Training Calibration-based Counterfactual Explainers for Deep Learning Models in Medical Image Analysis

Jayaraman J. Thiagarajan (jjthiagarajan@gmail.com)
Lawrence Livermore National Labs
Kowshik Thopalli
Arizona State University
Deepta Rajan
IBM Research AI
Pavan Turaga
Arizona State University

Research Article

Keywords: Deep Learning Models ,Medical Image Analysis, Artificial intelligence methods , healthcare

Posted Date: September 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-929235/v1

Version of Record: A version of this preprint was published at Scientific Reports on January 12th, 2022.
See the published version at https://doi.org/10.1038/s41598-021-04529-5.

Training Calibration-based Counterfactual Explainers for Deep Learning Models in Medical Image Analysis

Jayaraman J. Thiagarajan[1,*], **Kowshik Thopalli**[2], **Deepta Rajan**[3], **and Pavan Turaga**[2]

[1]Lawrence Livermore National Labs, Livermore, 94550, USA
[2]Geometric Media Lab, Arizona State University, Tempe, 85281, USA
[3]IBM Research AI, San Jose, 95035, USA
[*]jjayaram@llnl.gov

ABSTRACT

Artificial intelligence methods such as deep neural networks promise unprecedented capabilities in healthcare, from diagnosing diseases to prescribing treatments. While this can eventually produce a valuable suite of tools for automating clinical workflows, a critical step forward is to ensure that the predictive models are reliable and to enable a rigorous introspection of their behavior. This has led to the design of explainable AI techniques that are aimed at uncovering the relationships between discernible data signatures and decisions from machine-learned models, and characterizing strengths/weaknesses of models. In this context, the so-called counterfactual explanations that synthesize small, interpretable changes to a given query sample while producing desired changes in model predictions to support user-specified hypotheses (*e.g.*, progressive change in disease severity) have become popular. When a model's predictions are not well-calibrated (*i.e.*, the prediction confidences are not indicative of the likelihood of the predictions being correct), the inverse problem of synthesizing counterfactuals can produce explanations with irrelevant feature manipulations. Hence, in this paper, we propose to leverage prediction uncertainties from the learned models to better guide this optimization.

To this end, we present TraCE (Training Calibration-based Explainers), a counterfactual generation approach for deep models in medical image analysis, which utilizes pre-trained generative models and a novel uncertainty-based interval calibration strategy for synthesizing hypothesis-driven explanations. By leveraging uncertainty estimates in the optimization process, TraCE can consistently produce meaningful counterfactual evidences and elucidate complex decision boundaries learned by deep classifiers. Furthermore, we demonstrate the effectiveness of TraCE in revealing intricate relationships between different patient attributes and in detecting shortcuts, arising from unintended biases, in learned models. Given the wide-spread adoption of machine-learned solutions in radiology, our study focuses on deep models used for identifying anomalies in chest X-ray images. Using rigorous empirical studies, we demonstrate the superiority of TraCE explanations over several state-of-the-art baseline approaches, in terms of several widely adopted evaluation metrics in counterfactual reasoning. Our findings show that TraCE can be used to obtain a holistic understanding of deep models by enabling progressive exploration of decision boundaries, detecting shortcuts, and inferring relationships between patient attributes and disease severity.

Introduction

There is a growing interest in adopting artificial intelligence (AI) methods for critical decision-making, from diagnosing diseases to prescribing treatments and allocating resources, in healthcare[1–3]. However, in order to trust AI systems and to prioritize patient safety, it is imperative to ensure those methods are both accurate and reliable[4]. Examples of unreliable AI systems include a model that can produce highly confident predictions for patients presenting anomalies not seen in the training data, or a model accumulating evidence for a certain diagnosis based on uninformative regions in an image[5,6]. This has strongly motivated the need to both reliably assess a model's confidence in its predictions[7–9], and to enable rigorous introspection of its behavior[4,10–12]. To this end, uncertainty estimation methods are being adopted to determine the deficiencies of a model and/or the training data[13]. Meaningful uncertainties can play a crucial role in supporting practical objectives that range from assessing regimes of over (or under)-confidence and active data collection, to ultimately improving the predictive models themselves[14]. However, in practice, uncertainties are known to be challenging to communicate to decision-makers[15], and the robustness of decisions with respect to uncertainties can vary considerably between use-cases[16]. Consequently, it could sometimes be more beneficial to implicitly leverage uncertainties when performing introspective analysis of machine-learned models.

Model introspection approaches that attempt to explain the input-output relationships inferred by models are routinely used to understand and promote trust in AI solutions. While there has been a large body of work on building inherently

explainable models (*e.g.*, rule based systems), post-hoc explanation methods have become the modus operandi with modern deep learning systems[17]. In particular, *local* explanation methods are very popular as they provide a convenient way for users to introspect by generating local explanations specific to a given input (*e.g.*, health record of a patient) – elucidate what features in the input data maximally support the prediction. Broadly, local explanation methods can be categorized into approximation and example-based approaches. The former class of methods begin by sampling in the vicinity of a query example and fit an explainer to the chosen set of samples (*e.g.*, fit a linear model in LIME[18] or extract rules in ANCHORS[19]). In contrast, example-based methods synthesize data samples in the vicinity of a query, such that the predictions for those samples align with a user-specified hypothesis. The data samples from the latter approach are referred to as *counterfactual* explanations[20,21]. While counterfactual explanations provide more flexibility over feature importance estimation methods, user-studies have also demonstrated that counterfactuals can elicit meaningful insights into the data[22].

While it is common to utilize counterfactuals for causal reasoning, in the recent years, they have been found to be effective for scenario exploration even with predictive models[23,24]. In its most generic form, for a given query x, one can pose counterfactual generation based on a predictive model $\mathrm{F}: X \rightarrow Y$ as an optimization problem:

$$\arg\min_{\bar{\mathrm{x}}} d(\mathrm{x}, \bar{\mathrm{x}}) \quad \text{s.t.} \quad \mathrm{F}(\bar{\mathrm{x}}) = \bar{\mathrm{y}};\ \bar{\mathrm{x}} \in M(X) \tag{1}$$

where $\bar{\mathrm{x}}$ is a counterfactual explanation for the query x (*e.g.*, a medical image of a patient) and $\bar{\mathrm{y}}$ is the user-specified hypothesis about $\bar{\mathrm{x}}$ (*e.g.*, a certain diagnosis). Minimizing a suitable discrepancy $d(.,.)$ between x and $\bar{\mathrm{x}}$ ensures that the underlying semantic content of x is preserved in the counterfactual (*i.e.*, vicinity). Another important requirement to produce meaningful counterfactuals is that the generated $\bar{\mathrm{x}}$ should lie close to the original data manifold $M(X)$. When no tractable priors exist for $M(X)$, it is common to perform this optimization in the latent space of a pre-trained generative model (*e.g.* variational autoencoders (VAE)[25] or generative adversarial networks (GAN)[26]). Despite the effectiveness of such priors, when the model's predictions $\mathrm{F}(\bar{\mathrm{x}})$ are poorly calibrated, i.e., prediction confidences are not indicative of the actual likelihood of correctness[9,27], the optimization in eq. (1) can still lead to bad quality explanations. Though different variants of the formulation in eq. (1) have been considered in the literature[20], the fundamental challenge with uncalibrated predictions still persists. We propose to circumvent this challenge by integrating prediction uncertainties into the counterfactual generation process.

Proposed Work. In this work, we propose TraCE (*Training Calibration-based Explainers*), an introspection method for deep medical imaging models, that effectively leverages uncertainties to generate meaningful counterfactual explanations for clinical image predictors. As illustrated in Figure 1, our framework is comprised of three key components: (i) an auto-encoding convolutional neural network to construct a low-dimensional, continuous latent space for the training data; (ii) a predictive model that takes as input the latent representations and outputs the desired target attribute (*e.g.* diagnosis state, age etc.) along with its prediction uncertainty; and (iii) a counterfactual optimization strategy that uses an uncertainty-based calibration objective to reliably elucidate the intricate relationships between image signatures and the target attribute. While our approach is flexible to support the use of any uncertainty estimator or prediction models that use explicit regularization to produce well-calibrated predictions, TraCE builds upon the recent Learn-by-Calibrating (LbC) technique[28] to obtain prediction intervals for both classification and regression settings. LbC jointly trains an auxiliary interval estimator alongside the predictor model using an interval calibration objective, and has been shown to be effective at recovering complex function mappings in scientific datasets. We first adapt LbC for multi-class classification problems and subsequently propose a counterfactual generation approach based on the estimated prediction intervals. When compared to interpretability techniques that provide saliency maps or feature importance scores to explain a specific decision[29], TraCE enables progressive transition between different output states (*e.g.*, *normal* $\rightarrow$ *abnormal*) through appropriate image manipulations and more importantly, allows optimization with both categorical- and continuous-valued target variables.

Results. While recent advances in ML such as deep learning have produced disruptive innovations in many fields, radiology is a prominent example. Conventionally, trained physicians visually assess medical images for characterization and monitoring of diseases. However, AI methods have been showed to be effective at automatically identifying complex signatures in imaging data and providing quantitative assessments of radiographic characteristics[30]. Motivated by this wide-spread adoption of AI tools in radiology, our study focuses on detecting anomalies in chest X-ray (CXR) images. More specifically, we use images from the publicly available RSNA pneumonia detection challenge database (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge) in order to demonstrate the effectiveness of TraCE in performing introspective analysis. Given the diagnosis task of categorizing Chest X-ray (CXR) images into *normal* and *abnormal* groups (*i.e.*, pneumonia-related anomalies), one can adopt a variety of ML solutions including deep neural networks to build classifiers. However, in practice, purely data-driven AI solutions can learn unintended *shortcuts*[31] (e.g., superficial correlations) instead of meaningful decision rules. Such models typically perform well on the observed data, including passing standard cross-validation tests, yet fail when deployed in the real-world[5,6]. Hence, the foremost utility of TraCE is in validating that a predictive model has learned generalizable decision rules.

To this end, we use TraCE to progressively generate counterfactuals with different levels of likelihood in assigning a patient to the *abnormal* group. From our results, we find that using TraCE consistently produces highly meaningful counterfactual evidences, wherein severity of *abnormality* (e.g., pneumonia) is characterized primarily by changes in the lung opacity. More importantly, the image manipulations are highly concentrated in the chest region of the subject, thus showing that the models did not pick shortcut inductive biases (e.g., scanner-specific features or background image pixels). Using rigorous empirical studies, we also show that TraCE outperforms existing baseline methods, in terms of several widely adopted evaluation metrics in counterfactual reasoning. Furthermore, we find that TraCE can effectively detect shortcuts (or unintended biases) in trained models and infer relationships between different attributes (for example, age and diagnosis state), thus enabling a holistic understanding of deep clinical models.

Background and Related Work

Uncertainty Estimation. The growing interest in employing machine learning (ML) based solutions to design diagnostic tools and to gain new insights into a host of medical conditions strongly emphasizes the need for a rigorous characterization of ML algorithms. In conventional statistics, uncertainty quantification (UQ) provides this characterization by studying the impact of different error sources on the prediction[32–34]. Consequently, several recent efforts have proposed to utilize prediction uncertainties in deep models to shed light onto when and how much to trust the predictions[35–37]. Some of the most popular uncertainty estimation methods today include: (i) Bayesian neural networks[34,38]: (ii) methods that use the discrepancy between different models as a proxy for uncertainty, such as deep ensembles[39] and Monte-Carlo dropout that approximates Bayesian posteriors on the weight-space of a model[35]; and (iii) approaches that use a single model to estimate uncertainties, such as orthonormal certificates[40], deterministic uncertainty quantification[41], distance awareness[42], depth uncertainty[43], direct epistemic uncertainty prediction[44] and accuracy versus uncertainty calibration[45].

Prediction Calibration. It has been reported in several studies that deep predictive models need not be inherently well-calibrated[27], i.e., the confidences of a model in its predictions are not correlated to its accuracy. While uncertainties can be directly leveraged for a variety of downstream tasks including out-of-distribution detection and sequential sample selection, they have also been utilized for guiding models to produce well-calibrated predictions. In practice, these requirements are incorporated as regularization strategies to systematically adjust the predictions during training, most often leading to better performing models. For example, uncertainties from Monte-Carlo dropout[46] and direct error prediction[47] have been used to perform confidence calibration in deep classifiers. Similarly, the recently proposed Learn-by-Calibrating (LbC) approach[28] introduced an interval calibration objective based on uncertainty estimates for training deep regression models.

Counterfactual Generation in Predictive Models. Counterfactual (CF) explanations[20] that synthesize small, interpretable changes to a given image while producing desired changes in model predictions to support user-specified hypotheses (e.g., progressive change in predictions) have recently become popular. An important requirement to produce meaningful counterfactuals is to produce discernible local perturbations (for easy interpretability) while being realistic (close to the underlying data manifold). Consequently, existing approaches rely extensively on pre-trained generative models to synthesize plausible counterfactuals[20,21,48–50]. While the proposed TraCE framework also utilizes a pre-trained generative model, it fundamentally differs from existing approaches by employing uncertainty-based calibration for counterfactual optimization.

Results

Data. Our analysis uses CXR images available as public benchmark data for the tasks of predicting the diagnostic state and other patient attributes. In particular, our study uses the *RSNA pneumonia detection challenge database*, which is a collection of $30,000$ CXR exams belonging to the NIH CXR14 benchmark dataset[51], of which $15,000$ exams show evidence for lung opacities related to pneumonia, consolidation and infiltration, and $7,500$ exams contain no findings (referred as *normal*). The CXR images in the dataset were annotated by six board-certified radiologists and additional information on the data curation process can be found in[52]. In addition to the diagnostic labels, this dataset contains age and gender information of the subjects. Note that, for this analysis, we used healthy control subjects from the RSNA pneumonia dataset to define the *normal* group and designed predictive models to discriminate them from patients presenting pneumonia-related anomalies in their CXR scans. We refer to the latter as the *abnormal* group.

Evaluation Metrics. We used the following metrics for a holistic evaluation of the counterfactual explanations obtained using TraCE and other baseline methods:
(i) *Validity*: For categorical attributes (as in classification problems), this metric measures the ratio of the counterfactuals that actually have the desired target attribute to the total number of counterfactuals generated (higher the better). In the case of continuous-valued attributes we measure the mean absolute percentage error (MAPE) between the desired and achieved target

values (lower the better).

$$V_{cat} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\mathrm{F}(\mathrm{x}_i),\bar{\mathrm{y}}_i); \quad V_{cont} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\bar{\mathrm{y}}_i - \mathrm{F}(\mathrm{x}_i)}{\bar{\mathrm{y}}_i}\right|, \tag{2}$$

where $\mathbb{I}$ denotes the identity function that returns 1 when the arguments match, and N is the total number of query samples used for evaluation.
(ii) *Confidence*: In cases of categorical-valued targets (class labels), we compute the confidence $P(\bar{\mathrm{y}}_i|\bar{\mathrm{x}}_i;\mathrm{F})$ (from softmax probabilities) of assigning the desired class $\bar{\mathrm{y}}_i$ for a counterfactual $\bar{\mathrm{x}}_i$ (higher the better).
(iii) *Sparsity*: Since we perform optimization directly in the latent space, measuring the amount of change in the images is a popular metric in the literature. We compute the sparsity metric as the ratio of the number of pixels altered to the total number of pixels.

$$S(\mathrm{x}) = \frac{\sum_i\sum_j C(|\mathrm{x}_{ij} - \bar{\mathrm{x}}_{ij}|)}{T}, \text{ where } C(a) = \begin{cases} 1, & \text{if } a > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

and T denotes the total number of pixels in the query x. In general, sparser changes to an image are more likely to preserve the inherent characteristics of the query image.
(iv) *Proximity*: Recent works have considered the actionability of modified features by grounding them in the training data distribution. Following[53], we measure the average ℓ_2 distance of each counterfactual to the K-nearest training samples in the latent space (lower the better).

$$Prox(\bar{\mathrm{x}}) = \frac{1}{K}\sum_{\mathrm{x}\in \mathrm{N}_K(\bar{\mathrm{x}};X)} \|\mathrm{E}(\bar{\mathrm{x}}) - \mathrm{E}(\mathrm{x})\|_2, \tag{4}$$

where $\mathrm{N}_K(\bar{\mathrm{x}};X)$ denotes the set of K nearest neighbors of the counterfactual x from the training data X, and E denotes the encoder network (see Methods section for details) to compute the latent representation.
(v) *Realism score*: We also employ this metric from the generative modeling literature[54] to evaluate the quality of images obtained using TraCE. While standard metrics such as the FID (Frechét Inception Distance) score or the precision/recall metrics are used to evaluate the overall quality of a population of generated images, they are not sufficient to assess individual images. Hence, we utilize the realism score introduced in[55], which is high when the generated image is close to the true data manifold and decreases as the image moves further from the manifold. Denoting the feature vectors for the set of real images (used for training), obtained using a pre-trained classifier such as VGG-16, by the matrix $\Psi_r = [\psi_r^1, \cdots, \psi_r^N]$ and a generated image by ψ_g, the realism score can be computed as follows:

$$R(\psi_g, \Psi_r) = \max_{\psi_r^j}\left\{\frac{\|\psi_r^j - \psi_r^K\|_2}{\|\psi_r^j - \psi_g\|_2}\right\}, \tag{5}$$

where ψ_r^K refers to the feature vector corresponding to the K^{th} nearest neighbor (w.r.t to Ψ_r) from the set $N_K(\psi_r^j;\Psi_r)$.

TraCE Enables Progressive Exploration of Decision Boundaries. Given the rapid adoption of AI solutions in diagnosis and prognosis, it is critical to gain insights into black-box predictive models. In this study, we analyzed a predictive model that classifies CXR images into *normal* and *abnormal* groups, and used TraCE to synthesize counterfactuals for a given query image from the *normal* class to visualize the progression of disease severity. Such an analysis can reveal what image signatures are introduced by a predictive model to provide evidence for the *abnormal* class, and can be used by practitioners to verify if the model relies on meaningful decision rules or *shortcuts* (*e.g.,* changes to the background) that cannot generalize. In our implementation of TraCE, we first constructed a low-dimensional latent space (100 dimensions) for the dataset of CXR images using a Wasserstein auto-encoder[56]. We subsequently learned the predictive model F_D along with the interval estimator G_D, using a modified version of the LbC algorithm (details in the Methods section). The hyper-parameters η_1 and η_2 in Eq. (11) are critical to trade-off between preserving the inherent semantics from query x and achieving the desired prediction. Hence, one can progressively transition from the *normal* to the *abnormal* class by fixing η_2 and gradually relaxing η_1.

Figure 2 illustrates the counterfactuals obtained using TraCE for multiple different examples from our benchmark dataset. More specifically, the query samples x correspond to CXR images from the *normal* class and we varied η_1 between 0.5 and 0.05, while setting $\eta_2 = 0.5$ and $\eta_3 = 0.2$. These values were obtained using a standard hyper-parameter search based on 100 randomly chosen images. For each case from Figure 2, the different counterfactuals along with their estimated $P(state = abnormal)$ from the predictive model F_D are shown. It can be clearly observed from the results that the counterfactuals show increased opacity

in the lung regions (appearing as denser white clouds) as we progress towards the *abnormal* class, which strongly corroborates with existing studies on CXR-based image analysis. Furthermore, TraCE does not arbitrarily introduce irrelevant features into the image or make anatomical changes, thereby reliably preserving the inherent characteristics of the subject. By producing physically plausible evidences for crucial hypotheses, TraCE enables practitioners to effectively explore complex decision boundaries learned by deep predictive models.

Comparing TraCE to Baseline Methods. In order to perform a quantitative evaluation of TraCE, we obtained counterfactuals for 100 randomly chosen images from a held-out test set (not used for training) and Table 1 presents a detailed comparison of different baseline methods (see Methods section for details). Note that, we chose the hyper-parameters η_1, η_2, η_3 such that the average discrepancy in the latent space is similar across all methods. The first striking observation is that, despite using the same pre-trained latent space for counterfactual optimization, all methods that incorporate explicit calibration strategies or uncertainty estimation consistently outperform the *Vanilla* model. More specifically, for similar levels of discrepancy in the latent space, TraCE achieves a significantly higher validity score of 0.88 as opposed to 0.69 of the *Vanilla* model, while inducing similar or lower amount of changes to the query (indicated by the sparsity and proximity metrics). Furthermore, our approach outperforms the results obtained with state-of-the-art uncertainty estimators and calibration strategies (in all the metrics), thus demonstrating its efficacy in generating counterfactual explanations.

As discussed earlier, TraCE is applicable for predictive models outputting both categorical- and continuous-valued target variables. To demonstrate this, we considered only healthy control subjects from the RSNA dataset and designed a regressor to estimate their age attribute using their CXR images. Though the age prediction task is not necessarily relevant on its own in clinical diagnosis, as we will show next, such attribute estimators can be utilized for inferring relationships to the diagnosis state. For our evaluation, we used 100 randomly chosen test subjects whose age attribute was between 40 and 70 and set the desired value $\bar{y} = 20$. From Table 2, we notice that the proposed approach achieves lower validity (MAPE) scores, without compromising on the proximity metric, when compared to the other baselines. Interestingly, we find that changing the age attribute required the manipulation of much lesser number of pixels (low sparsity values) when compared to the diagnosis state.

TraCE Detects Shortcuts in Deep Models. An important challenge with purely data-driven methods is that they have the risk of inferring decision rules based on shortcuts, thereby limiting their utility in practice. Detecting such shortcuts is essential to both validate model behavior and to detect unintended biases (hospital-specific or device-specific information) in the training data. In order to demonstrate the use of TraCE in detecting such shortcuts, we synthetically introduced a *nuisance* feature into images from the *abnormal* class – overlaid the text PNEUMONIA in the top-left corner of each image, and used TraCE to check if the model's decision was based on this nuisance feature. After training the Wasserstein autoencoder and the LbC model using the altered images, we selected query images from the *normal* group and generated the corresponding counterfactual evidences for the *abnormal* group. As illustrated in Figure 3(a-d), TraCE exclusively manipulates the top-left corner to accumulate evidence for abnormality, thus revealing that the predictive model relies on the nuisance feature. Similarly, in Figure 3(e-h), one can transition from the *abnormal* (examples containing the nuisance feature) to the *normal* group by simply removing the synthetic text PNEUMONIA. This experiment clearly emphasizes the utility of TraCE in detecting model and data biases.

TraCE Reveals Attribute Relationships. Motivated by the effectiveness of TraCE in producing counterfactuals for different types of target attributes, we next explored how counterfactual optimization can be used to study relationships between patient attributes, such as age and gender, and the diagnosis state. Note, this analysis is based on the assumption that the patient attribute can be directly estimated from the CXR images, and the inferred relationship does not necessarily imply causality.

First, we study if the image signatures pertinent to the patient age attribute provides additional evidence for diagnosis state prediction. Given the age predictor, along with its interval estimator, $(\mathrm{F}_A, \mathrm{G}_A)$ and the diagnosis predictor $(\mathrm{F}_D, \mathrm{G}_D)$, we constructed counterfactuals based on two independent hypotheses. Note, both predictors were constructed based on the same low-dimensional latent representations. More specifically, we provided the hypotheses $\bar{y}_A = 70$ and $\bar{y}_D = \textit{abnormal}$ for the two cases, and used TraCE to generate counterfactuals $\bar{\mathrm{x}}_A$ and $\bar{\mathrm{x}}_D$ that adhere to our hypotheses. We then estimated the age-specific and diagnosis-specific signatures introduced by TraCE:

$$\Delta_A(\mathrm{x}) = \mathrm{x} - \bar{\mathrm{x}}_A; \quad \Delta_D(\mathrm{x}) = \mathrm{x} - \bar{\mathrm{x}}_D. \tag{6}$$

In order to check if there exists an apparent relationship between age and diagnosis state, we generated the hybrid counterfactual,

$$\bar{\mathrm{x}} = \mathrm{x} + \Delta_A(\mathrm{x}) + \Delta_D(\mathrm{x}). \tag{7}$$

Finally, we compared $\mathrm{F}_D(\bar{\mathrm{x}}) - \mathrm{F}_D(\bar{\mathrm{x}}_D)$ to quantify if incorporating age-specific features into $\bar{\mathrm{x}}_D$ increased the disease severity (*i.e.*, likelihood of being assigned to the *abnormal* class). An overview of this strategy is illustrated in Figure 4.

Figure 5 shows the results for 8 different *normal* subjects, wherein we find that there is an apparent increase in $P(state = abnormal)$ when age-specific signatures are incorporated. Using 100 randomly chosen *normal* subjects, we estimated an average change of 0.09 ± 0.08 in $\mathrm{F}_D(\bar{\mathrm{x}}) - \mathrm{F}_D(\bar{\mathrm{x}}_D)$, thus indicating that the diagnosis predictor is sensitive to age-specific patterns. In practice, if such a relationship is expected, it is a strong validation for the model's behavior. On the other hand, if the attribute is a confounding variable, it becomes critical to retrain the model wherein this sensitivity is explicitly discouraged. Interestingly, when we repeated this analysis with the gender attribute, such a relationship was not apparent (see results in Figure 6).

Methods

Constructing low-dimensional latent spaces. Given a set of samples from an unknown data distribution, our goal is to build a low-dimensional, continuous latent space that respects the true distribution, so that one can generate counterfactual representations in that space. A large class of generative modeling methods exist to construct such a latent space. In this work, we focus on Wasserstein autoencoders[56] since they have been found to outperform other variational autoencoder formulations, particularly in image datasets with low heterogeneity, *e.g.*, scientific images from physics simulations[57]. This network is a composition of an encoder network E that transforms the input x into its latent code z, and a decoder network D that reconstructs the image. Additionally the encoder has the objective of matching the latent distribution of the training samples $\mathbb{E}_{P_X}[\mathrm{E}(\mathrm{z} \mid \mathrm{x})]$ to a pre-specified prior P_Z. This helps us to sample from the prior as well as generate new unseen samples from the original data manifold $M(X)$ after training such auto-encoding models. Wasserstein autoencoders thus have to minimize: 1) Discrepancy cost D_x between the original data distribution and the generated; 2) Discrepancy cost D_z between the latent distribution of the encoded training samples to that of a prior. Following standard WAE consturction, we employ the mean squared error (MSE) for D_x and use the maximum mean discrepancy (MMD) to define D_z. As shown in Figure 7(a), we also find that including another loss term to maximize the structural similarity (SSIM)[58] between the original and reconstructed images led to higher quality reconstructions.

Training. All images were resized to 224×224 pixels, and treated as single channel images. With the latent space dimensionality fixed at 100, the encoder model was comprised of 4 convolutional layers, with the number of filters set to $[16, 32, 64, 32]$, followed by two fully connected layers with hidden units as 512 and 100. All convolutional layers used the kernel size (3,3) and stride 2. The decoder consisted of two fully connected layers with 512 and 6272 hidden units followed by 4 transposed convolutional layers with channels $[64, 32, 16, 1]$ respectively. ReLU non-linear activation was applied after every layer except for the last layer. We trained the models using the Adam[59] optimizer for 150 epochs with an initial learning rate of $1e-3$ and decreased it by factors 2, 5, 10 after 30, 50 and 100 epochs respectively. The three loss functions were assigned the weights $[1, 0.5, 0]$ for the first 20 epochs and subsequently changed to $[1, 0.1, 1]$ respectively until convergence.

Predictive Model Design using LbC. While conventional metrics such as cross entropy (for categorical-valued outputs) and mean squared error (for continuous-valued outputs) are commonly used, it has been recently found that interval calibration is effective for obtaining accurate and well-calibrated predictive models[28]. Hence, in TraCE, we adapt the Learn-by-Calibrating approach to train classifier (or regression) models that map from the CXR latent space to a desired target variable. By design, LbC provides prediction intervals in lieu of point estimates for the response y, i.e., $[\hat{\mathrm{y}} - \delta, \hat{\mathrm{y}} + \delta]$. Here, δ is used to define the interval. Suppose that the likelihood for the true response y to be contained in the prediction interval is $p(\hat{\mathrm{y}} - \delta \leq \mathrm{y} \leq \hat{\mathrm{y}} + \delta)$, the intervals are considered to be well-calibrated if the likelihood matches the expected confidence level. For a confidence level α, we expect the interval to contain the true response for $100 \times \alpha\%$ of realizations from $p(\mathrm{x})$.

Algorithm. The model is comprised of two modules F and G, implemented as neural networks, to produce estimates $\hat{\mathrm{y}} = \mathrm{F}(\mathrm{z})$ and $\delta = \mathrm{G}(\mathrm{z})$ respectively. For example, in the case of multi-class classification settings, $\hat{\mathrm{y}} \in \mathbb{R}^K$ is a vector of predicted logits for the K different classes. Since interval calibration is defined for continuous-valued targets, we adapt the loss function for training on the logits directly. To this end, we first transform the ground truth labels into logits. Note, for each sample, we allow a small non-zero probability (say 0.01) to all negative classes. As discussed earlier, suppose that the likelihood for the true $\mathrm{y}[k], k \in (1, \cdots, K)$ to be contained in the interval is $p(\hat{\mathrm{y}}[k] - \delta[k] \leq \mathrm{y}[k] \leq \hat{\mathrm{y}}[k] + \delta[k])$, the intervals are considered to be well-calibrated if the likelihood matches the confidence level. Denoting the parameters of the models F and G by θ and ϕ respectively, we use an alternating optimization strategy similar to[28]. In order to update ϕ, we use the empirical interval calibration error as the objective:

$$\phi^* = \arg\min_{\phi} \sum_{k=1}^{K} \left| \alpha - \frac{1}{N} \sum_{i=1}^{N} \mathbb{1} \left[(\hat{\mathrm{y}}_i[k] - \delta_i[k]) \leq \mathrm{y}_i[k] \leq (\hat{\mathrm{y}}_i[k] + \delta_i[k]) \right] \right|, \tag{8}$$

where $\delta_i = \mathrm{G}(\mathrm{z}_i; \phi)$, and the desired confidence level α (set to 0.9 in our experiments) is an input to the algorithm. When updating the parameters ϕ, we assume that the estimator $\mathrm{F}(.; \theta)$ is known and fixed. Now, given the updated ϕ, we learn the

parameters θ using the following hinge-loss objective:

$$\theta^* = \arg\min_{\theta} \sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} \Bigg[\max\Big(0, (\hat{y}_i[k] - \delta_i[k]) - y_i[k] + \tau\Big) + \max\Big(0, y_i[k] - (\hat{y}_i[k] + \delta_i[k]) + \tau\Big) \Bigg], \tag{9}$$

where $\hat{y}_i = F(z_i; \theta)$ and τ is the margin parameter (set to 0.05 in our experiments). Intuitively, for a fixed ϕ, obtaining improved estimates for $\hat{y}$ can increase the empirical calibration error in (8) by achieving higher likelihoods even for lower confidence levels. However, in the subsequent step of updating ϕ, we expect δ's to become sharper in order to reduce the calibration error. We repeat the two steps (eqns. (8) and (9)) until convergence.

Architecture. As showed in Figure 7(b), the predictor was designed as a $7-$layer fully connected network with hidden sizes $[512, 1024, 256, 128, 64, 16, K]$ and ELU activations, while the interval estimator was a $6-$layer network with sizes $[512, 1024, 256, 128, 64, K]$ and ReLU activations.

Uncertainty-Aware Counterfactual Generation. TraCE modifies the counterfactual generation process in Eq. (1) using the pre-trained predictor and interval estimator models from LbC. Our goal is to generate explanations to support a given hypothesis on the target variable – for example emulating high-confidence disease states given the CXR of a healthy subject. To this end, we first obtain the latent representation for the given query image x using the encoder, $z = E(x)$. We then use the pre-trained predictor (F) and interval estimator (G) models to generate the counterfactual $\bar{z}$. Finally, the generated counterfactuals in the latent space are passed to the decoder to obtain a reconstruction in the image space, $\bar{x} = D(\bar{z})$. We propose the following optimization to generate the counterfactual explanations:

$$\bar{z} = \arg\min_{\hat{z}} \eta_1 \|z - \hat{z}\|_2^2 + \eta_2 \mathscr{L}(\hat{y}, \delta, \bar{y}) + \eta_3 \delta, \text{ where } \hat{y} = F(\hat{z}), \delta = G(\hat{z}). \tag{10}$$

Here, $\bar{y}$ is the desired value for the target attribute (hypothesis), η_1, η_2, η_3 are hyper-parameters for weighting the different terms, and $\|.\|_2$ denotes the ℓ_2 norm. The first term ensures that the generated counterfactual is in the vicinity of the query sample x (in the latent space). The second term ensures that the expected target value is contained in the prediction interval (calibration), while the final term penalizes arbitrarily large intervals to avoid trivial solutions. The calibration objective $\mathscr{L}$ is implemented as a hinge-loss term:

$$\mathscr{L}(\hat{y}, \delta, \bar{y}) = \Big[\max\Big(0, (\hat{y} - \delta) - \bar{y} + \tau\Big) + \max\Big(0, \bar{y} - (\hat{y} + \delta) + \tau\Big) \Big], \tag{11}$$

where the margin was fixed at $\tau = 0.05$. Choosing η_1, η_2, η_3 is essential to controlling the discrepancy between z and $\bar{z}$, and ensuring that the prediction for the counterfactual is $\bar{y}$.

Baselines. We considered a suite of baseline approaches for our empirical study and they differ by the strategies used for training the classifier, and counterfactual optimization. In particular, we investigate approaches that produce explicit uncertainty estimators as well as those that directly build well-calibrated predictors. However, note that, all methods perform their optimization in the same latent space.

(i) *Vanilla*: In this approach, we train the classifier with no explicit calibration or uncertainty estimation, and use the following formulation to generate the counterfactuals:

$$\bar{z} = \arg\min_{\hat{z}} \eta_1 \|z - \hat{z}\|_2^2 + \eta_2 \mathscr{L}_{ce}\left[F(\hat{z}), \bar{y}\right], \tag{12}$$

where $\mathscr{L}_{ce}$ denotes the cross entropy loss.

(ii) *Mixup*: This is a popular augmentation strategy[60] that convexly combines random pairs of images and their labels, in order to temper overconfidence in predictions. Recently, in[61], it was found that mixup regularization led to improved calibration in the resulting model. In mixup, the model is trained not only on the training data, but also using samples in the vicinity of each training sample:

$$x = \lambda x_i + (1 - \lambda) x_j; \quad y = \lambda y_i + (1 - \lambda) y_j, \tag{13}$$

where x_i and x_j are randomly chosen samples with labels y_i and y_j. The parameter λ, drawn from a symmetric Beta distribution, determines the mixing ratio. Since this approach does not produce any uncertainty estimation, the counterfactual optimization is same as that of the *Vanilla* approach in Eq. (12).

(iii) *MC Dropout*: In this baseline, we train the classifier with dropout regularization and estimate the (epistemic) prediction uncertainty for any test sample by running multiple forward passes. Finally, we use the following heteroscedastic regression objective to implement uncertainty-based calibration during counterfactual optimization:

$$\bar{z} = \arg\min_{\hat{z}} \eta_1 \|z - \hat{z}\|_2^2 + \eta_2 \left[\frac{(\bar{y} - \mu_{\hat{z}})^2}{2\sigma_{\hat{z}}^2} + \frac{1}{2}\log(\sigma_{\hat{z}}^2) \right]. \tag{14}$$

Note, similar to the proposed approach, here we operate directly on the logits and the mean/variance estimates $(\mu_{\hat{z}}, \sigma_{\hat{z}}^2)$ are obtained using T (set to 5) forward passes with dropout.
(iv) *Deep Ensembles*: Deep ensembles form an important class of uncertainty estimation methods, wherein the model variance is used as a proxy for uncertainties. In this approach, we independently train M different models (with bootstrapping and different model initializations) with the same architecture. Subsequently, for any input sample x, we obtain the mean/variance estimates $(\mu_{\hat{z}}, \sigma_{\hat{z}}^2)$ by aggregating predictions from the M models. Finally, we employ the calibration objective in Eq. (14) to perform counterfactual optimization. While highly accurate and currently one of the best uncertainty estimation techniques, deep ensembles require training multiple models, which can become a computational bottleneck when training deep networks.
(v) *Uncertainty-Weighted Confidence Calibration (UWCC)*: The authors in[46] proposed to build calibrated classification models by augmenting a confidence-calibration term to the standard cross-entropy loss and weighting the two terms using the uncertainty measured via multiple stochastic inferences. Mathematically,

$$\frac{1}{N}\sum_{i=1}^{N} -(1-\alpha_i)\log(P(\hat{y}_i|z_i)) + \alpha_i D_{KL}(U(y)||P(\hat{y}_i|z_i)). \tag{15}$$

Here the first term denotes the cross-entropy loss, and the predictions $P(\hat{y}_i|z_i)$ are inferred using stochastic inferences for z_i, while the variance (α_i) in the predictions is used to balance the loss terms. More specifically, we perform T forward passes with dropout in the network and promote the softmax probabilities to be closer to an uniform distribution, i.e. high uncertainty, when the variance is large. The normalized variance α_i is given by the mean of the Bhattacharyya coefficients between each of the T predictions and the mean prediction. Since the model is inherently calibrated during training, we do not measure the uncertainties at test time and hence use the optimization in Eq. (12) for generating counterfactuals. For the case of continuous-valued targets (*i.e.*, regression tasks), we utilize the extension in[62] that performs heteroscedastic calibration of the MC dropout estimator during training.

Code Availability. The software associated with this paper will be hosted through a public code repository (github).

References

1. Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S. & Acharya, U. R. Deep learning for healthcare applications based on physiological signals: A review. *Comput. methods programs biomedicine* **161**, 1–13 (2018).
2. Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. medicine* **23**, 89–109 (2001).
3. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Briefings bioinformatics* **19**, 1236–1246 (2018).
4. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. The Royal Soc. Interface* **15**, 20170387 (2018).
5. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
6. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj* **369** (2020).
7. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1321–1330 (JMLR. org, 2017).
8. Leibig, C., Allken, V., Ayhan, M. S., Berens, P. & Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Sci. reports* **7**, 1–14 (2017).
9. Thiagarajan, J. J., Venkatesh, B., Rajan, D. & Sattigeri, P. Improving reliability of clinical models using prediction calibration. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, 71–80 (Springer, 2020).

10. Cabitza, F. & Campagner, A. Who wants accurate models? arguing for a different metrics to take classification models seriously. *arXiv preprint arXiv:1910.09246* (2019).

11. Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134* (2019).

12. Thiagarajan, J. J., Rajan, D. & Sattigeri, P. Understanding behavior of clinical models under domain shifts. *arXiv preprint arXiv:1809.07806* (2018).

13. Gawlikowski, J. *et al.* A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342* (2021).

14. Thiagarajan, J. J., Venkatesh, B., Sattigeri, P. & Bremer, P.-T. Building calibrated deep models via uncertainty matching with auxiliary interval predictors. *AAAI Conf. on Artif. Intell.* (2019).

15. Kompa, B., Snoek, J. & Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit. Medicine* **4**, 1–6 (2021).

16. Batteux, E., Avri, B., Johnson, S. G. & Tuckett, D. The negative consequences of failing to communicate uncertainties during a pandemic: The case of covid-19 vaccines. *medRxiv* (2021).

17. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R. *Explainable AI: interpreting, explaining and visualizing deep learning*, vol. 11700 (Springer Nature, 2019).

18. Ribeiro, M. T., 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, CA, USA, August 13-17, 2016*, 1135–1144 (2016).

19. Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).

20. Verma, S., Dickerson, J. & Hines, K. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

21. Singla, S., Pollack, B., Chen, J. & Batmanghelich, K. Explanation by progressive exaggeration. In *International Conference on Learning Representations* (2019).

22. Byrne, R. M. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, 6276–6282 (2019).

23. Cohen, J. P. *et al.* Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical Imaging with Deep Learning* (2021).

24. Narayanaswamy, V., Thiagarajan, J. J. & Spanias, A. Using deep image priors to generate counterfactual explanations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2770–2774 (IEEE, 2021).

25. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. In Bengio, Y. & LeCun, Y. (eds.) *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (2014).

26. Goodfellow, I. *et al.* Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680 (2014).

27. Kuleshov, V., Fenner, N. & Ermon, S. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263* (2018).

28. Thiagarajan, J. J. *et al.* Designing accurate emulators for scientific processes using calibration-driven deep models. *Nat. Commun.* **11** (2020).

29. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).

30. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).

31. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).

32. Smith, R. C. *Uncertainty quantification: theory, implementation, and applications*, vol. 12 (Siam, 2013).

33. Heskes, T. Practical confidence and prediction intervals. In *Advances in neural information processing systems*, 176–182 (1997).

34. Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584 (2017).

35. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059 (2016).

36. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 6402–6413 (2017).

37. Thiagarajan, J. J., Kim, I., Anirudh, R. & Bremer, P.-T. Understanding deep neural networks through input uncertainties. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2812–2816 (IEEE, 2019).

38. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural network. In *International Conference on Machine Learning*, 1613–1622 (PMLR, 2015).

39. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474* (2016).

40. Tagasovska, N. & Lopez-Paz, D. Single-model uncertainties for deep learning. *arXiv preprint arXiv:1811.00908* (2018).

41. Van Amersfoort, J., Smith, L., Teh, Y. W. & Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, 9690–9700 (PMLR, 2020).

42. Liu, J. Z. *et al.* Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv preprint arXiv:2006.10108* (2020).

43. Antorán, J., Allingham, J. U. & Hernández-Lobato, J. M. Depth uncertainty in neural networks. *arXiv preprint arXiv:2006.08437* (2020).

44. Jain, M. *et al.* Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501* (2021).

45. Krishnan, R. & Tickoo, O. Improving model calibration with accuracy versus uncertainty optimization. *arXiv preprint arXiv:2012.07923* (2020).

46. Seo, S., Seo, P. H. & Han, B. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9030–9038 (2019).

47. Thiagarajan, J. J., Narayanaswamy, V., Anirudh, R., Bremer, P.-T. & Spanias, A. Accurate and robust feature importance estimation under distribution shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 7891–7898 (2021).

48. Van Looveren, A. & Klaise, J. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584* (2019).

49. Dhurandhar, A. *et al.* Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *arXiv preprint arXiv:1802.07623* (2018).

50. Goyal, Y. *et al.* Counterfactual visual explanations. In *International Conference on Machine Learning*, 2376–2384 (PMLR, 2019).

51. Wang, X. *et al.* Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR* (2017).

52. Stein, A. Pneumonia dataset annotation methods (2018). https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/discussion/64723.

53. Dandl, S., Molnar, C., Binder, M. & Bischl, B. Multi-objective counterfactual explanations. In Bäck, T. *et al.* (eds.) *Parallel Problem Solving from Nature – PPSN XVI*, 448–469 (Springer International Publishing, 2020).

54. Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O. & Gelly, S. Assessing generative models via precision and recall. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 5234–5243 (2018).

55. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J. & Aila, T. Improved precision and recall metric for assessing generative models. *Adv. Neural Inf. Process. Syst.* **32**, 3927–3936 (2019).

56. Tolstikhin, I. O., Bousquet, O., Gelly, S. & Schölkopf, B. Wasserstein auto-encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (OpenReview.net, 2018).

57. Anirudh, R., Thiagarajan, J. J., Bremer, P.-T. & Spears, B. K. Improved surrogates in inertial confinement fusion with manifold and cycle consistencies. *Proc. Natl. Acad. Sci.* **117**, 9741–9746 (2020).

58. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).

59. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. & LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

60. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

61. Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T. & Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, 13888–13899 (2019).

62. Venkatesh, B. & Thiagarajan, J. J. Heteroscedastic calibration of uncertainty estimators in deep learning. *arXiv preprint arXiv:1910.14179* (2019).

63. Thiagarajan, J. J., Venkatesh, B. & Rajan, D. Learn-by-calibrating: Using calibration as a training objective. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2019).

Figure legends

Figure 1. An overview of TraCE applied for introspective analysis of chest X-ray (CXR)-based predictive models. Driven by a user-specified hypothesis, our approach takes an X-ray image as input and produces counterfactual explanations based on exploration in a pre-trained CXR image latent space. In this example, hypothesis on the diagnosis state (*normal* or *abnormal*) is used to generate the explanation. By controlling the level of discrepancy between z and $\bar{z}$, TraCE can progressively synthesize counterfactuals with varying likelihoods of being assigned to the *abnormal* group.

Figure 2. (a)-(n) Diagnosis-based counterfactual explanations generated using TraCE by progressively introducing relevant patterns into different query images (first image in each row) of healthy subjects to increase the likelihood of being assigned to the *abnormal* group ($P(state = abnormal)$).

Figure 3. Using TraCE to detect shortcuts in deep predictive models. In this experiment, we synthetically introduced a nuisance feature (overlaid the text PNEUMONIA in the top-left corner) into all images from the *abnormal* group, and used this data to train the predictive model. Given the entirely data-driven nature of machine-learned solutions, there is risk of inferring a decision rule based on this irrelevant feature in order to discriminate between *normal* and *abnormal* groups. (a-d) Here, we used randomly chosen query images from the *normal* class and generated counterfactuals for the *abnormal* class. In each case, we show the query image, the counterfactual explanation from TraCE and the absolute difference image between the two; (e-f) Here, we introduced the nuisance feature into CXR images from the *abnormal* group and synthesized counterfactuals for the *normal* class. We observe that TraCE can effectively detect such shortcuts – counterfactuals for changing the diagnosis state are predominantly based on manipulating the text on the top-left corner of the query images.

Figure 4. Using TraCE to infer relationships between a patient attribute (*e.g.*, age) and disease states. For this analysis, we construct two independent predictive models, *i.e.*, age and diagnosis state, and synthesize counterfactuals based on hypothesis on each of the predictions (*e.g.*, patient age should be predicted as 70 while the diagnosis state should be *abnormal*. Finally, we combine the changes induced in the two counterfactuals, $\Delta_A(\mathrm{x})$ and $\Delta_D(\mathrm{x})$ respectively, and check if incorporating age-specific patterns strengthens the evidence for the *abnormal* class

Figure 5. (a-h) Explanations generated using TraCE by introducing age-specific attributes into the counterfactuals synthesized for changing the diagnosis state of a *normal* subject to be *abnormal*. Interestingly, we find that there exists a correlation between the two attributes, as evidenced by the consistent increase in the likelihood $P(state = abnormal)$ when compared to counterfactuals that rely only on patterns from the diagnosis state predictor. In each case, we highlight the changes $\Delta_A(\mathrm{x}), \Delta_D(\mathrm{x}), (\Delta_A(\mathrm{x}) + \Delta_D(\mathrm{x}))$ and show the likelihood $P(state = abnormal)$.

Figure 6. (a-d) Explanations generated using TraCE by introducing gender-specific attributes into the counterfactuals synthesized for changing the diagnosis state of a *normal* subject to be *abnormal*. In contrast to the age attribute, image manipulations associated with change in gender (female $\rightarrow$ male) do not cause any apparent change to the likelihood of being assigned to the *abnormal* group. In each case, we highlight the changes $\Delta_A(\mathrm{x}), \Delta_D(\mathrm{x}), \Delta_A(\mathrm{x}) + \Delta_D(\mathrm{x})$ and show the likelihood $P(state = abnormal)$.

Figure 7. Framework Design for TraCE. (a) First, we train an auto-encoding neural network[56], and construct a low-dimensional, continuous latent space for CXR images. Note, we used a combination of maximum mean discrepancy (MMD), mean

squared error (MSE) and structural similarity (SSIM) losses to train the network parameters; (b) Next, we adapt the Learn-by-Calibrating[63] approach to train a classifier that takes as input the latent representation from the encoder and outputs a patient-specific attribute along with prediction intervals.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-JRNL-819406).

Author contributions statement

Jayaraman J. Thiagarajan (J.J.T);
Kowshik Thopalli (K.T);
Deepta Rajan (D.R);
Pavan Turaga (P.T)

J.J.T conceived the presented idea, and along with K.T developed the overall formulation. While P.T helped in setting the high-level objectives, D.R contributed substantially in technical discussions through the course of this work. J.J.T and K.T implemented an initial version of the proposed framework, and with help of D.R, set up the empirical studies. D.R was instrumental in preparing the dataset and validating the findings. While J.J.T led the manuscript writing efforts, all the other authors contributed significantly to different sections.

Competing interests

The authors declare no competing interests.

Data availability

All datasets used in this were obtained from publicly released databases and pre-processed using open-source tool chains. We have added appropriate links to obtain the data as well as access the scripts for pre-processing, wherever applicable.

Tables

Method	Validity ↑	Confidence ↑	Sparsity ↓	Proximity ↓	Realism ↑
Vanilla	0.69	0.61±0.08	0.28±0.19	4.46±0.62	1.18 ± 0.07
Mixup	0.8	0.71±0.13	0.25±0.13	4.03±0.49	1.24 ± 0.09
UWCC	0.81	0.76±0.11	0.23±0.19	4.11±0.59	1.19 ± 0.16
MC Dropout	0.74	0.7±0.09	0.31±0.12	4.33±0.44	1.22 ± 0.14
Deep Ensembles (5 models)	0.82	0.73±0.06	0.28±0.09	**3.69±0.51**	1.22 ± 0.09
TraCE	**0.88**	**0.78±0.15**	**0.22±0.15**	3.71±0.54	**1.31 ± 0.12**

Table 1. Performance evaluation of diagnosis-based counterfactual explanations obtained using different approaches. In each case, we report results averaged across 100 test samples.

Method	Validity ↑	Sparsity ↓	Proximity ↓	Realism ↑
Vanilla	2.4	0.07±0.06	3.96±0.43	1.29±0.05
Mixup	0.8	**0.04±0.04**	3.73±0.41	1.31±0.06
UWCC	0.7	0.08±0.04	3.69±0.45	1.32±0.03
MC Dropout	1.5	0.09±0.04	4.01±0.35	1.31±0.08
Deep Ensembles (5 models)	0.4	0.06±0.05	3.81±0.27	1.35±0.03
TraCE	**0.12**	**0.04±0.05**	**3.6±0.3**	**1.39 ± 0.04**

Table 2. Performance evaluation of age-based counterfactual explanations obtained using different approaches. In each case, we report results averaged across 100 test samples.

Figures

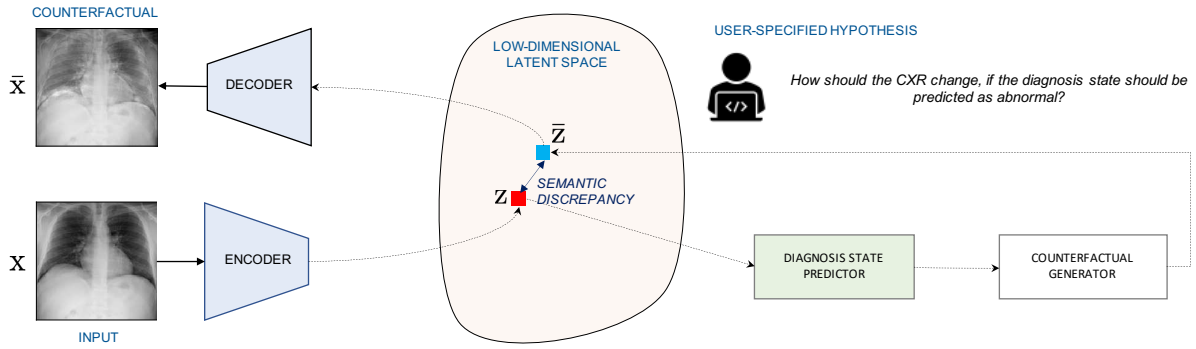


Figure 1. An overview of TraCE applied for introspective analysis of chest X-ray (CXR)-based predictive models. Driven by a user-specified hypothesis, our approach takes an X-ray image as input and produces counterfactual explanations based on exploration in a pre-trained CXR image latent space. In this example, hypothesis on the diagnosis state (*normal* or *abnormal*) is used to generate the explanation. By controlling the level of discrepancy between z and $\bar{z}$, TraCE can progressively synthesize counterfactuals with varying likelihoods of being assigned to the *abnormal* group.

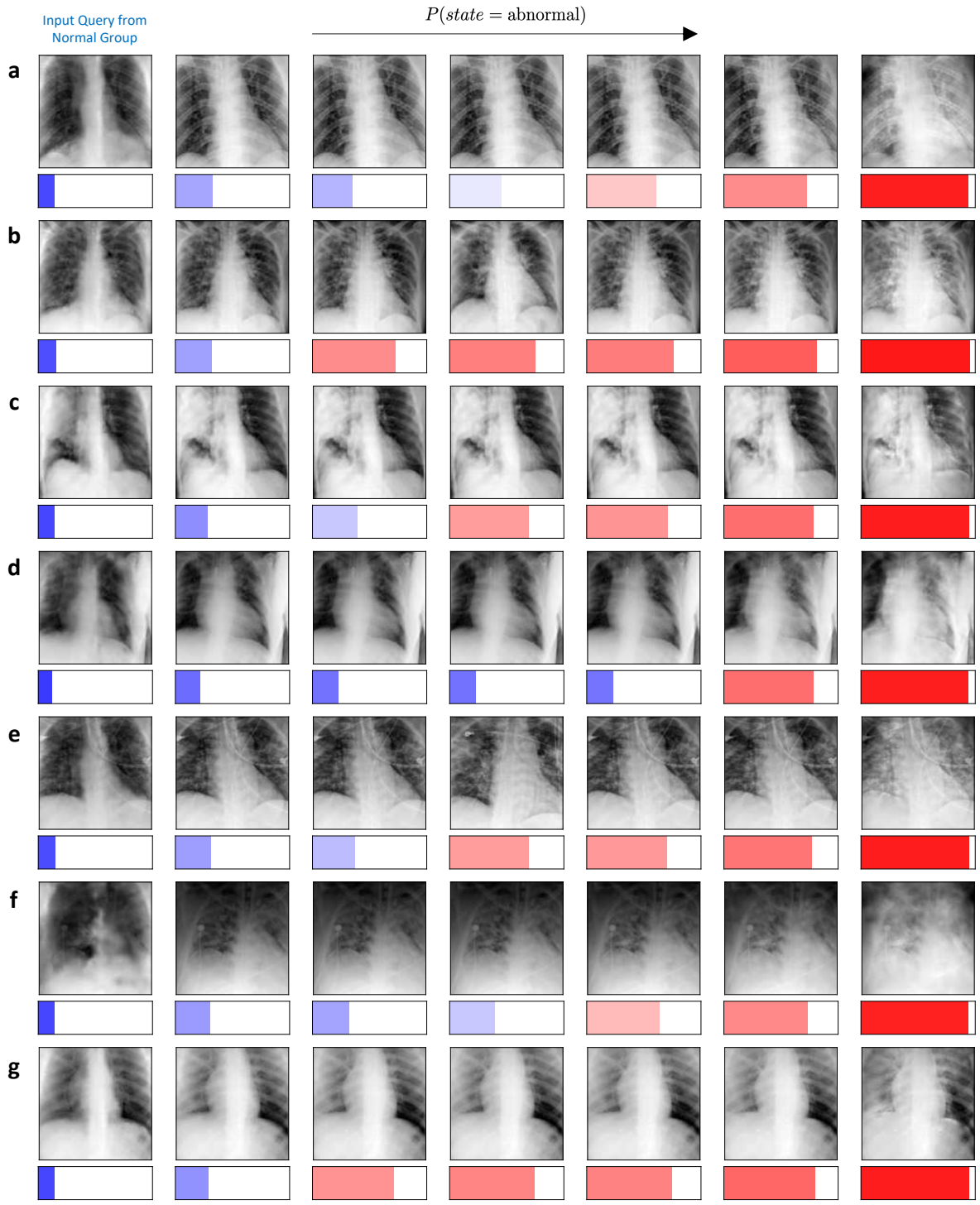


Figure 2. (a)-(g) Diagnosis-based counterfactual explanations generated using TraCE by progressively introducing relevant patterns into different query images (first image in each row) of healthy subjects to increase the likelihood of being assigned to the *abnormal* group ($P(state = abnormal)$).

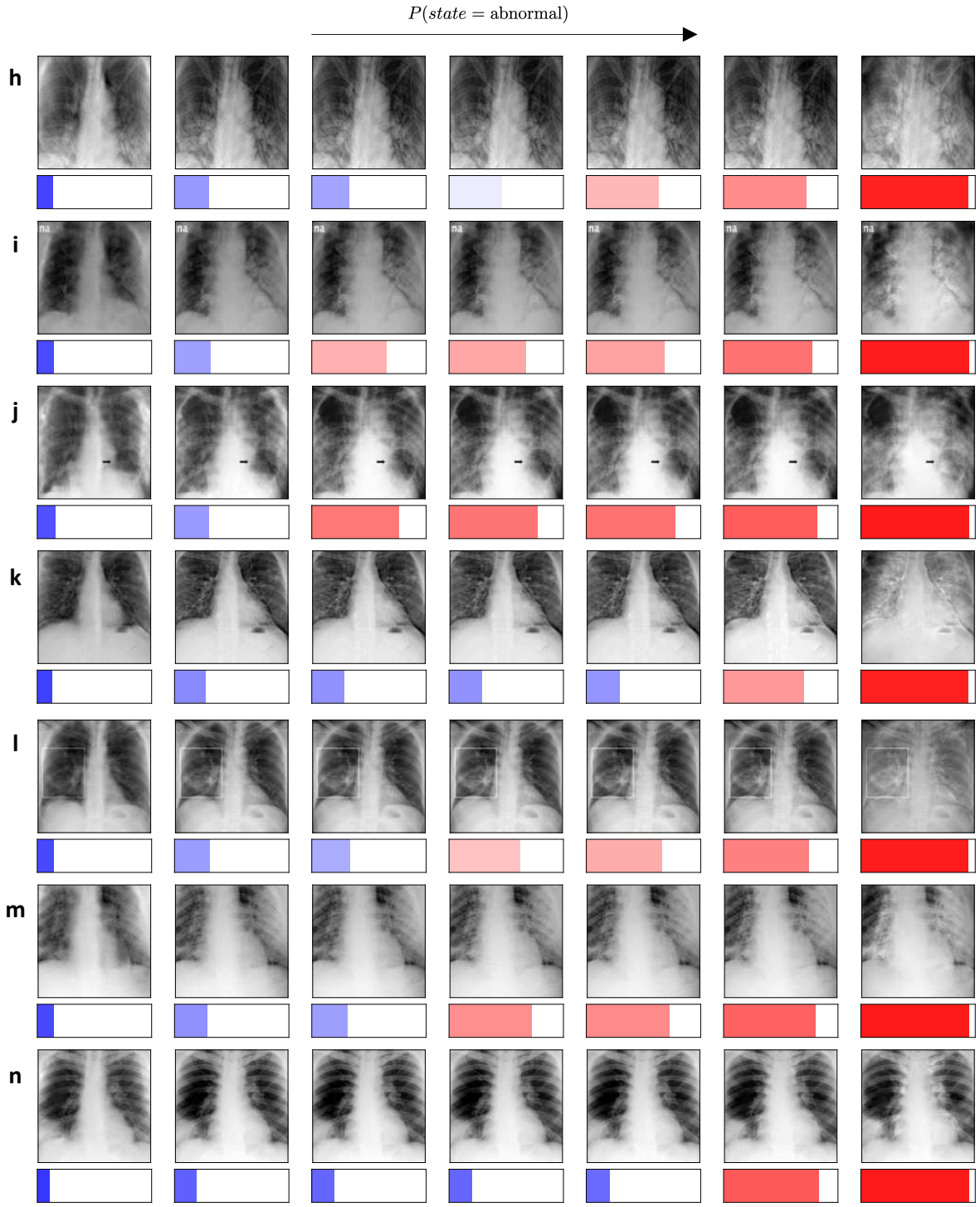


Figure 2. (h)-(n) Diagnosis-based counterfactual explanations generated using TraCE by progressively introducing relevant patterns into different query images (first image in each row) of healthy subjects to increase the likelihood of being assigned to the *abnormal* group ($P(state = abnormal)$).

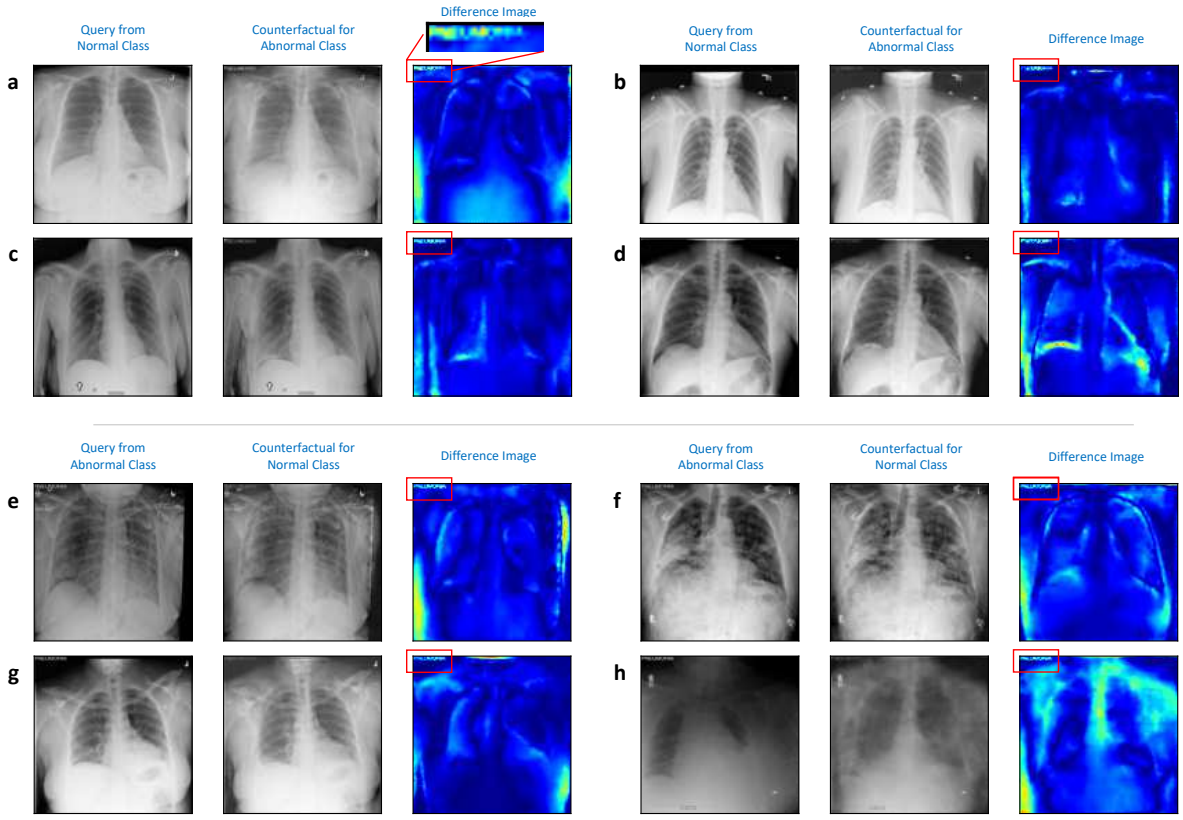


Figure 3. Using TraCE to detect shortcuts in deep predictive models. In this experiment, we synthetically introduced a nuisance feature (overlaid the text PNEUMONIA in the top-left corner) into all images from the *abnormal* group, and used this data to train the predictive model. Given the entirely data-driven nature of machine-learned solutions, there is risk of inferring a decision rule based on this irrelevant feature in order to discriminate between *normal* and *abnormal* groups. (a-d) Here, we used randomly chosen query images from the *normal* class and generated counterfactuals for the *abnormal* class. In each case, we show the query image, the counterfactual explanation from TraCE and the absolute difference image between the two; (e-f) Here, we introduced the nuisance feature into CXR images from the *abnormal* group and synthesized counterfactuals for the *normal* class. We observe that TraCE can effectively detect such shortcuts – counterfactuals for changing the diagnosis state are predominantly based on manipulating the text on the top-left corner of the query images.

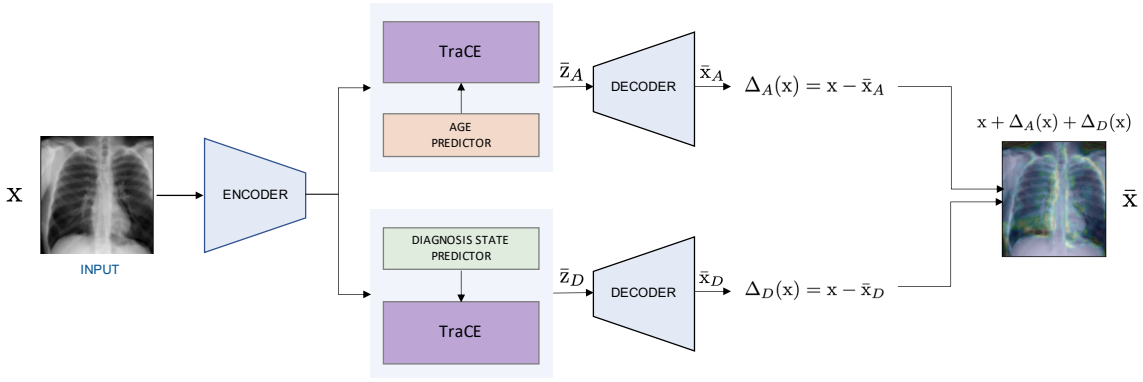


Figure 4. Using TraCE to infer relationships between a patient attribute (*e.g.*, age) and disease states. For this analysis, we construct two independent predictive models, *i.e.*, age and diagnosis state, and synthesize counterfactuals based on hypothesis on each of the predictions (*e.g.*, patient age should be predicted as 70 while the diagnosis state should be *abnormal*. Finally, we combine the changes induced in the two counterfactuals, $\Delta_A(\mathrm{x})$ and $\Delta_D(\mathrm{x})$ respectively, and check if incorporating age-specific patterns strengthens the evidence for the *abnormal* class

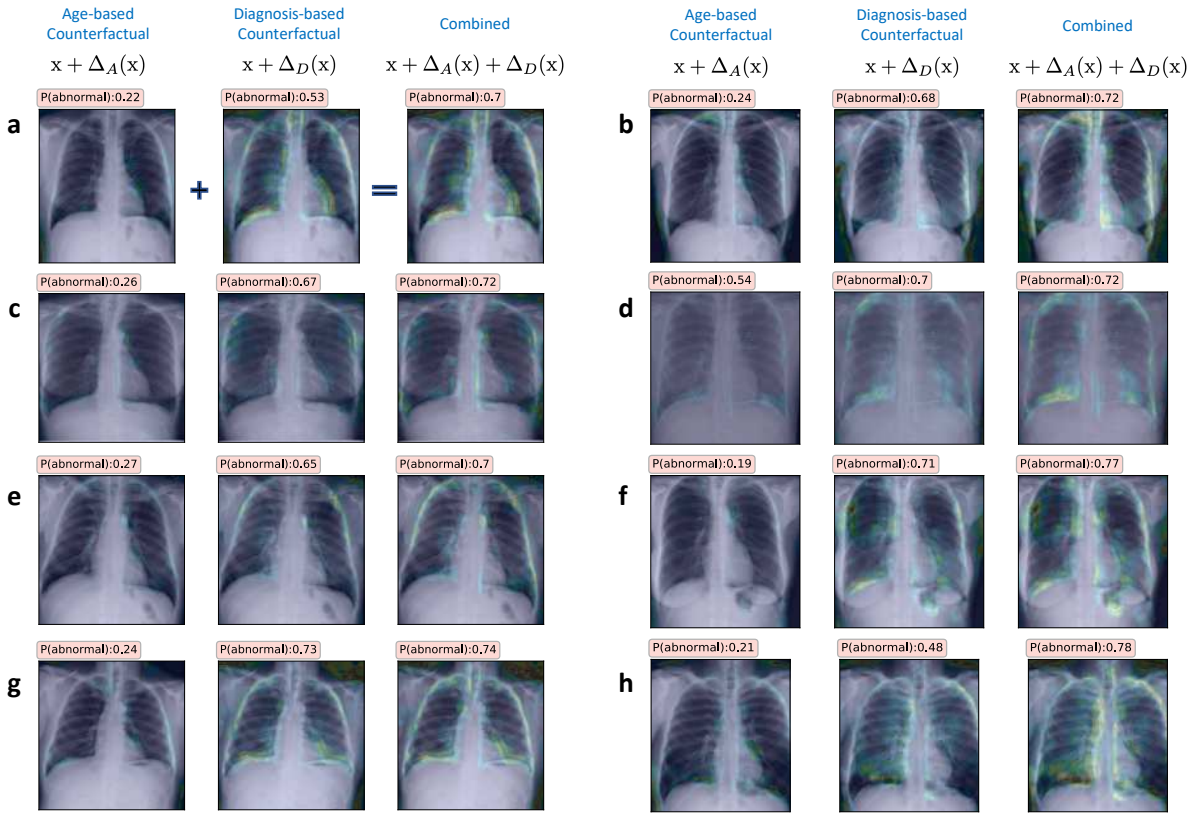


Figure 5. (a-h) Explanations generated using TraCE by introducing age-specific attributes into the counterfactuals synthesized for changing the diagnosis state of a *normal* subject to be *abnormal*. Interestingly, we find that there exists a correlation between the two attributes, as evidenced by the consistent increase in the likelihood $P(state = abnormal)$ when compared to counterfactuals that rely only on patterns from the diagnosis state predictor. In each case, we highlight the changes $\Delta_A(x), \Delta_D(x), (\Delta_A(x) + \Delta_D(x))$ and display the likelihood $P(state = abnormal)$.

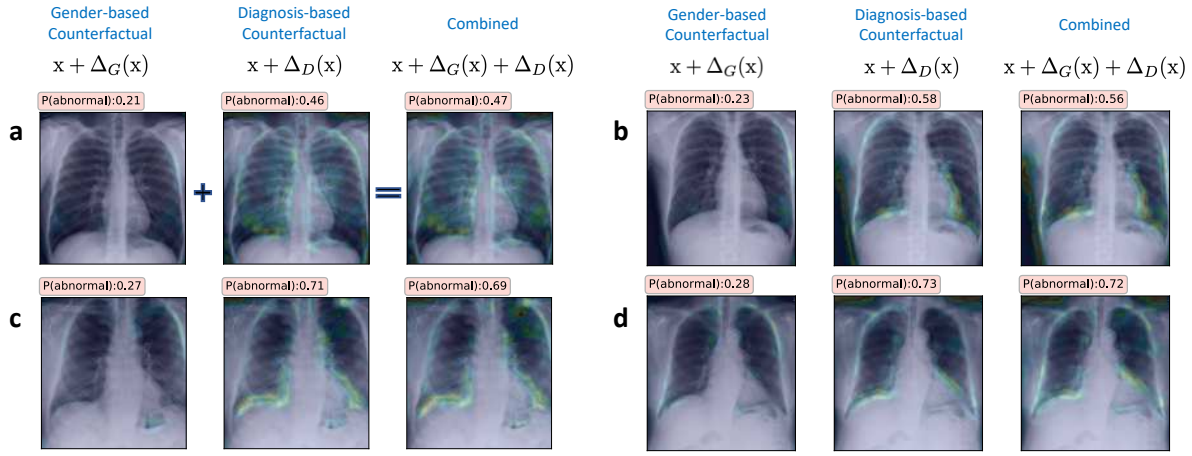
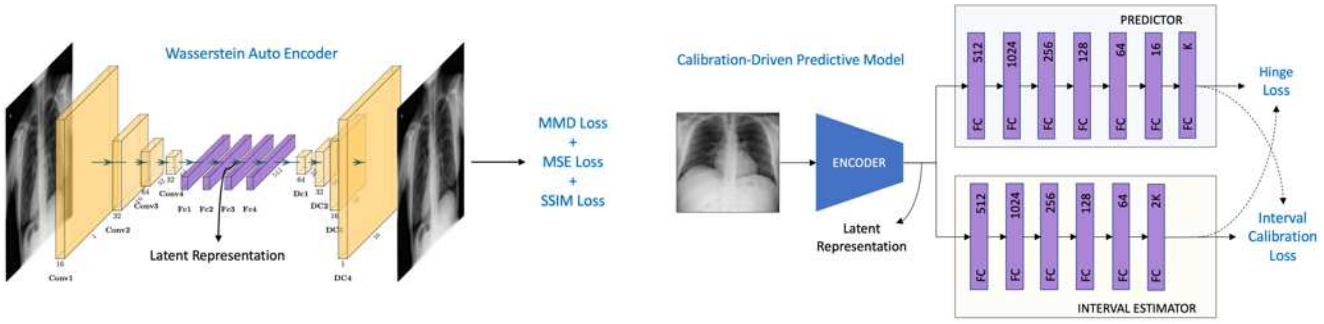


Figure 6. (a-d) Explanations generated using TraCE by introducing gender-specific attributes into the counterfactuals synthesized for changing the diagnosis state of a *normal* subject to be *abnormal*. In contrast to the age attribute, image manipulations associated with change in gender (female $\rightarrow$ male) do not cause any apparent change to the likelihood of being assigned to the *abnormal* group. In each case, we highlight the changes $\Delta_A(x), \Delta_D(x), \Delta_A(x) + \Delta_D(x)$ and show the likelihood $P(state = abnormal)$.



(a) (b)

Figure 7. Framework Design for TraCE. (a) First, we train an auto-encoding neural network[56], and construct a low-dimensional, continuous latent space for CXR images. Note, we used a combination of maximum mean discrepancy (MMD), mean squared error (MSE) and structural similarity (SSIM) losses to train the network parameters; (b) Next, we adapt the Learn-by-Calibrating[63] approach to train a classifier that takes as input the latent representation from the encoder and outputs a patient-specific attribute along with prediction intervals.