frontiers
in Digital Health

ORIGINAL RESEARCH
published: 03 August 2021
doi: 10.3389/fdgth.2021.671015

Detecting Spurious Correlations With Sanity Tests for Artificial Intelligence Guided Radiology Systems

Usman Mahmood[1], Robik Shrestha[2], David D. B. Bates[3], Lorenzo Mannelli[4], Giuseppe Corrias[5], Yusuf Emre Erdi[1] and Christopher Kanan[2]*

[1] Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, United States, [2] Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, United States, [3] Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, United States, [4] Institute of Research and Medical Care (IRCCS) SDN, Institute of Diagnostic and Nuclear Research, Naples, Italy, [5] Department of Radiology, University of Cagliari, Cagliari, Italy

Artificial intelligence (AI) has been successful at solving numerous problems in machine perception. In radiology, AI systems are rapidly evolving and show progress in guiding treatment decisions, diagnosing, localizing disease on medical images, and improving radiologists' efficiency. A critical component to deploying AI in radiology is to gain confidence in a developed system's efficacy and safety. The current gold standard approach is to conduct an analytical validation of performance on a generalization dataset from one or more institutions, followed by a clinical validation study of the system's efficacy during deployment. Clinical validation studies are time-consuming, and best practices dictate limited re-use of analytical validation data, so it is ideal to know ahead of time if a system is likely to fail analytical or clinical validation. In this paper, we describe a series of sanity tests to identify when a system performs well on development data for the wrong reasons. We illustrate the sanity tests' value by designing a deep learning system to classify pancreatic cancer seen in computed tomography scans.

Keywords: deep learning, computed tomography, bias, validation, spurious correlations, artificial intelligence

OPEN ACCESS

Edited by:
David Day-Uei Li,
University of Strathclyde,
United Kingdom

Reviewed by:
Hao Wu,
Anhui University of Technology, China
Dong Xiao,
University of Strathclyde,
United Kingdom

****Correspondence:***
Usman Mahmood
mahmoodu@mskcc.org

Specialty section:
This article was submitted to
Health Informatics,
a section of the journal
Frontiers in Digital Health

Received: *22 February 2021*
Accepted: *29 June 2021*
Published: *03 August 2021*

Citation:
Mahmood U, Shrestha R, Bates DDB, Mannelli L, Corrias G, Erdi YE and Kanan C (2021) Detecting Spurious Correlations With Sanity Tests for Artificial Intelligence Guided Radiology Systems.
Front. Digit. Health 3:671015.
doi: 10.3389/fdgth.2021.671015

1. INTRODUCTION

Artificially intelligent (AI) computer-aided diagnostic (CAD) systems have the potential to help radiologists on a multitude of tasks, ranging from tumor classification to improved image reconstruction (1–4). To deploy medical AI systems, it is essential to validate their performance correctly and to understand their weaknesses before being used on patients (5–8). For AI-based software as a medical device, the gold standard for analytical validation is to assess performance on previously unseen independent datasets (9–12), followed by a clinical validation study. Both steps pose challenges for medical AI. First, it is challenging to collect large cohorts of high-quality and diverse medical imaging data sets that are acquired in a consistent manner (13, 14). Second, both steps are time-consuming, and best practices dictate limited re-use of analytical validation data. The cost of failing the validation process could prohibit further development of particular applications.

One reason AI systems fail to generalize is that they learn to infer spurious correlations or covariates that can reliably form decision rules that perform well on standard benchmarks (15). For example, an AI system successfully trained to detect pneumonia from 2D Chest X-rays gathered from multiple institutions, but it failed to generalize when images from new hospitals outside of the training and assessment set were used to evaluate the system (16). The investigators found that the system had unexpectedly learned to identify metal tokens seen on the training and assessment images (16). In hindsight, the tokens were obvious spurious correlators, but in other cases, the covariates can be less obvious (15). For example, subtle image characteristics that may be unrelated to the target object, such as high-frequency patterns (17–19), object texture (20, 21), or intangible attributes of objects are known to cause AI systems to form decision rules that may not generalize (15, 22). Current research has focused on explaining or interpreting AI decisions using various visualization techniques (23), but these do not necessarily imply that a system will generalize (24–27).

Addressing system failures before clinical deployment is critical to ensure that medical AI applications are safe and effective. Identifying systems that are right for the wrong reasons during the development stages can expedite development by not wasting valuable validation data from multiple institutions or conducting doomed clinical validation studies.

The standard approach used to identify system failures involves testing with held-out development or generalization test datasets (28). However, development test sets are subsets of the training data, and their primary value lies in identifying systematic errors or bugs within the AI algorithm. Generalization test data are independent of the development data (i.e., their joint probability distribution of inputs and labels differ from training and development test data) (29). The generalization data's value is to assess how well a trained model may adapt to previously unseen data. However, neither type of test is sufficiently robust enough to declare when an AI system is ready for the clinic.

We provide a set of sanity tests that can demonstrate if a trained system is right for the wrong reasons. We developed a weakly supervised deep learning system for classifying pancreatic cancer from clinical computed tomography (CT) scans to illustrate their use. Our main contributions are:

1. We provide a set of sanity tests to determine if a system is making predictions using spurious correlations in the data.
2. We describe a system for using deep learning with CT images to detect pancreatic cancer, and we apply our set of sanity tests to both development and generalization test datasets. We train and assess four unique variants of this system to illustrate the pipeline and demonstrate that the system looks as if it performs well in many scenarios, but it is predicting using spurious correlations.
3. We illustrate how to use a method to generate noise images from the patients' volumetric CT scans. These can then be used to assess the influence of noise on the AI system's performance.

2. MATERIALS AND METHODS

2.1. Sanity Tests for AI Systems

There are various testing procedures employed in software engineering to determine if a system is working correctly, such as smoke and sanity tests (30). Smoke tests evaluate the critical functionality of a system before conducting additional tests. In AI, this is analogous to reaching an acceptable level of performance on the development test data, which matches the training data's distribution. Development test data is typically a random sample of the training data (e.g., 30% test and 70% train). The stopping point for many AI projects is when acceptable performance is achieved on the development test set, but in software engineering, the next step is to conduct 'sanity tests' that indicate if a system produces obvious false results. If the sanity tests fail, further development is done before conducting more time-consuming and rigorous tests, which for AI systems used in medical applications could correspond to analytical and clinical validation studies. For AI systems, sanity tests would identify if a system is achieving good results on the development test set for the wrong reasons (e.g., covariates or spurious correlations) and will therefore fail in other environments or on other datasets.

Sanity tests are occasionally used to identify if a system is unlikely to generalize (24, 31, 32). However, the tests are often designed to evaluate literature methods instead of being used as a crucial development tool. For example, Shamir et al. critiques the methods by which face datasets were designed and evaluated by showing that commonly used face recognition datasets were classified correctly even when no face, hair, or clothing features appeared in the training and testing datasets (33). As another example, in response to a report suggesting AI systems could diagnose skin cancer at the level of dermatologists (34), Winkler et al. evaluated the limits of the claim by testing a trained AI system using dermoscopic images where the covariate's, hand-drawn skin markings, were first present and then absent from pictures of the skin cancer. They observed that when skin markings were present, the probability that the AI system classified images as having skin cancer increased significantly. With the markings removed, the probability decreased, which led them to conclude that the AI system associated the markings with cancer instead of the actual pathology (31).

For AI-based medical devices, conducting sanity tests can prevent needless harm to the patient and save a considerable resources. However, without sufficiently large, well-annotated datasets, performing analytical validation to determine the root causes that drive AI systems to fail before deployment remains a challenge (5, 35). Moreover, after independent testing data is gathered, regulatory organizations advise that the data be used a limited number of times to prevent over-fitting (36). For example, the United States Food and Drug Administration "discourages repeated use of test data in the evaluation" of CAD systems (37). Clinical validation of deployed systems is likewise time-consuming to organize and often costly.

We propose a series of sanity tests to identify if an AI system may fail during the development phases and before conducting more extensive generalization tests. We also describe how the tests are used with a case study to detect pancreatic cancer from weakly labeled CT scans. The tests are as follows:

- **Train and test with the target-present and absent**. If an AI system is trained to distinguish between normal and abnormal diagnostic features (e.g., organ with cancer shown in **Figure 1A**), then it should fail when that target is removed from the development test data (e.g., **Figure 1B**). If the system still works effectively after removing the target from testing data, then that indicates it is confounded. In our case study,

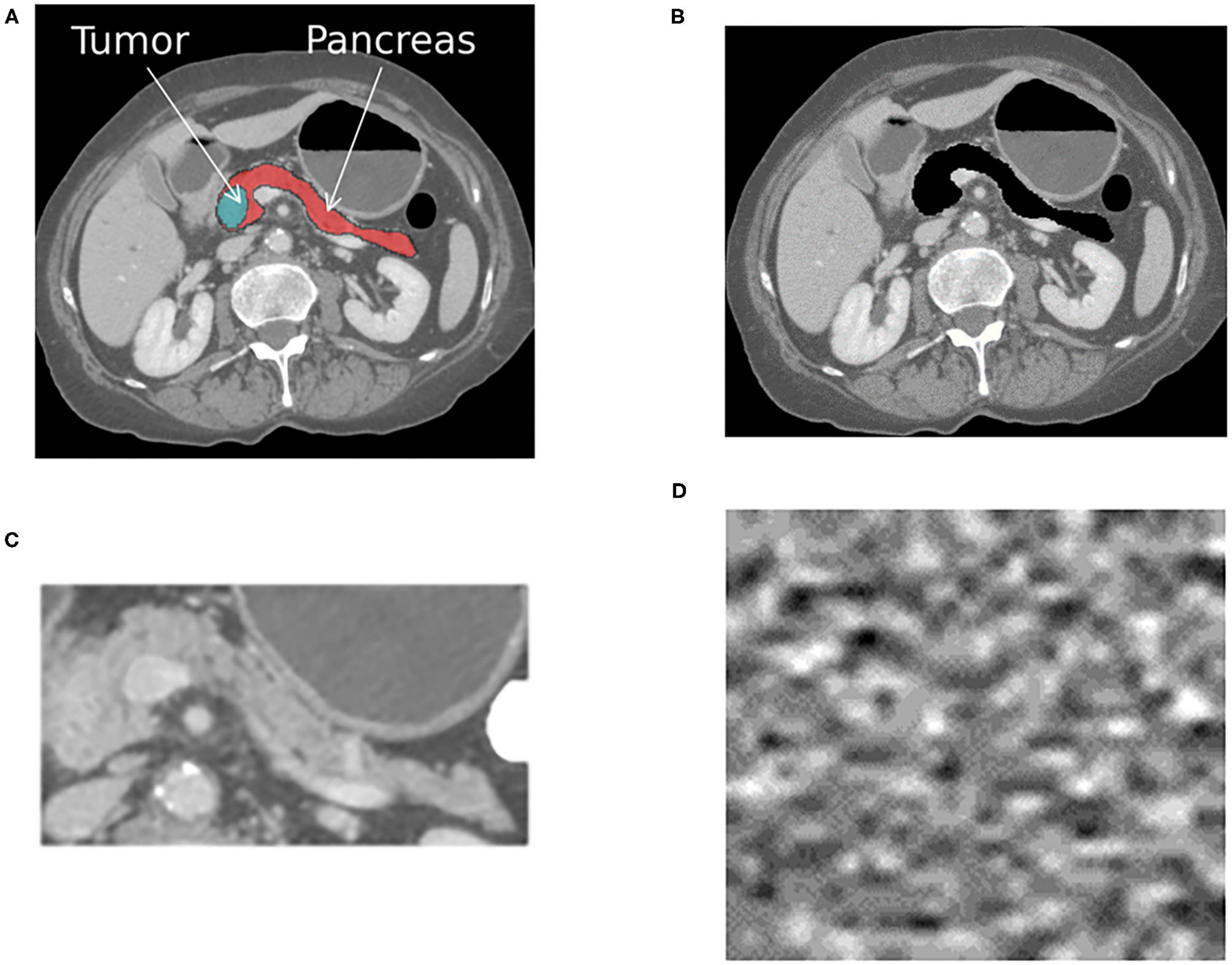


FIGURE 1 | The figure shows an example cross-sectional slice from a single patient with pancreatic cancer that was processed into four different input formats, as shown. A different AI system was trained with each input format, resulting in four different systems that were then tested on each format. **(A)** The original image with the pancreas and tumor present. **(B)** The original image with the pancreas and tumor removed. **(C)** The anatomy surrounding the pancreas is cropped out, and only the pancreas and tumor remain. **(D)** Noise image generated from the patient CT scans. The same processing was applied to all CT scans within the positive for pancreatic cancer and normal pancreas classes.

this corresponds to removing the pancreas from normal scans and pancreas with tumor from abnormal scans using a segmentation mask, as shown in **Figure 1C**. We removed the whole pancreas because the pancreatic tumor often distorts the contours of the surrounding anatomy (38).

- **Train and test the system with background patches or noise images**. Background patches consist of non-target regions of the image. Noise images can be generated from the volumetric CT scans in the development and generlization datasets. Both can determine if the different classes can be discriminated based on features unrelated to the target objects (33). If classes are discriminated against with high confidence using the noise image types, then the system is confounded, and it is using features of the image acquisition process to delineate classes. An example noise image generated from the patient CT scans is shown in **Figure 1D**.
- **Test with different regions of interests (ROIs)**. Training and testing AI systems on precisely outlined segments of images does not reflect real-world usage. Medical centers, private practices, or institutes where AI is deployed may not have the resources or expertise to precisely outline the anatomical area (39). Furthermore, similar to radiologists, AI systems may have to parse through anatomy they have never encountered during training. Therefore, it is desirable to ensure that when systems are trained on a select portion of images, as shown in **Figure 1C**, they can generalize to the original image shown in **Figure 1A**.

These sanity tests can be conducted solely using the development dataset, but ideally they would also be used in conjunction with another generalization dataset. They require four input formats, as shown in **Figure 1**, to be generated from the same development dataset.

TABLE 1 | Scan parameters and patient-specific characteristics for development and generalization data.

	Development Data: Train, Tune, and Test		
	TCIA - Pancreas CT	**Medical Image Segmentation Decathalon (MSD)**	**Generalization data**
Annotated	Yes	Yes	No
CT Vendor	Phillips and Siemens	General Electric	General Electric
CT Model	**	LS16 or HD750	HD750
Total # of Patients	82 (27 female/55 male)	281*	116 (61 female/55 male)
# used to train	58	60	NA
# used to tune	15	14	NA
# used to test	9	8	116 (58 without PC, 58 with PC)
Dataset Information:			
Average age (min to max)	46.8 (18 to 76)	**	63 (18–90)
Scan start time after contrast administration	∼70 s	80–85 s	∼40 s
Avg. # of total slices (min/max)	256 (181–466)	95 (37–751)	186 (102–278)
Avg. # of slices consisting of only pancreas (min/max)	85 (45–144)	30 (11–147)	NA
Scan parameters:			
Tube potential (kVp)	120	120	70 keV (80/140 kVp)
Slice thickness (mm)	1.5–2.5	2.5	2.5
Pixel dimensions (mm)	0.664 to 0.977	0.606 to 0.977	0.547 to 0.976
Tube current modulation index	**	Noise Index: 14 (HD750) / 12.5 (LS16)	NA
Tube current (mA) min to max range	**	220–380 mA	260–600
Rotation time (s)	**	****0.7 (HD750) / 0.8 (LS16)	****0.7 (HD750)
Pitch	**	****0.984 (HD750) / 1.375 (LS16)	****0.984 (HD750)
Reconstruction algorithm	**	**	***FBP/ASiR 20%
Reconstruction kernel	**	**	Standard
Iterative reconstruction strength	**	**	20%
# of data channels	**	**	64
Size of a single data channel (mm)	**	**	0.625
Bowtie filter	**	**	Large body
CT scan series released or used	Axial portal venous phase	Axial portal venous phase	Axial parenchymal phase

**A subset of the MSD dataset was randomly selected to train the model.*
***Not available in accompanied report or DICOM header.*
****FBP, Filtered Back Projection; ASiR, Adaptive Statistical Iterative Reconstruction.*
*****LS16, LightSpeed16; HD750, Discovery High Definition 750.*

2.2. Datasets

Institutional review board approval was obtained for this Health Insurance Portability and Accountability Act-compliant retrospective study. The requirement for informed consent was waived. The case study is designed as a binary classification problem with the aim of identifying patients who have pancreatic cancer vs. those who do not from the provided CT scans. We distinguish between the development dataset used for training, tuning and testing, and the generalization test data used to validate the efficacy of the system. The development dataset was processed into four different formats, as shown in **Figure 1**. The four formats were used to train four different AI systems. The input to each system was a volumetric CT scan that consisted of a normal pancreas or pancreas with tumor. The output from each system was a single classification score that indicated the probability of the patient having pancreatic cancer.

2.2.1. Development Data

The development dataset consisted of patient CT scans collected from two open-access repositories where detailed annotations were available. The normal pancreas CT scans were obtained from The Cancer Imaging Archive Normal (TCIA) Pancreas Dataset with 82 contrast-enhanced abdominal CT scans (40). Seventeen patients from the TCIA dataset were reported to be healthy kidney donors. The remaining patients were selected because they had no major abdominal pathology or pancreatic lesions (40). The abnormal pancreas CT scans were obtained from the Medical Image Segmentation Decathlon (MSD) dataset, consisting of abdominal CT scans from 281 patients. The MSD dataset contains patients who presented with intraductal mucinous neoplasms, pancreatic neuroendocrine tumors, or pancreatic ducal adenocarcinoma (41). They were originally used to predict disease-free survival or assess high-risk intraductal papillary mucinous neoplasms seen on the CT scans (41). We randomly selected 82 cases from the MSD dataset to match the TCIA dataset size to avoid class-imbalance issues. The development data were randomly split into a training (58 normal, 60 cancer), tuning (15 normal, 14 cancer), and held-out test (9 normal, 8 cancer) set. To ensure the number of positive and negative samples were balanced in each split, we used stratified five-fold cross-validation for training. **Table 1** shows the patient demographics and scanning parameters provided for each dataset.

2.2.2. Generalization Data: Dual Energy CT (DECT)

The generalization data consists of 116 patients (58 without PC, 58 with PC) who received routine DECT scans between June 2015 to December 2017 (see **Table 1**). The patients without pancreatic cancer received DECT CT Urography (CTU) exams and were selected based on the statement of a negative or unremarkable pancreas and liver in the radiologist report. Those with cancer were selected if they had undergone a DECT arterial phase CT scan and were histologically confirmed to have pancreatic cancer. All patients were scanned on a 64 slice CT scanner (Discovery CT750 HD, GE Healthcare, Milwaukee, WI, U.S.) with rapid switching DECT following the administration of 150 mL of iodinated contrast (Iohexol 300 mgI/mL, Omnipaque 300, GE Healthcare, Cork, Ireland), at 4.0 mL/s. The scan parameters are displayed in **Table 1**. With DECT, multiple image types can be generated, such as virtual monochromatic images (VMI) that depict the anatomy and physiology from the viewpoint of a monochromatic x-ray source (42). The VMI scans can be reconstructed at energies ranging from 40 to 140 keV. For this study, all scans were reconstructed at 70 keV because of its use in the clinic. The images were generated using the GSI MD Analysis software available on Advantage Workstation Volume Share 7 (GE Healthcare). Those patients who had a history of surgery and liver abnormalities were excluded from the test set, as were any patients who had metal adjacent to the pancreas or visible artifacts on the scans. This dataset was not used during the training or tuning stages.

2.3. AI System - CTNet

The prediction system is dubbed CTNet. It is designed to map a 3D CT scan to a probability estimate that indicates if pancreatic cancer is present or not. CTNet closely resembles systems in literature that use ImageNet pre-trained convolutional neural networks (CNNs) on radiology scans (31, 34, 43–47). The model architecture is shown in **Figure 2**.

Given a total set of s slices in a scan, where each individual slice t is a 299 × 299 image, an ImageNet pre-trained Inception v4 CNN was used to extract an embedding $\mathbf{h}_t \in \mathbb{R}^d$ from each slice. The embeddings were extracted from the penulmitate layer, which renders a $d = 1,536$ dimensional feature vector for each image (48). Because Inception v4 is designed to take as input a 299 × 299 × 3 RGB image, we replicated each slice to create faux RGB images. Following others (49), the CNN was not fine-tuned for CT data.

After extracting the embeddings from all scan slices within a CT scan volume, it is then fed into a neural network to make a final prediction, which is given by:

$$P\left(Cancer = 1|\mathbf{h}_1, \mathbf{h}_2, \ldots \mathbf{h}_s\right) = \sigma\left(b + \frac{1}{s}\mathbf{w}^T \sum_{t=1}^{s} \mathrm{ReLU}\left(\mathbf{U}\mathbf{h}_t + \mathbf{a}\right)\right), \quad (1)$$

where $\sigma(\cdot)$ denotes the logistic sigmoid activation function, $b \in \mathbb{R}$ is the output layer bias, $\mathbf{w} \in \mathbb{R}^{20}$ is the output layer's weight vector, $\mathbf{U} \in \mathbb{R}^{20\times1536}$ is the hidden layer weight matrix, $\mathbf{a} \in \mathbb{R}^{20}$ is the hidden layer bias, and ReLU is the rectified linear unit activation function. In preliminary studies, we found that using 20 hidden units sufficed to achieve strong performance.

The model was trained using the binary cross-entropy loss function with a mini-batch size of 1. The weights were initialized using the Kaiming method. For all systems trained in this study, we used the Adam optimizer with (50) a base learning rate of $1e^{-4}$, L_2 weight decay of $1e^{-6}$, and bias correction terms, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was reduced by a factor of 2 over the course of training when the validation loss had stopped improving. Each system was trained for 100 epochs. Since our training dataset was balanced with positive and negative cases, we did not scale the loss for any particular class's prevalence. During training, no data augmentation techniques were applied. The model was implemented in Python 3.8 with PyTorch 1.6.0 on a computer with a 12 GB NVIDIA Titan V GPU.

2.4. Scan Preprocessing

Since the voxel size varied from patient to patient, the CT scans were first resampled to an isotropic resolution of 1.0 × 1.0 × 1.0 mm using SINC interpolation. They were then resized to a height and width of 299 × 299 pixels using bilinear interpolation, which is the original input image size used to train the Inception v4 network. The voxel Hounsfield unit (HU) value was clipped to be between $\pm300HU$ and normalized to have zero mean and unit variance (i.e., $[0, 1]$). Normalization was performed by subtracting the mean and dividing by the standard deviation computed from the training dataset. This processing was applied to both the development and generalization datasets.

2.5. Noise Image Generation

We derived noise images from the actual scans within each class to determine if the institutional scanning practices or noise characteristics of the imaging systems confound the classification results. As a result, they are composed of unrecognizable or hidden patterns that are a byproduct of the scanner image processing schemes or X-ray detection characteristics. A key attribute of the noise image is that it must be uniform and devoid of any perceptible patterns or structured anatomy. We generated noise images from each patient's CT scan using an approach similar to (51, 52), and as shown in **Figure 4**. For a scan with s sequential slices, where each slice t is an image $I_t \in \mathbb{R}^{299\times299}$, we subtract adjacent slices to produce $s - 1$ difference images D_I, where $D_I = I_t - I_{t-1}$ and $1 \leq t < s$. The subtraction process eliminates most of the anatomical features seen in the scan. We then apply a Sobel edge enhancing filter to each D_I to identify and remove any remaining anatomical patterns. Then we loop through each D_I

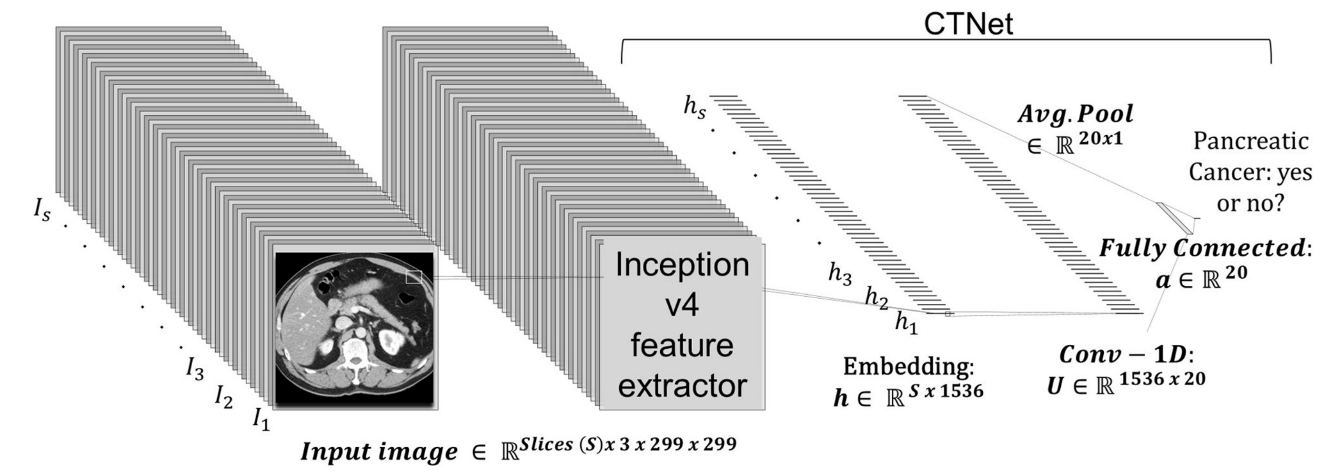

CTNet
h_s
$Avg. Pool$ $\in \mathbb{R}^{20x1}$
Pancreatic Cancer: yes or no?
I_s
$Fully\ Connected$: $a \in \mathbb{R}^{20}$
Inception v4 feature extractor
h_3
h_2
h_1
I_3
I_2
I_1
Embedding: $h \in \mathbb{R}^{S\,x\,1536}$
$Conv - 1D$: $U \in \mathbb{R}^{1536\,x\,20}$
$Input\ image \in \mathbb{R}^{Slices\ (S) x\, 3\, x\, 299\, x\, 299}$

FIGURE 2 | CTNet architecture. CTNet takes as input a volumetric CT scan and outputs a classification prediction. Features are extracted from each slice of the CT scan by the Inception v4 network. The output feature vector is then reduced in dimension with a single convolutional layer, followed by an adaptive average pooling operation applied over the number slices. The resulting vector is fed into a fully connected layer, which has a single output.

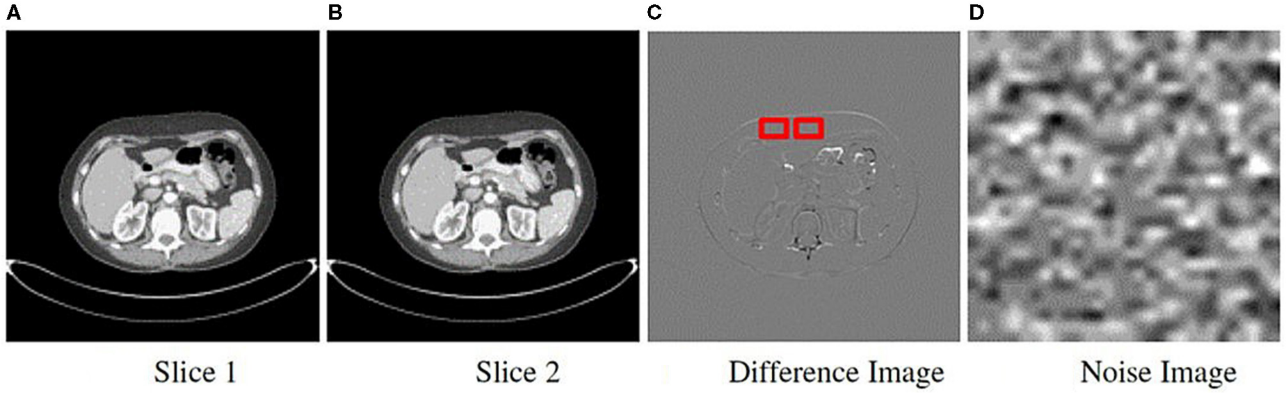


FIGURE 3 | Method to obtain noise maps from sequentially acquired images. Two sequential slice images are subtracted from each other **(A,B)**. **(C)** Difference image resulting from subtraction showing the sliding 30 × 30 pixel window used to extract uniform patches. **(D)** All patches were averaged to generate a single noise image.

to extract non-overlapping patches of size 30 × 30 pixels. The patch size was selected to minimize the impact of the non-uniformity of the CT HU values within the region of interest (e.g., due to streaking or beam hardening artifacts) (52). However, patches of transitional boundary areas (i.e., interface between different tissue types) consisted of discernible patterns that could be spuriously correlated with the class labels. Consequently, to identify and exclude boundary patches, we generated and analyzed each patch's histogram. First-order statistical measures, such as skewness, kurtosis, and standard deviation, and the number of peaks within the histogram were used to identify and exclude boundary patches. Histograms with a skewness value within ±0.1, kurtosis of 3.0 ± 0.5, a standard deviation less than 16, and those with a single peak were included. Published descriptions of the noise image generation method do not provide choices for each of the parameters, so we chose them via visual inspection to eliminate transition areas or edges. The patches that met the criteria were then averaged together to create a single noise image representation of size 30 × 30 for the D_I, as shown in **Figure 3**. Finally, the $s - 1$ noise images for the patient were upsampled using SINC interpolation to a dimension of 299 × 299.

2.6. Experiments

To employ the sanity tests, we processed four representations or input formats of the same training, tuning, and held out development test sets. Representative images for a single patient from the cancer positive class are shown in **Figure 1**. The first format we evaluated were the pancreas-only scans, as shown in **Figure 1C**. We used the provided annotations to exclude the organs surrounding the pancreas and pancreas with tumor for this format. The cropped portion shown in **Figure 1C** was resized to an input dimension of 299 × 299 before being fed into CTNet. The second format, referred to as the original with the pancreas (WP) scans and shown in **Figure 1A**, consisted of the uncropped patient CT scans where the normal or abnormal pancreas was present. For the third format shown in **Figure 1B**, the pixels that composed the normal or abnormal pancreas were replaced with zeros. These are referred to as the original without a pancreas (WOP) scans. The fourth format, shown in **Figure 1D**, consisted of the noise images. We trained four systems, one for each input format, and tested each of them with the held-out test sets of the other formats. Since annotations were not available for the generalization test set, we generated two formats: (1) the original uncropped images, which are referred to as DECT original WP, and (2) the noise images, which are referred to as DECT noise scans. We performed stratified five-fold cross-validation with the same division of scans across the four systems. For this study, we consider the baseline against which all results are compared to be the system trained with the pancreas-only scans shown in **Figure 1C**, as it should be the representation that maximizes the classification performance.

2.7. Statistical Analysis

Each system's classification performance was assessed using the area under the receiver operator characteristic curve (AUC). We report the average AUC and corresponding 95% confidence interval (CI) across cross-validation runs. An average AUC score of 1.0 represents perfect classification performance. The average AUC across runs and the corresponding confidence intervals were determined using R (Rstudio version 3.6.2) with the package cvAUC for cross-validated AUC (53). In addition to confidence intervals, statistically significant differences between test runs was confirmed with the DeLong test statistic for AUCs (54). The level of significance was set at $P \leq 0.05$.

3. RESULTS

Table 2 provides an overview of how the sanity tests should be interpreted and implemented in practice. **Figure 4** shows the performance of each trained system on the held-out tests and the generalization test set. The diagonal elements for the development tests correspond to training and testing on the same input format (i.e., self-tests), while the others represent AUC scores from training on one format and testing on the other (i.e., non-self tests). We expect a system trained on one format to perform the best on test data processed in an identical manner, which is consistent with the self-test results along the diagonal of **Figure 4**. For instance, the system trained with the pancreas-only images achieved an AUC of 0.82 (95% CI: 0.73–0.92) on its self-test format. If the system was considered to pass the sanity tests, we would expect it to have the highest AUC across self-test results and the original WP test format. However, instead, it is the lowest among the self-tests. Its performance is significantly lower ($P < 0.001$) than systems trained on the original WP and WOP, 0.95 AUC (95% CI: 0.89–1.0) and 0.97 AUC (95% CI: 0.93–1.0), respectively.

The second and third rows of **Figure 4** show the performance of the systems trained with the original WP and WOP formats. Both systems performed exceptionally well on their self-test sets and each other's test format, but they both saw a drop in performance when they were evaluated with the pancreas-only and noise image test formats ($P < 0.001$). The performance across test formats suggests that the confounding variables within the development data are readily associated with the image-level labels. Several sources of bias may be responsible for the observed results, such as the differences in scan parameters, types of scanners, choice of reconstruction algorithms, and the distribution of contrast within the pancreas.

The noise-only system achieved the highest AUC of 0.98 (95% CI: 0.96, 1.0) among the self-test sets, suggesting that discriminative but unrecognizable features unique to each dataset were used to distinguish each class. The results observed on the development data are confirmed on the generalization tests where each system's performance is significantly lower than on the development data self-tests ($P < 0.001$). For example, the average AUC achieved by the pancreas-only system on the self-test was 0.82 (95% CI: 0.73, 0.92), but its performance was significantly lower on the DECT original WP generalization test format ($P < 0.001$). One reason for the reduced performance on the generalization dataset is the difference in scan parameters with the development datasets. The DECT scans are synthesized images that depict anatomy from the viewpoint of a monochromatic X-ray source. However, within the clinic, AI systems may be tasked with assessing scans acquired on any type of CT system.

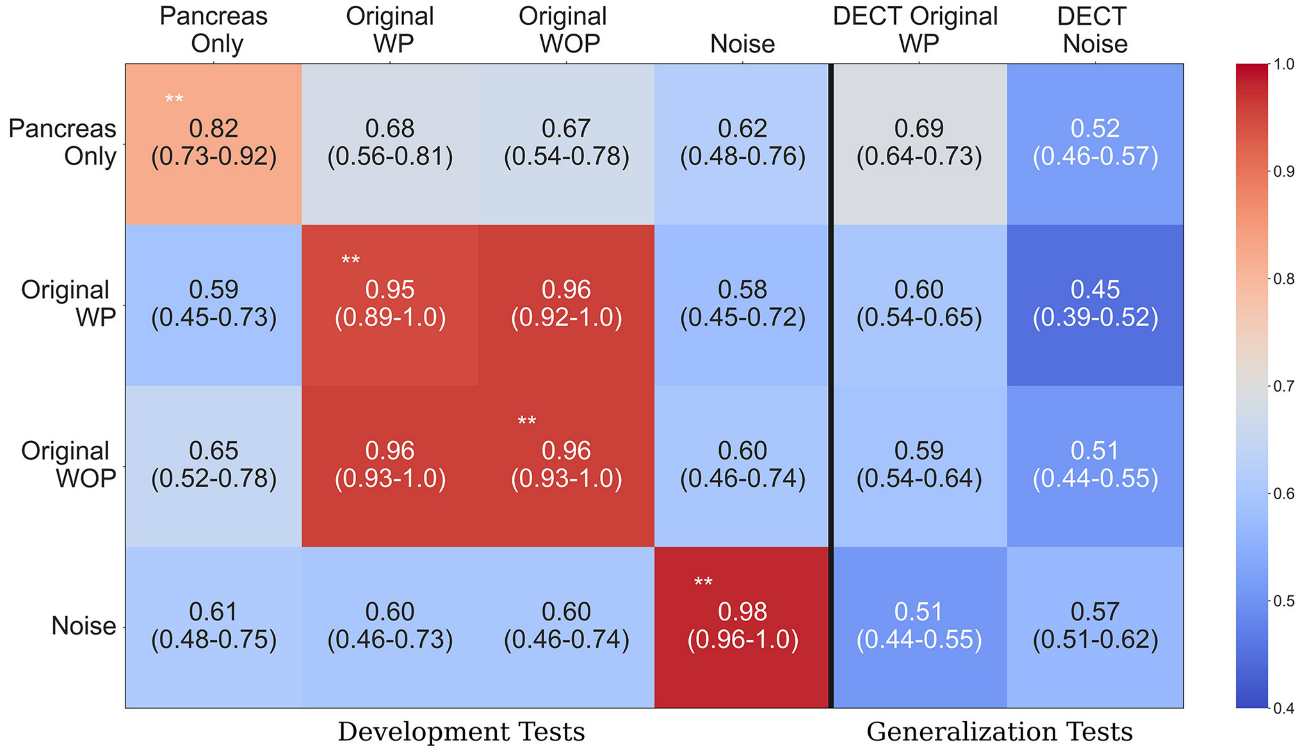
4. DISCUSSION

Identifying covariates that cause unintended generalization or those that cause machines to fail unexpectedly in deployment remains a challenge across deep learning applications. We described sanity tests that could reveal if covariates drive classification decision-making and tested them with a case study designed to classify pancreatic cancer from CT scans. Failing these sanity tests provides an early indicator of potential biases being responsible for the observed performance and that further in the development process, a system will unintentionally generalize or have much lower performance when deployed. We argue that others should routinely use these tests in publications. For industry, these tests could save time and money. Failing them indicates that the target objects' attributes are not being used by the systems undergoing analytical and clinical validation studies. Hence, as we show, relying only on conventional testing strategies with development data will not provide adequate assurances of generalization. Our sanity tests can be used with development data as long as ROIs are available or a background noise image can be generated. While we focused on binary classification, the sanity tests apply to the multi-class classification and regression problems, with appropriate statistical analysis modifications.

The proposed sanity tests are designed to identify early when an AI system reaches the correct classification for the wrong reasons, but they are not designed to identify the reason for the incorrect decision. As far as what those reasons are, it may not be possible to tease them apart given the limitations of public datasets where private and non-private meta-data are removed. For instance, both NIH (40) and MSD (41) reports did not indicate if iterative reconstruction (IR) was used, the size or

TABLE 2 | The proposed sanity tests to assess the reliability of medical AI systems.

Sanity test	Implications of failing the test	Does CTNet pass the test?
Train and test with and without the target: The system should achieve an AUC of around 0.5 when tested without the target in test images.	Images contain spurious covariates that can be exploited by the model.	✗
Train and test using noise images: The system should achieve an AUC of around 0.5 on test data.	Classification performance cannot be attributed to recognition of the target (i.e., covariates contribute to the learned classification decision rule).	✗
Test system with different sized ROIs: The additional or reduced context should not alter the performance.	The system cannot decorrelate features of the target from its co-occurring context [i.e., Contextual Bias (55)].	✗



	Pancreas Only	Original WP	Original WOP	Noise	DECT Original WP	DECT Noise
Pancreas Only	** 0.82 (0.73-0.92)	0.68 (0.56-0.81)	0.67 (0.54-0.78)	0.62 (0.48-0.76)	0.69 (0.64-0.73)	0.52 (0.46-0.57)
Original WP	0.59 (0.45-0.73)	** 0.95 (0.89-1.0)	0.96 (0.92-1.0)	0.58 (0.45-0.72)	0.60 (0.54-0.65)	0.45 (0.39-0.52)
Original WOP	0.65 (0.52-0.78)	0.96 (0.93-1.0)	** 0.96 (0.93-1.0)	0.60 (0.46-0.74)	0.59 (0.54-0.64)	0.51 (0.44-0.55)
Noise	0.61 (0.48-0.75)	0.60 (0.46-0.73)	0.60 (0.46-0.74)	** 0.98 (0.96-1.0)	0.51 (0.44-0.55)	0.57 (0.51-0.62)

FIGURE 4 | Area under the curve (AUC) heatmap across models for each input format type. Each row indicates the cross-validated mean AUC with 95% confidence intervals for the systems trained with a given input format and evaluated with all other formats from the development dataset (first four columns). The last two columns show the performance of each system on the generalization dataset. The diagonal elements on the development tests correspond to training and testing with the same input format. Red indicates the highest AUC values, while light blue indicates the lowest AUC values. The non-significant difference on the original with pancreas (WP) and without pancreas (WOP) development test sets indicates that spurious correlations drive the performance observed on the self-test sets, instead of features specific to the pancreas or pancreatic cancer. **Development test images processed identically to the data used for training that model. WP, With pancreas; WOP, Without pancreas; DECT, Dual Energy CT.

number of data channels used to acquire images, and the Bowtie filter. These parameters were also not present in the DICOM meta-data. Modern CT scanners often use varying strengths of IR to suppress image noise, but with the application of IR, images appear smoother, and depending on the strength, the noise texture becomes finer or more coarse (56–58). Another source of bias is the quality of the annotations and accuracy of information released about a public dataset (59). For example, Suman et al. found that parts of the pancreas were absent in the provided segmentation's for the NIH pancreas-CT dataset (59). In addition, we observed a discrepancy between the reported slice thickness from the MSD dataset (41) and existing information in the DICOM meta-data. The report indicates that all scans were acquired using a slice thickness and reconstruction width of 2.5, but the information derived from the DICOM meta-data shows that the slice thickness's for some scans was: 0.70, 1.25, 1.5, 2.0, 2.5, 3.75, 4.0, 5, and 7.5 mm. A common slice thickness could prevent resampling errors (e.g., aliasing) that may arise from down or up-sampling the CT scans. The discrepancy and missing meta-data motivates the need for data-reporting standards and standardized study designs with more rigorous validation procedures.

We did not attempt to use techniques to mitigate the impact of spurious correlations. These include adversarial regularization (19, 60, 61), model ensembling (62, 63), invariant risk minimization (64, 65) and methods that encourage grounding on causal factors instead of spurious correlations (66–69). However, as shown by Shrestha et al. (70), methods that were thought to overcome spurious correlations were behaving as regularizers instead of overcoming the issues that stemmed from the covariates. Our sanity tests could be used with these mitigation methods to measure their true impact, in that we would expect them to only be able to provide significant benefit when the target is present.

There are some limitations to this study. As with most AI studies involving medical image analysis, we trained and tested with a small dataset and did not account for spectral or disease prevalence biases within development or generalization data. Results stemming from small-data may not always transfer to scenarios where larger datasets are used to train systems, but this is in part why sanity tests when using smaller datasets are critical since it is likely easier for spurious correlations to impact them. Also, since the goals of this study were to define sanity tests and illustrate their application, we did not investigate the reasons behind the reduced performance on the generalization data. However, the reduced performance could be a byproduct of the divergent scan parameters and difference in scan type (i.e., SECT vs. DECT and scan phase) between the development and generalization datasets. In general, our sanity tests help reveal when an AI model predicts the right answer for the wrong reasons and will therefore have a large gap between development and external generalization tests. A complementary approach uses visualization methods to understand if a system is not looking at the target to perform its classification.

In conclusion, we demonstrated how our proposed sanity tests could identify spurious confounds early, using development data solely. While the methods are simple, we argue that sanity tests similar to these should be performed wherever possible, especially with smaller datasets, and if no external dataset is available. Otherwise, study results can be very misleading and fail to generalize on other datasets. In safety-critical AI domains, such as healthcare, sanity tests could prevent harm to patients, and they could better prepare novel medical AI systems for regulatory approval. We present a workflow and practical sanity tests that can reliably reveal error-prone systems before influencing real-world decision-making.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because Institutional policy and privacy laws. Requests to access the datasets should be directed to mahmoodu@mskcc.org.

AUTHOR CONTRIBUTIONS

UM and CK conceived the study. UM implemented the algorithms and carried out the experiments. UM, CK, and RS wrote the paper. DB, LM, GC, and YE helped gather the data, provided the advice, and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded in part through the National Institutes of Health/National Cancer Institute Cancer Center support grant P30 CA008748 as well as by NSF award #1909696.

REFERENCES

1. Jin D, Harrison AP, Zhang L, Yan K, Wang Y, Cai J, et al. Chapter 14 - Artificial intelligence in radiology. In: Xing L, Giger ML, Min JK, editors. *Artificial Intelligence in Medicine*. Cambridge, MA: Academic Press (2021). p. 265–89. Available online at: https://www.sciencedirect.com/science/article/pii/B9780128212592000144. doi: 10.1016/B978-0-12-821259-2.00014-4
2. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. *Brit J Radiol*. (2020) 93:20190855. doi: 10.1259/bjr.20190855
3. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*. (2019) 292:60–6. doi: 10.1148/radiol.2019182716
4. Antonelli M, Johnston EW, Dikaios N, Cheung KK, Sidhu HS, Appayya MB, et al. Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. *Eur Radiol*. (2019) 29:4754–64. doi: 10.1007/s00330-019-06244-2
5. Voter A, Larson M, Garrett J, Yu JP. Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of cervical spine fractures. *Am J Neuroradiol*. (2021). doi: 10.3174/ajnr.A7179
6. Laghi A. Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence. *Lancet Digital Health*. (2020) 2:e225. doi: 10.1016/S2589-7500(20)30079-0
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x
8. The Lancet. Is digital medicine different? *Lancet*. (2018). 392:95. doi: 10.1016/S0140-6736(18)31562-9
9. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers–from the radiology editorial board. *Radiology*. (2020) 294:487–9. doi: 10.1148/radiol.2019192515
10. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology*. (2019) 290:590–606. doi: 10.1148/radiol.2018180547
11. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol*. (2019) 20:405–10. doi: 10.3348/kjr.2019.0025
12. El Naqa I, Ruan D, Valdes G, Dekker A, McNutt T, Ge Y, et al. Machine learning and modeling: data, validation, communication challenges. *Med Phys*. (2018) 45:e834–40. doi: 10.1002/mp.12811
13. Recht MP, Dewey M, Dreyer K, Langlotz C, Niessen W, Prainsack B, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur Radiol*. (2020) 30:3576–84. doi: 10.1007/s00330-020-06672-5
14. Parmar C, Barry JD, Hosny A, Quackenbush J, Aerts HJ. Data analysis strategies in medical imaging. *Clin Cancer Res*. (2018) 24:3492–9. doi: 10.1158/1078-0432.CCR-18-0385
15. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. *Nat Mach Intell*. (2020) 2:665–73. doi: 10.1038/s42256-020-00257-z

16. 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. *PLoS Med*. (2018) 15:e1002683. doi: 10.1371/journal.pmed.1002683
17. Jo J, Bengio Y. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*. (2017).
18. Kafle K, Shrestha R, Kanan C. Challenges and prospects in vision and language research. *Front Artif Intell*. (2019) 2:28. doi: 10.3389/frai.2019.00028
19. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A. Adversarial examples are not bugs, they are features. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R, editors. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. NeurIPS*. Vancouver, BC (2019). p. 125–36. Available online at: https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html
20. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: *7th International Conference on Learning Representations, ICLR 2019*. New Orleans, LA (2019). Available online at: https://openreview.net/forum?id=Bygh9j09KX
21. Baker N, Lu H, Erlikhman G, Kellman PJ. Deep convolutional networks do not classify based on global object shape. *PLoS Comput Biol*. (2018) 14:e1006613. doi: 10.1371/journal.pcbi.1006613
22. Sinz FH, Pitkow X, Reimer J, Bethge M, Tolias AS. Engineering a less artificial intelligence. *Neuron*. (2019) 103:967–79. doi: 10.1016/j.neuron.2019.08.034
23. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, Tengg-Kobligk HV, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol Artif Intell*. (2020) 2:e190043. doi: 10.1148/ryai.2020190043
24. Adebayo J, Gilmer J, Muelly M, Goodfellow IJ, Hardt M, Kim B. Sanity checks for saliency maps. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems. NeurIPS 2018*. Montréal, QC (2018). p. 9525–36. Available online at: https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html
25. Kim B, Wattenberg M, Gilmer J, Cai CJ, Wexler J, Viégas FB, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Dy JG, Krause A, editors. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*. Stockholm: PMLR (2018). p. 2673–82. Available online at: http://proceedings.mlr.press/v80/kim18d.html
26. Ghorbani A, Abid A, Zou JY. Interpretation of neural networks is fragile. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI. The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019. The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. Honolulu, HI: AAAI Press (2019). p. 3681–8. doi: 10.1609/aaai.v33i01.33013681
27. Lakkaraju H, Bastani O. "How do I fool you?" Manipulating user trust via misleading black box explanations. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY (2020). p. 79–85. doi: 10.1145/3375627.3375833
28. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. (2018) 286:800–9. doi: 10.1148/radiol.2017171920
29. Teney D, Kafle K, Shrestha R, Abbasnejad E, Kanan C, van den Hengel A. On the value of out-of-distribution testing: an example of Goodhart's law. In: *Neural Information Processing Systems (NeurIPS)*. (2020).
30. Gupta V, Saxena V. Software testing: smoke and sanity. *Int J Eng Res Technol*. (2013) 2:1674–8.
31. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol*. (2019) 155:1135–41. doi: 10.1001/jamadermatol.2019.1735
32. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. New York, NY (2020). p. 151–9. doi: 10.1145/3368555.3384468
33. Shamir L. Evaluation of face datasets as tools for assessing the performance of face recognition methods. *Int J Comput Vis*. (2008) 79:225. doi: 10.1007/s11263-008-0143-7
34. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056
35. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. *Radiology*. (2020) 295:4–15. doi: 10.1148/radiol.2020192224
36. Petrick N, Sahiner B, Armato SG III, Bert A, Correale L, Delsanto S, et al. Evaluation of computer-aided detection and diagnosis systems A. *Med Phys*. (2013) 40:087001. doi: 10.1118/1.4816310
37. US Food and Drug Administration. *Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data–Premarket Approval (PMA) and Premarket Notification [510 (k)] Submissions*. Silver Spring, MD (2020).
38. Galvin A, Sutherland T, Little AF. Part 1: CT characterisation of pancreatic neoplasms: a pictorial essay. *Insights Imaging*. (2011) 2:379–88. doi: 10.1007/s13244-011-0102-7
39. Price WN II. Medical AI and contextual bias. *Harvard J Law Technol*. U of Michigan Public Law Research Paper No. 632 (2019) 33. Available online at: https://ssrn.com/abstract=3347890
40. Roth HR, Farag A, Turkbey E, Lu L, Liu J, Summers RM. *Data from Pancreas-CT*. The Cancer Imaging Archive (2016).
41. Simpson AL, Antonelli M, Bakas S, Bilello M, Farahani K, van Ginneken B, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint*. (2019) arXiv:1902.09063.
42. Hsieh J. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. vol. 114. SPIE Press (2003).
43. Draelos RL, Dov D, Mazurowski MA, Lo JY, Henao R, Rubin GD, et al. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Med Image Anal*. (2021) 67:101857. doi: 10.1016/j.media.2020.101857
44. Raghu M, Zhang C, Kleinberg JM, Bengio S. Transfusion: understanding transfer learning for medical imaging. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R, editors. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. NeurIPS 2019*. Vancouver, BC (2019). p. 3342–52.
45. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med*. (2018) 15:e1002699. doi: 10.1371/journal.pmed.1002699
46. Liu J, Wang D, Lu L, Wei Z, Kim L, Turkbey EB, et al. Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. *Med Phys*. (2017) 44:4630–42. doi: 10.1002/mp.12399
47. Paul R, Hawkins SH, Balagurunathan Y, Schabath MB, Gillies RJ, Hall LO, et al. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography*. (2016) 2:388. doi: 10.18383/j.tom.2016.00211
48. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Singh SP, Markovitch S, editors. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco, CA: AAAI Press; (2017). p. 4278–84.
49. Van Ginneken B, Setio AA, Jacobs C, Ciompi F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. Brooklyn, NY: IEEE (2015). p. 286–9. doi: 10.1109/ISBI.2015.7163869
50. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y, editors. In: *3rd International Conference on Learning Representations, ICLR 2015*. San Diego, CA (2015).
51. Christianson O, Winslow J, Frush DP, Samei E. Automated technique to measure noise in clinical CT examinations. *Am J Roentgenol*. (2015) 205:W93–9. doi: 10.2214/AJR.14.13613

52. Tian X, Samei E. Accurate assessment and prediction of noise in clinical CT images. *Med Phys*. (2016) 43:475–82. doi: 10.1118/1.4938588
53. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron J Stat*. (2015) 9:1583. doi: 10.1214/15-EJS1035
54. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. (1988) 44:837–45. doi: 10.2307/2531595
55. Singh KK, Mahajan D, Grauman K, Lee YJ, Feiszli M, Ghadiyaram D. Don't judge an object by its context: learning to overcome contextual bias. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. Seattle, WA: IEEE (2020). p. 11067–75. doi: 10.1109/CVPR42600.2020.01108
56. Barrett HH. Is there a role for image science in the brave new world of artificial intelligence? *J Med Imaging*. (2019) 7:012702. doi: 10.1117/1.JMI.7.1.012702
57. Solomon JB, Christianson O, Samei E. Quantitative comparison of noise texture across CT scanners from different manufacturers. *Med Phys*. (2012) 39:6048–55. doi: 10.1118/1.4752209
58. Reiazi R, Abbas E, Famiyeh P, Rezaie A, Kwan JY, Patel T, et al. The impact of the variation of imaging parameters on the robustness of Computed Tomography Radiomic features: a review. *Comput Biol Med*. (2021) 133:104400. doi: 10.1016/j.compbiomed.2021.104400
59. Suman G, Patra A, Korfiatis P, Majumder S, Chari ST, Truty MJ, et al. Quality gaps in public pancreas imaging datasets: implications & challenges for AI applications. *Pancreatology*. (2021). doi: 10.1016/j.pan.2021.03.016
60. Ramakrishnan S, Agrawal A, Lee S. Overcoming language priors in visual question answering with adversarial regularization. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems. NeurIPS 2018*. Montréal, QC (2018). p. 1548–58.
61. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New Orleans, LA (2018). p. 335–40. doi: 10.1145/3278721.3278779
62. Cadéne R, Dancette C, Ben-younes H, Cord M, Parikh D. RUBi: reducing unimodal biases for visual question answering. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R, editors. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. NeurIPS 2019*. Vancouver, BC (2019). p. 839–50.
63. Clark C, Yatskar M, Zettlemoyer L. Don't take the easy way out: ensemble based methods for avoiding known dataset biases. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics (2019). p. 4069–82. doi: 10.18653/v1/D19-1418
64. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*. (2019).
65. Choe YJ, Ham J, Park K. An empirical study of invariant risk minimization. *arXiv preprint arXiv:2004.05007*. (2020).
66. Selvaraju RR, Lee S, Shen Y, Jin H, Ghosh S, Heck LP, et al. Taking a HINT: leveraging explanations to make vision and language models more grounded. In: *2019 IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE (2019). p. 2591–600. doi: 10.1109/ICCV.2019.00268
67. Qi J, Niu Y, Huang J, Zhang H. Two causal principles for improving visual dialog. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA: IEEE (2020). p. 10857–66. doi: 10.1109/CVPR42600.2020.01087
68. Agarwal V, Shetty R, Fritz M. Towards Causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA: IEEE (2020). p. 9687–95. doi: 10.1109/CVPR42600.2020.00971
69. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. *Nat Commun*. (2020) 11:1–10. doi: 10.1038/s41467-020-17478-w
70. Shrestha R, Kafle K, Kanan C. A negative case analysis of visual grounding methods for VQA. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics*. Stroudsburg, PA (2020). p. 8172–81. doi: 10.18653/v1/2020.acl-main.727

Conflict of Interest: CK was employed at Paige, a commercial company, during the preparation of this manuscript. This company played no role in the sponsorship, design, data collection and analysis, decision to publish, or preparation of the manuscript.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.