TITLE

Separating Hope from Hype: Artificial Intelligence Pitfalls and Challenges in Radiology

AUTHORS:

Name: Jared Dunnmon
Degree: PhD
Affiliation: Stanford University, Department of Biomedical Data Science
Email: jdunnmon@cs.stanford.edu
Mailing Address: 353 Jane Stanford Way, Stanford, CA, 94043
Corresponding Author: Yes

DISCLOSURE STATEMENT

The author declares no commercial conflicts of interest, financial conflicts of interest, or funding sources related to this work.

CORRESPONDING AUTHOR

Jared Dunnmon

SYNOPSIS

While recent scientific studies suggest that Artificial Intelligence (AI) could provide value in many radiology applications, much of the hard engineering work required to consistently realize this value in practice remains to be done. In this chapter, we summarize the various ways in which AI can benefit radiology practice, identify key challenges that must be overcome for those benefits to be delivered, and discuss promising avenues by which these challenges can be addressed.

KEYWORDS

Radiology; Artificial Intelligence; Pitfalls and Challenges

KEY POINTS

- AI systems can provide value to radiologists in a number of ways, ranging from reduced time on task to discovery of new knowledge
- Potential challenges in deploying AI systems for radiology include myriad technical issues, difficulties mitigating algorithmic bias, and poor alignment between measured performance and clinical value
- Promising directions to address these challenges include improved software engineering practices, close clinician involvement in model development, and robust post-deployment monitoring

While recent scientific studies suggest that Artificial Intelligence (AI) could provide value in many radiology applications, much of the hard engineering work required to consistently realize this value in practice remains to be done. In this chapter, we summarize the various ways in which AI can benefit radiology practice, identify key challenges that must be overcome for those benefits to be delivered, and discuss promising avenues by which these challenges can be addressed.

How Can AI Provide Value to Radiologists?

Though headlines often gravitate towards AI systems that claim to perform as well as or better than humans on a particular task, AI can provide value to radiologists in several specific ways. These include automated information extraction from imaging exams, increased diagnostic certainty, decreased time on task, faster availability of results, reduced cost of care, better clinical outcomes, discovery of new knowledge, and improved patient access to radiological expertise.[1,2] While other chapters in this volume describe such applications in detail, we provide a brief overview here.

Leveraging information contained within an image to make prognostic and diagnostic decisions is a core component of radiology practice; AI systems can provide value to radiologists by enabling them to do so more effectively. For instance, while a clinician's diagnostic ability is defined by a combination of first principles knowledge and experience with specific cases, an AI system can leverage information contained in millions or billions of data points to refine how image features are mapped to prognostic or diagnostic outputs. Recent analyses of AI models trained on large radiology datasets demonstrate the potential not only to improve diagnostic sensitivity or specificity,[3,4] but also to yield novel image features that correspond more directly

to the outcome of interest than those that comprise existing standards.[5] Furthermore, the fact that AI systems can perform such analysis with high levels of standardization across patients[5] – and without being vulnerable to fatigue or cognitive biases – can yield substantial value in the real world.[2,6] AI-based approaches can augment human analysis both by surfacing information that is not readily apparent and by improving the utility of reconstructed images for human readers.[7]

AI systems can also provide value to radiologists by increasing the speed with which imaging results are processed and by reducing required clinician effort. Automated optimization of worklists, for instance, can reduce time-to-treatment for life-threatening and severe conditions while still ensuring human review of all cases.[1,2,8,9] With appropriate algorithmic design and human factors engineering, the integration of AI-based triage and second read systems into clinical workflows holds the potential to decrease the time required per case. This would simultaneously increase patient access, lower costs, and improve outcomes by enabling radiologists to spend more of their time on cases that require substantial analysis.[1] Decreased time requirements would also help to alleviate the workforce shortage that radiology is expected to experience in the coming years as demand for services continues to increase.[1]

Finally, the consistent use of AI systems in radiology practice can yield new knowledge that improves patient care. The development of "radiomic" features that are not discernable by the human eye, but may nonetheless be predictive of outputs ranging from diagnosis to prognosis to treatment response, represents a particularly promising area of research.[10] AI can also play a supporting role in such tasks as patient selection, tumor tracking, and adverse event detection that can inform the clinical trials necessary to create new forms of diagnosis and treatment.[11]

AI systems that provide value in each of these ways have been conceptualized – and in some cases evaluated for clinical use – across a wide variety of applications, many of which have

been detailed in this volume.[11] The balance of this chapter will describe important pitfalls in development and deployment of these systems that must be addressed in order for AI systems to provide widespread value for radiologists.

What Challenges Must Be Overcome for AI to Provide Value to Radiologists?

Translating the potential that academic studies and early clinical trials have shown into concrete improvements in radiology practice will require that researchers and practitioners alike be aware of the challenges that can accompany the development and deployment of AI systems in radiology applications. This section provides an overview of the major pitfalls that AI systems face in radiology, and the subsequent section will outline compelling approaches for addressing these challenges.

Meaningful Performance Measurement

The first, and perhaps most important challenge in developing an AI system for radiology is ensuring that the task of interest is sufficiently well-posed that performance can be meaningfully measured. Defining a suitable clinically relevant task is not always as easy as it might seem. Consider the example of chest X-ray (CXR) classification, a commonly studied application of AI. Much work in this area has focused on developing deep learning models that classify CXRs into one of the 14 different classes used in Rajpurkar et al.,[3] but it is clear that several of these classes (e.g. atelectasis, consolidation, infiltration) can be inconsistently understood across different clinicians. As a result, models trained for this particular task may confuse these three classes, or may provide outputs with which certain clinicians would agree more than others. Such ambiguity in task definition reduces our ability to effectively measure

performance.[12] Furthermore, it is critical to ensure that the measure of AI system performance is directly related to the outcome of interest. It is not immediately clear, for instance, that high levels of performance on a 14-class CXR abnormality classification task will translate into one of the types of value described previously (e.g. reduced radiologist time, improved diagnostic certainty). In fact, one could argue that framing this task in a slightly different way – binary normal versus abnormal CXR triage for worklist prioritization[4,8] – could provide more direct clinical value because its place in the clinical workflow is clear, and metrics like turnaround time for high-priority cases can be immediately computed. Collaboration between radiology domain experts and AI developers will remain key to ensuring that AI systems are developed for tasks that are meaningful, and that performance is measured in ways that directly correlate with clinical value.

Even when a clinically useful task has been defined, inappropriately chosen performance metrics can hinder model development (see the chapter by Dr. Kalpathy-Kramer for more detail). While sensitivity and specificity may be familiar to many clinicians, multi-class classification, segmentation, and reconstruction tasks are evaluated quite differently than binary classification, and metrics suitable to the task must be used. Equity considerations are also important in designing suitable metrics. For instance, it is often the case that deep learning models for classification perform well on classes that make up the majority of the training set, but perform poorly on classes that are small. In some situations, this could be acceptable – in which case unweighted metrics are commonly used – but in others it would not, meaning that class-weighted metrics should be reported. Furthermore, the common use of Area Under the Receiver Operating Characteristic curve (AUROC) or Area Under the Precision Recall Curve (AUPRC) as figures of merit should be viewed with caution; while useful in describing overall classification

performance, these metrics can be misleading because they do not indicate how a model will perform at the specific operating points that must be chosen in practice.

In addition to computing an appropriate metric, evaluation procedures must be designed in a way that yields meaningful results. A common error is assuming that models that are internally validated – i.e. that perform well on the same population on which they were developed – will continue to perform well when applied externally (i.e. to a different population).[13] Evaluation datasets must represent the population upon which a model is intended to be used, otherwise performance computed thereon will be misleading. When comparing multiple algorithms, performance should also be evaluated on a common dataset in order to provide meaningful information.[14] Finally, a common pitfall in AI performance measurement occurs when the task schema is insufficiently granular to capture important variations in performance. A common example of this phenomenon – which has been termed "hidden stratification"[15] – occurs in classification problems when performance variation occurs due to a variable that the original dataset curators did not consider. As shown in Figure 1, for instance, Oakden-Rayner et al.[15] recently demonstrated that while a common CXR classification model yields an overall AUROC value of 0.87 for detecting pneumothoraces, that performance increases to 0.95 on images that display a chest drain and drops to 0.77 on images that do not. Thus, if this model had been deployed in practice, it would have performed much worse on the very population – pneumothoraces without a chest drain – that would be of clinical interest. Similar issues can cause models to be biased and perform poorly on a given subclass (e.g. non-

Caucasian patients) because it just so happens that (a) that subclass makes up a minority of a dataset and (b) the dataset was not labeled with subclass information.

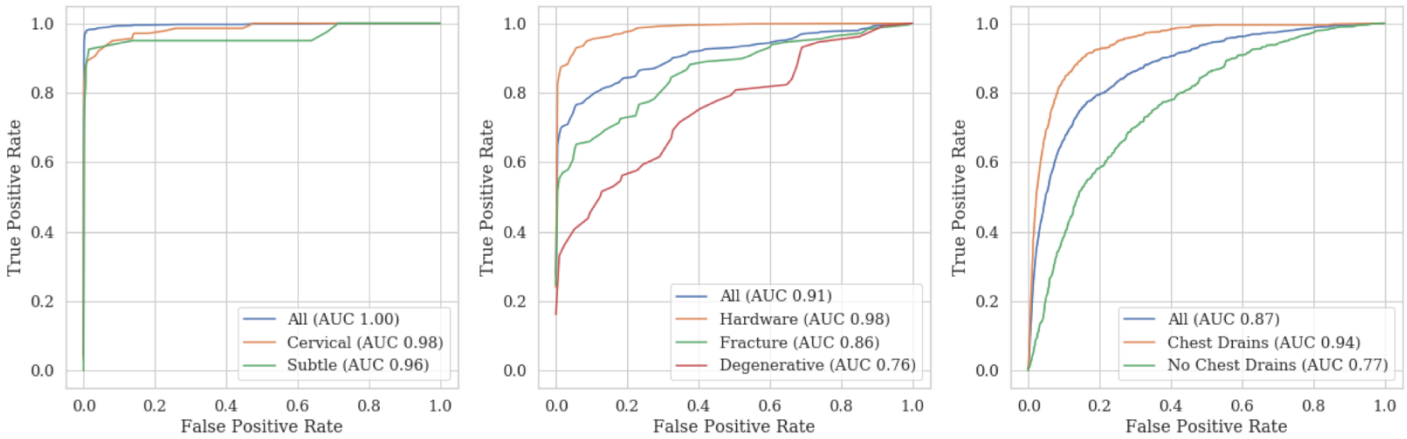

Figure 1. ROC curves for subclasses of models trained on multiple datasets. Panel (a) shows model performance on different subclasses of the "abnormal" class for a model designed to detect abnormalities on radiographs from the Adelaide Hip Fracture dataset, panel (b) shows model performance on different subclasses of the "abnormal" class for a model designed to detect abnormalities in musculoskeletal radiographs from the MURA dataset, and panel (c) shows model performance on different subclasses of the "pneumothorax" class for a multi-class CXR classification model designed to detect 14 different pathologies on the CXR-14 dataset. From Oakden-Rayner et al.,[15] with permission.

Creating Training Datasets

Once an appropriate task and measurement metric have been defined, creating an AI system to perform that task generally requires constructing a dataset on which a model will be trained. In supervised learning, which dominates current applications in radiology, this requires curating labeled training data. Unfortunately, the cost of creating these labeled datasets can limit the application of AI systems in clinical practice. Using the work of Gulshan et al.[16] as an example, 3-7 physicians, most of whom are licensed ophthalmologists, were reported to have graded every single one of 128,175 retinal fundus photographs. Conservatively assuming 3 labelers per image, 15 seconds per image, and a $100 per hour rate, this comes out to a cost

estimate in excess of $150,000 and 180 clinician-days for a single iteration of data labeling; in practice, multiple data labeling efforts are often necessary.

Importantly, even meticulously labeled training sets are not guaranteed to support models that generalize across different diseases, modalities, imaging systems, classification ontologies, clinical protocols, and medical guidelines, all of which change over time and with different application contexts.[5,13,17,18] This concept is known as *distribution shift*, and often causes model performance to degrade when used outside of the exact population on which the training set was constructed. This behavior has been observed in a variety of medical applications including pneumonia detection on CXR,[13] diabetic retinopathy detection on retinal fundus photographs,[18] and dermatology image classification,[17] and remains arguably the dominant challenge in applying AI systems in practice. While various mechanisms for handling distribution shift exist, this problem cannot be considered solved and mitigating it can remain a major cost driver for AI systems in radiology.

A final reason that the burden of creating training datasets can be problematic for AI systems in radiology is that it can lock in outdated standards of care or treatment protocols.[19] For instance, if an AI system for image triage was trained on a dataset that did not contain cases from a newly discovered disease such as COVID-19, it could spuriously deprioritize individuals with those infections. Furthermore, continued use of models trained with expensive datasets that may someday reflect outmoded practice (e.g. x-ray scoring systems that disadvantage marginalized patient subpopulations[5]) would result in patients receiving medical recommendations that are below the modern standard of clinical care. For radiologists, this issue may be particularly apparent for imaging protocols, which evolve over time and may be inconsistently implemented. As an example, widespread use of AI systems for CT analysis developed using a particular

protocol for contrast timing may result in the continued use of that protocol even though it may be suboptimal for other reasons.[1]

In summary, creating appropriately representative labeled datasets is likely to remain a challenge for widespread use of AI systems in radiology, both because of the cost associated and the inherent difficulty of ensuring that a dataset represents all important axes of variation, including variations caused by changes in radiology technology and practice in the future.

Mitigating Algorithmic Bias

A major challenge for both users and developers of clinical AI systems is ensuring that they do not create or amplify biases in the provision of care that would disadvantage particular groups of patients. In technical parlance, this involves building models that are "robust" to important variations in the patient population such as gender, ethnicity, socioeconomic status, and other protected factors. As described above, creating representative datasets for training and evaluation of AI models is an important component of mitigating model bias, and it is worth further discussing specific error modes that can lead to biased datasets. First, data from Electronic Health Records (EHRs) are often not meant for algorithm development, meaning that models developed using cohorts and labels drawn from EHRs may contain a variety of inherent biases such as those resultant from the use of billing codes rather than pathological descriptions for diagnoses.[2] Second, because it can be difficult to access patient data (even for patients themselves), standard strategies for enrolling diverse populations in clinical development efforts can be difficult to apply.[1,2] Some health systems also suffer from selection bias, where information that would be useful for data labeling is only recorded for cases of particular academic or clinical interest. Furthermore, even with appropriate cohort design, data may either

be missing[20] or only available in certain segments of the population. A particularly striking example of this situation was highlighted recently by the work of Kaushal et al.,[21] which showed that the majority of AI studies in imaging performed in the United States leveraged data from only three states. Prospective users of AI systems must be constantly vigilant for these types of dataset curation issues that can result in biased algorithms.

Common training approaches that do not account for such issues as hidden stratification or class imbalance can also result in biased models. For instance, models are often trained to optimize average performance; such procedures result in models that perform well on majority classes (or subclasses) at the expense of less common groups in the population.

Finally, it is worth pointing out that unintended bias can also occur in algorithms aimed at improving elements of the image reconstruction process in volumetric imaging. For example, while both tomographic protocols and magnetic resonance imaging (MRI) could benefit from AI-based steps in the calibration, signal conditioning, denoising, and reconstruction processes, it is not always clear that mathematical transformations learned on a particular set of data or population will provide similar utility on other datasets.[7] Common axes of variation that should be considered in dataset curation and algorithm design for such applications include scanner or hardware type, exam protocol, tracer type, patient characteristics, and other parameters that could affect image acquisition and reconstruction.

Measuring Correlation Instead of Causation

A particularly concerning pitfall in deep learning systems has been their ability to make accurate predictions based on features that are correlated with the outcome, but which are *non-causal*. In radiology, examples include algorithms that predict severe disease when they

recognize a portable scanner was used instead of a fixed x-ray machine (which would require the patient to be well enough to travel to the radiology department for the image), and those that rely on the presence of chest drains to predict pneumothorax.[13,15] In dermatology, a prime example is a recent algorithm that used the presence of surgical markings to recognize melanoma in dermoscopic images.[22] Because deep learning systems are usually optimized to maximize a specific performance metric without considering causality, they are prone to mistakes such as these, predicting outcomes based on confounding, non-causal features.

Technical and Engineering Issues

Even if the risks described to this point are appropriately mitigated, a variety of common technical issues can result in AI systems that do not perform as designed. One such problem is overfitting, which occurs when models perform well on a training set but poorly on a held-out evaluation set; this is often the result of insufficient regularization during training or distribution shift between training and evaluation sets. Data leakage between training and evaluation sets occurs when samples that are in the evaluation set also appear in the training set, and leads to overly optimistic performance metrics on the evaluation set because the model was exposed to very similar examples during training (and it can memorize them rather than learn generally useful image features). While the exact same examples can be included in both sets by accident, a more subtle version of this same error can occur when examples from the same patient are included in both training and evaluation sets.

Poorly calibrated models can also be problematic. A "calibrated" model is one in which the quantitative values output from the model reflect true probabilities; for example, if a well-calibrated diagnostic algorithm predicts that each of four patients have a disease with 75%

probability, one should expect that three out of those four patients would actually have the disease. If a model is not calibrated, clinicians could erroneously interpret model outputs in a manner that would negatively affect patient care.

AI systems can also simply fail; because AI systems are a type of software, bugs are unfortunately a fact of life. In radiology applications, particularly important types of engineering errors include images that are corrupted in transmission/storage and cause erroneous predictions; preprocessing differences between datasets or institutions that result in distribution shift; or simple coding errors that cause model weights to be incorrectly loaded or output to be incorrectly computed. These errors can have real-world consequences, like a critically ill patient being deprioritized or benefits being withheld from needy individuals.[23]

Finally, for image enhancement and reconstruction applications, a major technical challenge involves ensuring that as AI algorithms generate images that are more suitable for human interpretation, they do not insert spurious information that was not in the original image. The difference between *imputation* (the recovery of lost or imperfectly acquired information), *enhancement* (making better use of existing information), and *hallucination* (the creation of information that was not in the original image) is often subtle, and it can be difficult even for domain experts to evaluate.[7] As this area of the field – sometimes referred to as "upstream AI" – matures further, it will be critical to develop robust metrics and evaluation procedures to ensure that AI-enabled image processing techniques can provide value by improving image analysis without inserting spurious information.

Post-Deployment Monitoring

Post-deployment monitoring represents an additional challenge for deployment of AI systems. To mitigate issues related to distribution shift and model bias – as well as to continuously evaluate whether a model is providing the anticipated operational benefit – it is critical that models be constantly under assessment while deployed. Various strategies for post-deployment monitoring exist, including manual human audits of model output, automated algorithmic evaluation of distribution shift or hidden stratification, out-of-distribution (OOD) sample detection, and continued evaluation protocols, but many academic studies that demonstrate initial viability of an AI system do not consider how post-deployment monitoring should be implemented.[15] Furthermore, when considering whether to deploy a given AI system, the cost of continuous monitoring – which includes subject matter expert time, additional data curation, and even the expense of taking a model out of service if it begins performing poorly – must be considered.

Deployment Details

In addition to technical and functional issues, deploying AI algorithms in radiology practice raises a number of ethical, medicolegal, economic, and logistical questions that have not yet been convincingly resolved.

First, if an outside developer creates a model, it must be decided how liability from mistakes that occur in the course of practice should be divided amongst the radiologist, the algorithm developer, the device manufacturer, and other relevant parties.[6] Furthermore, it is often not clear how model output is explained to a patient, whether patients should be informed that AI algorithms were used in their care, and what recourse might be available toward

disputing treatment decisions made based on model output. These issues become even more fraught if models have been fine-tuned for a particular site or deployment environment, and may depend on whether a given model has been developed internally on custom or open-source tooling, has been developed internally using a commercial platform, or is provided via a software-as-a-service or model-as-a-service agreement.

Second, AI models and deployment hardware must be co-optimized to ensure that model execution time is sufficiently rapid to provide anticipated value. In particular, if users of models deployed to edge devices (e.g. laptops, phones, etc.), on extremely large images (e.g. volumetric scans), or in time-critical contexts like interventional radiology do not ensure that sufficient compute capability and network bandwidth are available to support proposed use cases, the resulting slowdown in computing model outputs could have negative clinical consequences. The alternative, however, may be the deployment of expensive new hardware at clinical sites or the use of cloud processing, each of which involves its own risks and benefits.

Third, in order for models to be used ethically, policies regarding the use of and access to patient data by the patients, the treatment center, and any external parties must be explicitly delineated. Unfortunately, in many contexts, public policy has not yet provided sufficient guidance for users to know exactly what procedures should be observed on this front.

Fourth, security considerations in deployment must be appropriately addressed. Were bad actors to gain access to a model or the training data, various attacks can be envisioned that could reveal patient identity, interfere with treatment, or exfiltrate valuable data to which various parties (including the patient) have exclusive rights as well as expectations of privacy. Proposed AI deployments in radiology often do not fully consider the scope of potential attack vectors on both data and models, and do not explicitly guard against such attacks as data poisoning

(affecting model performance by altering training data) or model inversion (reconstructing training data from model parameters). Remaining robust to these sorts of attacks is heavily related to post-deployment monitoring described above, and may benefit from specific approaches to model training and evaluation.[24]

User Trust

In order for AI to provide value in radiology practice, these systems must gain the confidence of both patients and clinicians. Concerns about the deleterious effects of automated assistance on radiologist performance, lack of interpretability in clinical decisions, and the potential for reinforcement of existing biases or outmoded practice must be overcome.[19,25] Automation bias is a serious problem wherein the very fact that human readers have algorithmic support causes them to trust the automated result even when it is flawed. Deep neural networks have well-documented difficulties establishing exactly what reasoning led to a given model output. The danger of introducing models that disadvantage particular patient groups is ever-present. As a result, to make effective and equitable use of AI in radiology, it is critical to design workflows that incorporate not only algorithmic input and broad clinical domain expertise, but also the individualized expertise that doctors have about the situation of each patient and the intimate knowledge that each patient has of their own body.[20]

Regulatory Approval

Deployment of AI algorithms for clinical use cases will rarely occur outside the bounds of governmentally stipulated regulatory structures. As a result, regulations for AI systems in radiology must be designed to balance potential improvements in patient care with the risks that

such systems can pose if deployed incorrectly. Though both governmental agencies[26] and independent bodies[27,28] have recently made substantial progress towards defining constructive paths forward, the evolving regulatory environment will likely mean that certain applications will move faster than others (e.g. computer assisted detection vs. computer assisted diagnosis), and that it will be particularly important for clinicians to understand exactly what models can and cannot do before using them in practice. While substantial discussion of regulation for clinical AI models is handled in a separate chapter, it suffices to say that clinicians intending to use AI in practice should remain up to date on regulations governing system use, processes for approval, and associated reporting requirements.

How Can These Challenges Be Overcome?

While the challenges described above are substantial, technical and operational approaches to mitigate nearly all of them either exist or are in development. The degree to which AI algorithms can provide meaningful value in radiology practice will likely be determined by the effectiveness with which these techniques are implemented in practice and rigorously analyzed in the context of real-world operational data.

Meaningful Performance Measurement

Several concrete steps could help to improve performance measurement of AI models in radiology.

First, common, widely available datasets suitable for evaluating performance on tasks of clinical interest should be constructed and continuously updated by objective bodies such as professional societies, academic consortia, or government agencies. Importantly, these

evaluation datasets should be labeled in a way that closely reflects the intended workflow into which the model will be deployed, as opposed to using arbitrary academic schema. Existing efforts like datasets released by the Radiological Society of North America (RSNA), The Cancer Imaging Archive (TCIA), and others should be expanded.[14,29–31] Furthermore, each task of clinical interest should have evaluation datasets that are *frequently updated* so that models can be evaluated on the latest imaging technologies and not be allowed to overfit to a particular evaluation set.

Second, datasets should be labeled with important subclasses in order to enable analysis of potential model bias and reduce the impact of hidden stratification. Recent unsupervised methods can also be used to algorithmically identify subclasses of interest.[32]

Third, it may sometimes be beneficial to define the scope of model functionality more narrowly in order to enable sharper measurements of performance.[8,33] Instead of aiming for a single model that can generalize across data from different institutions (i.e. multiple distributions), for instance, modelers could consider developing multiple different single-institution models and avoid having to constantly measure relative performance across potentially different populations. Conceptually, this idea resembles recent approaches from precision medicine.[33] If applied carefully, such a strategy could improve the utility of performance measurements for AI models in radiology.

Finally, assessing model performance on downstream clinical tasks – rather than on intermediate performance metrics like accuracy – will help to ensure that performance is measured in a way that is clinically meaningful. Ideally, direct comparison to existing baseline systems should be performed via randomized controlled trials wherein the AI system should be directly integrated into a clinician workflow and the downstream clinical outcome is measured.[25]

The more realistic the setting is, and the closer that we can come to measuring *clinical value* rather than *algorithmic performance*, the more likely we are to arrive at a useful assessment of AI system utility.

Creating Training Datasets

Recent technical progress on methods that can relieve the burden of creating and updating datasets has been promising. First, methods from *weak supervision* have enabled large datasets with weaker, noisier labels to support AI models that perform similarly to those trained on hand-labeled datasets of similar size.[34–36] Many of these methods directly leverage human expertise in a way that enables rapid relabeling and retraining to combat model performance and distribution shift issues. Automated, NLP-based labelers have also shown promise in building labeled datasets, though adapting them to new domains can be labor-intensive.[37,38]

Other technical approaches have focused on leveraging additional sources of signal within the model training process. Modern data augmentation techniques enable users to increase the effective size of training datasets by applying transformations to existing images without disrupting the meaningful features within those images. Common examples include applying rotations to labeled images or synonymy swaps to labeled text in language modeling tasks.[39,40] Multitask learning – building models that learn to perform multiple, related tasks simultaneously – can also help to decrease the number of labeled examples required by leveraging additional information from the dataset. Transfer learning applies a similar approach, but usually involves two steps: (1) pre-training a model on a task that is related to the final task of interest and (2) fine-tuning that pre-trained model by continuing to train it on the task of interest.[41] In medical computer vision applications, for instance, it is particularly common to use

models that are pre-trained on the ImageNet database as a starting point upon which to train models for clinical use cases.[8,11,33,42,43] Recent approaches from self-supervision and contrastive learning that leverage large, unlabeled datasets for model pre-training have also shown promise in reducing the required size of labeled datasets.[44]

In clinical applications, another way that the data curation burden can be reduced is by standardizing protocols. Instead of having to train models over images acquired via a wide variety of protocols – e.g. tube currents, voltages, and reconstruction settings in computed tomography – it can be advantageous to train models obtained using a standard protocol and then ensure that such models are only applied to images obtained using that standard protocol. Similar to the precision medicine perspective presented above, this approach trades off generalizability for a narrow task definition.

Mitigating Algorithmic Bias

Combating algorithmic bias is one of the single most important tasks required to deploy AI models ethically and equitably within radiology practice. In addition to constructing training data in as non-biased a way as possible, there exist several additional approaches that can help to mitigate this problem.

First, a variety of training algorithms focused on reducing the worst-case subgroup performance – that is, ensuring that there exists no subgroup of data on which a model performs substantially worse than another – have been the focus of recent research.[32,45–47] As these and additional approaches for improving algorithmic fairness are developed, they should be considered for clinical translation.[48]

Second, because these training algorithms are often used during model development rather than model deployment, clinical users may rarely interact with them. However, clinical users will routinely be exposed to model output, and as a result tooling designed to clearly and dynamically evaluate model robustness will become an increasingly important part of successful AI deployments in radiology.[49,50] Research and development studies focused on enabling clinical users to reliably determine which model features are most responsible for a given output, to quickly assess model performance on a wide variety of subclasses or subgroups, and to rapidly evaluate the effect of such variations on clinical outcomes would improve our ability to deploy models equitably.

Finally, direct participation from physician and patient communities in model development and deployment can help to ensure that individuals are best served by these models in practice. Indeed, as pointed out by Esteva et al. in their recent review article,[11] community participation recently enabled the discovery of dataset bias and identified demographics underserved by a model for population health management.[51] A similar case occurred when evaluating models for detecting diabetic retinopathy in Southeast Asia, where socioeconomic factors heavily impacted model efficacy.[18] If radiologists are able to deploy AI models in cooperation with their clinical communities – while ensuring that non-AI backups are used when appropriate – these capabilities stand a much better chance of having a clinical impact that is both positive and equitable.

Measuring Correlation Instead of Causation

Ensuring that models do not rely on confounding variables in making their predictions requires many of the same strategies described above. Model auditing by human actors can help to discover cases where models make the right prediction for the wrong reason. External

validation can be a particularly helpful tool in ensuring that dataset artifacts are not responsible for model performance. Encouraging models to respect important invariances via data augmentation strategies can further reduce the possibility of non-causal features driving model predictions. Finally, interpretability analyses such as heatmaps that identify which structures informed the algorithmic decisions[52] and other visualization methods can help radiologists to identify such behavior before it becomes a problem.

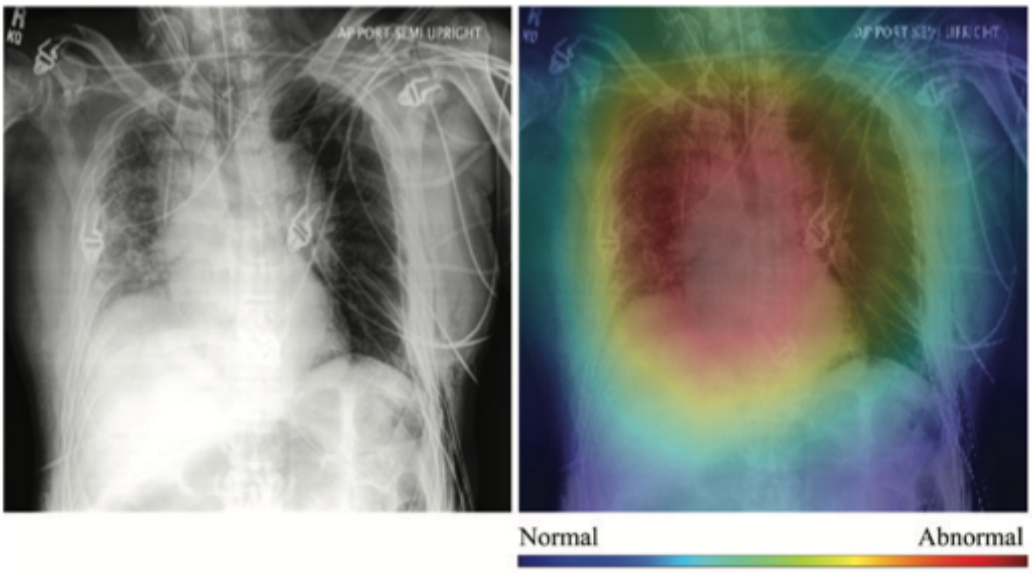

Figure 2. Image (left) and class activation map (right) showing the area that most heavily influenced a neural network designed for binary radiograph triage to provide an "abnormal" prediction. Red indicates areas of relatively high contribution to an abnormal score, while blue areas indicate the opposite. From Dunnmon et al., with permission.[8]

Technical and Engineering Issues

Many of the technical issues described here should be identified and addressed by applying best practices from software engineering. Clearly defining testing strategies before model development, ensuring that systems are routinely tested during deployment, and integrating the entire data processing pipeline into those procedures can reduce the probability of

unintended errors making their way into critical software paths. In radiology, the data processing pipeline includes data ingestion from hardware, image reconstruction, transfer to and egress from a Picture Archiving and Communications System (PACS), conditioning operations such as histogram equalization, and model inference.

Post-Deployment Monitoring

Post-deployment monitoring can be accomplished in several ways, as described by Oakden-Rayner et al.[15] First, if clinicians are able to define subgroups or performance tests of interest before model development, tests based on these definitions can be implemented and continuously evaluated for anomalous behavior during deployment. Algorithmic auditing, where human experts periodically inspect model output to identify concerning trends, is often viable in cases where it is not possible to write a comprehensive set of tests before development. Finally, recently developed algorithmic measures for assessing worst-case subgroup performance can provide value by identifying poorly performing groups without human intervention.[32]

An important aspect of post-deployment monitoring is ensuring that cases on which the model was not intended to be executed – for instance, a lateral chest X-ray for a model that was trained on frontal exams – is not erroneously provided to a model for analysis. The increasing amount of research dedicated to the task of identifying samples that are outside the distribution on which a model was intended to operate, commonly called "out-of-distribution (OOD) detection," has provided encouraging evidence that OOD samples can be automatically identified and flagged. OOD detection should become a standard tool in post-deployment monitoring suites, and should inform both deployment practice and future model development.

Finally, consistent use of adverse event registers for AI systems in radiology could help to provide high-level monitoring for undesired outcomes. Such registers are standard practice for deployed medical products, and for AI systems would simply record any untoward medical occurrence that happened while that system was in use. Though they do not provide causal information, observational information from adverse event registers could be useful in post-deployment monitoring for broadly deployed AI systems in radiology.

Deployment Details

To address deployment challenges described above, additional development work is required on a number of fronts. On medicolegal issues of liability, responsibility, and data rights, the larger volumes of case law that should be expected in the near future should help to directly resolve some of these questions. On the hardware-software codesign front, effective systems engineering and modular design should become standard practice from model developers as the industry matures. Hospitals may ultimately desire to invest in their own inference hardware (e.g. dedicated CPUs, GPUs, mobile devices), or even to run computation in a secure cloud environment; each of these decisions has advantages and disadvantages, and it is not clear what approach will become dominant. Finally, we expect a similar trend in model cybersecurity. As it becomes clear that both models and associated data have substantial economic value (and possibly legal protections), penetration testing and other traditional cybersecurity protocols will likely become an even more important part of medical information technology systems than they already are. Practitioners can improve the chances of a successful AI deployment by accounting for the associated engineering and compliance costs up front, and ensuring that they weigh these costs against the expected value provided by the AI system.

User Trust

To improve user trust in AI systems for radiology, involving clinical and patient users in model and workflow development from the beginning is essential. To be able to use a system confidently in practice, clinician users must have trained with it, internalized its strengths and weaknesses, and become comfortable with both integrating its output into their decision processes and explaining those processes to patients. Like any clinical investigation, patient awareness and education will be paramount for effective engagement and improvement of care. Any improvements that can be made to model interpretability will assist clinician users in bridging this gap, and incorporating the possibility of a follow-up exam to confirm the predictions of an AI system would likely have positive outcomes in many cases. In the end, user trust will only be developed insomuch as the benefit of the AI systems for concrete clinical decisions can be directly observed by clinicians and clearly communicated to patients.

Regulatory Approval

Clinicians have an opportunity to work directly with the public policy community to create regulatory structures that incentivize innovation while maintaining appropriate safety standards. Linking regulatory guidance and approval to standardized reporting for model development and performance such as the SPIRIT-AI and CONSORT-AI guidelines would not only provide clarity for regulatory approvers, but also ensure that users of a given AI-based system are well-informed about exactly how it was developed, precisely what population it was intended for, and any other items that would be important for post-deployment monitoring and clinical use. While much work in this area remains to be done, progress in recent years has been rapid, and we expect that the regulatory environment will continue to mature in the near future.

Regulatory issues for AI systems in radiology are discussed in further detail in a separate chapter from Harvey et al.

Acknowledgements

The author is grateful to the following individuals for their helpful feedback on this work: Daniel Rubin, Matt Lungren, Luke Oakden-Rayner, Sarah Hooper, Khaled Saab, Neel Guha, Swetava Ganguli, and Adele Xu.

References

1. Thrall JH, Li X, Li Q, et al. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. *J Am Coll Radiol*. 2018;15:504-508. doi:10.1016/j.jacr.2017.12.026
2. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med*. 2019;380(14):1347-1358. doi:10.1056/nejmra1814259
3. Rajpurkar Id P, Id JI, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. Published online 2018. doi:10.1371/journal.pmed.1002686
4. Tang YX, Tang YB, Peng Y, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit Med*. 2020;3(1). doi:10.1038/s41746-020-0273-z
5. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med*. 2021;27(1):136-140. doi:10.1038/s41591-020-01192-7
6. Chiwome L, Okojie OM, Rahman AKMJ, Javed F, Hamid P. Artificial Intelligence: Is It Armageddon for Breast Radiologists? Published online 2020. doi:10.7759/cureus.8923
7. Chaudhari AS, Sandino CM, Cole EK, et al. Prospective Deployment of Deep Learning in <scp>MRI</scp> : A Framework for Important Considerations, Challenges, and Recommendations for Best Practices. *J Magn Reson Imaging*. Published online August 24, 2020:jmri.27331. doi:10.1002/jmri.27331
8. Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs. *Radiology*. 2019;290(2):537-544. doi:10.1148/radiol.2018181422
9. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med*. 2018;24(9):1337-1341. doi:10.1038/s41591-018-0147-y
10. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. Published online 2017.

doi:10.1038/nrclinonc.2017.141
11. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. *npj Digit Med*. 2021;4(1):5. doi:10.1038/s41746-020-00376-2
12. Exploring the ChestXray14 dataset: problems – Luke Oakden-Rayner. Accessed January 24, 2018. https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/
13. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. Sheikh A, ed. *PLOS Med*. 2018;15(11):e1002683. doi:10.1371/journal.pmed.1002683
14. Sawyer Lee R, Dunnmon JA, He A, Tang S, Ré C, Rubin DL. Comparison of segmentation-free and segmentation-dependent computer-aided diagnosis of breast masses on a public mammography dataset. *J Biomed Inform*. 2021;113:103656. doi:10.1016/j.jbi.2020.103656
15. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Published online September 26, 2019. Accessed November 4, 2019. http://arxiv.org/abs/1909.12475
16. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 2016;316(22):2402. doi:10.1001/jama.2016.17216
17. Kamulegeya LH, Okello M, Bwanika JM, et al. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. *bioRxiv*. Published online October 31, 2019:826057. doi:10.1101/826057
18. Beede E, Baylor E, Hersch F, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery; 2020:1-12. doi:10.1145/3313831.3376718
19. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: The case of electrocardiograms. *J Am Med Informatics Assoc*. 2003;10(5):478-483. doi:10.1197/jamia.M1279
20. Medicine's Machine Learning Problem | Boston Review. Accessed January 9, 2021. https://bostonreview.net/science-nature/rachel-thomas-medicines-machine-learning-problem
21. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA - J Am Med Assoc*. 2020;324(12):1212-1213. doi:10.1001/jama.2020.12067
22. Winkler JK, Fink C, Toberer F, et al. Association between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*. 2019;155(10):1135-1141. doi:10.1001/jamadermatol.2019.1735
23. A healthcare algorithm started cutting care, and no one knew why - The Verge. Accessed January 18, 2021. https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy
24. Cohen J, Rosenfeld E, Kolter JZ. Certified adversarial robustness via randomized smoothing. In: *36th International Conference on Machine Learning, ICML 2019*. ; 2019.
25. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial

intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-237. doi:10.1136/bmjqs-2018-008370
26. Harvey HB, Gowda V. Special Review How the FDA Regulates AI. Published online 2020. doi:10.1016/j.acra.2019.09.017
27. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7
28. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x
29. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J Digit Imaging*. Published online 2013. doi:10.1007/s10278-013-9622-7
30. 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*. Published online 2020. doi:10.1148/ryai.2020190211
31. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. Published online 2019. doi:10.1038/s41597-019-0322-0
32. Sohoni N, Dunnmon JA, Angus G, Gu A, Ré C. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems.
33. Thrall JH, Fessell D, Pandharipande P V. Rethinking the Approach to Artificial Intelligence for Medical Image Analysis: The Case for Precision Diagnosis. *J Am Coll Radiol*. 2021;18:174-179. doi:10.1016/j.jacr.2020.07.010
34. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid training data creation with weak supervision. *Proc VLDB Endow*. 2017;11(3):269-282.
35. Dunnmon J, Ratner A, Khandwala N, et al. *Cross-Modal Data Programming Enables Rapid Medical Machine Learning*.; 2019. Accessed April 20, 2019. http://arxiv.org/abs/1903.11101
36. Fries JA, Varma P, Chen VS, et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nat Commun*. 2019;10(1):3111. doi:10.1038/s41467-019-11012-3
37. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proc AAAI Conf Artif Intell*. 2019;33(01):590-597. doi:10.1609/aaai.v33i01.3301590
38. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *Proc Am Med Informatics Assoc Summits Transl Sci*. 2018;2017:188.
39. Ratner AJ, Ehrenberg H, Hussain Z, Dunnmon J, Ré C. Learning to compose domain-specific transformations for data augmentation. In: *Advances in Neural Information Processing Systems*. ; 2017:3236-3246.
40. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le Q V. Autoaugment: Learning augmentation strategies from data. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. ; 2019. doi:10.1109/CVPR.2019.00020
41. Eyuboglu S, Angus G, Patel BN, et al. *Multi-Task Weak Supervision Enables Automated*

Abnormality Localization in Whole-Body FDG-PET/CT.
42. Jia Deng, Wei Dong, Socher R, Li-Jia Li, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *IEEE CVPR*. Published online June 2009:248-255. doi:10.1109/CVPR.2009.5206848
43. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056
44. Azizi S, Mustafa B, Ryan F, et al. *Big Self-Supervised Models Advance Medical Image Classification*.
45. Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. Published online November 20, 2019. Accessed December 3, 2019. http://arxiv.org/abs/1911.08731
46. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant Risk Minimization. Published online July 5, 2019. Accessed November 23, 2019. http://arxiv.org/abs/1907.02893
47. Pfohl S, Marafino B, Coulet A, Rodriguez F, Shah NH, Pala-Niappan L. Creating Fair Models of Atherosclerotic Cardiovascular Disease ACM Reference Format. doi:10.1145/3306618.3314278
48. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med*. 2020;26(1):16-17. doi:10.1038/s41591-019-0649-2
49. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. ; 2019. doi:10.1145/3287560.3287596
50. Goel K, Rajani N, Vig J, et al. Robustness Gym : Unifying the NLP Evaluation Landscape. :1-25.
51. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-)*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
52. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. In: *IEEE CVPR*. ; 2016:2921-2929. Accessed March 30, 2018. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Zhou_Learning_Deep_Features_CVPR_2016_paper.pdf