Title

Lessons Learned in Building Expertly Annotated Multi-Institution Datasets and Hosting the RSNA AI Challenges.

Journal

Radiology: Artificial Intelligence, 6(3)

Authors

Kitamura, Felipe
Prevedello, Luciano
Colak, Errol
et al.

Publication Date

2024-05-01

DOI

10.1148/ryai.230227

Peer reviewed

Radiology: Artificial Intelligence

SPECIAL REPORT

Lessons Learned in Building Expertly Annotated Multi-Institution Datasets and Hosting the RSNA AI Challenges

Felipe C. Kitamura, MD, PhD • Luciano M. Prevedello, MD, MPH • Errol Colak, MD, FRCPC • Safwan S. Halabi, MD • Matthew P. Lungren, MD • Robyn L. Ball, PhD • Jayashree Kalpathy-Cramer, PhD • Charles E. Kahn, Jr, MD, MS • Tyler Richards, MD • Jason F. Talbott, MD, PhD • George Shih, MD • Hui Ming Lin, HBSc • Katherine P. Andriole, PhD • Maryam Vazirabad, MSc • Bradley J. Erickson, MD, PhD • Adam E. Flanders, MD • John Mongan, MD, PhD**

From the Department of Applied Innovation and AI, Dasa, São Paulo, Brazil (F.C.K.); Department of Diagnostic Imaging, Universidade Federal de São Paulo (Unifesp), Av Prof Ascendino Reis, 1245, 131, São Paulo, SP, Brazil 04027-000 (F.C.K.); Department of Radiology, The Ohio State University Wexner Medical Center, Columbus, Ohio (L.M.P.); Department of Medical Imaging, University of Toronto, Toronto, Canada (E.C.); Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, Ill (S.S.H.); Microsoft HLS, Redmond, Wash (M.P.L.); Department of Biomedical Data Science, Stanford University, Stanford, Calif (M.P.L.); The Jackson Laboratory, Bar Harbor, Maine (R.L.B.); Department of Ophthalmology, University of Colorado Denver School of Medicine, Aurora, Colo (J.K.C.); Department of Radiology, University of Pennsylvania, Philadelphia, Pa (C.E.K.); Department of Radiology, University of Utah, Salt Lake City, Utah (T.R.); Department of Radiology and Biomedical Imaging (M.P.L., J.F.T., J.M.) and Center for Intelligent Imaging (J.M.), University of California San Francisco, San Francisco, Calif; Department of Radiology, Weill Cornell Medical College, New York, NY (G.S.); Department of Medical Imaging, Unity Health Toronto, Toronto, Canada (H.M.L.); Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, MGB Data Science Office, Boston, Mass (K.P.A.); Informatics Department, Radiological Society of North America, Oak Brook, Ill (M.V.); Department of Radiology, Mayo Clinic, Rochester, Minn (B.J.E.); and Department of Radiology, Thomas Jefferson University, Philadelphia, Pa (A.E.F.). Received June 26, 2023; revision requested September 29; revision received February 17, 2024; accepted February 27. **Address correspondence to** F.C.K. (email: *kitamura.felipe@gmail.com*).

* A.E.F. and J.M. are co–senior authors.

Authors declared no funding for this work.

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

Radiology: Artificial Intelligence 2024; 6(3):e230227 • https://doi.org/10.1148/ryai.230227 • Content code: AI

The Radiological Society of North America (RSNA) has held artificial intelligence competitions to tackle real-world medical imaging problems at least annually since 2017. This article examines the challenges and processes involved in organizing these competitions, with a specific emphasis on the creation and curation of high-quality datasets. The collection of diverse and representative medical imaging data involves dealing with issues of patient privacy and data security. Furthermore, ensuring quality and consistency in data, which includes expert labeling and accounting for various patient and imaging characteristics, necessitates substantial planning and resources. Overcoming these obstacles requires meticulous project management and adherence to strict timelines. The article also highlights the potential of crowdsourced annotation to progress medical imaging research. Through the RSNA competitions, an effective global engagement has been realized, resulting in innovative solutions to complex medical imaging problems, thus potentially transforming health care by enhancing diagnostic accuracy and patient outcomes.

The Radiological Society of North America (RSNA) has organized at least one artificial intelligence (AI) competition per year for the past 6 years, engaging a worldwide audience to raise awareness about real-world medical imaging problems that can be addressed with AI (1–4). The goals of these competitions include the following: *(a)* to advance the state of the art through innovative solutions to the contest task; *(b)* to provide a head-to-head comparison of the performance of different approaches on the same data; *(c)* to develop and publicize high-quality and publicly available imaging datasets; *(d)* to present the technologies, standards, and challenges of radiology to the AI and data science community; and *(e)* to provide an opportunity for physicians and data scientists to develop AI skills applicable to medical imaging and exposure to cutting-edge AI techniques. These competitions offer a unique opportunity for trainees to learn from both their own experiences and other competitors' approaches and for raising awareness about radiology and AI among medical students. In each competition, contestants are given a clinically oriented task, such as identifying and locating fractures or hemorrhage, as well as an expert-curated dataset on which to train their models. The participants whose model scores the highest on a predetermined performance metric (5) (eg, accuracy, F1 score, mean absolute error [MAE]) are declared the competition winners. Organizing such competitions depends on the availability of high-quality datasets (6).

The purpose of this article is to provide an in-depth analysis of the challenges and processes encountered when organizing medical imaging AI competitions, with a specific focus on the creation and curation of high-quality datasets. By detailing the critical aspects of use-case definition, dataset construction, data extraction, and de-identification, we aim to offer valuable insights and guidance to researchers, competition organizers, and institutions engaged in advancing the state of the art in medical imaging through AI. We also emphasize the importance of ethical considerations, patient privacy, and data security in the context of medical imaging competitions.

Overview of Challenges

Creating and curating high-quality datasets for medical imaging AI challenges require substantial coordination and collaboration. Obtaining medical imaging data can be difficult because of patient privacy and data security concerns and the resources required to safely extract and

Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, DICOM = Digital Imaging and Communications in Medicine, HIPAA = Health Insurance Portability and Accountability Act, MAE = mean absolute error, PHI = protected health information, RSNA = Radiological Society of North America

Summary

Organizing artificial intelligence competitions for medical imaging, such as those held by the Radiological Society of North America, requires intricate processes for creating and curating high-quality datasets. These competitions have successfully fostered global collaboration, advanced medical imaging research, and have the potential to transform health care.

Keywords

Use of AI in Education, Artificial Intelligence

de-identify imaging data from clinical archives (7). Collecting a sufficiently diverse and representative data cohort often requires participation from multiple health care facilities involving different geographic locations, diverse patient populations, and disparate imaging modalities.

Another challenge in creating high-quality datasets for medical imaging AI is the shortage of expert annotators to label the data with "ground truth"—the output that we expect the model to produce when provided with the data as input. Medical imaging data are complex, and accurate labeling requires specialized domain knowledge. This is compounded by the fact that producing high-quality annotations can be both time-consuming and costly, as well as necessitating the effort to organize the logistics of annotation. That is further complicated when nonimaging data are a critical element of the dataset, such as the clinical outcome, laboratory, pathology, or genomic markers. The reason for the added effort in dealing with these other modalities is twofold: gathering data from other silos (laboratory information system, electronic health record) and the usual lack of standardization in generation of the data (staining of pathologic specimens, genomic methods).

Creating a high-quality dataset requires substantial attention to the quality and consistency of data in the dataset. Medical imaging data can vary in ways that are significant to AI algorithms, depending on factors such as the modality, manufacturer, imaging protocol, patient age, sex, ethnicity, and comorbidities. Substantial planning is required to construct a database that minimizes bias and is appropriately diverse in representation. Because of heterogeneity in how imaging data are collected and organized across and even within different medical centers, creating a high-quality dataset requires careful curation, normalization, and cleaning of the data to ensure consistency and comparability of each element of the dataset. On the other hand, it is key that datasets still represent the real-world clinical setting to increase the chances of generalizability to daily clinical use.

Planning and hosting an AI competition are very different from planning and executing a multi-institutional collaborative research project. For instance, research projects are most interesting when they tackle a task that has not yet been adequately addressed with AI and success is uncertain; competitions are generally unsatisfying when the outcome is that no one can develop a successful algorithm and generally focus on tasks that are known to be achievable, like predicting bone age from hand and wrist radiographs, identifying intracranial hemorrhage on head CT scans, and detecting cancer on mammograms. Research models should be evaluated using datasets that match clinical truth as closely as possible so that the performance metrics are predictive of what would be observed in clinical practice. In contrast, because the main goal of the competition is education, the private test sets used in competitions should be optimized to be similar to the provided training and validation data so that the selection of competition winners is as fair as possible and minimizes the element of randomness. The disadvantage of that approach is that the best models are not selected on their generalizability to data from other institutions. Moreover, an AI competition is generally guided by a strict and unwavering timeline; the lure of increasing the complexity of the competition (predicting more diseases, annotating a larger dataset, or with a more detailed annotation scheme) must be tempered by time limitations.

We cannot overemphasize the importance of meticulous project management and the adherence to carefully constructed timelines in the orchestration of AI competitions so they can be launched on time. As we reflected on our experiences in organizing competitions, we found that the deceptive vastness of time seemed to envelop us in the early stages, creating an illusion of endless preparation time. This complacency, however, often gave way to a startling reality check as we found ourselves considerably behind schedule, rushing to make up for lost time. Let this serve as a cautionary tale, underscoring the importance of diligent planning, including a detailed timeline with clear milestones and continuous progress checks. The orchestration of these competitions is a marathon, not a sprint, and success hinges heavily on consistent pacing and unwavering attention to the timeline. A punctual launch happens not by accident but by the culmination of deliberate, daily efforts stringently aligned with a clear timeline.

Figure 1 demonstrates the overall processes to organize medical imaging AI competitions.

AI Challenge Task

An initial step in designing an AI competition is to specify the task or tasks. One of the key differentiators between an AI research project and an AI competition is that in a research project, one often chooses problems where it is initially unclear whether it is possible to successfully address the problem with AI, like predicting survival from imaging studies or predicting who will develop a disease based on a normal examination. In contrast, competitions are generally built around tasks where there is high confidence that existing AI methods can address the defined use case or problem, like localizing consolidations on chest radiographs and fractures on cervical spine CT scans. In addition, as a medical specialty society, the RSNA tries to select competition tasks that are directly clinically relevant. That means considering the clinical needs, relevance to patient outcomes, and trade-off between task specificity and complexity. For health care professionals creating machine learning

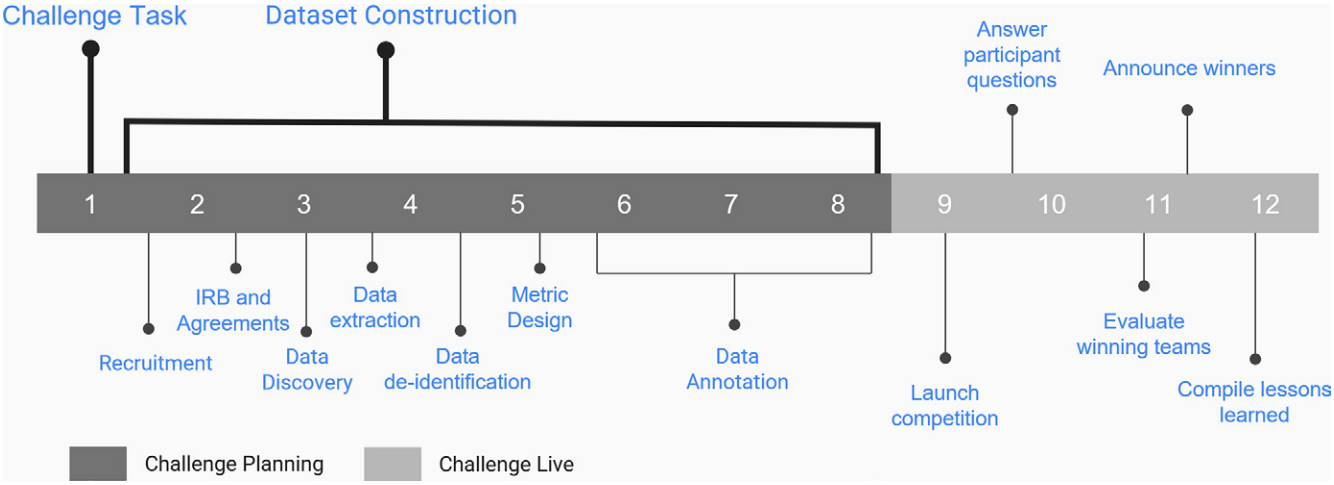

Figure 1: Overall processes to organize medical imaging artificial intelligence competitions, from months 1 to 12. IRB = institutional review board.

competitions, selecting the right tasks for a medical imaging dataset is key. Focus on identifying areas where machine learning can notably improve such aspects as diagnostic accuracy or disease progression. Ensure these tasks directly influence patient outcomes, such as through early disease detection. Balance the task's clinical specificity with its complexity, aiming for tasks that address precise medical issues while also pushing the boundaries of what is currently achievable in medical diagnostics and treatment planning. This approach ensures that tasks are both clinically relevant and technically challenging.

It is necessary to consider whether the specific imaging data required to achieve the task exist and are obtainable in sufficient quantity, including appropriate controls. This is done by asking contributing sites to perform discovery on their local data. Another key aspect of a use-case definition is whether expert annotation is required to establish ground truth for the proposed dataset, which is decided by the organizing committee based on the difficulty of the task (eg, it is safe to rely on trainees to identify body part, but neuroradiologists are needed to diagnose multiple sclerosis). This involves determining whether it is feasible to annotate the data and estimating how many examinations in each class are needed to achieve reasonable performance in addressing the task. Differentiating the performance of entrants requires a minimum number of positive cases in the test set, which helps estimate the total size of the test set based on the prevalence of the disease. To determine the feasibility of annotating the data, the committee members propose annotation schemes and annotate a small sample themselves to measure the time per study and also the cognitive effort to avoid inducing annotator burnout and quitting.

Dataset Construction

Recruitment

The first step in creating a dataset for medical imaging AI is engaging with institutions that may become prospective data donors to gauge their interest in contributing data. This often involves recruiting clinical champions from each contributing site who can provide introductions to relevant staff, such as IT personnel, compliance officers, and privacy officers, and can help facilitate submission of appropriate documentation, supervise data discovery, data de-identification, and data transfer.

Institutional Review Board and Agreements

Once initial interest has been established, the next step is to negotiate a data usage agreement with local compliance and privacy officers, as well as obtain approval from the local institutional review board. This process establishes the terms under which data will be shared and ensures that all necessary legal and ethical issues are considered. This will often include specified guarantees regarding de-identification and protection of confidentiality, as in the real world; each country has different data protection laws, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, the General Data Protection Regulation in Europe, and the General Personal Data Protection Law in Brazil. It is important to allow for sufficient lead time for local approvals because this process may take several months. These processes can often be expedited when participating institutions share their data usage agreement, data sharing agreement, and institutional review board documents with one another.

Data Discovery

Once the legal and ethical considerations have been addressed, the next step is for each site to perform discovery on the local data stores or clinical archives at each site and inventory examinations that meet the requirements provided by the challenge organizers. The time involved can vary depending on the search tools available, the complexity of the search, the local prevalence of the classes, and the consistency of imaging protocols (eg, scanning parameters, such as sequences and section thickness). Locating appropriate controls can be almost as time-intensive as identifying the target class; it is important that negative cases differ from positive cases only in the presence of the finding of interest but otherwise be drawn from the same distribution of patient demographic characteristics and reason for examination. It has been helpful for the organizers to provide "pseudo code" search queries

to achieve more consistent search results across sites. Control groups may be difficult to acquire for such tasks as mammography for breast cancer detection (definition of negative includes negative follow-up or negative biopsy results), classification of brain tumor radiogenomics (genetic methods to define positive *MGMT* methylation vary across institutions), and detection of cervical spine fracture and abdominal trauma (injuries can happen in multiple anatomic locations per patient, creating association biases).

Several iterations of data discovery and inventory may be necessary before the site can proceed with data extraction and de-identification. If a site has limited experience in this area, it may be helpful to identify the tools and techniques that are available for finding, accessing, and preparing the necessary data and training personnel to perform the steps. If there is known nonuniformity in how the imaging data are acquired at a site, it may be necessary to obtain a summary of imaging protocols at each site to ensure that the dataset is representative of the types of examinations that are commonly performed and to confirm with the challenge organizers that the data will be appropriate to be included in the final dataset. Identifying appropriate examinations that fit the specific use case often involves the use of natural language processing techniques to search radiology reports or electronic medical record databases to identify relevant examinations. For example, donating institutions used natural language processing to identify positive cases for the following competitions: intracranial hemorrhage detection, pulmonary embolism detection, brain tumor radiogenomic classification, cervical spine fracture detection, screening mammography breast cancer detection, and abdominal trauma detection. Alternatively, if the required number of cases is not overwhelmingly high, some sites might consider manual report review.

There is no objective method to determine the dataset size for tasks that have never been tackled. If there is literature describing AI models for the task, we can use that information to estimate the achievable accuracy for a given dataset size. If no literature about the task exists, one option is to quickly assemble a dataset with available data and train a model. That not only allows for dataset size estimation but also guarantees that the task is solvable by a machine learning model. Having defined the size of the dataset, one way to secure sufficient data is to grow vertically (ask for more studies from each donating institution) and/or horizontally (enroll more donating institutions).

It is also necessary to establish a local quality assurance strategy to assess the quality of the data at the study level. This could involve checking images and reports for expected findings to ensure that the dataset is complete and accurate. Once a cohort of candidate examinations has been extracted, it is useful to randomly audit a subsample of examinations to ensure the desired results. We have learned from past challenges that donating institutions are heterogeneous in their ability to follow the instructions to deliver data in the defined format. We frequently receive examinations with nonpertinent body parts (eg, chest CT in the cervical spine fracture dataset), with wrong contrast phase (eg, angiography series while the requirement was nonenhanced series only for the cervical spine fracture dataset), with wrong section thickness, or containing diseases that should have been excluded (eg, tumors and postoperative changes in the degenerative lumbar spine MRI dataset).

Data Extraction

Data extraction heavily depends on the IT infrastructure and technical expertise of the contributing sites. Some sites have off-the-shelf software to automatically extract batches of studies, whereas others must extract data manually using less sophisticated tools. Not all institutions have all studies available in online clinical archives, so some studies may have to be retrieved from backups. Data stored in long-term storage are often compressed using unique formats across multiple storage media, which adds complexity to the process of mapping, extracting, and decompressing the studies.

Another concern with extracting large batches of studies from the production picture archiving and communication system is the risk of instability in clinical viewing systems that could jeopardize clinical workflows. It is advised to schedule large data extractions during times of low clinical workload or use a research picture archiving and communication system.

Handling large numbers of files involves unique difficulties. Uploading hundreds of thousands of files to cloud storage using a web browser frequently causes some files to be dropped in the transfer process. Tools such as *rsync* (typically used at the command line) help to avoid this issue; even if any loss of connection happens, *rsync* will seamlessly restart the process such that no files are lost (8). Compressing multiple files into a single larger file increases the transfer rate over the internet, not only because the total size is reduced but because transferring a single file is faster than multiple files, even if the total size is the same.

Coding skills are valuable to automate extraction, compression, series selection, and de-identification of Digital Imaging and Communications in Medicine (DICOM) studies. Often, it is necessary to customize data curation with methods and tools that are not readily available off the shelf. For example, it is possible to code a script to automatically select the required series by filtering DICOM studies based on the metadata using complex rules to select such features as specific planes and section thicknesses. Figure 2 provides an example of such a script. Even with these approaches, some series may be incomplete or cover the wrong anatomic region. Manual quality assurance is essential to ensure a high-quality dataset.

Finally, it is important to allow enough time for the technical steps involved in creating the dataset and to provide guidance on-the-go as necessary. This could involve providing training and support to local staff to ensure that the data are extracted and prepared correctly.

Data De-identification

Protected health information (PHI) refers to any data created for the purpose of providing care service to a patient that can be used to identify an individual. If not properly managed, the unauthorized release of such information could lead to serious privacy violations and potential patient harm. Furthermore, various laws and regulations, such as the HIPAA in the United States, mandate stringent protection of PHI, imposing heavy fines and penalties for breaches.

```python
#Import package to read metadata from DICOM files
from dcmtag2table import dcmtag2table

#Define the tags we want to read from DICOM files
list_of_tags = [
                "PatientID",
                "StudyInstanceUID",
                "SeriesInstanceUID",
                "SOPInstanceUID",
                "SliceThickness",
                "Modality"
                ]

#Define the folder where DICOM files are stored
folder = "/media/felipe/easystore/Datasets/RSNA2019/mdai/epm/"

#Read tags from DICOM files into a Pandas Dataframe
df = dcmtag2table(folder, list_of_tags)

#Select only CTs
df = df[df["Modality"] == "CT"]

#Select only images with Slice Thickness of 1.0 mm or less
df = df[df["SliceThickness"] <= 1.0]
```

Figure 2: Python script demonstrates how to read metadata from Digital Imaging and Communications in Medicine (DICOM) files and then use it to select a subset of images. In the example shown in this figure, we selected only CT images with a section thickness of 1.0 mm or less, but any combination of filters can be used to select the desired images. However, some DICOM metadata can be inconsistent, producing undesired results.

The concept that drives data de-identification is simple: remove protected PHI from shared data. However, in practical terms, the de-identification of DICOM images is a complex task. Most publicly available de-identification tools are not HIPAA-compliant, and automated DICOM de-identification is not infallible; thus, manual checks are necessary (9).

The RSNA Artificial Intelligence Committee has created a process considered less prone to PHI leakage. It includes three steps: allow-listing, dumping of unique strings, and manual check. Allow-listing consists of recreating a DICOM file that includes only allowed tags. It helps guarantee that no unintended content is contained in the resultant file. Dumping unique values for each element in the metadata reduces the workload of the manual checks while keeping all values that appear at least once (10). This ensures that any PHI will be shown in the manual check. The RSNA has also developed a popular open access tool called Anonymizer, which will simultaneously extract DICOM data from an archive and de-identify the data (11). This solution is both convenient and efficient for extracting large datasets. A step-by-step guide on how to use Anonymizer can be found elsewhere (12).

After the data are de-identified, some hosting platforms, such as Kaggle, do a second round of de-identification to make sure the remaining tags are consistent and not a source of target leakage. Metadata in a dataset can lead to data leakage if they contain information that indirectly reveals sensitive data or gives away the answer to the prediction task. For example, a timestamp may reveal the order of events or a unique identifier might be correlated with a target variable, both leading to unrealistically high predictive performance during model training but poor performance in real-world predictions.

It should be anticipated that difficulties pertaining to DICOM tags may arise, particularly upon interacting with an unfamiliar modality. Modalities such as mammography and tomosynthesis may present more issues because of their substantial use of private tags and metadata. Although DICOM documentation is comprehensive, the management of private DICOM tags using Python scripts often involves numerous undisclosed methods. Some of these techniques are inadequately documented in standard resources, such as StackOverflow, or lack documentation entirely.

When curating the dataset, even technically adept contributors will make mistakes. Consider adding a second layer of data quality or protocol check after receiving data from sites. Visual inspection (by human reviewers) for secondary capture objects, dose reporting sheets with PHI, and burned-in pixel data are necessary. Coding can be used to implement optical character recognition on pixel data to remove embedded PHI that varies in location inside the image (13). For PHI that is fixed for a given manufacturer, RSNA CTP DICOM Pixel Anonymizer is an appropriate tool (14). However, this process is usually not flawless, necessitating manual reviews for accuracy.

Metric Design

The metric choice should consider the purpose of the competition. Binary classification competitions can work well with metrics such as log-loss, accuracy, the area under the receiver operating characteristic curve (AUC), precision, recall, F1, and others. Class imbalance can overestimate accuracy and AUC, and F1 is often a better choice in that scenario (15). In segmentation tasks, metrics such as Dice similarity coefficient and intersection over union (IoU or Jaccard similarity) are more appropriate than accuracy and AUC because there is often a large class imbalance.

Some metrics are more clinically oriented (more familiar to health care professionals), such as sensitivity, specificity, positive and negative predictive values, and accuracy. Other metrics, such as AUC and log-loss, are more difficult to generalize to expected clinical performance but can reflect finer variations in performance, allowing for differentiation of submissions. Regardless of the choice of metric to rank participants in the leaderboard of the competition, the model output from each participant can be used to calculate all clinically oriented metrics in the postcompetition analysis; however, the ordering of participants by different metrics will usually be different.

Another consideration for the choice of metrics is the difficulty of the task. Using AUC for easy tasks might create a leaderboard where the top participants achieve very similar results. For example, classifying sex from chest radiographs will result in most participants being in the 0.99 AUC range, making the results statistically similar and undermining the discrimination power. In that context, more rigorous metrics, such as F1, may help discriminate between the top performers.

Metrics that rely on binarized predictions (precision, recall, specificity) incur a higher risk of ties (participants achieving the same value for the metric), particularly when the test set has a low absolute number of cases. Increasing the test set size or using probabilistic metrics (AUC, log-loss, MAE) reduces the risk of

ties. MAE was used in the bone age competition; because the predicted age in months is a continuous variable, the chances of ties among participants is minimal.

In general, the more complex and novel the metric, the more likely there are to be unanticipated and unintended consequences of the metric. These consequences of the metric often create opportunities for contestants to "game" the metric—that is, to create entries that score well on the metric but do not accomplish the task in a clinically useful or meaningful way. Metrics that attempt to capture findings at multiple levels are particularly subject to this because they may be gamed by entrants submitting logically inconsistent results (eg, fracture at C3, no fracture overall for the study) to hedge difficult cases. On the other hand, metrics that consider more than one target variable can be fairer because they represent the model capabilities more holistically. Examples of these holistic metrics include the use of weighted log loss to account for fractures in multiple vertebral levels in the cervical spine fracture detection competition and to determine whether a pulmonary embolism is acute or chronic, its side, and the presence or absence of right ventricular overload in the pulmonary embolism detection competition. Simple, well-proven metrics generally result in the fewest unpleasant surprises.

Data Annotation

The use case should guide the definition of the labeling scheme, which is usually a trade-off between the richness of information and the annotation effort. The metric of the competition highly depends on the labeling scheme (eg, there is no way to use a segmentation Dice score if the region of interest on the image was not labeled with masks).

Receiving data prelabeled from the contributing site can provide key information regarding class distribution (eg, normal or abnormal) and can help expedite and organize the annotation process. Even with this information, the organizing team must perform a quality assurance process to validate the accuracy of the prelabel. This is often done by manually checking a sample of the data from each donating institution. Prelabeled data are useful because it is usually less time-consuming and less error-prone to annotate a batch one knows is (or should be) entirely composed of positive cases. The same holds true for batches with only negative cases.

The schema for annotation encompasses three primary elements. First, the categorization of "readers" incorporates aspects such as the quantity of readers, their particular subspecialty training, and accumulated years of experience. Second, the "type of annotation" includes options such as point, line, bounding box, polygon, region of interest or mask, classification, and regression. Last, the "level of annotation" can be dissected into image-level, series-level, study-level, and patient-level.

The orchestration of the annotation process necessitates multiple procedural stages. These include the recruitment of domain experts, preferably from subspecialty societies (subspecialty trained). We usually do not require a minimum number of years of experience. The process further necessitates the identification of an ample pool of willing experts to ensure redundancy. The delineation of a manageable task is crucial, such as a task requiring approximately 10 hours of work that does not overly strain cognitive resources. Motivating annotators is an art. This has been done by offering a contributor position (group authorship) in the author list of an article describing the dataset, provided a minimum number of cases is annotated. Communicating a clear and feasible scope of work and deadline also helps keep annotators motivated.

The selection of an appropriate annotation platform is essential. Some ideal characteristics of an annotation platform for labeling data to train medical imaging AI models would include user-friendly interfaces, high-precision tools for detailed annotations, support for multiple imaging formats, integration capabilities with existing health data systems, the ability to manage large datasets efficiently, built-in quality control mechanisms, and collaboration support for multiple annotators. Figure 3 depicts an example of a platform used to annotate datasets from previous competitions. The provision of multimedia instructions (including examples of positive and negative cases) for the annotation process is a key requirement, as is the initiation of a pilot annotation. This preliminary annotation aims to identify and select those who can accurately perform the required task. This is done by asking annotators to label a small batch of cases (10–20) after reading the instructions and having those labels reviewed by the organizing team. Annotators who do not follow the instructions have a chance to try again after feedback. It is a transparent process used to select who will annotate the dataset (Fig 4).

The quality assessment of annotation is conducted based on outputs from multiple annotators. Metrics are computed between each annotator and the "ground truth" defined by the organizers if a common pilot set is used for all annotators. In scenarios where different but overlapping pilot sets are assigned to annotators, the agreement between each annotator is calculated. A minimal competency metric is established that annotators must reach before gaining access to the actual data.

Those annotators who do not meet expectations are required to undergo retraining and retesting. Upon the successful launch of the dataset annotation, it is critical to ensure adherence to set deadlines. Concurrent monitoring of annotator progress and the quality of annotations is performed. Figure 4 shows a process to perform quality assurance during annotation. The overall annotation process takes around 2–3 months if there is no unanticipated delay.

Batches are reassigned between annotators as needed to compensate for incomplete assignments. Finally, the process culminates in the execution of sanity checks on the final annotation.

Frequent communication with annotators is recommended, including reiteration of annotation instructions, because misunderstandings are frequent, even after a round of practice cases.

Although some label noise in the training set (the publicly available data) is acceptable, the competition benefits from high-quality curated labels in the test set (the privately held data used for evaluating each submission) so that participants are ranked according to a labeling accuracy metric. A minimum of three annotators are assigned to the test set, and a consensus method for labeling and region of interest is established (16). For example, in the bone age challenge (1), test set labels were averaged over six different radiologist labelers, making the ground truth a continuous variable, whereas the clinical Greulich-Pyle bone

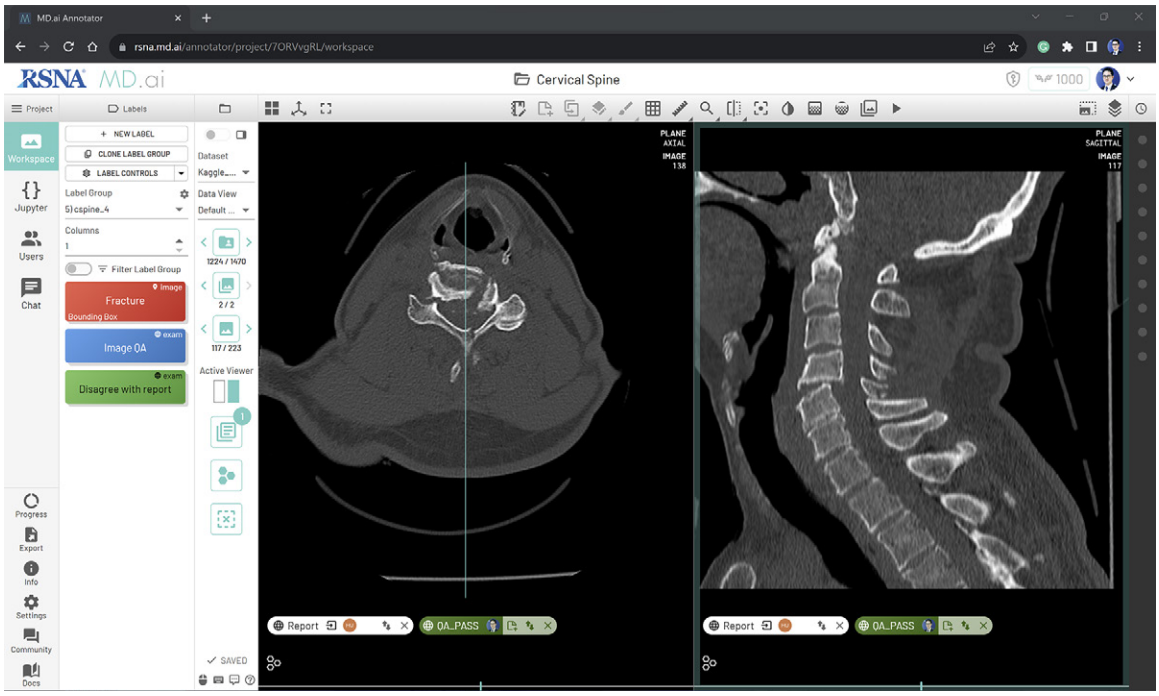

Figure 3: Annotation platform used to annotate datasets in prior competitions. QA = quality assurance.

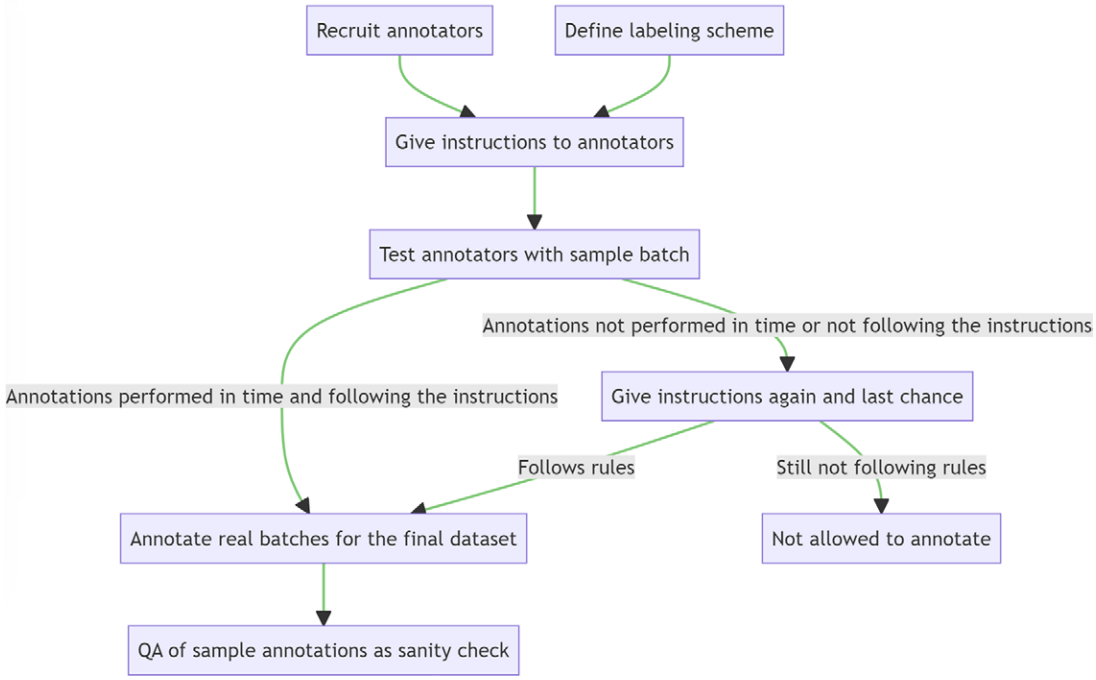

Figure 4: Proposed process to mitigate errors in dataset annotation. QA = quality assurance.

age assessed by a single reader was a discrete variable. The MAE between the winning model and the averaged ground truth was approximately 4 months, whereas the MAE between the winning model and each of the six radiologists was greater than 4 months, suggesting that the averaged ground truth was more accurate than any of the six individual readers.

However, combining multiple labels is not always straightforward. In the pneumonia detection challenge, the ground truth of the test set was defined as the intersection of the annotators' bounding boxes, which made the test set bounding boxes consistently smaller than those in the training set drawn by a single annotator. Participants had to identify and correct for this difference to obtain the highest scores on the metric.

In summary, using multiple readers generally increases the accuracy of annotations, but combining label data from multiple readers must be done with careful consideration to avoid

introducing systematic differences between single-reader and multiple-reader annotations.

General Considerations

Team Building

In exploring the factors that contribute to the success of AI competitions, one key element emerges as pivotal: the assembly of an effective and diverse team. This team should not only initiate and oversee the competition but also participate in its progression, from inception to completion. This includes monitoring the drafting of manuscripts and conducting post hoc analyses when appropriate. Drawing from experiences within the RSNA, we believe that such team composition and function is the "secret sauce" behind the success of AI competitions. Teams should encompass a diverse range of individuals, including veterans with experience hosting competitions and subject matter experts in the clinical subspecialty, AI, and statistics. Incorporating novices into the team can help develop talent for future projects and spread the workload. This blend of skills and experiences can provide a balanced approach to managing and conducting such ventures, fostering innovation, and ensuring successful outcomes.

Dataset Size

Large datasets in AI can greatly enhance the performance and accuracy of models because of the increased diversity and representativeness of the data. They can enable complex models to generalize better, reduce overfitting, and allow for more intricate patterns to be learned. However, there are also drawbacks. Larger datasets demand more computational resources and time for processing and training, which is a barrier for participants to join. Larger datasets may also increase the risk of privacy violations.

Data Preprocessing

AI models are heavily influenced by preprocessing techniques. That is why it is usually better to keep the images in their original format, avoiding transformations such as resizing, reformatting, and windowing. Transformations can be lossy or lossless. Some researchers convert DICOM to other formats (eg, NifTI, PNG, or JPEG). If the transformation is lossless, it has no negative effect on model training. Training on JPEG is not necessarily a bad practice because JPEG is frequently the image format used in the pixel data in DICOM files. If preprocessing is necessary, make sure to provide the preprocessing code so that the final models created in the competition can be applied to external datasets.

Data Splitting

Some competition platforms have their own policies on how to split the data into training and test sets, which can change over time. Decide that early on, especially if you plan to refine only the test set labels. The data split policy includes choosing the percentage of data in the test set and the data distribution (the test set could have a similar or different distribution than the training set).

Licensing

Publicly releasing a dataset requires assigning a license to it. Otherwise, the dataset is not truly open because users will not have permission to use it under copyright laws, even if they have access to it. Licenses define how the data can be used and shared, and they should be as simple as possible (17). An example on how to apply a license to your code can be found elsewhere (18).

Patient Privacy

When a dataset is released publicly, the major concern is the risk of exposing PHI. Creating many automated layers of deidentification followed by manual checks is advised as a safety rule. One should never assume that automated tools work perfectly. Always manually double-check the dumped unique strings from DICOM metadata for the entire dataset before releasing the data. Visual inspection of images is advised to avoid leakage of pixel-embedded PHI. Because this is a tedious task, it might not be possible to check the entire dataset, particularly in larger ones. In that case, a sample of the data should be checked with particular attention to secondary capture and dose information.

Clinical Use of Winning Models

One of the deliverables of an AI competition is a set of usually state-of-the-art models that can be used in the next step of model validation: clinical testing. Although clinical use requires regulatory approval, clinical research about the effect of AI in radiologic workflows can be done with AI models from competitions. Those studies tackle the human–machine interaction; most (if not all) use cases so far have human-in-the-loop (or human in charge, AI-in-the-loop). We should not take for granted that AI models will improve clinical outcomes (19). Underreliance and overreliance of humans in AI can undermine the benefit of AI.

Conclusion

AI competitions have engaged a global community to effectively address real-world medical problems. RSNA AI competitions have demonstrated the potential of crowdsourced ground-truth annotation to advance medical imaging research and promote awareness about the need for innovative solutions. The strategies and methods used to create datasets for these competitions have been described in detail, highlighting the importance of careful dataset design to avoid potential pitfalls. Through the RSNA Artificial Intelligence Committee, these competitions have successfully brought together diverse perspectives and expertise to tackle complex challenges in medical imaging to set a standard for future competitions and collaborations in the field. Readers who want to participate in future challenges should watch out for the regular communications RSNA sends, usually by email. If you want to volunteer as an annotator, the call for annotators is usually done via subspecialty societies such as American Society of Emergency Radiology, Society of Abdominal Radiology, American Society of Neuroradiology, American Society of Spine Radiology, or

any other specialty, depending on the task of the competition. AI researchers who want to help organize new AI competitions can volunteer for the AI committee (20). Ultimately, these efforts have the potential to transform health care by improving diagnostic accuracy and patient outcomes.

Acknowledgments: We would like to extend our heartfelt appreciation to the RSNA staff for their indispensable assistance and guidance throughout our research. In particular, we owe our profound gratitude to Christopher Carr, Michelle Riopel, and Jamie Dulkowski, whose dedication and expertise significantly contributed to the success of this report. Their continuous support, insightful suggestions, and unwavering belief in our work were pivotal during the research process. The professionalism and commitment exhibited by the entire RSNA team have greatly enhanced the quality of our work and for that, we are deeply grateful. The authors were the only source of information and ideas for this article. The authors used GPT4 (OpenAI; *http://openai.com*) (in March 2023) for content editing to effectively communicate their work in the first draft of the manuscript. This draft was then revised by all coauthors, who provided substantial edits to the manuscript. Figure 4 was created using GPT4 (OpenAI; *http://openai.com*) with Diagrams: Show Me plugin to create a diagram from the complete description of a process provided entirely by the author.

Author contributions: Guarantor of integrity of entire study, **F.C.K.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **F.C.K., L.M.P., E.C., S.S.H., G.S., A.E.F.**; experimental studies, **F.C.K., L.M.P., J.F.T., K.P.A.**; statistical analysis, **R.L.B., K.P.A., M.V.**; and manuscript editing, **F.C.K., L.M.P., E.C., S.S.H., J.K.C., C.E.K., T.R., J.F.T., G.S., H.M.L., K.P.A., B.J.E., A.E.F., J.M.**

Disclosures of conflicts of interest: **F.C.K.** Consultant for MD.ai and GE HealthCare; speaker for Sharing Progress in Cancer Care; Co-chair of the ML Education Subcommittee at SIIM (no payment); early career consultant to the Editor of the journal *Radiology* (no payment); medical director at Bunkerhill Health (after this paper was written); associate editor of *Radiology: Artificial Intelligence*. **L.M.P.** Associate editor of *Radiology: Artificial Intelligence*. **E.C.** No relevant relationships. **S.S.H.** No relevant relationships. **M.P.L.** Leadership role in RSNA (no payment, not related to this work); employee of Microsoft (not related to this work). **R.L.B.** Support from RSNA, consulting fees from RSNA, RSNA, AI Committee. **J.K.C.** Grants/contracts from NIH, EU, GE, and Genentech; software licensed to Boston AI; consulting fees from Siloam Vision; stock/stock options Siloam Vision; deputy editor of *Radiology: Artificial Intelligence*. **C.E.K.** Travel support from Sectra USA (travel expenses paid for service on advisory board); Editor of *Radiology: Artificial Intelligence* (salary support paid to employer). **T.R.** No relevant relationships. **J.F.T.** No relevant relationships. **G.S.** SIIM board member, MD.ai shareholder and board member. **H.M.L.** No relevant relationships. **K.P.A.** Senior consultant to the editor for *Radiology: Artificial Intelligence*. **M.V.** No relevant relationships. **B.J.E.** Research committee chair for SIIM; consultant to the editor for *Radiology: Artificial Intelligence*. **A.E.F.** RSNA Board of Directors. **J.M.** Grant form Siemens paid to institution; royalties from GE paid to institution; payment for expert testimony from Covington and Gibson-Dunn; support for travel from RSNA and Nuance (Nuance travel support for attending Nuance advisory meetings); RSNA AI Committee Chair; associate editor of *Radiology: Artificial Intelligence*.

References

1. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA pediatric bone age machine learning challenge. Radiology 2019;290(2):498–503.
2. Shih G, Wu CC, Halabi SS, et al. Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. Radiol Artif Intell 2019;1(1):e180041.
3. Flanders AE, Prevedello LM, Shih G, et al. Construction of a machine learning dataset through collaboration: the RSNA 2019 Brain CT Hemorrhage Challenge. Radiol Artif Intell 2020;2(3):e190211. [Published correction appears in Radiol Artif Intell 2020;2(4):e209002.]
4. Colak E, Kitamura FC, Hobbs SB, et al. The RSNA pulmonary embolism CT dataset. Radiol Artif Intell 2021;3(2):e200254.
5. Erickson BJ, Kitamura F. Magician's corner: 9. Performance metrics for machine learning models. Radiol Artif Intell 2021;3(3):e200126.
6. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. Radiology 2020;295(1):4–15.
7. Prevedello LM, Halabi SS, Shih G, et al. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. Radiol Artif Intell 2019;1(1):e180031.
8. rsync documentation. https://download.samba.org/pub/rsync/rsync.1. Updated May 22, 2023. Accessed November 19, 2023.
9. 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.
10. Kitamura FC, Colak E. DICOM tag to table Python helper code. GitHub Repository. https://github.com/kitamura-felipe/dcmtag2table. Published May 28, 2020. Updated November 19, 2023. Accessed November 19, 2023.
11. Perry J. Anonymizer. http://mirc.rsna.org/download. Published April 24, 2022. Updated April 24, 2022. Accessed November 19, 2023.
12. RSNA COVID-19 DICOM data anonymizer. https://rsna.org/-/media/Files/RSNA/Covid-19/RICORD/RSNA-Anonymizer-Program-Instructions.pdf?_gl=1*jg5a2u*_ga*MTA5NjY4NTc4OC4xNjc5NDQxNDIx*_ga_EQ32SZ84M3*MTY5ODkzMDQwOS4yNTAuMS4xNjk4OTMwNTA0LjYwLjAuMA. Published May 30, 2020. Updated May 30, 2020. Accessed November 19, 2023.
13. Khosravi B, Mickley JP, Rouzrokh P, et al. Anonymizing Radiographs Using an Object Detection Deep Learning Algorithm. Radiol Artif Intell 2023;5(6):e230085.
14. RSNA CTP DICOM pixel anonymizer. https://mircwiki.rsna.org/index.php?title=The_CTP_DICOM_Pixel_Anonymizer. Published December 30, 2018. Updated December 30, 2018. Accessed November 19, 2023.
15. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: pitfalls and recommendations for image analysis validation. arXiv 2206.01653 [preprint] https://arxiv.org/abs/2206.01653. Published June 3, 2022. Accessed April 9, 2024.
16. Lampert TA, Stumpf A, Gancarski P. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. IEEE Trans Image Process 2016;25(6):2557–2572.
17. How to apply a license to your open source software project. https://fossa.com/blog/apply-license-open-source-software-project/#:~:text=If%20your%20project%20did%20not%20have%20any%20license%20up%20until%20this%20point%2C%20nobody%20can%20legally%20use%20it%2C%20even%20if%20it%E2%80%99s%20public%20and%20visible%20to%20the%20entire%20world. Accessed February 17, 2024.
18. How to use GNU licenses for your own software. https://www.gnu.org/licenses/gpl-howto.html. Published January 21, 2023. Updated January 21, 2023. Accessed November 19, 2023.
19. Gaube S, Suresh H, Raue M, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. Sci Rep 2023;13(1):1383.
20. Volunteer to serve on an RSNA committee. https://www2.rsna.org/timssnet/About/volunteer.cfm?_ga=2.31532918.992809885.1602009162-1331479168.1602009162&_gl=1*vpk28p*_ga*MTA5NjY4NTc4OC4xNjc5NDQxNDIx*_ga_4699REKRC5*MTY5OTk2MjA0My4yOC4xLjE2OTk5NjIxMDMuNjAuMC4w*_ga_EQ32SZ84M3*MTY5OTk2MjA0My4yNjQuMS4xNjk5OTYyMTAzLjYwLjAuMA. Accessed November 19, 2023.