Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385

Review

Clinical Cancer Research

# Data Analysis Strategies in Medical Imaging

Chintan Parmar[1], Joseph D. Barry[2], Ahmed Hosny[1], John Quackenbush[2,3], and Hugo J.W.L. Aerts[1,4]

## Abstract

Radiographic imaging continues to be one of the most effective and clinically useful tools within oncology. Sophistication of artificial intelligence has allowed for detailed quantification of radiographic characteristics of tissues using predefined engineered algorithms or deep learning methods. Precedents in radiology as well as a wealth of research studies hint at the clinical relevance of these characteristics. However, critical challenges are associated with the analysis of medical imaging data. Although some of these challenges are specific to the imaging field, many others like reproducibility and batch effects are generic and have already been addressed in other quantitative fields such as genomics. Here, we identify these pitfalls and provide recommendations for analysis strategies of medical imaging data, including data normalization, development of robust models, and rigorous statistical analyses. Adhering to these recommendations will not only improve analysis quality but also enhance precision medicine by allowing better integration of imaging data with other biomedical data sources. *Clin Cancer Res; 24(15); 3492–9.* ©2018 AACR.

## Introduction

Large-scale radiographic imaging of diseased tissue offers an incredibly rich data resource for scientific and medical discovery. Because imaging data are collected during routine clinical practice, large datasets are potentially readily available for medical research. Buoyed by advancements in artificial intelligence (AI), statistical methodology, and image processing capabilities, the number of publications related to big data analysis of radiographic datasets is growing at an exponential pace (1, 2). Indeed, the automated quantification of radiographic characteristics of tissues can be helpful in the detection, characterization, and monitoring of diseases. This process, referred to as "radiomics" (3–6), uses either a set of predefined engineered features (7) that describe radiographic aspects of shape, intensity, and texture or alternatively, features that can be automatically "deep learned" directly from example images (8, 9). Early success of radiomics for assisting clinical decisions related to the diagnosis and risk stratification of different cancers (4, 10–14) has spurred rapid expansion in this field and has opened new avenues of investigating the clinical utility of medical imaging in radiology (15).

The analysis of radiologic data presents many challenges. Although some of these challenges are specific to the imaging field, many are generic and have already been addressed in quantitative fields such as genomics and biostatistics. Microarray data analysis field, for example, although now a mature field, initially struggled with many obstacles related to data normalization (16), batch effects (17, 18), replicability (19), and the use of gene expression profiles for disease subtyping (20) and classification (21). In its current youthful state, data analysis in radiology faces similar challenges (3, 5, 15, 22). However, many researchers in radiology are unaware of some commonly observed and avoidable data analysis pitfalls. It is our contention that the quality of radiology studies could be greatly improved by following a simple set of data science "best practices" specific to the radiology field. These start with basic experimental design principles, including data normalization and standardization protocols, and expand to data analysis and validation strategies with appropriate reporting. In this review, we provide examples of typical data analysis pitfalls observed in radiomics studies and suggest strategies for avoiding them.

## Data Analysis Strategies

Although data analysis strategies may vary considerably between studies in medical imaging, certain common strategies related to study design, analysis, and reporting (see Fig. 1 and Table 1) could enhance the validity and generalizability of the study. Here, we discuss these strategies in detail.

### Design: Research definition, data curation, and strategic decisions

Experimental design should be defined from the outset of a study. An experienced statistician should be consulted from the beginning of the study. By anticipating big picture challenges, the research question should be defined, required resources and data should be identified and curated, and ultimately, high-level decisions related to analysis strategies should be made (see Fig. 1 and Table 1).

[1]Department of Radiation Oncology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts. [2]Department of Biostatistics & Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts. [3]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts. [4]Department of Radiology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.

**Corresponding Author:** Hugo J.W.L. Aerts, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Institutes of Medicine, 450 Brookline Avenue, Boston, MA 02215. Phone: 617-525-7156; Fax: 617-582-6037; E-mail: Hugo_Aerts@dfci.harvard.edu

**doi:** 10.1158/1078-0432.CCR-18-0385

Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385



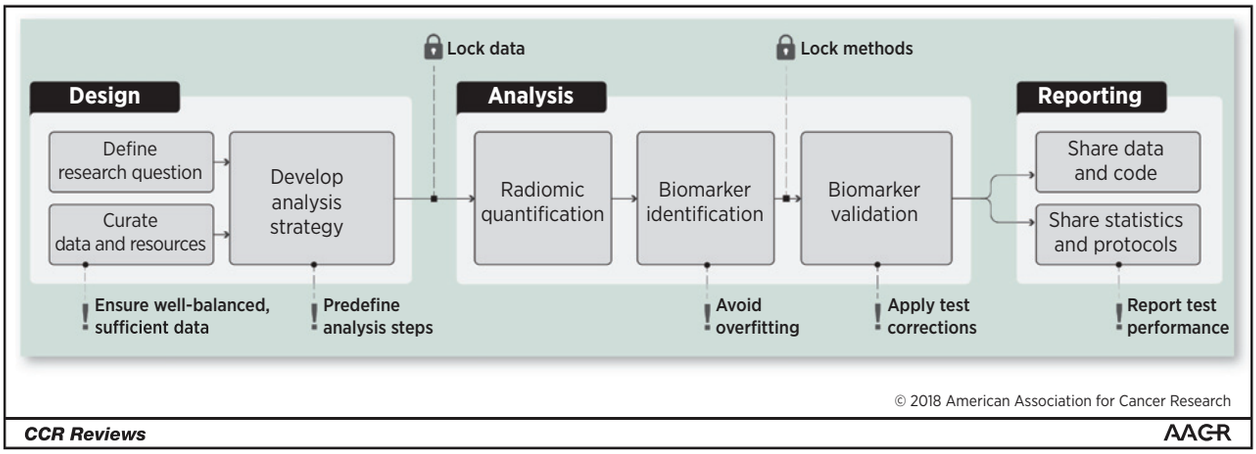


**Figure 1.**

Data analysis stages in medical imaging. Design involves defining research scope and questions, curating the required data and resources, and exploring and developing the analysis strategies. Analysis begins with radiomic quantification, followed by biomarker identification and validation. Data, code, and other details related to the experiment are reported and shared during the reporting stage.

### Define research question

To begin with, the overall scope of the research and the potential impact in the field should be assessed by reviewing relevant scientific literature and consulting domain experts. Subsequently, feasibility of the study should be assessed, and a tentative timeline should be established. If possible, effect sizes should be anticipated to assess whether additional data and resources would be required. Finally, the research questions to be investigated and the corresponding resource requirements should be defined.

### Curate data and resources

During this step, imaging and clinical data should be gathered and curated, and access to computational resources and methods should be established. Careful curation and annotation of the imaging and clinical data are important for quality control of the data and analysis. Validated open-source software tools should be preferred over their proprietary commercial counterparts to increase reproducibility and interpretability. Moreover, if multiple analysts are working on the analysis, standardized computational platforms and software versions should be used to increase consistency and reproducibility. Researchers should strive for balance to assure that different phenotypic groups are represented appropriately in the training and validation datasets. Here, balance refers to the balanced proportion of different classes of outcome or target variables. In cases where class imbalance is inevitable, appropriate strategies like augmentation or bootstrapping can be utilized.

### Develop analysis strategy

Once the scope is defined and the required data and resources are curated, important decisions related to analysis strategies should be made. Different computational approaches should be reviewed to identify suitable methods for the analysis of radiographic data. For example, feature quantification strategies (i.e., engineered features or deep learning), image preprocessing methods, data normalization approaches, dimensionality reduction and feature selection methods, as well as different supervised and unsupervised modeling approaches, should all be explored for the exploratory empirical analysis. Another important aspect in this step is to define and lock (fix) the training and validation cohorts. Within the analysis stage, the training cohort can be used for the exploratory empirical analysis to further select and lock the computational methods. However, the validation data should be kept locked and untouched until all methods are fixed in the analysis phase (see Fig. 1). Locked validation data will prevent information leakage from training to validation and will limit the possibility of overfitting.

A commonly observed pitfall (see Table 2) that can be avoided at the design stage is when the number of samples is realistically too low to attain significance or to train a model. Insufficient training data could diminish the learning capability of a model, whereas insufficient validation data hinder the true evaluation of the underlying hypotheses. In this case, it might be best to gather more data samples, investigate only conservative questions, or postpone the study until a later time. Although seemingly undesirable, calling off a study at the design stage is preferable to investing time and effort in a study that is too premature or underpowered to achieve statistical significance. Such a premature analysis can lead to overfitting as researchers scramble to find a combination of analysis choices that give "publishable" numbers.

### Analysis: Preprocessing

In the second phase, the analysis is initiated with the goal of quantifying radiologic characteristics and ultimately, establishing and validating imaging biomarkers (see Fig. 2). When analyzing radiologic cohorts, some preprocessing is required to reduce technical variability across images. Different sources of batch variability should be investigated, including difference in scanning instrumentations, signal drifts, and other calibration-related issues as well as longitudinal effects and changes in imaging protocols. Statistical methods to correct for such batch effects should be applied. These steps ensure true assessment and validation of the underlying hypotheses.

Examples of image preprocessing prior to feature quantification include resampling of dimensions to isometric voxels to

Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385



**Table 1.** Recommendations for analyzing medical image data in radiology[a]

| | |
|---|---|
| Design: Define research question | – Define research questions<br>– Review related literature and assess the scope of the research<br>– Requirements gathering (required datasets, tools, etc.)<br>– Feasibility assessment and timeline<br>– Refine and finalize research questions and resource requirements |
| Design: Data and resource curation | – Gather and curate the required resources (data and tools)<br>– Check the quality of imaging and clinical data and perform the appropriate selection |
| Design: Develop analysis strategy | – On the basis of available data, define suitable analysis strategies<br>– Explore and decide suitable methods and computational approaches:<br>  – Feature quantification (engineered vs. deep learning)<br>  – Image preprocessing<br>  – Data normalization<br>  – Supervised vs. unsupervised<br>  – Dimensionality reduction<br>  – Statistical modeling<br>– Define the analysis flow and timeline<br>– Fix the hypotheses and their evaluation strategies<br>– Fix the training and validation cohorts and ensure no data leakage<br>– Fix the resources to be used |
| Design: Strive for balance | – Assure that different phenotypic groups are represented appropriately in the training data |
| Design: Lock data | – Lock training and validation cohorts to avoid information leakage<br>– Ensure that validation data remain locked (unused) until the exploratory analysis and biomarker identification are done on training cohorts |
| Analysis: Preprocessing | – Perform the required image preprocessing<br>– Assess potential batch effects between cohorts and apply correction methods if needed |
| Analysis: Radiomic quantification | – Explore and fix the feature quantification methods. Radiologic characteristics can be quantified, for example, using engineered or deep learned features<br>– Accordingly fix the feature transformation and data normalization approaches |
| Analysis: Biomarker identification | – Explore the feature selection/reduction and machine learning/deep learning modeling approaches using the cross-validation of training cohorts<br>– Explore different parameter settings and tuning strategies using cross-validation of training data |
| Analysis: Lock methods | – All methods and parameters should be locked (fixed) before applying them to validation data<br>– A report, testifying that validation data were not seen during the training and exploratory analysis stage, should be sent to the institutional review board, along with the list of locked methods that will be applied on the validation cohort |
| Analysis: Biomarker validation | – Evaluate performance of previously fixed methods in the validation data<br>– Perform multiple test corrections if applicable<br>– Statistically compare the performance of identified biomarker with conventional clinical markers<br>– Also, evaluate the complementary additive effect of the identified biomarker on conventional clinical markers and test whether there is a significant increase in the performance |
| Reporting: Statistics and protocols | – Report all relevant information and parameters related to each statistical test, such as the number of features tested, uncorrected and corrected *P* values, effect size, and a rationale for the choice of a model or test<br>– Report acquisition protocols of the imaging data. Also, report segmentation protocols if these were performed<br>– List all the software and tools used<br>– If space is an issue, then these details can be provided as supplementary information to the article or report |
| Reporting: Share data and methods | – Share the data and methods with the scientific community if feasible. This often brings more reproducibility in science and also increases the overall impact of a publication |

[a]These recommendations cover different steps like design, analysis, and reporting.

homogenize image resolutions. Images with isometric voxel dimensions can be either reconstructed from the raw Digital Imaging and Communications in Medicine (DICOM) data or interpolated from the image data. Moreover, some modalities, such as MRI, require normalization of image intensity values. Sources of variability and image normalization steps are critical correction measures for imaging-related batch effects.

### Analysis: Radiomic quantification

Within radiology, AI methods can perform comprehensive quantifications of tissue characteristics (see Fig. 2). These methods can convert 3D radiologic images into high-dimensional phenotype descriptors. This approach, called "radiomics," uses engineered features and/or deep learning. Here, we will describe these two approaches independently.

***Engineered features.*** Engineered features are predefined and hard-coded algorithms designed to quantify specific radiographic characteristics of diseased tissues (7). Domain-specific expertise can be used to identify important phenotypic characteristics of diseased tissues, which could then be mathematically defined and programmatically extracted in an automated manner. For example, nodular ground-glass opacity is considered as one of the vital factors for the management of pulmonary nodules (23, 24). This domain-specific knowledge has been used to define different statistical features (mean, median, variance, range, etc.) from the intensity distributions of the nodules (7). It has been shown that the ground-glass nodules have significantly lower median intensity values than partly solid or solid nodules (25). Similarly, it has been demonstrated that different radiomic features based on the shape, texture, and regional heterogeneity of the diseased tissues are associated with several clinical endpoints in oncology (4, 10, 11, 13, 25–32) and in other domains such as cardiology (14).

Additional data normalization steps like feature transformation and standardization are needed for engineered features due to the intrinsic differences of range, scale, and statistical distributions of these features. Untransformed features may have high levels of skewness, which tend to result in artificially low *P* values

Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385



**Table 2.** Commonly observed pitfalls: A list of commonly observed pitfalls in data analysis of medical image data

| | | |
|---|---|---|
| Pitfall 1: No predefined analysis protocol | – | Not defining the analysis protocols beforehand could result in testing a large number of different design strategies to optimize the performance, which often does not generalize to other independent datasets. |
| Pitfall 2: Insufficient data for training and validation | – | Insufficient training data could diminish the learning capability of a model. Insufficient validation data hinder the true evaluation of the underlying hypotheses. |
| Pitfall 3: No multiple test correction | – | When a large number of features are statistically evaluated, often there is a chance of false "discovery." Several statistical techniques, known as multiple testing corrections, are available to prevent that. No (or incorrect) multiple testing correction could induce erroneous statistical inferences and results. |
| Pitfall 4: No feature reduction | – | A large number of features increase the chance of false "discovery." Moreover, it could induce multicollinearity and overfitting. Therefore, to avoid the curse of dimensionality, feature reduction or selection methods are required. |
| Pitfall 5: Overfitting | – | Although an overfitted model works extremely well in the initial training dataset, its performance degrades substantially on other independent datasets. Proper strategies should be applied to reduce the chance of overfitting on training data. |
| Pitfall 6: No locked data and information leakage | – | Often, validation (parts of) data are used within training procedure. For example, features are selected on the basis of the complete dataset, including validation data, and then, these features are used for training a model in the training data. Validating this model in the validation cohort would be incorrect. As features are selected using the validation cohort, there is a possibility of information leakage and overfitting. Similarly, validation data cannot be used for tuning the hyperparameters. Validation data should always be locked and untouched during training. |
| Pitfall 7: Not reporting appropriate performance metrics | – | Often, using a single performance metric for the evaluation of a model is not sufficient. For example, accuracy of a classifier is sensitive to the event ratio (class distribution) of a population. Accurate evaluation can be achieved by reporting multiple performance metrics. For example, in classification studies, AUC, sensitivity, specificity, PPV, NPV, etc., should be reported, along with the accuracy of a classifier. |
| Pitfall 8: Training performance is incorrectly reported | – | Often, it has been observed that model performance on the training data is reported as results. Although this could provide information regarding the learning capability and convergence of a model, it does not give any information regarding the generalizability of a model and hence does not allow valid evaluation. Results should only emphasize the model performance in the independent validation cohorts. |

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

in downstream statistical tests for low sample size cohorts. Neglecting feature standardization may lead to individual features being over- or underrepresented in statistical models and eventually introduce bias into the analysis. Methods like standardization and logarithmic transformations (33) can transform features into zero-centered distributions having identical variances and symmetric distributions.

***Deep learning.*** Deep learning enables the extraction of multiple feature levels from data directly without explicit definition (8, 34, 35). It provides a higher level of feature abstraction, thus potentially providing better prediction performance. Although deep learning has recently achieved impressive successes in practice, a strong theoretical backing is yet to be established (36). The lack of rules of thumb makes choosing an appropriate deep learning network architecture challenging. The problem at hand: classification, detection, segmentation, registration, or reconstruction, in addition to the type and size of data, all hint at the appropriate architecture to utilize. Starting from published architectures that have proven successful in their respective tasks is common practice. Convolutional neural networks are the most prevalent deep learning architectures in medical imaging today (37). Transfer learning, or using pretrained networks, is often an attractive option when dealing with scarce data (38). Data normalization is an essential data preprocessing step for deep learning. It ensures increased numerical stability and quicker and stable convergence. This could be achieved through sample-wise, feature-wise, or principal components analysis (PCA) whitening normalization depending on the data type. It is important to note, however, that even with inputs being normalized, their distributions are highly susceptible to change as they propagate through the network where parameters are constantly optimized during training—a problem referred to as "internal covariance shift"—and can be mitigated using batch normalization (39) or layer normalization (40). Normalization is hence made to be an integral part of the network architecture, as opposed to merely being a preprocessing step, and is performed on inputs to each layer based on the distribution of the given batch of cases.

A few important points should be considered before choosing the quantification method. Although deep learning is able to perform quantifications in an automated manner, it generally requires fairly large datasets. Often with quantitative imaging, it could be challenging to gather and curate large cohorts of patients with similar clinical, imaging, and demographic characteristics. Engineered features are less sensitive to the cohort size but need to be defined manually by experts. Moreover, they also require the segmentation of the diseased tissue. Recent efforts utilizing transfer learning have broadened the applicability of deep learning approaches to smaller cohorts (38, 41). Hence, these points should be carefully considered prior to choosing a particular quantification method.

### Analysis: Biomarker identification

Quantitative analysis of medical image data involves mining large number of imaging features, with the goal of identifying highly predictive/prognostic biomarkers. Here, we discuss these concepts for engineered features and deep learning methods separately.

***Engineered features.*** A large number of engineered features are extracted from the medical image data, which might be highly redundant. The use of appropriate feature selection/reduction

Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385



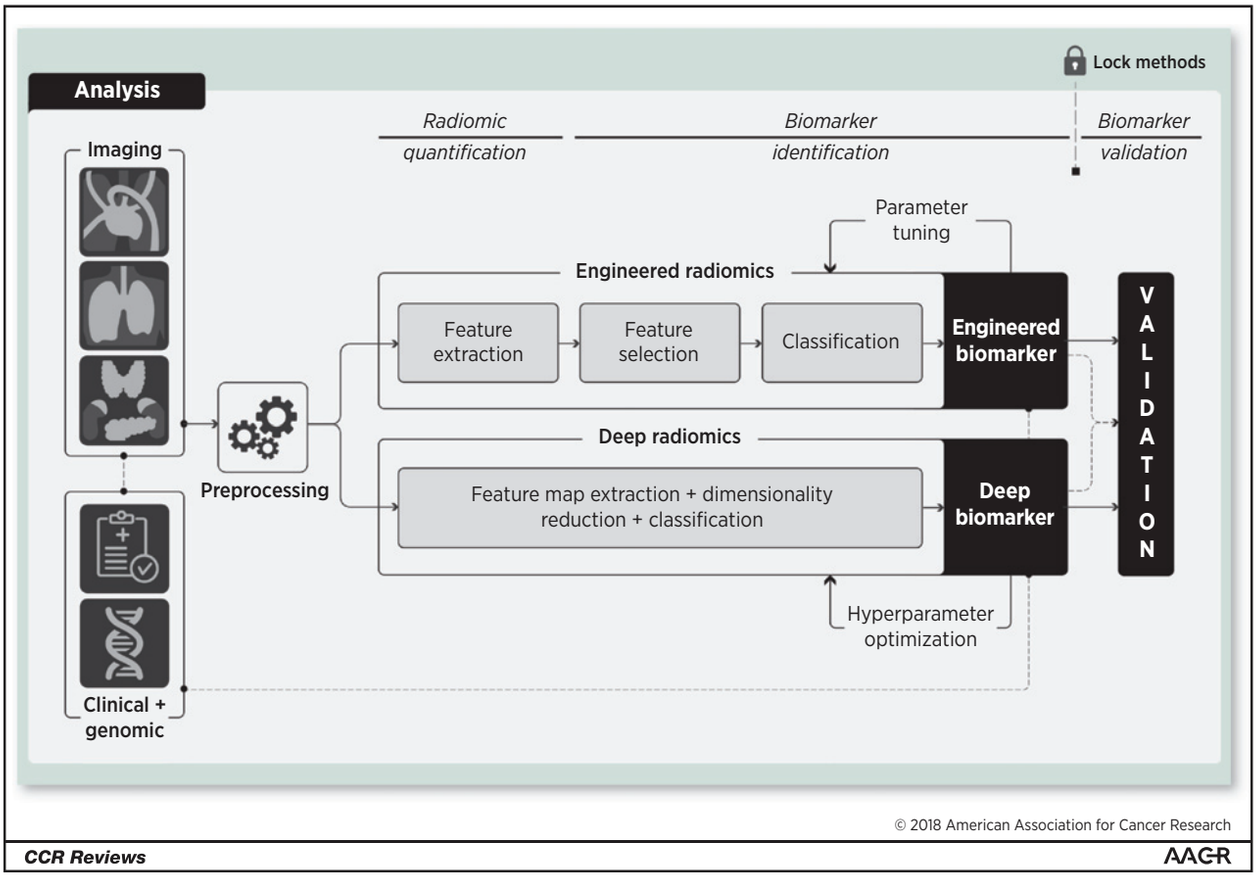


**Figure 2.**
Detailed analysis stage. Analysis begins with the preprocessing of medical images to avoid different technical variability and batch effects. After that, in the radiomic quantification step, radiomic descriptors capturing different phenotypic characteristics of diseased tissues are quantified. Radiomic quantification can be done using either engineered features or deep learning methods. According to the quantification method, in the biomarker identification step, appropriate analysis methods are explored, and suitable methods are applied to develop biomarkers. Finally, in the biomarker validation step, the developed biomarker is validated in the locked and independent validation cohort.

strategies can minimize the feature redundancies and mitigate the "curse of dimensionality" (42). Many unsupervised and supervised feature selection methods have been described in the literature (10). Unsupervised methods based on PCA analysis, independent component analysis, or correlation can reduce the feature space without using a target outcome. Alternatively, supervised methods can be used to select features that are relevant to the particular target outcome (43, 44). These supervised methods can be separated into three categories: wrapper, embedded, and filter methods. Computationally expensive wrapper and embedded methods use stricter model structure assumptions and are likely to have low generalizability, whereas model-independent filter methods are relatively efficient and provide better generalizability (43, 44).

Predictive and prognostic models with high accuracy, reliability, and efficiency are vital factors driving the success of quantitative imaging approaches (10). Myriad supervised and unsupervised methods exist in the machine learning literature derived using different statistical assumptions (42, 45, 46). Several investigations have compared these methods for quantitative image analyses (10, 47–49,). These studies have demonstrated that the variability of prediction scores is highly influenced by method choice (10, 42). Therefore, the choice of feature selection and machine learning modeling methods should be cautiously made.

One of the most commonly observed pitfalls during the biomarker identification stage is overfitting. Researchers have a tendency to exhaustively search through different modeling methods and parameter configurations to obtain high and publishable performance, which can result in poor biomarker generalizability. To avoid this, use of standard settings for hyperparameters (50) or use of locked and independent validation data (51) is recommended.

***Deep learning.*** In deep learning, dimensionality reduction and classification are performed alongside feature extraction in an integrated manner (see Fig. 2). However, the quality and output of these cascaded layers depend on different hyperparameters such as the number of layers and feature maps, layer arrangement and architecture, size of the receptor field, etc. Many different network architectures have been described using different sets of hyperparameters (52, 53). Prediction performance could be

Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385



influenced by the choice of these hyperparameters and architectures. As in the case of engineered features, optimizing the hyperparameters by exhaustively searching the parameter space is a common pitfall in deep learning as well. The deep learning architecture should be selected on the basis of the underlying scope and application of the research, the statistical properties of data in hand, and the effective data size.

Overfitting represents a major challenge in deep learning and can drastically affect a network's ability to generalize to unseen data. Often, deep learning methods are treated as a black box, and not enough attention is given to understanding the actual methods and technical concepts. This is particularly an issue when working with limited training data. Deeper and more complex networks trained on limited data could induce overfitting. The use of shallower networks could avoid overfitting but may result in insufficient learning, also known as underfitting. One solution could be data augmentation, where the training data are expanded by applying label-preserving image transformations (54) like cropping, reflections, and rotations. Dropout or other regularization methods could also be used to reduce overfitting (55, 56). Dropout makes the network less sensitive to individual weights and hence increases the robustness of a network. Other regularization methods allow penalizing large parametric weights to make the network more robust and generalizable. During training, the network performance should be evaluated and monitored using cross-validation. A cross-validation–based early-stopping (57) approach can also be utilized to avoid overfitting. These important factors should be considered during the biomarker identification stage.

After identifying biomarkers, a document should be created listing all the cross-validation and analysis steps taken during the exploratory analysis of training data, the final hypotheses to be validated, the training data used and corresponding inclusion criteria, as well as a list of identified and locked computational approaches. Furthermore, it should also be declared that the validation data were not used during the exploratory analysis or biomarker identification stage. The same document can later be used to incorporate validation results and reported alongside the study findings.

**Analysis: Biomarker validation**

As described in the previous sections, avoiding overfitting and data leakage is essential when working with machine learning and deep learning models. It should be ensured that locked validation cohorts remain blinded during the training and hyperparameter tuning. Only after fixing the models, network architectures, computational methods, and corresponding hyperparameters should the validation be carried out on the locked validation cohorts. These steps ensure the true evaluation of the underlying hypotheses.

Appropriate performance metrics should be used, especially when dealing with highly unbalanced classes. For example, the accuracy of a classifier is potentially sensitive to the event ratio (class distribution) of a population, resulting in deceivingly overoptimistic results. Accurate evaluation can be achieved by reporting multiple performance metrics like AUC, sensitivity, specificity, positive predictive value, negative predictive value, etc. Moreover, robustness of the biomarkers with respect to data perturbation is also a vital aspect of biomarker validation (58). An essential and often overlooked step before choosing candidate biomarkers is performing multiple testing corrections. When testing hundreds of features, it is expected that some associations with clinical outcomes can be found entirely by chance. To avoid this, correction methods like Bonferroni (59) and Benjamini and Hochberg (60) should be applied.

To gauge the true clinical impact, it is also essential to statistically compare the developed biomarkers to standard clinical markers during validation. Furthermore, the additive increase in performance should also be evaluated by combining the developed biomarkers and standard clinical markers in a computational model. Moreover, emphasis should be given to multiple external validation sets, which can ensure the robustness and generalizability of a biomarker. It is our view that studies demonstrating biomarker robustness through reproducibility should be looked upon favorably.

**Sharing and reporting**

In the final phase, a detailed report should be made about the data, analysis methods, and results. Furthermore, enough emphasis should be given to sharing the analysis protocols, documented code, original as well as processed data, along with detailed descriptions.

***Share data and code.*** Several technical and privacy issues should be considered while sharing biomedical imaging data. Patient confidentiality must be protected throughout the entire process. A system similar to that of the database of Genotypes and Phenotypes (dbGaP) should be adopted, where raw imaging data can be archived and distributed in a manner that enables data sharing without compromising patient confidentiality. The publicly available online medical imaging repository The Cancer Imaging Archive (http://www.cancerimagingarchive.net/) is a great place for sharing and archiving medical image data, as it uses a standards-based approach to de-identification of DICOM images to ensure that images are free of protected health information. For analysis, repositories like GitHub (http://github.com) allow for relative ease in sharing code. Furthermore, computational tools such as Jupyter Notebooks and Sweave can make "executable documents," which should be included as supplements to scientific articles to promote reproducibility and generalizability. Experiments housed in software environment containers with locked versioning such as Docker (www.docker.com) allow for swift reproducibility and portability.

***Share statistics and protocols.*** As building imaging biomarkers from radiologic data involves multiple analytic steps, sharing comprehensive descriptions of statistical methods is essential for reproducibility. In the shared document, we recommend reporting the number of samples and features tested, feature definitions, statistical tests and algorithms used during the analysis, list of model and hyperparameters used, details regarding optimization methods, nominal and corrected *P* values, and effect size assessments. Once imaging data have gone through the quantification steps outlined above, the result is normally a matrix of features versus samples. As long as sample labels are properly de-identified, this feature matrix can usually be shared without any risk to patient confidentiality. Ideally, all code required to reproduce study findings from the feature matrix should also be shared, with the exception of sections that require the use of protected clinical metadata.

An effective strategy often adopted for data sharing is to group data, analysis reports, and runnable code blocks into a single object such as R data packages with vignettes. For studies that use

Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385



deep learning approaches, the type of network architecture and hyperparameters should be shared. In addition, sharing entire networks with trained weights can be highly beneficial for transfer-learning efforts. With an increasing number of deep learning libraries, sharing trained networks in universal formats such as ONNX (www.onnx.ai) can facilitate straightforward cross-platform compatibility. It is our experience that sharing methods and code in this manner also has the natural effect of raising the overall quality of data analyses in addition to ensuring reproducibility.

Data science and big data are rapidly becoming major components of health care applications in both the industrial and academic settings. However, performing accurate and meaningful data analyses is challenging and potentially plagued with errors and incorrect conclusions. In this article, we addressed data analysis practices specific to radiographic medical images. We identified commonly observed pitfalls and recommended vital guidelines. Recommendations given here are not exclusive solutions and do not guarantee completely error-free unbiased experimentation. However, they do act as guidelines for the management and handling of research-compromising pitfalls. Mindful implementation of these guidelines will enhance the quality of radiomic analyses as well as allow for better integration of imaging data with other patient-specific data for precision medicine efforts.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Acknowledgments

The authors acknowledge financial support from the NIH (NIH-USA U24CA194354 and NIH-USA U01CA190234 to H.J.W.L. Aerts).

Received January 31, 2018; revised February 26, 2018; accepted March 22, 2018; published first March 26, 2018.

## References

1. Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. JMIR Med Inform 2014;2:e1.
2. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights 2016;8:1–10.
3. Aerts HJWL. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncol 2016;2:1636–42.
4. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.
5. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging 2012;30:1234–48.
6. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441–6.
7. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104–7.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
9. Rusk N. Deep learning. Nat Methods 2015;13:35.
10. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. Sci Rep 2015;5:13087.
11. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RTH, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114:345–50.
12. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. Front Oncol 2016;6:71.
13. Huynh E, Coroller TP, Narayan V, Agrawal V, Romano J, Franco I, et al. Associations of radiomic data extracted from static and respiratory-gated CT scans with disease recurrence in lung cancer patients treated with SBRT. PLoS One 2017;12:e0169172.
14. Kolossváry M, Kellermayer M, Merkely B, Maurovich-Horvat P. Cardiac computed tomography radiomics: a comprehensive review on radiomic techniques. J Thorac Imaging 2018;33:26–34.
15. O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol 2017;14:169–86.
16. Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002;32:496–501.
17. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;11:733–9.
18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118–27.
19. Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci U S A 2000;97:9834–9.
20. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 2006;10:515–27.
21. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65.
22. Aerts HJ. Data Science in radiology: a path forward. Clin Cancer Res 2018;24:532–4.
23. Lee C-T. What do we know about ground-glass opacity nodules in the lung? Transl Lung Cancer Res 2015;4:656–9.
24. de Hoop B, Gietema H, van de Vorst S, Murphy K, van Klaveren RJ, Prokop M. Pulmonary ground-glass nodules: increase in mass as an early indicator of growth. Radiology 2010;255:199–206.
25. Yip SSF, Liu Y, Parmar C, Li Q, Liu S, Qu F, et al. Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer. Sci Rep 2017;7:3519.
26. Nie K, Chen J-H, Yu HJ, Chu Y, Nalcioglu O, Su M-Y. Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. Acad Radiol 2008;15:1513–25.
27. Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, et al. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. Tomography 2016;2:430–7.
28. Jain R, Poisson LM, Gutman D, Scarpace L, Hwang SN, Holder CA, et al. Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. Radiology 2014;272:484–93.
29. Bae JM, Jeong JY, Lee HY, Sohn I, Kim HS, Son JY, et al. Pathologic stratification of operable lung adenocarcinoma using radiomics features extracted from dual energy CT images. Oncotarget 2017;8:523–35.
30. Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife 2017;6:e23421.
31. Rios Velazquez E, Parmar C, Liu Y, Coroller TP, Cruz G, Stringfield O, et al. Somatic mutations drive distinct imaging phenotypes in lung cancer. Cancer Res 2017;77:3922–30.
32. Parmar C, Leijenaar RTH, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Sci Rep 2015;5:11044.

Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385




33. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002;18:S96–104.
34. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2017 May 6. [Epub ahead of print].
35. Kevin Zhou S, Greenspan H, Shen D. Deep learning for medical image analysis. Cambridge (MA): Academic Press; 2017.
36. Wang G. A perspective on deep imaging. IEEE Access 2016;4: 8914–24.
37. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
38. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35: 1285–98.
39. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Ithaca (NY): Cornell University; 2015 [cited 2018 Jun 10]. Available from: http://arxiv.org/abs/1502.03167.
40. Ba JL, Kiros JR, Hinton GE. Layer normalization. Ithaca (NY): Cornell University; 2016 [cited 2018 Jun 10]. Available from: http://arxiv.org/abs/1607.06450.
41. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016;3:9.
42. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer Science + Business Media; 2013.
43. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.
44. Brown G, Pocock A, Zhao M-J, Luján M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 2012;13:27–66.
45. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. Cambridge (MA): MIT Press; 2012.
46. Fernández-Delgado M, Cernadas E, Barro S. Do we need hundreds of classifiers to solve real world classification problems. J Mach Learn Res 2014;15:3133–81.
47. El Naqa I, Li R, Murphy MJ. Machine learning in radiation oncology: theory and applications. Cham (Switzerland): Springer; 2015.
48. Wang J, Wu C-J, Bao M-L, Zhang J, Wang X-N, Zhang Y-D. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. Eur Radiol 2017;27:4082–90.
49. Zhang B, He X, Ouyang F, Gu D, Dong Y, Zhang L, et al. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. Cancer Lett 2017;403:21–7.
50. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. Front Oncol 2015;5:272.
51. Skocik M, Collins J, Callahan-Flintoft C, Bowman H, Wyble B. I tried a bunch of things: the dangers of unexpected overfitting in classification. BioRxiv [Preprint]. 2016 bioRxiv 078816 [posted 2016 Oct 3; cited 2018 Jun 10]: [19 p.]. Available from: https://www.biorxiv.org/content/early/2016/10/03/078816. doi: https://doi.org/10.1101/078816.
52. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35:1798–828.
53. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117.
54. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25. Red Hook (NY): Curran Associates; 2012. p.1097–105.
55. Bell RM, Koren Y. Lessons from the netflix prize challenge. SIGKDD Explor Newsl 2007;9:75–9.
56. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
57. Prechelt L. Early stopping - but when? In: Orr GB, Müller K-R, editors. Neural networks: tricks of the trade. Berlin/Heidelberg (Germany): Springer; 1998. p. 55–69.
58. Beck AH, Knoblauch NW, Hefti MM, Kaplan J, Schnitt SJ, Culhane AC, et al. Significance analysis of prognostic signatures. PLoS Comput Biol 2013; 9:e1002875.
59. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936;8:3–62.
60. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57:289–300.

Published OnlineFirst March 26, 2018; DOI: 10.1158/1078-0432.CCR-18-0385



Clinical Cancer Research

# Data Analysis Strategies in Medical Imaging

Chintan Parmar, Joseph D. Barry, Ahmed Hosny, et al.

*Clin Cancer Res* 2018;24:3492-3499. Published OnlineFirst March 26, 2018.