requiring greater effort to annotate each image. The increased possibility of model overfitting for high-dimensional data in the absence of large-scale datasets is also a major challenge affecting model generalizability. Due to these challenges, following best practices and avoiding methodological pitfalls is essential for developing generalizable models with the potential to be deployed in clinical settings. Developing methods tailored to medical images remains an active research area.

Medical image analysis is also interdisciplinary, requiring contributions from imaging, computational, and medical experts. Lack of expertise in one of these domains might lead to developing models that are not generalizable. For example, if medical expertise is available, it is unlikely that a comparison between pediatric samples and adult patients would be considered a valid experimental design given the substantial differences in anatomic and imaging components between these two groups. Furthermore, a model built to classify COVID-19 versus healthy lung using lung radiographs would immediately be recognized by a medical expert as requiring a more rigorous evaluation to ensure the model does not falsely detect other lung abnormalities as COVID-19. Collaboration and cooperation among various experts at each stage of medical image analysis are essential for the development of models that can ultimately be applied in a clinical setting.

This study had some limitations. While there are several DL architectures available that one could customize for medical image analysis, the models used in our study were limited to well-established architectures. This study could be further augmented by considering other model architectures and how susceptible they are to the violation of the independence assumption. Another important factor affecting the generalizability of ML and DL models, but not covered in this study, is the sample size used for model training and evaluation. Although there is general consensus on the positive impact of larger sample sizes and more varied datasets for training ML models, the literature could benefit from additional empirical analyses of the effect of sample size on model generalizability for future research. Another factor that has been reported to affect the performance of DL models (33,34), but not covered in this study, is image resolution. Using the National Institutes of Health ChestX-ray14 dataset, Sabottke and Spieler (33) studied the effect of image resolution on the performance of two widely used DL model architectures, ResNet34 (35) and DenseNet121 (33,36). Investigating different image resolutions ranging from 32 × 32 to 600 × 600 pixels, they showed that the optimal selection of image resolution is task dependent and essential for increasing model performance in several classification tasks. Last, in this investigation, we mainly focused on the methodological errors resulting from violation of the independence assumption. However, another critical requirement for developing generalizable models is that the datasets used for model training and evaluation should represent the distribution of the data in a real-world setting (ie, data at the deployment phase), which may not always be possible. For example, samples from minority classes might be absent or poorly represented in the test set. In such a case, a model that works well on prevalent classes but poorly on minority classes still achieves high performance measures when evaluated using such a test set. Techniques such as stratified data splitting could be used to ensure similar representations for only known classes or covariates, not unidentified covariates, across training, validation, and test sets. The problem is more pronounced when using small sample sizes, highlighting the need for developing large and diverse datasets for medical image analyses. One must ensure that best practices in ML model development are followed despite restricted access to patient data and source code.

In conclusion, we studied several methodological pitfalls in developing ML and DL models. These pitfalls are undetected during the internal evaluation of models, leading to overoptimistic estimations of model performance and consequent lack of generalizability. Awareness of these pitfalls and consideration of the suggested guidelines for avoiding them are important for developing generalizable ML and DL health care models.
Disclosures of conflicts of interest: F.M. No relevant relationships. K.O. No relevant relationships. R.G. No relevant relationships. C.R. Grants or contracts from Imaging Science and Innovation. R.G. Research grant from McGill University Health Centre Foundation, TD Bank, GE Healthcare, and Intel for Artificial Intelligence for Urgent Care: Providing Better Care to Remote Communities, payments made to institution (Research Institute of the McGill University Health Centre); Canadian Cancer Society/Canadian Institutes of Health Research/Brain Canada Foundation Spark Grant: Novel Technology Applications in Cancer Prevention and Early Detection (SPARK-21), research grant, payments made to institution (McGill); Fonds de recherche en santé du Québec and Fondation de l’Association des radiologistes du Québec grant providing research salary support and time (payments to author) and general research operating grant (payments to institution), not directly funding this work.

References
1. McErlean A, Panicek DM, Zabor EC, et al. Intra- and interobserver variability in CT measurements in oncology. Radiology 2013;269(2):451–459.
2. Obuchowicz R, Oszust M, Piorkowski A. Interobserver variability in quality assessment of magnetic resonance images. BMC Med Imaging 2020;20(1):109.
3. Federau C, Christensen S, Scherrer N, et al. Improved segmentation and detection sensitivity of diffusion-weighted stroke lesions with synthetically enhanced deep learning. Radiol Artif Intell 2020;2(5):e190217.
4. Dadoun H, Rousseau AL, de Kerviler E, et al. Deep learning for the detection, localization, and characterization of focal liver lesions on abdominal US images. Radiol Artif Intell 2022;4(3):e211010.
5. Liu K, Li Q, Ma J, et al. Evaluating a fully automated pulmonary nodule detection approach and its impact on radiologist performance. Radiol Artif Intell 2019;1(5):e180084.
6. Kadiouhi H, Duong L, Joncas J, et al. Convolutional neural networks for automatic risser stage assessment. Radiol Artif Intell 2020;2(3):e180063.
7. Li H, Chen M, Wang J, Illipani VSP, Parikh NA, He L. Automatic segmentation of diffuse white matter abnormality on T2-weighted brain MR images using deep learning in very preterm infants. Radiol Artif Intell 2021;3(3):e200166.
8. Blansat K, Retson T, Masutani E, Bahrami N, Hsiao A. Deep learning-based prescription of cardiac MRI planes. Radiol Artif Intell 2019;1(6):e180069.
9. Commandeur F, Goeller M, Razipour A, et al. Fully automated CT quantification of epicardial adipose tissue by deep learning: a multicenter study. Radiol Artif Intell 2019;1(6):e190945.
10. Kelly CJ, Kandikasalimgan A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195.
11. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalizability in clinical research and machine learning in health care. Lancet Digit Health 2020;2(9):e489–e492.
12. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13(1):1.
13. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14(12):749–762.
14. Morgan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. Radiol Artif Intell 2020;2(2):e200029.
15. Whiting PF, Rutjes AWS, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529–536.
16. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. J Clin Epidemiol 2017;82:1–10.
17. Aerts HJWJ, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1):4006.
18. Wee L, Dekker A. Data from head-neck-radiomics-HN1. The Cancer Imaging Archive; 2019.
19. 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(6):1045–1057.
20. Yang J, Veeraraghavan H, Armato SG 3rd, et al. Augmentation and synthetic propagation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. Med Phys 2018;45(10):4568–4581.
21. Yang J, Sharp G, Veeraraghavan H, et al. Data from Lung CT Segmentation Challenge. The Cancer Imaging Archive. https://search.datacite.org/works/10.7937/k8/teca.2017.3r5vot8. Published 2017. Accessed July 24, 2021.
22. Wei JW, Tafe LJ, Linink KYA, Vaidkus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci Rep 2019;9(1):3358.
23. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med 2021;27(5):775–784.
24. Kermamy DS, Goldbaum M, Gai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122–1131.e9.
25. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failure in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. 2020.
26. Le WT, Maleki F, Rompero FP, Forgiani R, Kadoury S. Overview of machine learning: Part 2: deep learning for medical image analysis. Neuroimaging Clin N Am 2020;30(4):417–431.
27. Kokhlikyan N, Miglani V, Martin M, et al. Captum: A unified and generic model interpretability library for PyTorch. arXiv 2019;07.0896 [preprint] https://arxiv.org/abs/2019.07896. Posted September 16, 2020. Accessed January 27, 2022.
28. Kao PY, Shailja S, Jiang J, et al. Improving patch-based convolutional neural networks for MRI brain tumor segmentation by leveraging location information. Front Neurosci 2020;13:1449. [Published correction appears in Front Neurosci 2020;14:328.]
29. Howard FM, Dolezal J, Kochanny S, et al. The impact of digital histopathology batch effect on deep learning model accuracy and bias. bioRxiv 2020;2020.12.03.410845 [preprint]. Posted December 5, 2020. Accessed January 27, 2022.
30. Ligero M, Jordi-Olvero O, Bernatowicz K, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol 2021;31(3):1460–1470.
31. Roberts M, Drigo D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3(3):199–217.
32. Partnership on AI. How to be responsible in AI publication. Nat Mach Intell 2021;3(5):367.
33. Sabbatke CF, Spieler BM. The effect of image resolution on deep learning in radiography. Radiol Artif Intell 2020;2(1):e190015.
34. Lakhan P. The importance of image resolution in building deep learning models for medical imaging. Radiol Artif Intell 2020;2(1):e190177.
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
36. Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017. Piscataway, NJ: IEEE, 2261–2269.
37. van Griethuysen JM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–e107.
38. Deng J, Dong W, Socher R, Li L, Kai L, Li FF. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, June 20–25, 2009. Piscataway, NJ: IEEE, 2009:248–255.
39. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RJA. Improving neural networks by preventing co-adaptation of feature detectors. arXiv 2012.07580 [preprint] https://arxiv.org/abs/1207.07580. Posted July 3, 2012. Accessed September 1, 2018.
40. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, Mass: MIT Press; 2016.
41. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv 1412.6980 [preprint] https://arxiv.org/abs/1412.6980. Posted December 22, 2014. Accessed September 29, 2018.
42. Bualaa E, Iglovikov Vl, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AAJ. Alumabations: fast and flexible image augmentations. Information (Basel) 2020;11(2):125.