Radiology

Chest Radiographs in Congestive Heart Failure: Visualizing Neural Network Learning

Jarrel C. Y. Seah, MBBS (Hons), BMedSci (Hons) • Jennifer S. N. Tang, MBBS (Hons) • Andy Kitchen, BSc • Frank Gaillard, MBBS (Hons), MMed, FRANZCR • Andrew F. Dixon, MBBS (Hons), FRANZCR

From the Department of Radiology, Alfred Hospital, 55 Commercial Rd, Melbourne, Victoria 3004, Australia (J.C.Y.S., A.F.D.); Department of Radiology, Royal Melbourne Hospital, Melbourne, Australia (J.S.N.T., F.G.); Melbourne, Australia (A.K.); and Department of Radiology, Melbourne University, Melbourne, Australia (F.G.). Received April 17, 2018; revision requested May 22; final revision received September 3; accepted September 7. **Address correspondence to** J.C.Y.S. (e-mail: *J.Seah@alfred.org.au*).

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

See also the editorial by Ngo in this issue.

Radiology 2019; 290:514–522 • https://doi.org/10.1148/radiol.2018180887 • Content codes: CA CH IN

Purpose: To examine Generative Visual Rationales (GVRs) as a tool for visualizing neural network learning of chest radiograph features in congestive heart failure (CHF).

Materials and Methods: A total of 103 489 frontal chest radiographs in 46 712 patients acquired from January 1, 2007, to December 31, 2016, were divided into a labeled data set (with B-type natriuretic peptide [BNP] result as a marker of CHF) and unlabeled data set (without BNP result). A generative model was trained on the unlabeled data set, and a neural network was trained on the encoded representations of the labeled data set to estimate BNP. The model was used to visualize how a radiograph with high estimated BNP would look without disease (a "healthy" radiograph). An overfitted model was developed for comparison, and 100 GVRs were blindly assessed by two experts for features of CHF. Area under the receiver operating characteristic curve (AUC), κ coefficient, and mixed-effects logistic regression were used for statistical analyses.

Results: At a cutoff BNP of 100 ng/L as a marker of CHF, the correctly trained model achieved an AUC of 0.82. Assessment of GVRs revealed that the correctly trained model highlighted conventional radiographic features of CHF as reasons for an elevated BNP prediction more frequently than the overfitted model, including cardiomegaly (153 [76.5%] of 200 vs 64 [32%] of 200, respectively; $P < .001$) and pleural effusions (47 [23.5%] of 200 vs 16 [8%] of 200, respectively; $P = .003$).

Conclusion: Features of congestive heart failure on chest radiographs learned by neural networks can be identified using Generative Visual Rationales, enabling detection of bias and overfitted models.

Online supplemental material is available for this article.

Deep learning is increasingly used in medical imaging; however, advancement in efficacy has outpaced the ability to visualize what these models are actually learning (1). Difficulty in isolating learned image features currently limits the safe application of deep learning in radiology.

This study examined a technique developed by the authors using generative learning (2), a form of unsupervised deep learning, to create Generative Visual Rationales (GVRs). GVRs are a visual output that displays features used to classify an image. Unlike existing interpretability techniques, GVRs use Wasserstein generative adversarial networks (GANs) (2,3) to synthesize visual reconstructions that answer the question "How would this patient's image need to change to appear without the disease?"

The example of congestive heart failure (CHF) prediction from chest radiographs was chosen by the authors to test the utility of GVRs. Conventional radiographic features of CHF include cardiomegaly, pulmonary venous congestion, septal lines, airspace opacification, and pleural effusions (4). In other chest diseases such as lung cancer, features are often localized to one part of the image, and hence conventional visualization methods such as heat maps can discern what the model is focusing on. However, in CHF, features tend to affect large parts of the image, making heat maps not as effective, while GVRs can demonstrate how the image as a whole contributes to the prediction of disease.

Serum B-type natriuretic peptide (BNP) is secreted by the heart to regulate fluid balance and is commonly used to diagnose and monitor CHF. At the threshold of 100 ng/L, serum BNP has a sensitivity of 0.95 and specificity of 0.63 for acute CHF (5). Thus, at this cutoff, BNP can reliably exclude acute CHF independent of clinical gestalt and radiologic interpretation (6).

Using BNP levels instead of radiology reports as labels enables the training of a deep learning model free of human bias. It allows the comparison of features that the neural network model has learned de novo with features of CHF that radiologists have traditionally identified.

Our study assessed chest radiograph GVRs that the neural network model estimated to have a BNP above 100 ng/L to determine the frequency at which CHF features are highlighted, with the primary hypothesis that a correctly trained deep learning model will

Abbreviations

AUC = area under the ROC curve, BNP = B-type natriuretic peptide, CHF = congestive heart failure, GAN = generative adversarial network, GVR = Generative Visual Rationale, PACS = picture archiving and communication system, ROC = receiver operating characteristic

Summary

Generative Visual Rationales can identify imaging features learned by a model trained to predict congestive heart failure from chest radiographs, allowing radiologists to better identify faults and biases.

Implication for Patient Care

Generative Visual Rationales can identify image features learned by neural networks when estimating their degree of heart failure; radiologists can then examine these features to uncover hidden biases, enabling the safer application of deep learning to imaging studies.

produce GVRs that do so more frequently than an intentionally overfitted model.

Materials and Methods

Data Set and Image Handling

Approval was obtained from the relevant research ethics committee for use of a de-identified data set, without the requirement for patient consent. All available frontal chest radiographs (supine or erect) were extracted from the authors' institution's picture archiving and communication system (PACS) from January 1, 2007, to December 31, 2016, with 103 489 images obtained in 46 712 unique patients. All extracted radiographs were included in the study data set, with no exclusion criteria applied. No financial or material support was obtained for this project, and no conflicts of interest are known to the authors. Although the same data set was used previously (7) to explain the technical parameters of the GVR algorithm, no statistical analysis was performed, and hence no data in the 46 712 patients were reported. In contrast to earlier work (7), this study conducted an experiment using GVRs to identify biases and overfitting in a model by quantifying and statistically analyzing image features that neural networks learn.

The REASON Cohort Discovery Tool (8) identified 7390 radiographs from 5232 unique patients with a paired BNP result within 36 hours of image acquisition, comprising the labeled data set. A total of 96 099 radiographs did not have a corresponding BNP result and comprised the unlabeled data set. Radiographs were scaled to 128 × 128 pixels. The labeled data set was split by patient, with data in 4185 (80%) of 5232 patients used for training and data in 1047 (20%) of 5232 used for testing.

Deep Learning Model and GVR Creation

Model training used Python 3.4 (Python Software Foundation), PyTorch 0.2.0 (9), and NumPy 1.13.1 (10). Model construction was performed by J.C.Y.S., a 1st-year radiology resident with 6 years of machine learning experience. First, a GAN was trained on the unlabeled data set. A deep convolutional GAN architecture was used, with 64 feature maps in the layer prior to the generated image (11). This produced a generator (decoder) capable of producing artificial radiographs indistinguishable from real radiographs in the unlabeled data set.

Next, a neural network encoder was trained to reproduce latent space representations from input radiographs. The encoder and decoder together form an autoencoder (12), enabling the conversion of radiographs to and from this latent space. A latent space representation is a set of numbers that represents the salient features of that radiograph. Last, simple statistical models, including a linear regression as well as a multilayer perceptron, were trained on the latent representations of the labeled data set. A radiograph with a high predicted BNP (the original radiograph showing heart failure) was then processed by encoding it into the latent space and then decoding it to produce what we refer to as a "diseased" radiograph, which is similar but subtly different from the original radiograph, as some information is lost during this encoding and decoding process. The encoding and decoding processes are important steps for the model to learn how to recreate what we refer to as a "healthy" radiograph. We changed the latent space representation of the original radiograph such that disease (as defined by a BNP level $>$ 100 ng/L) was no longer predicted and decoded this representation into a healthy radiograph, which is what the model predicts the radiograph will look like without heart failure. It is important to note that the diseased and healthy radiographs are not actual radiographs but are synthetically generated images that indicate what the model is seeing.

The healthy radiograph was subtracted from the diseased radiograph with the difference superimposed in color (orange representing density removed and purple representing density added) over the original radiograph to produce the GVR (Fig 1). A technique to produce inverse GVRs is also shown, where predicted healthy radiographs are permuted until they appear diseased.

Qualitative Assessment of GVRs

To evaluate the usefulness of GVRs, a comparison was made between a correctly trained BNP prediction model and an overfitted model. As mentioned in the Data Set and Image Handling section, the labeled data set is split by patient into a training and testing data set at a ratio of 80:20. In the correctly trained model, latent representations from the training data set were fitted to the BNP results, and the subsequent model was evaluated on the testing set. This is standard practice in most machine learning setups to prevent the model from memorizing patient-specific features.

In the overfitted model, the training and testing data sets were deliberately combined during training, with one-fourth of the training data set withheld as a further validation set. The correctly trained model implements early stopping based on the results of the held-out testing set, while the overfitted model is unable to do so because it has access to the testing set during training and is trained to convergence.

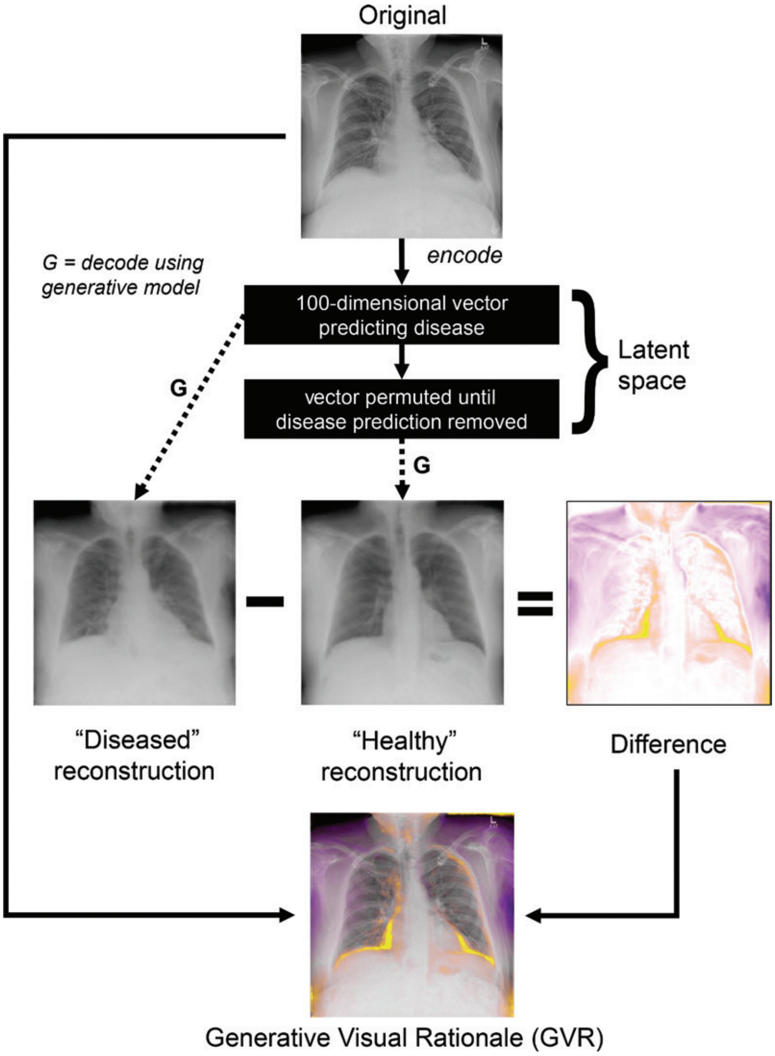
Original
encode
G = decode using generative model
100-dimensional vector predicting disease
vector permuted until disease prediction removed
Latent space
G
G
"Diseased" reconstruction
"Healthy" reconstruction
Difference
Generative Visual Rationale (GVR)

Figure 1: Generative Visual Rationale (GVR) creation process. The original image is encoded into a 100-dimensional vector and then permuted until the disease is no longer predicted. The original and permuted vectors are decoded by using the generative model into "diseased" and "healthy" reconstructions. The subtracted difference is superimposed over the original image to produce the GVR, with the removed density in orange and the added density in purple. The GVR essentially answers the question "How would this patient's image need to change to appear without the disease?"

Because the training and testing split is created to prevent the model from memorizing patient-specific features, we hypothesized that without this split, the model would choose to memorize these features rather than using disease-specific features. This enables testing of the GVR method's ability to identify the features used by a particular model, as the overfitted model should produce GVRs with less conventional radiographic features of CHF as compared with the correctly trained model. Two readers, one radiology resident (J.S.N.T., with 2 years of experience [reader A]) and one radiologist (A.F.D., with 10 years of experience [reader B]), were independently given 100 GVRs with predicted high BNPs and were blinded to the originating model and all other information about each GVR. Fifty GVRs were from the correctly trained model and 50 were from the overfitted model.

Readers were tasked with identifying whether GVRs were highlighting CHF features, including cardiomegaly, pleural effusion, and airspace opacity. Pulmonary venous congestion and septal lines were not assessed because of resolution limitations. To evaluate intraobserver consistency, each GVR appeared in a random order in two rounds, which were performed within a day of each other. Pooled results from both readers were analyzed for statistically significant differences between features identified on GVRs created by the correctly trained model and those created by the overfitted model. Pooled results from both readers were preferred to reduce the risk of observer bias and to obtain a more accurate estimate of the prevalence of each feature generated by the correctly trained and overfitted models, because different readers have different thresholds for naming features within a GVR.

Statistical Analyses

Area under the receiver operating characteristic (ROC) curve (AUC) was used to assess the predictive accuracy of the models. Intra- and interobserver agreement were assessed with Cohen κ coefficients. A mixed-effects logistic regression model was used to test the null hypothesis that the frequency of highlighted features was independent of the model (correctly trained or overfitted) from which the GVR originated, taking into account fixed effects from different readers and patients. Simple statistical models, including a fully connected multilayer perceptron, as well as a linear regression, were used to predict BNP values from the latent space representations of radiographs. Mann-Whitney *U* tests were used to analyze population demographics. $P < .05$ was considered to indicate a statistically significant difference. Statistical calculations were performed with SciKit-Learn 0.19.0 (13) and R Studio (version 3.5.1 for Windows, R Core Team, Vienna, Austria) with the lme4 package (14).

Results

Mean age, BNP levels, and sex ratios were similar between the training and testing sets (Table 1). The unlabeled set had an equivalent sex ratio but younger mean age than the labeled sets. Female patients had a mean age of 60 years (range, 0–107 years), and male patients had a mean age of 55 years (range, 0–105 years). The entire study population had a mean age of 57 years (range, 0–107 years). The average age of the female patients was significantly higher than average age of male patients ($P < .001$).

Performance of the Deep Learning Model

The neural network's predictive accuracy for BNP from chest radiographs was analyzed on the testing set without further

Table 1: Demographics of the Data Sets

Characteristic	Test Set	Training Set	Unlabeled Set
Per radiograph (103 489 total radiographs)			
Total no.	1518	5872	96 099
Median BNP (ng/L)*	350 (121–978)	374 (127–1010)	…
No. of radiographs with BNP < 100 ng/L†	301/1518 (20)	1244/5872 (21)	…
Per patient (46 712 unique patients)			
Total no.	1047	4185	41 480
Mean age ± standard deviation (y)	74 (16)	74 (16)	57 (22)
Age range (y)	14–104	17–104	0–107
No. of female patients†	492/1047 (47)	1884/4185 (45)	18 549/41 480 (45)
No. of male patients†	555/1047 (53)	2301/4185 (55)	22 931/41 480 (55)

Note.—There was no difference in age, sex ratio, or B-type natriuretic peptide (BNP) level between the testing and training sets used for the BNP prediction models ($P > .1$). The unlabeled set, which was used to train the generative model to reconstruct radiographs, had a younger mean age ($P < .001$). Each patient had one or more frontal chest radiographs in the set, leading to more radiographs than patients. Importantly, testing and training sets were separated at a patient level.

* Data in parentheses are interquartile ranges.

† Data in parentheses are percentages.

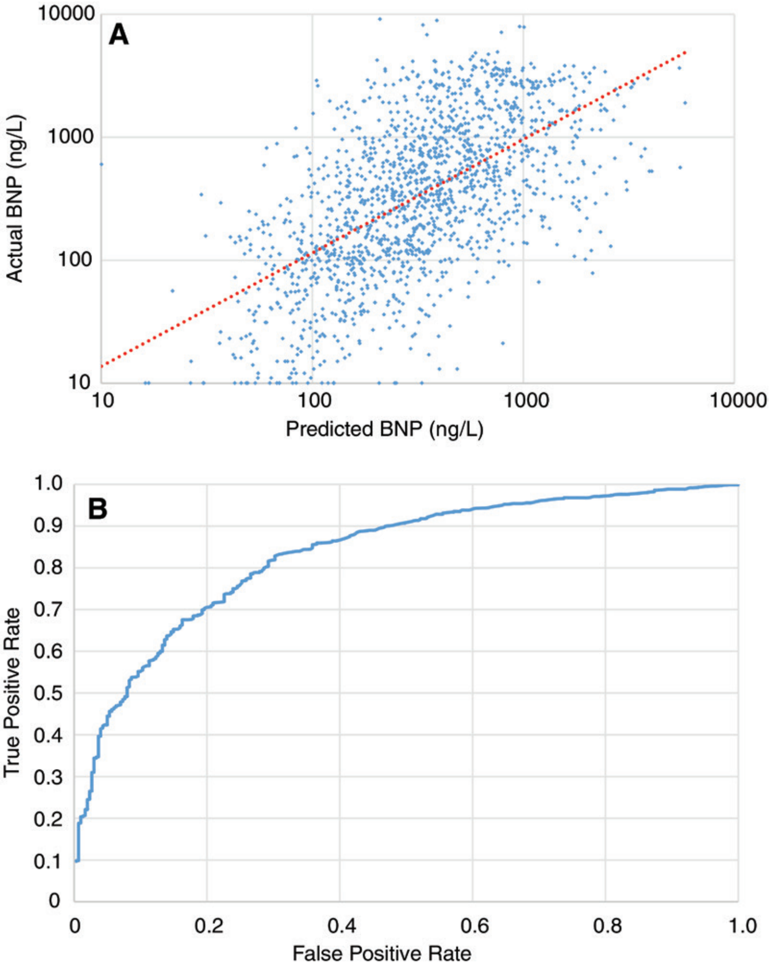

Figure 2: *A,* Scatterplot of predicted B-type natriuretic peptide *(BNP)* versus actual BNP on a logarithmic scale for the correctly trained model. *B,* Receiver operating characteristic (ROC) curve. At a cutoff of 100 ng/L, this model achieved an area under the ROC curve, or AUC, of 0.82 for predicting BNP from frontal chest radiographs.

ensembling. Images were encoded into the latent space, and a multilayer perceptron (15) was applied. At a cutoff BNP of 100 ng/L as a marker for the presence of CHF, the correctly trained model obtained a good AUC of 0.82 (Fig 2). The linear regression model applied to the latent space achieved the same AUC of 0.82. The overfitted model achieved an AUC of 0.99 because it memorized each radiograph during training. When tested on the previously unseen validation set, the overfitted model subsequently achieved an AUC of 0.75, indicating poor ability to generalize as a result of overfitting.

GVRs for Predicted "Diseased" Radiographs

A selected sample of GVRs from the correctly trained and overfitted models are shown in Figure 3. Results of the blinded GVR feature assessments are presented in Table 2. Reader A was generally less likely to assign features than reader B and had lower intraobserver agreement, particularly for airspace opacity. Interobserver agreement was highest for cardiomegaly. There was a higher frequency of cardiomegaly (153 [76.5%] of 200 vs 64 [32%] of 200; $P < .001$) and pleural effusion (47 [23.5%] of 200 vs 16 [8%] of 200; $P = .003$), highlighted by the correctly trained model compared with the overfitted model. While airspace opacity was more frequently highlighted by GVRs from the correctly trained model, this was not statistically significant (35 [17.5%] of 200 vs 12 [6%] of 200; $P = .38$). Overall, 127 (63.5%) of 200 GVRs from the overfitted model failed to highlight any features of CHF, compared with 40 (20%) of 200 for the correctly trained model ($P < .001$).

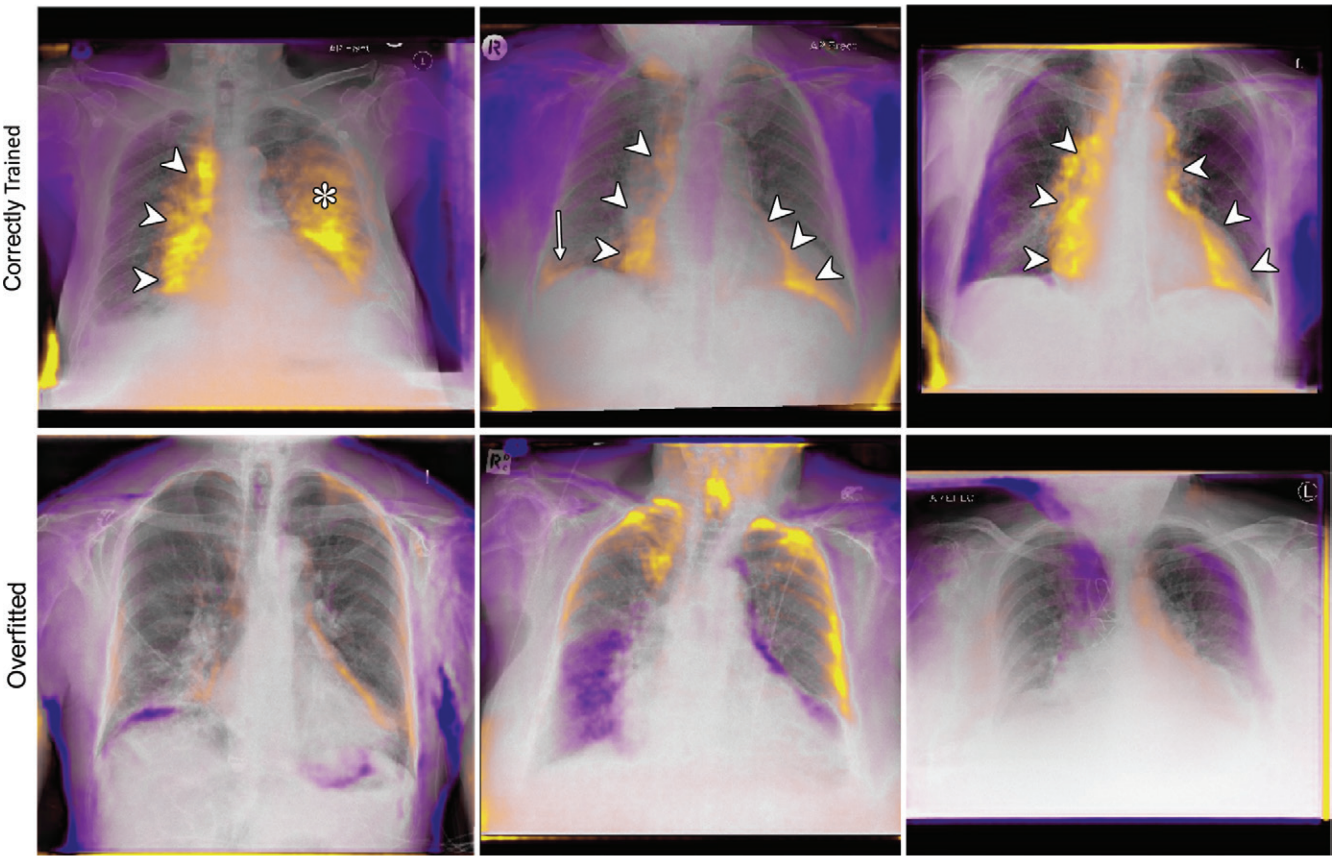

Figure 3: Chest radiographs predicted to indicate a high B-type natriuretic peptide (BNP) level. From top left to bottom right, images obtained in, respectively, an 80-year-old man, a 69-year-old woman, an 69-year-old woman, a 48-year-old man, a 68-year-old man, and an 86-year-old woman. Selected Generative Visual Rationales (GVRs) from the correctly trained model (top) and overfitted model (bottom) are shown. GVRs highlight features that the model believes need to be removed (orange) or added (purple) to remove the disease prediction. In the example of BNP prediction, the GVRs from the correctly trained model highlight congestive heart failure features, including cardiomegaly (arrowheads), pleural effusions (arrow), and airspace opacity (*). These features are statistically assessed in a larger blinded set in Table 2.

Table 2: Number of Features Identified by Blinded Readers on GVRs

	Heart Failure Features							
	Cardiomegaly		Pleural Effusion		Airspace Opacity		No CHF features	
Model and Reader	First Reading	Second Reading	First Reading	Second Reading	First Reading	Second Reading	First Reading	Second Reading
Correctly trained model								
Reader A	35/50	34/50	6/50	3/50	4/50	3/50	12/50	15/50
Reader B	41/50	43/50	19/50	19/50	14/50	14/50	5/50	8/50
Pooled result	153/200 (76.5)		47/200 (23.5)		35/200 (17.5)		40/200 (20)	
Overfitted model								
Reader A	18/50	14/50	3/50	1/50	0/50	0/50	31/50	36/50
Reader B	16/50	16/50	7/50	5/50	6/50	6/50	31/50	29/50
Pooled result	64/200 (32)		16/200 (8)		12/200 (6)		127/200 (63.5)	

Note.—Reader A was a radiology resident, and reader B was a radiologist. Each reader assessed 100 Generative Visual Rationales (GVRs)—50 from a correctly trained model and 50 from an overfitted model. This process was repeated (in reading 1 and reading 2) with GVRs displayed in random order. For instance, reader A identified 35 GVRs demonstrating cardiomegaly as a feature from the correctly trained model on their first read. This demonstrates a higher frequency of cardiac failure features highlighted by the correctly trained model. CHF = congestive heart failure. Data in parentheses are percentages. The intraobserver κ for reader A was 0.78 for cardiomegaly, 0.43 for pleural effusion, 0.26 for airspace opacity, and 0.76 for no CHF features. The intraobserver κ for reader B was 0.75 for cardiomegaly, 0.79 for pleural effusion, 0.75 for airspace opacity, and 0.81 for no CHF features. The interobserver κ was 0.64 for cardiomegaly, 0.31 for pleural effusion, 0.20 for airspace opacity, and 0.67 for no CHF features. The *P* value for the correct versus overfitted difference was .001 for cardiomegaly, .003 for pleural effusion, .38 for airspace opacity, and .001 for no CHF features.

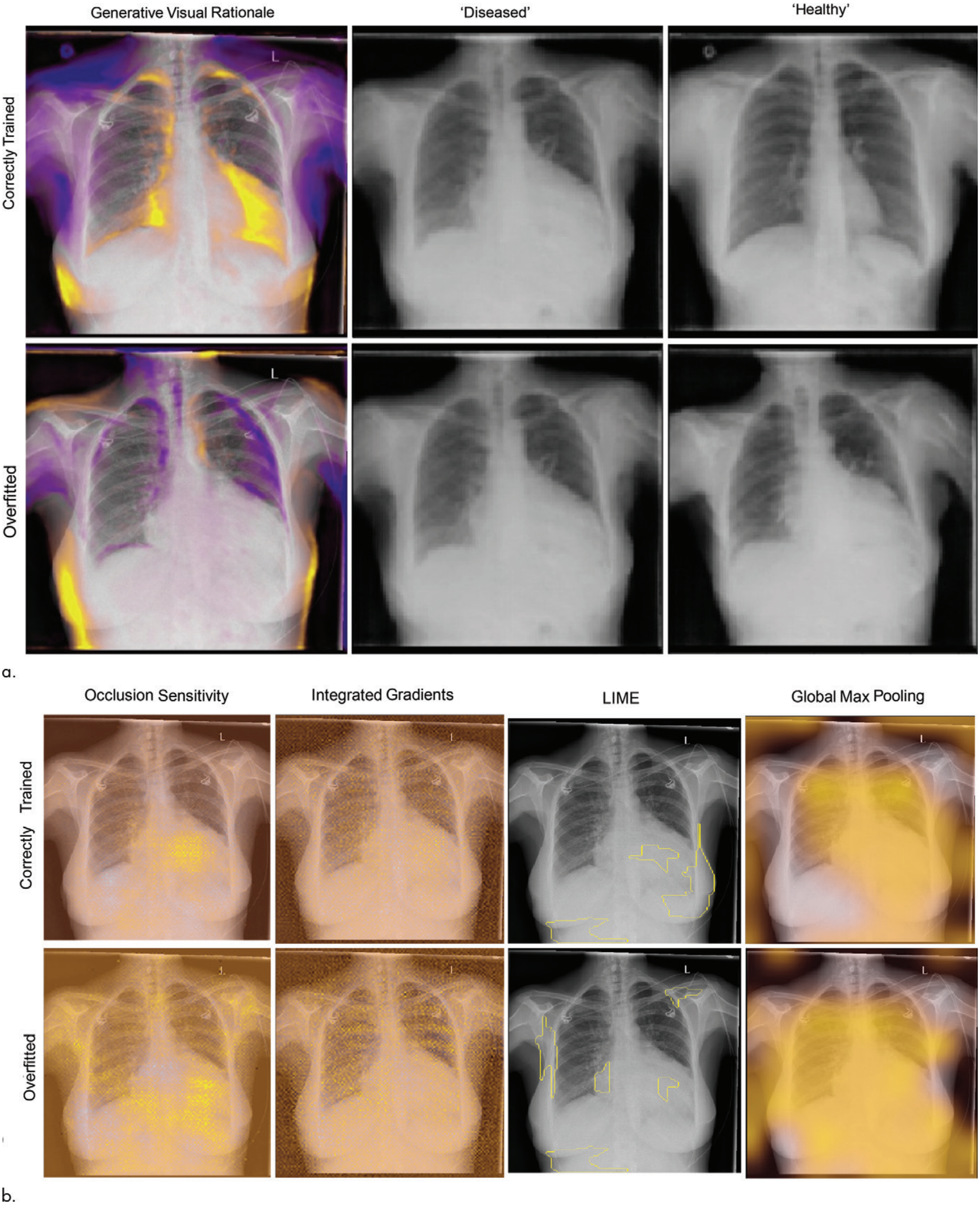

Figure 4: Chest radiographs in a 36-year-old woman. **(a)** Generative Visual Rationale (GVR) using the correctly trained model (top) and overfitted model (bottom) for the same input radiograph. For GVR, the correctly trained model highlights cardiomegaly (a sign of congestive heart failure), as well as low body mass (independent B-type natriuretic peptide, or BNP, associations), while the overfitted model highlights edges, suggesting memorization of the radiograph. **(b)** Comparison with other available interpretability methods. The other methods produce less intelligible visual reasoning. An animated movie showing this radiograph being permuted from "diseased" to "healthy" is available as Movie 1 (online). *LIME* = Local Model-Agnostic Explanations.

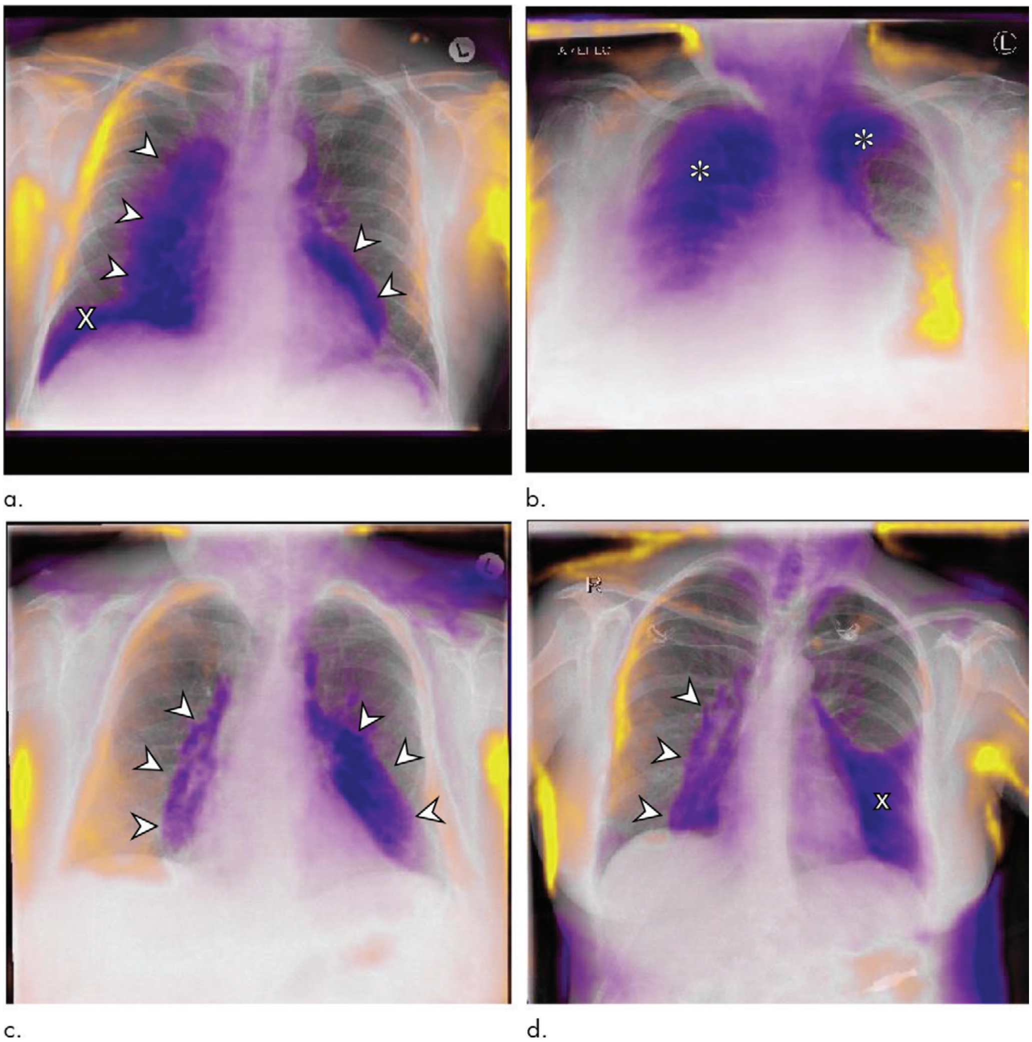

Figure 5: Normal chest radiographs in **(a)** 73-year-old, **(b)** 86-year-old, **(c)** 81-year-old, and **(d)** 94-year-old women show selected inverse Generative Visual Rationales for radiographs with predicted normal B-type natriuretic peptide (BNP) that answer the question "How would this patient's radiograph change with an elevated BNP?" Purple = density added and shows that the model creates the expected features for congestive heart failure, including cardiomegaly (arrowheads), pleural effusions (×) and airspace opacity (*).

The readers noted during their evaluation of the GVRs that reduced body mass (axillary/supraclavicular/chest regions highlighted purple) was unexpectedly highlighted frequently in GVRs. Because this was not a conventional radiographic feature of increased BNP, readers had not been asked to specifically identify this. However, in light of the fact that obesity is known to reduce BNP values, we believed that it was biologically plausible that the model was utilizing the apparent body mass as a feature for identifying radiographs with low BNP values.

Hence, readers were asked to reevaluate the blinded GVRs and found that reduced body mass was more frequently highlighted as a reason for elevated BNP by the correctly trained model (145 [72.5%] of 200 vs 32 [16%] of 200; $P < .001$).

Figure 4a compares a GVR produced by the correctly trained model and one produced by the overfitted model for the same radiograph. It demonstrates the density changes that the correct model predicts would need to occur to normalize BNP: removal of cardiomegaly and higher body mass elsewhere. The GVR from the overfitted model simply alters edges, indicating memorization of the radiograph. Other commonly used interpretability techniques, including occlusion maps, integrated gradients, Local Interpretable Model-Agnostic Explanations, or LIME, and global max pooling (16,17) provide far less useful information about the prediction (Fig 4b).

Inverse GVRs for Predicted "Healthy" Radiographs

To further validate the image features learned, inverse GVRs were generated on predicted healthy radiographs to answer the question "How would this radiograph change if it had an elevated BNP?" A selected sample of inverse GVRs is shown in Figure 5 and confirms that the model adds CHF features. Inverse GVRs from a single healthy radiograph at progressively higher BNP levels are shown in Figure 6. At 4000 ng/L, the model adds a pacemaker, indicating that it has learned this as an additional feature of raised BNP.

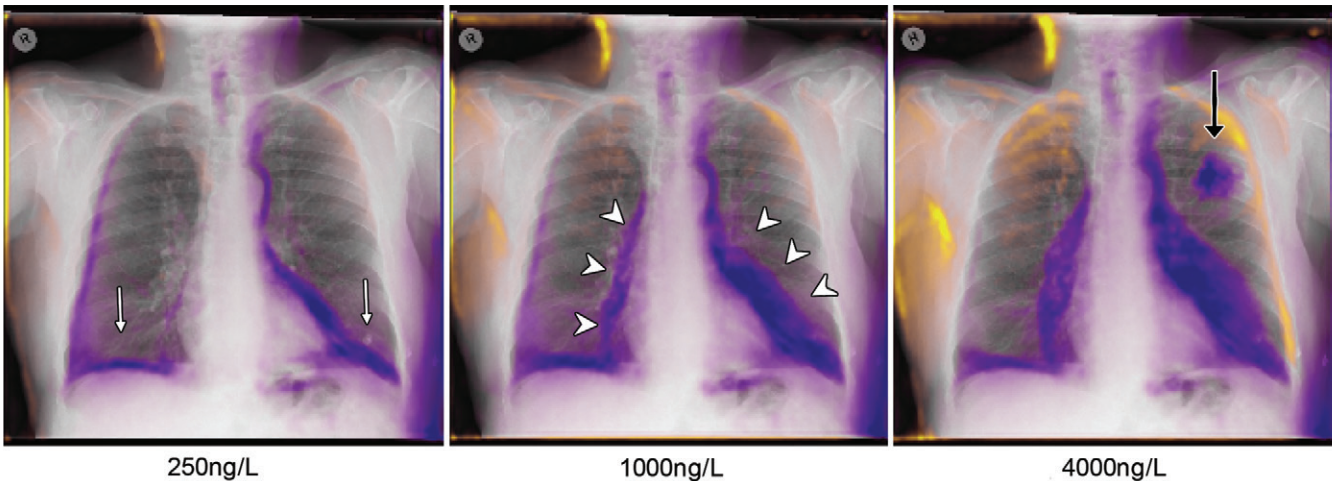

Figure 6: Normal chest radiograph in a 91-year-old man analyzed with the inverse Generative Visual Rationale (GVR) technique. Top row from left to right: inverse GVRs for a radiograph with predicted normal B-type natriuretic peptide (BNP) visualized at BNP levels of 250, 1000, and 4000 ng/L, respectively. The model progressively adds cardiomegaly (arrowheads) and pleural effusions (white arrow), and, at 4000 ng/L, it adds a focal left upper zone density representing a pacemaker (black arrow). An animated movie showing this radiograph being permuted from "healthy" to "diseased" is available as Movie 2 (online).

Discussion

This study shows that GVRs can reveal imaging features learned by a deep learning model trained to predict BNP values as a marker of CHF from frontal chest radiographs. Current techniques available to localize findings on medical images include occlusion maps, global max pooling, and LIME. These methods examine the contribution of individual pixels or small patches (groups of pixels) to the final predication, with more important pixels or patches highlighted. In contrast, GVRs examine how the image as a whole has to alter to change the prediction and are therefore able to identify non-local image features such as soft-tissue thickness or cardiomediastinal contour, which are not necessarily found within one region of an image. Given the ability of GVRs to highlight features that make the image abnormal, this helps discriminate between correctly trained models, which would use expected features, and overfitted models, which would not.

GVRs in Medical Imaging

Unlike recognizing handwritten digits, where there is a clear answer, many problems within radiology deal with images that cannot be confidently classified. Therefore, when applying machine learning to radiology problems, incorrect predictions based on incorrect features are difficult to detect, especially with existing interpretability techniques. Most efforts to avoid bias therefore focus on rigorous data preparation and training methods (18), with little current attempt to isolate faults after the model has been trained.

GVRs are predicated on the idea that the ability to generate realistic images similar to those in the original data set implies understanding of the data set itself. By predicting disease from the latent space representation of an image, the model can use image features learned during its unsupervised generative training. This allows GVRs to give an intuitive visual explanation of what a deep learning model has learned and to rationalize individual predictions. This could be useful in the emerging role of deep learning models as second readers—by providing GVRs to justify individual predictions. The specific application to chest radiographs is interesting because of their wide use and variety of pathologic features, with many groups tackling this problem specifically with deep learning models such as ChexNet (19).

Our study results demonstrate that GVRs can identify some confounding image features learned by a deep learning model. GVRs not only showed that CHF features had been learned de novo by the correctly trained model but also demonstrated an ability to expose the overfitted model and revealed that the correctly trained model was unexpectedly highlighting chest wall soft tissues when predicting BNP. This prompted a reevaluation of the blinded GVRs, which found that reduced body mass was more frequently highlighted by the correctly trained model. This unexpected feature caused the authors to review published literature and learn of a negative association between BNP and obesity (20,21). While adding an extra feature for assessment is unconventional, we believe that the ability of GVRs to identify this unexpected feature is a strong argument for the utility of GVRs and that the unusual step of asking readers to review an additional feature does not invalidate these results.

Although we have demonstrated the GVR technique only in our specific data set and problem, GVRs have also been shown to be effective in nonmedical data sets such as the benchmark Modified National Institute of Standards and Technology digit recognition problem (7), which suggests the technique may be able to be extended into other medical imaging modalities.

While a number of techniques have recently been developed to visualize predictions, these typically identify the predictive contribution of individual image patches. In contrast, the GVR examines how the image as a whole would need to change to elicit a different prediction. With this global visual understanding, radiologists and engineers can better exclude undesirable biases and potentially even identify previously unknown imaging features of disease.

Limitations of the Technique

A technical challenge for the GVR technique is in generating realistic images larger than 128 × 128 pixels at current memory limitations. In our study, limited resolution prevented the assessment of fine image details. However, much higher resolution generative networks (up to 1024 × 1024 pixels) may mitigate this limitation in the future (22).

Another fundamental limitation of GVRs is that they use a different architecture from existing deep learning models that cannot be applied to already trained models. Adoption of a GVR-friendly architecture would be required to use GVRs more broadly.

In conclusion, GVRs can identify imaging features learned by a model trained to predict CHF from chest radiographs, allowing radiologists to better identify faults and biases. Future work will explore the application of generative learning to further improve deep learning interpretability in medical imaging.

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

Disclosures of Conflicts of Interest: **J.C.Y.S.** disclosed no relevant relationships. **J.S.N.T.** disclosed no relevant relationships. **A.K.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: was loaned one NVIDIA GPU card for evaluation. Other relationships: disclosed no relevant relationships. **F.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: is the founder and CEO of Radiopaedia.org, a collaborative online radiology resource. **A.F.D.** disclosed no relevant relationships.

References

1. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
2. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. http://arxiv.org/abs/1406.2661. Published 2014. Accessed March 11, 2018.
3. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017; 214–223.
4. Chait A, Cohen HE, Meltzer LE, VanDurme JP. The bedside chest radiograph in the evaluation of incipient heart failure. Radiology 1972;105(3):563–566.
5. Roberts E, Ludman AJ, Dworzynski K, et al. The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. BMJ 2015;350:h910.
6. Lokuge A, Lam L, Cameron P, et al. B-type natriuretic peptide testing and the accuracy of heart failure diagnosis in the emergency department. Circ Heart Fail 2010;3(1):104–110.
7. Seah JCY, Tang J, Kitchen A, Seah JCN. Generative Visual Rationales. https://arxiv.org/abs/1804.04539. Published 2018. Accessed DATE.
8. Bain C, MacManus C, Seah JCY. Web based cohort identification across large healthcare data sets – opening the treasure chest. In: International Institute of Engineers, Bangkok, Thailand, April 20–21, 2015.
9. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. In: Proceedings of the Conference on Neural Information Processing Systems, 2017.
10. Oliphant T. Guide to NumPy. 2nd ed. USA: CreateSpace, 2015.
11. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. http://arxiv.org/abs/1511.06434. Published 2015. Accessed March 11, 2018.
12. Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela T, Duch W, Girolami M, Kaski S, eds. Artificial Neural Networks and Machine Learning – ICANN 2011. ICANN 2011. Lecture Notes in Computer Science, vol 6791. Berlin, Germany: Springer, 2011; 52–59.
13. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–2830.
14. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1–48.
15. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Collins A, Smith EE, eds. Readings in Cognitive Science, 1988; 399–421.
16. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. http://arxiv.org/abs/1703.01365f. Published 2017. Accessed March 11, 2018.
17. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, Calif, August 13–17, 2016. New York, NY: ACM, 2016; 1135–1144.
18. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data 2012;6(4):1–21.
19. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. https://arxiv.org/abs/1711.05225. Published 2017. Accessed May 29, 2018.
20. Clerico A, Giannoni A, Vittorini S, Emdin M. The paradox of low BNP levels in obesity. Heart Fail Rev 2012;17(1):81–96.
21. Hsich EM, Grau-Sepulveda MV, Hernandez AF, et al. Relationship between sex, ejection fraction, and B-type natriuretic peptide levels in patients hospitalized with heart failure and associations with inhospital outcomes: findings from the Get With The Guideline-Heart Failure Registry. Am Heart J 2013;166(6):1063–1071.e3.
22. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In: Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, April 29–May 3, 2018.