TSR ARTIFICIAL INTELLIGENCE APPLICATIONS GUIDE

TSR ARTIFICIAL INTELLIGENCE APPLICATIONS GUIDE

1. Introduction

We are entering an era of profound changes in healthcare and radiology, as in many other areas of life, with the increasing number of successful applications developed within the scope of artificial intelligence (AI) and the introduction of learning algorithms. Although it is accepted that we are in the infancy stage of this inevitable change, it is seen that the change will be faster than expected.

Our association, while foreseeing that this change carries a potential threat to our profession, aims to focus on the power of AI as a game-changing technology product to provide more new opportunities and to ensure that applications are carried out for the benefit of patients and within the framework of ethical rules. This guide has been developed to guide our colleagues in the problems they may encounter in the field of AI, and to contribute to the use of AI in a way that will increase the quality of radiological evaluations, both in the algorithm development phase and in clinical applications.

Our main goal is the continuous development and widespread use of the guide, which is a joint product of the TSR Imaging Informatics Working Group.

2. Related Definitions

This section provides definitions of frequently used terms in core resources to aid in the understanding of AI and related concepts in radiology. Terms are presented with their English equivalents. For more detailed definitions, the TSR Imaging Informatics Dictionary at https://safari.net.tr/trd-gbs/ can be consulted.

Here are the detailed explanations of the terms included in this section:

- **Algorithm:**
  - The term used for the arithmetical path that determines the steps and the goal followed for the realization of a process.
- **Algorithmic Bias:**
  - The situation where AI models make incorrect or misleading predictions for certain patient groups or disease types due to imbalances in the training data.
- **Artificial Intelligence:**
  - The name given to studies that aim to realize the skills that are thought to be unique to humans and generally used for communication and problem solving by machines.
- **Attribute:**
  - The general name of the features in images.
- **Convolutional Neural Networks:**
  - A type of multilayer artificial neural network in which the convolution operation is used instead of matrix multiplication in at least one of its layers.
- **Data Mining:**
  - The effort to extract meaningful results from patterns within big data.
- **Deep Learning:**
  - A machine learning field that operates with artificial neural networks and similar algorithms containing one or more hidden layers.
- **Expert Systems:**
  - A computer program developed to solve a problem, created by utilizing expert knowledge, and often using rules in the form of "if/then".
- **Explainable AI (XAI):**
  - A set of approaches that make the decision-making mechanisms of AI models transparent.
  - Enables users to understand how the model works.
- **Federated Learning:**
  - A method of training an algorithm using data from different endpoints instead of collecting and using data in a central location in machine learning.
- **Generative Adversarial Networks:**
  - Networks capable of making productions similar to real images by comparing generated data with real images and using received feedback.
- **Labeling:**
  - The process of naming or numbering objects with marking tools in informatics applications.
- **Large Language Models:**

    - AI models trained with very large-scale text data, capable of generating human-like text, extracting meaning, and performing a wide variety of language tasks.
    - They learn language structure and context, enabling functions like text completion.
- **Machine Learning:**
    - An activity area that forms a subset of artificial intelligence applications and aims to learn from experiences or extract information from examples.
- **Model Training:**
    - The process by which an artificial intelligence model develops the ability to solve a specific problem by working on the given training data.
- **Natural Language Processing:**
    - The entirety of operations that enable the interpretation of texts, usually in plain text, according to certain procedures.
- **Neural Networks:**
    - A computer program or system that produces solutions by mimicking neuron behavior.
- **Overfitting:**
    - A situation where an AI model performs poorly on new data due to excessive adaptation to the training data.
- **Parsing:**
    - The process of separating text according to grammar rules and determining the relationships between words in natural language processing studies.
- **Radiogenomics:**
    - A field that aims to develop personalized diagnosis and treatment methods by using data obtained from medical imaging and genetic information together.
- **Radiomics:**
    - A field of study that aims to extract a large number of quantitative features from medical images using data characterization algorithms [1] and to find the clinical equivalents of these features.
- **Segmentation:**
    - The separation of a desired region or organ from the main structure using various algorithms.
- **Supervised Learning:**
    - A learning method in machine learning that requires the input of image features.
- **Test Set:**
    - An independent data set used to measure how a model will perform in the real world after the training and validation phases of a model are completed.
- **Texture Analysis:**
    - The mathematical evaluation of spatial heterogeneity in a selected region of an image.
- **Training Set:**
    - The data set used in the learning process of machine learning and deep learning models.
- **Transfer Learning:**
    - The use of all or part of the connection coefficients and activation values of a machine learning algorithm for solving other problems after learning the solution to a problem for classification, regression, or prediction.
- **Unsupervised Learning:**
    - A learning method in machine learning where the features used for image evaluation are extracted by the algorithms used.

- **Validation Set:**
    - The data set used to monitor the performance of machine learning models and make hyperparameter adjustments during the training process.

3. General Principles

The general principles regarding the development, accessibility, trade, use, and responsibility sharing of artificial intelligence applications are also valid and similar for applications to be developed for the field of radiology. In accordance with the recommendations of relevant international institutions, it would be appropriate to pay attention to the following principles:

- The effects of the use of AI models in the healthcare field on the patient, such as the role they play in diagnosis and treatment decisions, should be explainable and shared with the patient.
- The decision to continue or discontinue the use of artificial intelligence models should be made by authorized persons, and models should be able to be deactivated when necessary.
- Guiding principles should be determined by organizations with sanctioning power, taking into account the views of all stakeholders and the unique strengths and weaknesses of countries.
- The protection of personal data to be used in artificial intelligence studies should be a top priority. Individuals should be informed about which data will be used, for what purpose, and how.
- The development stages of artificial intelligence applications should be closely monitored, awareness should be raised about their potential dangers as well as their benefits, and necessary precautions should be taken at an early stage.
- Before the use of artificial intelligence models, radiology specialists and other users should be provided with training on basic issues such as the operating conditions, benefits and potential harms of the system, when to prefer it, and when it may be insufficient.
- It should be guaranteed that all segments of society can benefit equally and fairly from the potential benefits of AI models in the healthcare field.
- In cases where AI models are supported in academic studies, the results obtained should be meticulously checked, and the authors should bear the ethical responsibility of the final product. Copyrights should be observed, the benefit obtained from the AI model should be stated, and data security and confidentiality should be ensured.
- The algorithms to be developed should be free from biased elements that are contrary to human and patient rights and that restrict accessibility.
- Developers of models that are rapidly being used by the public, especially large language models, should have ethical values in the healthcare field.
- As the use of artificial intelligence applications increases and becomes widespread, recurring periodic studies should be carried out on possible ethical and legal problems that may be encountered.
- The purpose of using artificial intelligence in the processing and analysis of medical images is to enable experts to better interpret images and make more accurate diagnoses, increase clinical utilization, and make the work more economical in terms of time and cost. Artificial intelligence will not be a competitor or alternative to radiologists, but rather a helpful and useful tool they use. To avoid a wide variety of irreparable medical

errors, systemic failures, and chaos, artificial intelligence applications should not be put into use without radiologist supervision.

- Artificial intelligence studies should not pursue goals that will take away the jobs of radiologists.
- Radiology specialists and other healthcare professionals should develop their knowledge and skills in AI and be open to possible new job divisions.
- Decisions should be made by determining priority needs in the stages of developing artificial intelligence applications and including them in the workflow.
- The role of AI applications in different clinical scenarios, their place in the workflow, and how AI-human collaboration will be should be determined in cooperation, and the place of AI applications in the workflow should be determined.
- Before artificial intelligence applications are used, the institution's infrastructure capabilities should be reviewed, and trial versions should be requested.
- Improving patient-physician communication, enriching the exchange of ideas between physicians, increasing the working efficiency of imaging units, continuous professional development, and directing specialist students' education towards research that can generate added value should be among the goals of radiology specialists who gain extra time with AI applications.
- Studies to be conducted on artificial intelligence should aim to protect and improve the patient's health level and improve the deteriorated health condition.
- Before artificial intelligence models are put into use, their usefulness and limitations on the target audience should be tested. Attention should be paid to applications that disrupt the workflow, are time-consuming, and do not provide benefits that will make a difference in routine practice.
- In the use of artificial intelligence-supported decision models, care should be taken not to restrict the patient's freedom to make decisions on matters related to themselves.
- Producing and using high-quality data should be the main goal. The 10V rule with its English equivalents can be taken as an example in evaluating data quality: Vagueness, Variability, Venue, Variety, Value, Volume, Veracity, Validity, Vocabulary, Velocity.
- The diversity of data sets used for the purpose of developing artificial intelligence applications, image quality, and the accuracy of labels, if any, should be checked.

4. Principles to be Followed in Application Development Processes

4.1. Clinical Problem and Context Determination

Before developing an artificial intelligence-supported model, the first step should be to identify the clinical problem. After identifying situations in practice that are incomplete, complex, require numerous calculations, and have accuracy issues, existing AI applications are evaluated. Accordingly, a plan is made to develop a new model in line with the identified deficiencies. Technical and clinical success criteria (sensitivity/specificity targets, performance metrics, shortening of examination interpretation time, improvement of patient survival, etc.) should be determined in advance, and the expected benefit of the AI-supported model to be developed for the healthcare system should be clearly demonstrated. It should be evaluated whether the benefits outweigh the costs of potential negative results that may arise with the model. To avoid problems in product conversion, the guidelines and regulations of regulatory institutions in the target market (e.g., Software as a Medical Device (SAMD): Clinical Evaluation for the United States, EU Artificial Intelligence Act for the European Union) should be carefully reviewed.

The American College of Radiology (ACR) Data Science Institute (DSI) can also be utilized through the Define-AI Directory, which defines target use scenarios for the development of effective models suitable for clinical application.

4.2. Data Collection and Data Management

4.2.1. Data Collection Strategies

Clinical information, laboratory findings, radiological images, DICOM metadata, or expert labels used in the model development process are called data. Data sources can be PACS (Picture Archiving and Communication Systems), electronic health records, radiology reports, pathology reports, research and imaging databases. The performance of artificial intelligence models is affected by data diversity and quality as well as data size.

It is a common observation that model performance is directly proportional to the amount of data. However, it is currently not possible to accurately estimate the ideal sample size for obtaining a successful and generalizable AI model. The minimum number of patients required to develop a model largely depends on the complexity of the problem we are trying to solve, the targeted model performance, and the modeling method to be used. For example, disease detection problems are considered simpler than prognostic problems based on predicting treatment response, two-group classification problems are considered simpler than three-group classification problems, and radiomics methods are considered simpler than artificial neural networks, so they can be planned with a smaller patient cohort.

From a purely statistical perspective, in a binary classification problem, 10-15 patients are required in the training set for each feature participating in the radiomics signature. Several basic approaches have been defined that can be used as a guide in determining the data set to be used in the training of artificial neural networks. The first approach is to determine the training data set according to the number of classes, in which case the recommended minimum number of data is 50-1000 times the number of classes. The second approach is based on the total number of attributes used. Here, the recommended minimum number of data is between 10-100 times the number of attributes. However, the most commonly used rule is usually based on the number of weights in the network. According to this rule, the recommended minimum number of data should be at least 10 times the number of weights. However, these rules are generally simplified approaches to facilitate ease of use in real-world applications, and in specific cases, the sample size needs to be optimized with the learning curve drawn against the success rate.

It is important that the collected data represents the target population and reflects the heterogeneity and diversity in the real world as much as possible. Artificial intelligence systems rely on training data and lack context. It is important that the data used to train the artificial intelligence system represents the patient population in which the system will be used. In other words, if a high-performance model is desired, a training set with a wide variety of cases covering various diseases, anatomical variations, imaging protocols, reconstruction methods, scanner models, and demographic factors is essential. If the training set lacks diversity, the performance of the AI model will decrease when it encounters new cases. The representativeness of the data should be comprehensively evaluated during the algorithm development process. Possible differences between the collected data and the target population characteristics (e.g., which groups are not represented or underrepresented) should be identified

and reported. In this way, the data can be interpreted correctly and the validity of the results can be better evaluated.

The size and representativeness of the training set can be increased by using multicenter studies or open-source image archives. However, multicenter research designs also bring challenges, especially in terms of data privacy and security. Open-source image archives, on the other hand, may be insufficient, especially for accessing images of rare diseases. Excluding classes consisting of rare diseases will bring selection bias, so the target population representativeness of the data can be increased and class imbalance can be reduced by grouping all of these diseases under the "other" class.

There are many obstacles in large-scale medical imaging data collection and sharing. Differences in DICOM meta tags and naming of imaging examinations are among the main problems affecting data integrity. The same body region can be labeled with different titles in different clinics or devices. Human errors and lack of training on data management can also create variability and errors in the data. Solving these problems requires increasing cooperation between device manufacturers and strengthening data management awareness. In addition, making the data entered into electronic medical record systems analyzable should be the goal of all healthcare systems. The difficulty of obtaining structured data from free-text radiology reports can be partially overcome with natural language processing techniques.

It may not always be possible to create large data sets due to patient privacy concerns and the difficulty of accessing high-quality labeled data. Three basic strategies are used to develop high-performance models with limited data: (i) data augmentation: The data set is enlarged by creating synthetic images with methods such as rotation, noise addition, scaling, cropping, brightness and contrast adjustment; (ii) semi-supervised learning: In cases where full labeling is costly, images with "pseudo-labeling" by a partially trained model are also included in the training; (iii) transfer learning: A model pre-trained on a large data set is used as a starting point and fine-tuned to adapt to the relevant task.

4.2.2. Data Improvement (Data Curation)

Using larger data sets does not always guarantee obtaining higher-performance models. This is because poor quality or incorrectly labeled data can also increase the error rates of the model. One of the prerequisites for obtaining high-performance models is to use well-organized and correctly labeled data sets. However, challenges are experienced in creating such data sets due to labor-intensive labeling processes.

Data curation refers to methods of ensuring the quality, consistency, and overall integrity of data. It includes activities such as identifying inconsistencies and duplicates in the database; detecting missing data, data in different image formats; finding low-quality images, images planned with incorrect examination windows or imaging protocols.

Detailed evaluation of data quality is frequently recommended. It should be clearly stated how, when, and with which tools each variable was measured. Limitations regarding data quality must be reported. For data quality and mislabeling control, data should be manually evaluated for mislabeling by randomly selecting subgroups, if possible, and the error rate should be reported.

In the literature, significant emphasis is placed on the quality of reference standard data in particular. How the data was collected (biopsy, radiology reports, laboratory tests, etc.) and potential problems during collection should be discussed. If manual labeling was performed, the experience of the labelers must be stated. In order to eliminate bias in the evaluation of model performance, it is important that the labeling is done independently and that experts do not directly participate in the evaluation of model performance. Inter-observer variability should be calculated and reported to measure the quality of labels.

Commonly used preprocessing techniques include data augmentation, outlier removal, variable transformation (e.g., scaling, standardization), and imputation. It is important to clearly state the reasons for the transformations applied to the data and the preprocessing steps, along with the software used.

Two solutions are generally recommended for dealing with missing data: imputation and exclusion. In addition, using machine learning methods that can work with missing data is also an option. Generally, imputation is a preferred solution method compared to completely excluding the data.

Image features may vary depending on the use of different scanners, imaging protocols, or reconstruction parameters. These systemic and technical differences that mask biological and diagnostic information are also called batch effects. Batch effects can be a more prominent factor, especially in multicenter studies or scenarios where magnetic resonance (MR) images are used. This is because, in magnetic resonance imaging (MRI), the signal intensity of pixels is determined in a non-standardized way, unlike "Hounsfield Unit" (HU), and varies between manufacturers. Solutions aimed at eliminating batch effects are called data harmonization. Harmonization can be applied at the image or data level. ComBat, an effective solution used at the data level, makes data obtained from different devices or centers compatible, making analytical results more reliable and comparable. The most commonly used approach at the image level is normalization, which includes a series of techniques where pixel values are shifted and/or scaled, and can be applied to different image features (such as spatial normalization or intensity normalization). Spatial normalization is any process that changes the spatial properties of the image (pixel size, field of view-FOV, series orientation, etc.). With intensity normalization, the gray values of the pixels are rescaled to the same range, and the effect of numerically large values is balanced. Thus, the training time of the model is usually reduced and its performance is improved.

Image resampling is the general name for all geometric transformation operations of digital images. This includes operations such as creating new data points with image rotation or spatially aligning images obtained at different time points. Medical images are generally larger in volume and three-dimensional, so they are more complex than non-medical image tasks. Since convolutional neural networks are usually trained with smaller-sized (e.g., 300 x 300 pixels) two-dimensional images, processing medical images may require more computing power. In this case, it may be necessary to reduce the resolution of medical images or perform patch-based evaluation. Patch-based evaluation reduces the computational load by dividing the image input into sub-sections and enables classification based on a specific region (e.g., an algorithm aimed at localizing prostate cancer focuses only on pixels containing the prostate gland). Quantitative image analysis methods also require image resampling. Since radiomics tries to express tissue heterogeneity with mathematical formulas, it can be affected by voxel size. Therefore, it is important to make voxels isotropic in three-dimensional radiomics studies.

Model performance can be improved with image preprocessing, which includes improving data quality by removing noise and artifacts. Spatial mapping can be easily performed with four-dimensional data, such as in lung or cardiac computed tomography (CT), dynamic imaging, and diffusion-weighted imaging, with motion correction procedures. Filtering and motion correction procedures are methods that can also distort diagnostic information in images and should be avoided as much as possible, especially in prospective studies. MR images are mostly positively affected by noise that distorts biological tissue properties due to inhomogeneity in the magnetic field, so the use of "bias-field" correction procedures is recommended.

4.2.3. Ensuring Data Privacy

One of the main obstacles in the development of artificial intelligence tools is ensuring data privacy and overcoming ethical problems. Before starting data collection, it must be ensured that the project complies with local personal data privacy legislation. Data protection experts should be consulted to take appropriate privacy measures when necessary. Especially in prospective studies, patients should be informed about the rationale for using patient data and how the data will be used, and informed consent should be obtained. In retrospective studies, informed patient consent may not be required since patients do not need to undergo an additional procedure. However, whether there is a valid reason for using the data must be evaluated by ethics committees in both cases. Especially in retrospective studies, ethics committee approval may be easier in cases where obtaining explicit consent from patients is not possible, the risks associated with data sharing are minimal, and data controllers can be trusted.

In the process of developing AI-based solutions that can be used in the real world, it may be necessary to cooperate with the private sector. Since artificial intelligence developers do not have direct access to PACS ("Picture Archiving and Communication System"), the data needs to be prepared and transferred. Before data transfer, patient privacy must be protected by de-identification. De-identification can be done in two different ways: anonymization and pseudo-anonymization. Anonymization refers to the irreversible removal of patient-related information from records. This method is the preferred approach for sharing medical data. Pseudo-anonymization refers to replacing patient information with artificial values so that the original data can only be revealed with a secret key. All sensitive health data to be transferred must be removed from both DICOM (“Digital Imaging and Communications in Medicine”) metadata and images. The basic anonymization methods offered by most PACS during export may not be sufficient. There are multiple tools to automatically and freely remove identifiable information from DICOM metadata. DICOM Library and RSNA Clinical Trials Processor can be used for this purpose. These are two free and proven tool sets. Even anonymized and metadata-stripped images can allow access to patients' identity information, especially in head imaging, by detecting facial features. Tools such as Pydeface (https://github.com/poldracklab/pydeface) or mridefacer (https://github.com/mih/mridefacer) can help automatically remove such facial features from medical images.

After de-identification, data can be stored for a certain period by transferring it to physical servers or the cloud. While cloud-based data storage can bring security concerns, it is an expensive solution that requires high internet speed. However, it can greatly benefit multicenter studies by facilitating data sharing. To protect privacy during the data storage process, data encryption, limiting data access to authorized persons, securely logging transaction records of all actions performed on the data, regularly conducting security tests of the environments where the data is located, and transparently reporting data breaches are required. In addition, the

duration for which data can be stored should be determined to minimize risks from unauthorized access.

In our country, the Personal Data Protection Law No. 6698 (KVKK) has been published, thus regulating activities such as the protection, recording, processing, and sharing of personal data. In addition, with the "Regulation on Personal Health Data" that entered into force by being published in the Official Gazette on June 21, 2019, the Ministry of Health has determined the general framework covering the processing, access, concealment, correction, destruction, transfer, and security of health data. According to Article 16 of the regulation, it is stated that scientific studies can be conducted with personal health data, provided that it is anonymized by the data controller. The transfer of personal health data is regulated by Article 15 of the regulation. Accordingly, health data can only be transferred without seeking explicit consent from the data subject by persons or authorized institutions and organizations under the obligation of confidentiality for the purpose of protecting public health, [1] preventive medicine, conducting medical diagnosis, treatment and [2] care services, and planning and managing health services and their financing. In other words, regulations governing the confidentiality of health data state that explicit patient consent is not required for the processing of completely anonymized data and that misuse of data can theoretically be prevented in this way. Article 12 of the KVKK defines obligations regarding data security. Accordingly, the data controller is obliged to take all necessary technical and administrative measures to ensure the security of personal data and prevent unlawful access and processing, and to fulfill this obligation jointly with data processors. Therefore, researchers may need to take security and privacy measures regarding the storage, access, sharing, and destruction of data in scientific research projects and record them in a data management plan.

To overcome concerns about data privacy and regulatory restrictions, federated learning has emerged, where data is not transferred outside the hospital, but instead, the AI model is sent to hospitals and trained there. The fact that debugging is difficult due to the developers being inherently blind to the data may cause the model performance to fall below expectations.

4.2.4. Image Labeling and Annotation

Supervised machine learning involves creating algorithms to match medical images or clinical variables with "label" data. The development of an effective model depends on the number and quality of labeled data. After being trained on labeled data, the algorithm tries to predict the label of a new image according to the patterns it has learned. The reference standard represents the information that the model needs to learn. Reference standards may vary depending on the model's purpose. For example, bounding boxes for localization tasks, pixel-based masks created by experts for segmentation tasks, images with category labels for classification tasks, measurement markings for tracking tasks, or clinical outcomes (survival, metastasis, progression, etc.) for prognosis tasks.

Depending on the model's purpose, labels can originate from radiology reports, expert reviews, clinical or pathological data. The necessity of labeling large data sets for training high-capacity neural networks and the labor-intensive nature of the labeling task necessitate alternative solutions. These include AI-supported automatic labeling tools (e.g., automatic segmentation tools or automatic label extraction from radiology reports), crowd-sourced labeling, and weakly supervised learning (training with incomplete labels or low-quality labels). While AI-supported automatic labeling methods enable rapid labeling of large data sets, automatic labeling solutions suitable for every scenario may not exist, and the human factor is required to control label

quality. One of the automatic labeling methods is information extraction from radiology reports with natural language processing or recurrent neural networks. Information extracted from unstructured radiology reports may contain human-induced errors (spelling errors, interpretation errors, etc.) or AI-induced errors (individual differences in language usage style, etc.). It is estimated that 2-20% of radiology reports contain demonstrable errors. Structured reporting can be a solution to obtain higher quality data from reports and make data more easily accessible. Distributing the labeling task to a larger number of people can reduce the individual labeling burden, but in this case, harmony and consistency problems may arise between labelers. Today, non-experts can be used to examine and label large amounts of data to support automated services. Recent studies have reported that non-expert observers can be used in labeling some medical images. Although these methods simplify the labeling process, the generally low quality of the data obtained is a significant limitation. However, it is known that models trained using large data sets can show good results even with low-quality data. Model training in these types of situations where data is incomplete or unreliable is called weakly supervised learning. Two basic weakly supervised learning approaches that can be used to deal with incompletely labeled data are active learning and semi-supervised learning. Active learning aims to train the model with minimal human labeling effort by iteratively selecting the most informative samples from the data pool by experts for the labeling task.

Determining the reference standard is critical for the reliability of a study and the accuracy of its results. In the labeling process, priority should be given to re-evaluation and labeling by expert radiologists as much as possible, rather than weak labels extracted from reports. If there are multiple labelers, the reliability of the labels should be measured by methods such as inter-observer agreement (e.g., Cohen's kappa statistic or intra-class correlation-ICC). Establishing a standard labeling guide before labeling (e.g., which findings will be considered "positive" or "negative"), training observers on this, and performing trial labeling on a small test group will increase consistency. Labeling quality and agreement can be monitored by cross-checking randomly selected cases at regular intervals. In cases where the manual labeling method is used, the details of the labeling process should be clearly reported. This includes the labeler's experience, the strategy in handling difficult cases, and the measures taken to minimize bias. In addition, how inter-observer variability is handled in the manual labeling process should also be reported.

4.3. Modeling and Validity Test

A model is a program or algorithm trained to recognize specific patterns. These algorithms produce outputs such as predictions, detections, classifications, segmentations, or recommendations from input data. Modeling approaches used in image analysis are generally divided into two groups: traditional machine learning (requires predefined features) and deep learning (automatically learns features from data). Determining the most appropriate method to be used in the modeling phase depends on the data type (e.g., text, image, categorical, or numerical), problem type (e.g., classification, regression, or survival analysis), available computing resources (CPU or GPU), problem complexity, and the preferred balance between model accuracy and interpretability. In traditional machine learning methods, predefined and formulated features (e.g., radiomics texture features) are extracted from images, and models are created using these features with various machine learning algorithms. This approach usually offers more useful and interpretable results in small data sets. However, it may be limited in performance for complex problems such as image analysis and text recognition. Examples of frequently used algorithms in traditional machine learning include support vector machine (SVM), random forest (RF), k-nearest neighbor (kNN), and Naive Bayes. There is no single

and universal algorithm that will give the best result for every problem. Therefore, it is recommended to train and compare different algorithms separately and determine the most appropriate algorithm for each problem.

Deep learning is a subfield of machine learning based on multi-layered neural networks. This method automatically learns complex features from raw data during the training process, thus eliminating the need to manually design features. However, deep learning also has disadvantages: the need for a high amount of labeled data, the long and costly training process, and low interpretability due to its "black box" structure are the main ones. Today, many deep learning architectures specialized for solving different problems have been developed. The most important of these architectures are convolutional neural networks (CNN), recurrent neural networks (RNN), transformers, and generative models. For example, CNN is the most frequently used deep learning architecture, especially in the field of image analysis. As a result, traditional machine learning methods can be preferred in limited and well-structured data sets or when fast and interpretable results are needed. In contrast, deep learning methods are more suitable when dealing with large and complex data sets, especially in problems such as image analysis or text recognition.

For AI models produced in the field of medical imaging, there is confusion between the terms "validation" used in machine learning and medical literature. From a machine learning perspective, validation refers to selecting and adjusting the best model. However, from a medical perspective, validation usually refers to the process of verifying the performance of a model on unseen data, similar to the "test set" in machine learning. This difference can cause confusion, so the terms "development set" or "validation test set" are sometimes used instead of "validation set" in the medical context.

Data is ideally divided into three sets: training, validation (development), and test. The optimal separation ratio varies for each problem, and there is no single solution. As a basic rule, it is common practice to split the data as 80% training, 10% validation, and 10% test. However, in some sources, it is seen that the validation (development) set is not used. In this case, the training:test ratio is usually chosen between 60:40 and 90:10. The validation set is used to determine the optimal hyperparameters and select the best model to obtain the best result. The test set is used only once to measure the generalizability of the model. In smaller data sets, the split validation scheme can prevent the production of strong models due to the inability to provide sufficient diversity in the training set and the misleading performance measurement in the test set. In such cases, cross-validation can be used if it is not possible to obtain more data. In this method, after the data set is divided into k equal layers, the algorithm is trained on almost all layers for each training and tested on the excluded layer. The final performance is recorded as the average of the k performances measured.

Independent validation on an external data set should be preferred over internal validation to accurately assess the potential and generalizability of the model. In the medical literature, external validation is often considered the final test to definitively assess the safety, reliability, and generalizability of a model. However, it is seen that the number of studies with external validation in the literature is around 6%. Ideally, the data used in external validation should reflect the populations for which the final use of the model is intended. However, these data are often selected based on suitability and accessibility. As a rule, external validation must be carried out by independent researchers from different institutions. Although prospective validation is quite rare, it is preferred by the literature because it can give a better idea of the actual usability of models.

Technical performance validity testing is the measurement of the model's performance in training and test sets using various quantitative metrics. Performance validation metrics may vary depending on the model's purpose. For example, discrimination area under the receiver operating curve (ROC), sensitivity, specificity, positive and negative predictive values, calibration, and decision curve analysis for classification problems; intersection over union (IoU) and/or mean average precision (mAP) for detection problems; IoU and/or Dice score for segmentation problems can be used. All performance metrics should be reported separately for both training and test sets. This is important as it is informative in evaluating the overfitting of the model. In addition, it is important to report the scores of more than one performance metric suitable for the model's purpose with confidence intervals and to perform error analyses of misclassified cases. Comparing the model's performance with the best current radiological practices or alternative AI algorithms is a necessary method in the performance evaluation process.

Clinical validity studies may have the potential to accelerate the integration of models, but standards have not yet been established for such studies. However, it is extremely important for radiologists to play an active role in the creation and implementation of these standards. In clinical validation, in addition to the accuracy of the model, the increased rate of pathology detection with the use of AI, the acceleration in reporting time, the change in patient survival, or cost-effectiveness analyses can be performed. With the integration of the model, potential changes in patient care quality or hospital operations can be analyzed. Possible error scenarios can be determined by examining the model's incorrect predictions (false positives and false negatives) on a case basis.

One of the biggest obstacles to creating an effective model that can be adapted to the real world and generalized is overfitting. Another reason for this is training with a data set that contains sampling errors and has low diversity. When the model over-adapts to the features in the training set, it becomes overly sensitive to noise or random changes in the data and cannot be generalized to new data sets. This can lead to a decrease in model performance. The adjustment set is very important in understanding overfitting. If the model performs well in the training set compared to the validation set, it has probably overfitted the training data. Steps such as reducing the complexity of the model or training with more data should be taken to reduce overfitting.

Unlike training data sets, keeping validation data sets centrally and having model validation done by unbiased third parties can increase confidence in the performance results of models and facilitate clinical integration. Competition (challenge) events organized to guide the development of artificial intelligence models, make the testing phase reliable and effective by providing a central and comprehensive validation set, and increase the confidence of radiologists and regulatory institutions by selecting the best of the models with the same purpose can be useful.

It is important that the reference standards of validation data sets have higher standards than the data sets used for training. Radiologists should not only ensure that validation data sets are as close to the target population as possible, but also play a critical role in defining the criteria used for algorithm validation.

4.3.1. Performance Evaluation

When evaluating the success of a model's results, not only task success but also transparency (explainability), clinical benefit, safety, and stability should be considered. In addition, the model's errors should be analyzed, and the model's limitations and potential development opportunities should be explored.

4.3.1.1. Prediction Performance Evaluation

In simple binary classification problems, the success of diagnostic tools or procedures can be evaluated based on metrics such as sensitivity, specificity, positive and negative predictive values. Since multiple sensitivity and specificity pairs are generated with changing thresholds, there is a need for a single performance metric that can be used for comparison. Receiver Operating Curve (ROC) analysis is an important metric that can effectively evaluate the model's discrimination performance in this scenario. The most common summary measure of the ROC curve is the C statistic or AUC, known as the area under the curve. AUC reflects the probability that a patient with the disease will receive a higher risk score from the model than a healthy patient. AUC ranges from 0 to 1 and determines the discriminative performance of the diagnostic test. However, since an AUC value of 0.5 is equivalent to random prediction, it is practically accepted as the lower limit. Since AUC alone may not give accurate results in the case of class imbalance, it is important to report other performance metrics such as accuracy, sensitivity, specificity, F1 score, and Matthews correlation coefficient. Especially in classification tasks, it is important to evaluate confusion matrices, which can provide more detailed information about the competencies of models. Performance metrics should be carefully selected according to the characteristics of the data, clinical scenario, and the model's purpose. Clear and measurable success criteria and performance thresholds should be determined in advance with clinicians. In multi-category classifications (e.g., multiple pathological diagnoses), performance statistics should be reported separately for each class. The average of these statistics can be taken or weighted according to the prevalence of each class in the test set. In addition, error matrices must be given to evaluate whether certain classes are frequently confused.

In regression tasks, an attempt is made to predict a continuous and numerical dependent variable. Metrics such as R2, mean square error, root mean square error, root mean square logarithmic error, and mean absolute error are used for regression models.

There are a wide variety of metrics for evaluating object segmentation success. However, overlap-based metrics are commonly used for this purpose. Object detection and segmentation require metrics that define how well an area marked by the model's prediction matches the reference area, usually determined by a radiologist. Intersection is the overlap between the reference area and the area predicted by the model, while union is the total area covered by the predicted area and the reference area. Intersection over Union (IoU) and Dice score are two commonly used metrics to measure model performance in segmentation tasks. However, both measurement methods also have some limitations and pitfalls. For example, disadvantages such as being too sensitive to segmentation errors of small structures, obtaining lower than necessary scores in the presence of erroneous reference segmentations, not being sensitive to the shape of predicted segmentations, and not being equally affected by too many or too few segmentations are among the most well-known. In the object detection task, it can be examined whether the detection is correct by determining a threshold value for IoU, as in classification problems.

4.3.1.2. Calibration Performance Evaluation

Although ROC curves are widely used, they have some limitations. Basically, AUC is a ranking metric and shows how well the model separates patients. However, it does not provide information about the accuracy of probability estimates. Therefore, calibration is also very important in prediction model evaluation. Calibration is a method that evaluates the similarity between predicted probability values and actual probabilities.

In the calibration curve, the x-axis is the predicted probabilities, while the y-axis is the realized probabilities. Predicted probabilities are obtained from the numerical outputs given by the model. However, realized probabilities are not observable (e.g., samples are either disease positive or negative). Therefore, in order to determine how compatible model predictions are with reality, patients with similar predicted probability calculations are grouped (typically according to the decimal fractions of the predicted values) and the x-axis is placed from the low probability group to the high probability group. A graph is drawn from the points obtained by using the average predicted value in each group as the x-coordinate and the actual probability (i.e., the ratio of diseased samples to total samples in the group) as the y-coordinate. Perfect calibration is parallel to the 45-degree line. Statistical tests such as the Hosmer-Lemeshow goodness-of-fit test complement visual evaluation. Non-significant results (e.g., $p>.05$) suggest good calibration, but insufficient data size can lead to false good calibration results.

4.3.1.3. Clinical Benefit Analysis

The prediction-oriented progress of artificial intelligence has led to the neglect of clinical impact and outcomes. However, a model with high prediction success may not always be clinically beneficial. Because clinical benefit is related to calibration performance as well as the model's discrimination performance. For example, even if we divide the risk prediction scores of a test by ten, the classification performance of the test does not change because the optimal threshold value will change accordingly. However, if the patient's probability of having cancer drops from 30% to 3%, the patient's or doctor's decision to biopsy is affected, and the risk of missing an aggressive cancer arises. At this point, decision curve analysis (DCA) is valuable as a method that evaluates the clinical benefit of tests by considering both the discrimination and calibration performance of diagnostic tools. Its use is also recommended by the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guide.

The decision curve is a tool that graphically shows the net benefits of various clinical strategies according to different risk thresholds. The horizontal axis contains predetermined possible risk thresholds, while the vertical axis contains the net benefit. The area under the decision curve (DCA) is a measure of the clinical benefit of the prediction model. Each DCA has at least two reference lines, one horizontal and one diagonal, showing two possible option approaches. The dotted horizontal line shows the net benefit of a strategy where no patient is treated. The net benefit of not treating any patient is always zero. The diagonal dashed line shows the net benefit of a clinical strategy where all patients are treated. For risk thresholds below disease prevalence, "treat all patients" has a higher net benefit than "treat no patients." For risk thresholds above prevalence, the opposite is true and the "treat all patients" approach will create a negative net benefit. These two lines represent the two most extreme strategies possible. In reality, any other clinical strategy will involve treating certain patients and not treating others. For a model to be considered clinically beneficial at a certain threshold, it must have a higher net benefit than "treat all" or "treat none."

4.3.2. Explainability and Interpretability

Explainable AI is a fundamental part of responsible AI applications. The terms interpretability and explainability are two concepts that are confused in machine learning resources. Interpretability refers to our ability to directly understand the internal workings and outputs of the model (without the need for an additional analysis technique). Explainability, on the other hand, emphasizes the need to use additional tools or methods such as LIME, SHAP, and Grad-CAM, especially to show why complex or "black box" models reach a certain decision. Therefore, while interpretability is an inherent feature of the modeling method, explainability usually includes additional solutions developed after the model is trained. Explainable AI serves various purposes such as supporting human decision-making, increasing transparency between AI systems and humans, enabling debugging when unexpected behaviors occur, facilitating auditing to meet legal requirements, bringing trust in AI to appropriate levels and facilitating clinical integration, and verifying generalization ability.

There are many explainable AI techniques recommended in the literature. Three different elements are based on to classify explainable AI techniques: (i) local and global, (ii) model-specific and model-independent, (iii) pre-model (natural) and post-model.

Natural explainability refers to machine learning models (decision trees, linear regression, etc.) that are considered interpretable due to their simple structures. Post-model explainability refers to making the model explainable with various methods applied later (permutation feature importance, integrated gradients, etc.). Model-independent methods are flexible techniques that can work with different AI models, whereas model-specific methods, as the name suggests, can only work with certain models. Global explainability methods analyze the entire data set to understand the general patterns that support the model's predictions. This method provides explanations of which patterns in the data are important for the model's predictions. On the other hand, local explainability methods focus on explaining a single prediction made by the model. In other words, it aims to provide insight into why the model gave a single output in that way.

The explainability of models can be improved with three different methods: feature, text, and example-based. Feature-based methods can mark which input has the greatest impact on the model result on the original image. In the text-based method, semantic descriptions are used to explain the model's decision (for example, an algorithm with an object detection and classification task in mammography can mention the spiculated edge feature or high density that affects the malignancy probability of the detected lesion). In addition, influence function methods are used to analyze which training samples are affected by a certain prediction result of an AI model.

Visualization methods can help uncover disease-related confounding factors (e.g., pneumothorax diagnosis with the aid of a chest tube or increased prevalence of pneumonia in portable chest films) that can affect the system's performance, as well as increase users' trust in the system. Attention maps are frequently used visualization methods and create heat maps by marking the pixels that are important for correctly classifying images. However, these maps have some risks and limitations. These methods are primarily one of the local explainers used only to explain individual predictions rather than analyzing the overall behavior of the model. In addition, it can sometimes be difficult to understand what an emphasized area in an image actually means.

In summary, simple modeling methodologies (linear regression, decision trees, etc.) are naturally interpretable and do not require an additional method. However, it is essential to use additional tools to understand the decision processes of complex models such as deep learning.

5. Principles to be Followed in Clinical Applications

5.1. Application Selection

When selecting artificial intelligence applications for radiology practice, "patient benefit" and requirement should be the determining factors. For this purpose, a needs analysis should be conducted to determine the problems encountered in the clinical environment and the "necessity and most appropriate AI solutions" for these problems. The selected applications should significantly affect patient management in the clinic and facilitate workflow and increase productivity. In addition, features of the application such as "remote access", "license purchase", "pay-as-you-go model" should be discussed by institution managers and practicing radiologists, and cost/effectiveness should be considered.

The validity and effectiveness of the selected AI application should be supported by qualified scientific research. These studies should provide reasonable and solid evidence in terms of clinical accuracy, sensitivity, and reliability. General and local standards should be met in terms of safety, privacy, and effectiveness when using applications. The selected application should be approved by CE, FDA, or similar local health authorities. These approvals guarantee that the application meets the necessary legal and ethical standards for clinical use. The selected AI applications should be user-friendly to accelerate integration in clinical practice. The application should integrate seamlessly with existing systems and work in a way that does not disrupt and facilitate the clinical workflow. Therefore, integration with information systems and PACS should be preferred if possible. This will facilitate its safe and effective use. There should be user-friendly interfaces that will enable healthcare personnel to adopt the application quickly and effectively.

5.2. Application Usage

All measures should be taken for the effective and safe use of AI applications in the clinic. For this purpose, the potential weaknesses and safety criteria of AI applications should be known and adequately shared with the relevant parties.

Users should be provided with sufficient training in basic information systems and informatics topics, as well as AI applications and the selected AI application specifically. These training programs should be effective and able to use all functions of the application, and in this way, the application should be used correctly and effectively.

Ensuring the privacy and security of patient data should be guaranteed. In terms of data security, patient information should be protected from unauthorized access. Compliance with ethical and legal obligations should be guaranteed with standard protocols. Consistency and security should be ensured in clinical applications.

Mechanisms should be established to continuously monitor and evaluate the performance of artificial intelligence applications, and in this way, possible errors and areas for improvement should be identified. This continuous evaluation aims to increase the reliability and effectiveness of the application.

Mechanisms should be established where users can provide feedback. In this way, it is aimed to continuously improve the application, and user experiences and problems encountered should be reported and used for this purpose.

5.3. Informing Patients About Applications

Informing patients about the use of artificial intelligence applications should be done primarily by considering patient rights and ethical values. For this purpose, patients should be informed that applications will be used, and explicit and informed consent should be obtained regarding their use.

Informed consent should clearly include how the application works, what purpose it is used for, and what data is collected, enabling patients to make informed decisions in a transparent manner. In this way, patients should have a full understanding of the application.

Patients should be ensured to have appropriate knowledge about the advantages-disadvantages and weak-strong points of artificial intelligence uses. The effects of using artificial intelligence should not be exaggerated and turned into a commercial promotion.

The possible risks and potential benefits of the application should be explained to patients. With these explanations, all measures should be taken for patients to easily learn about the advantages and disadvantages.

Patients should be informed about which of their data will be used and how, and their data privacy rights. It should be stated how the data is protected and for what purposes it is used. Conditions should be provided where patients' questions about the application will be answered satisfactorily.

In written or verbal information of patients about the application, patients' cultural levels should be considered, and solutions should be provided to reach each patient (or their responsible relatives).

5.4. Reliability Levels of Applications

Preferably, multiple validity tests should have been performed on how the applications to be used perform in the clinical environment with real data sets after they are developed, which increases their reliability in the clinical environment. In addition to published scientific publications on this subject, studies to be conducted by practitioners in their own clinical environments are important. By subjecting the applications to regular quality control tests, it should be ensured that their performance remains at a consistently high level.

In addition to the general performance and accuracy criteria of the applications, it is important to monitor the response times. Response times should be short enough to meet specific-clinical expectations. It is important that the evaluation of the applications is carried out by independent institutions. Applications should be continuously updated and improved at regular intervals to be compatible with the latest technological innovations.

5.5. Continuous Improvement and Quality Audit

As in all systems, continuous monitoring and improvement of AI algorithms is necessary. For this purpose, it is recommended to create working groups at the institutional level and to monitor AI tools and measure their performance with the parameters to be determined. It is expected that the identified problems will help improve with corrective actions.

5.6. Reimbursement

If AI tools that will be added to workflows and increase image and diagnostic quality are used, they should be entitled to a refund. This should be carried out fairly with regulations to be made by the authorized institutions.

5.7. Responsibilities

Who will be responsible for the problems that may arise from the use of artificial intelligence tools should be defined by legal regulations. Since there is no national regulation yet, applications must be carried out in accordance with the applicable laws. In the specific case of radiology, the radiology specialist is the sole responsible for the evaluation and report, and it will be appropriate to state in the report if AI tools have been used.

6. References

1. *Britannica Dictionary. Address: https://www.britannica.com*
2. *Hsu HH, Ko KH, Chou YC, et al. Performance and reading time of lung nodule identification on multidetector CT with or without an artificial intelligence-powered computer-aided detection system. Clin Radiol. 2021;76(8):626.*
3. *Larsen M, Hoff SR, Auensen S. AI risk score on screening mammograms preceding breast cancer diagnosis. 2023;(20):19–24.*
4. *Dreizin D, Staziaki PV, Khatri GD, et al. Artificial intelligence CAD tools in trauma imaging: a scoping review from the American Society of Emergency Radiology (ASER) AI/ML Expert Panel. Emerg Radiol [Internet]. 2023;30(3):251–65.*
5. *Moawad AW, Fuentes DT, Elbanan MG, et al. Artificial intelligence in diagnostic radiology: where do we stand, challenges, and opportunities. J Comput Assist Tomogr. 2022;46(1):78–90.*
6. *Galbusera F, Cina A. Image annotation and curation in radiology: an overview for machine learning practitioners. Eur Radiol Exp. 2024 Feb 6;8(1):11.*
7. *Turkish Language Institution Dictionary. Address: https://sozluk.gov.tr/*
8. *Eklund A, Dufort P, Forsberg D, LaConte SM. Medical image processing on the GPU - Past, present and future. Med Image Anal [Internet]. 2013;17(8):1073–94.*
9. *Walston SL, Seki H, Takita H, Mitsuyama Y, Sato S, Hagiwara A, Ito R, Hanaoka S, Miki Y, Ueda D. Data set terminology of deep learning in medicine: a historical review and recommendation. Jpn J Radiol. 2024 Jun 10.*
10. *Varghese BA, Cen SY, Hwang DH, Duddalwar VA. Texture analysis of imaging: what radiologists need to know. AJR Am J Roentgenol. 2019 Mar;212(3):520-528.*
11. *Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42(December 2012):60–88.*
12. *Chae A, Yao MS, Sagreiya H, et al. Strategies for implementing machine learning algorithms in the clinical practice of radiology. Radiology. 2024;310(1).*
13. *Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. Korean J Radiol. 2021;22(3):442–53.*
14. *Moezzi SAR, Ghaedi A, Rahmanian M, Mousavi SZ, Sami A. Application of deep learning in generating structured radiology reports: a transformer-based technique. J Digit Imaging [Internet]. 2023;36(1):80–90.*
15. *Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol. 2024 Mar 6;30(2):80-90.*
16. *Özkan M, Kar G. Multi-classification of scientific texts written in Turkish language using deep learning technique. Mühendislik Bilim ve Tasarım Derg. 2022;10(2):504–19.*
17. *Karataş E, Karaca U. Protection of intellectual products generated by artificial intelligence according to the Law No. 5846 on Intellectual and Artistic Works. Maltepe Üniversitesi Hukuk Fakültesi Derg. 2022;(1):17–50.*
18. *Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology. 2024;310(1).*
19. *Baheti AD, Thakur MH, Jankharia B. Informed consent in diagnostic radiology practice: Where do we stand?. Indian J Radiol Imaging 2017;27:517-20.*
20. *Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol [Internet]. 2024;42(1):3–15.*

21. *Park HJ. Patient perspectives on informed consent for medical AI: A web-based experiment. Digit Heal. 2024;10.*
22. *Rowell C, Sebro R. Who will get paid for artificial intelligence in medicine? Radiol Artif Intell. 2022 Aug 3;4(5):e220054.*
23. *Artificial Intelligence European Commission Act. Address: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#ai-act-different-rules-for-different-risk-levels-0*
24. *AI European Commission Decision. Address: https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en*
25. *America AI Laws Draft Guideline. Address: https://www.whitehouse.gov/ostp/ai-bill-of-rights*
26. *Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: Practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. J Med Imaging Radiat Oncol. 2024 Feb;68(1):7-26.*
27. *Cambridge Dictionary Address: https://dictionary.cambridge.org/dictionary/english/regulator*
28. *Najjar R. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. Diagnostics (Basel). 2023 Aug 25;13(17):2760.*
29. *Graziani M, Dutkiewicz L, Calvaresi D, Pereira J. A global taxonomy of interpretable AI : unifying the terminology for the technical and social sciences [Internet]. C. 56, Artificial Intelligence Review. Springer Netherlands; 2023. 3473–3504 s.*
30. *Mehta N, Pandit A. International Journal of Medical Informatics Concurrence of big data analytics and healthcare : A systematic review. Int J Med Inform [Internet]. 2018;114(January):57–65.*
31. *Chen R, Ch I, Hatabu H, Valtchinov VI, Siegelman J. Detection of unwarranted CT radiation exposure from patient and imaging protocol meta-data using regularized regression. Eur J Radiol Open [Internet]. 2019;6(April):206–11.*
32. *Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. Insights Imaging. 2020 Feb 4;11(1):10.*
33. *Datta S, Godfrey-Stovall J, Roberts K. RadLex Normalization in Radiology Reports. AMIA . Annu Symp proceedings AMIA Symp. 2020;2020:338–47.*
34. *Şenol U, Aktaş A, Saka O. Radiology Information System. Akad Bilişim'07 - IX Akad Bilişim Konf Bildir 31 Ocak - 2 Şubat 2007 Dumlupınar Üniversitesi, Kütahya. 2007;431–3.*
35. *Digital Hospital Address: https://dijitalhastane.saglik.gov.tr/TR-4877/dicom-digital-imaging-and-communications-in-medicine---tipta-dijital-goruntuleme-ve-iletisim.html*
36. *Subaşıoğlu, F. (2024). "Ethics in electronic health records", Electronic Health Records (11-25), Nobel Akademik Yayıncılık, Ankara.*
37. *Tıp Bilişimi Derneği HL7 Summary Information Document. Address: https://turkmia.net/file/HL7-ozet-bilgi.pdf*
38. *Tang A, Tam R, Cadrin-ch A, et al. Canadian Association of Radiologists white paper on artificial intelligence in radiology. 2018;69:120–35.*
39. *Sutherland J, Belec J, Sheikh A, ve ark. Applying modern virtual and augmented reality technologies to medical ımages and models. J Digit Imaging. 2019 Feb;32(1):38-53.*
40. *Gupta S, Johnson EM, Peacock JG, ve ark. Radiology, Mobile Devices, and Internet of Things (IoT). J Digit Imaging. 2020;33(3):735–46.*
41. *Türkiye Bilişim Derneği Bilişim Terimleri Sözlüğü. Adres:http:///bilisimde.ozenliturkce.org.tr/onerilen-tum-terimler-ingilizce-turkce*
42. *Regulatory considerations on artificial intelligence for health. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.*
43. *Rowell C, Sebro R. Who will get paid for artificial ıntelligence in medicine? Radiol Artif Intell. 2022 Aug 3;4(5):e220054.*
44. *YZ Avrupa Komisyonu Yasası. Adres: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#ai-act-different-rules-for-different-risk-levels-0*
45. *YZ Avrupa Komisyonu Kararı. Adres: https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en*
46. *Amerika YZ Yasaları Yönerge Taslağı. Adres: https://www.whitehouse.gov/ostp/ai-bill-of-rights*
47. *Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: Practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. J Med Imaging Radiat Oncol. 2024 Feb;68(1):7-26.*
48. *Ueda D, Kakinuma T, Fujita S, ve ark. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol [Internet]. 2024;42(1):3–15.*
49. *Tang A, Tam R, Cadrin-ch A, ve ark. Canadian Association of Radiologists white paper on artificial ıntelligence in radiology. 2018;69:120–35.*
50. *Yükseköğretim kurumları bilimsel araştırma ve yayın faaliyetlerinde üretken yapay zekâ kullanımına dair etik rehber. Adres: https://www.yok.gov.tr/Documents/2024/yapay-zeka-kullanimina-dair-etik-rehber.pdf*
51. *Koçak B. Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: focus on radiomics. Diagn Interv Radiol. 2022;28(5):450-462.*
52. *Define-AI Use Case Directory | American College of Radiology. Accessed March 28, 2024. https://www.acrdsi.org/DSI-Services/Define-AI*
53. *Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563-577.*
54. *Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. Radiology. 2020;295(1):4-15.*

55. *Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. Proceedings of the IEEE International Conference on Computer Vision. 2017;2017-October:843-852.*
56. *Ranschaert ER, Morozov S, Algra PR. Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks.;2019.*
57. *Shur JD, Doran SJ, Kumar S, et al. Radiomics in Oncology: A Practical Guide. RadioGraphics. 2021;41(6):1717-1732.*
58. *Papanikolaou N, Matos C, Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging. 2020;20(1):33.*
59. *Alwosheel A, van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. Journal of Choice Modelling. 2018;28:167-182.*
60. *Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? Published online November 19, 2015.*
61. *Jain AK, Chandrasekaran B. Dimensionality and sample size considerations in pattern recognition practice. In: ; 1982:835-855.*
62. *Baum EB, Haussler D. What Size Net Gives Valid Generalization? Neural Comput. 1989;1(1):151-160.*
63. *Haykin S. Neural Networks and Learning Machines. Vol 3.; 2008.*
64. *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.*
65. *ProstateNET | Prostate Imaging Archive. Accessed March 28, 2024. https://prostatenet.eu/*
66. *Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. J Cogn Neurosci. 2007;19(9):1498-1507.*
67. *Candemir S, Nguyen X V., Folio LR, Prevedello LM. Training Strategies for Radiology Deep Learning Models in Data-limited Scenarios. Radiol Artif Intell. 2021;3(6).*
68. *Cheng PM, Montagnon E, Yamashita R, et al. Deep learning: An update for radiologists. Radiographics. 2021;41(5):1427-1445.*
69. *Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611-629.*
70. *Galbusera F, Cina A. Image annotation and curation in radiology: an overview for machine learning practitioners. Eur Radiol Exp. 2024;8(1).*
71. *McMahan B, Ramage D. Federated Learning: Collaborative Machine Learning without Centralized Training Data. Google AI Blog. Published April 6, 2017. Accessed March 28, 2024. https://blog.research.google/2017/04/federated-learning-collaborative.html*
72. *Matheny M, Israni ST, Whicher D, Ahmed M. Artificial intelligence in health care: The hope, the hype, the promise, the peril. In: Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril (2019). ; 2023.*
73. *de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med. 2022;5(1).*
74. *Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. 2022;5(1).*
75. *Diaz O, Kushibar K, Osuala R, et al. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. Physica Medica. 2021;83:25-37.*
76. *Castiglioni I, Rundo L, Codari M, et al. AI applications to medical images: From machine learning to deep learning. Phys Med. 2021;83:9-24.*
77. *Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging. 2004;22(1):81-91.*
78. *Montagnon E, Cerny M, Cadrin-Chênevert A, et al. Deep learning workflow in radiology: a primer. Insights Imaging. 2020;11(1):22.*
79. *Aryanto KYE, Oudkerk M, van Ooijen PMA. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol. 2015;25(12):3685-3695.*
80. *Library D. DICOM Library - Anonymize, Share, View DICOM files ONLINE. Accessed April 28, 2024. https://www.dicomlibrary.com/*
81. *RSNA. CTP-The RSNA Clinical Trial Processor. Accessed April 28, 2024. https://mircwiki.rsna.org/index.php?title=MIRC_CTP*
82. *Kişisel Verilerin Korunması Kanunu. 2016;57(29677). Accessed April 27, 2024. https://www.resmigazete.gov.tr/eskiler/2016/04/20160407-8.pdf*
83. *Levy A, Agrawal M, Satyanarayan A, Sontag D. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In: Conference on Human Factors in Computing Systems - Proceedings. ; 2021.*
84. *Lakhani P, Kim W, Langlotz CP. Automated extraction of critical test values and communications from unstructured radiology reports: An analysis of 9.3 million reports from 1990 to 2011. Radiology. 2012;265(3).*
85. *Chen MC, Ball RL, Yang L, et al. Deep learning to classify radiology free-text reports. Radiology. 2018;286(3).*
86. *Brady A, Laoide RÓ, McCarthy P, McDermott R. Discrepancy and error in radiology: Concepts, causes and consequences. Ulster Medical Journal. 2012;81(1).*
87. *Heim E, Roß T, Seitel A, et al. Large-scale medical image annotation with crowd-powered algorithms. Journal of Medical Imaging. 2018;5(03).*
88. *Mehta P, Sandfort V, Gheysens D, Braeckevelt GJ, Berte J, Summers RM. Segmenting the kidney on CT scans via crowdsourcing. In: Proceedings - International Symposium on Biomedical Imaging. Vol 2019-April. ; 2019.*

89. *Wahid KA, Fuentes D. Weak Supervision, Strong Results: Achieving High Performance in Intracranial Hemorrhage Detection with Fewer Annotation Labels. Radiol Artif Intell. 2024;6(1).*
90. *Zhou ZH. A brief introduction to weakly supervised learning. Natl Sci Rev. 2018;5(1):44-53. doi:10.1093/nsr/nwx106*
91. *Boehringer AS, Sanaat A, Arabi H, Zaidi H. An active learning approach to train a deep learning algorithm for tumor segmentation from brain MR images. Insights Imaging. 2023;14(1).*
92. *McCague C, Ramlee S, Reinius M, et al. Introduction to radiomics for a clinical audience. Clin Radiol. 2023;78(2):83-98.*
93. *Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. Korean J Radiol. 2019;20(3).*
94. *Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1).*
95. *Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58(5).*
96. *Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40(19).*
97. *Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9).*
98. *Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. Insights Imaging. 2024;15(1):16.*
99. *Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J Radiol. 2019;20(7):1124.*
100. *Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights Imaging. 2023;14(1). doi:10.1186/s13244-023-01415-8*
101. *Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell. 2020;2(2).*
102. *Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology. 2018;286(3):800-809.*
103. *Pencina MJ, D'Agostino RB. Evaluating discrimination of risk prediction models: The C statistic. JAMA - Journal of the American Medical Association. 2015;314(10).*
104. *Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed.; 2019. http://www.springer.com/series/2848*
105. *Kocak B, Kus EA, Kilickesmez O. How to read and review papers on machine learning and artificial intelligence in radiology: a survival guide to key methodological concepts. Eur Radiol. 2021;31(4):1819-1830.*
106. *Han K, Song K, Choi BW. How to develop, validate, and compare clinical prediction models involving radiological parameters: Study design and statistical methods. Korean J Radiol. 2016;17(3):339-350.*
107. *Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging. 2015;15(1).*
108. *Reinke A, Tizabi MD, Sudre CH, et al. Common Limitations of Image Processing Metrics: A Picture Story. Published online April 12, 2021.*
109. *Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: ACM International Conference Proceeding Series. Vol 148. ; 2006.*
110. *Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12(1).*
111. *Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res Notes. 2022;15(1):210.*
112. *Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3(1):18.*
113. *Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Ann Intern Med. 2015;162(1).*
114. *Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. Eur Urol. 2018;74(6).*
115. *Piovani D, Sokou R, Tsantes AG, Vitello AS, Bonovas S. Optimizing Clinical Decision Making with Decision Curve Analysis: Insights for Clinical Investigators. Healthcare (Switzerland). 2023;11(16).*
116. *Kurdziolek M. Explaining the hard to explain: An overview of Explainable AI (XAI) for UX. Published online 2022. Accessed April 13, 2024. https://www.youtube.com/watch?v=JzK_SBhakUQ*
117. *Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiol Artif Intell. 2020;2(3).*
118. *de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med (Lausanne). 2023;10.*
119. *Neri E, Aghakhanyan G, Zerunian M, et al. Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology. Radiologia Medica. 2023;128(6).*
120. *Groen AM, Kraan R, Amirkhan SF, Daams JG, Maas M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? Eur J Radiol. 2022;157.*

7. Contributors

This guide was prepared between March 2024 and March 2025 under the leadership of Dr. Oğuz Dicle, with contributions from the members of the TSR Imaging Informatics Working Group: Dr. Mustafa Özmen, Dr. Utku Şenol, Dr. Fırat Atak, Dr. Nur Hürsoy, and Dr. Sinem Gezer.