Using generative AI to investigate medical imagery models and datasets

Oran Lang,[a,**] Doron Yaya-Stupp,[a] Ilana Traynis,[b] Heather Cole-Lewis,[a] Chloe R. Bennett,[c] Courtney R. Lyles,[a,d] Charles Lau,[a] Michal Irani,[e] Christopher Semturs,[a] Dale R. Webster,[a] Greg S. Corrado,[a] Avinatan Hassidim,[a] Yossi Matias,[a] Yun Liu,[a] Naama Hammel,[a] and Boris Babenko[a,*]

[a]Google, Mountain View, CA, USA
[b]Work Done at Google Via Advanced Clinical, Deerfield, IL, USA
[c]Work Done at Google via Pro Unlimited, Folsom, CA, USA
[d]University of California San Francisco, Department of Medicine, San Francisco, CA, USA
[e]Weizmann Institute of Science, Israel

eBioMedicine
2024;102: 105075
Published Online 1 April 2024
https://doi.org/10.1016/j.ebiom.2024.105075

Summary

Background AI models have shown promise in performing many medical imaging tasks. However, our ability to explain what signals these models have learned is severely lacking. Explanations are needed in order to increase the trust of doctors in AI-based models, especially in domains where AI prediction capabilities surpass those of humans. Moreover, such explanations could enable novel scientific discovery by uncovering signals in the data that aren't yet known to experts.

Methods In this paper, we present a workflow for generating hypotheses to understand which visual signals in images are correlated with a classification model's predictions for a given task. This approach leverages an automatic visual explanation algorithm followed by interdisciplinary expert review. We propose the following 4 steps: (i) Train a classifier to perform a given task to assess whether the imagery indeed contains signals relevant to the task; (ii) Train a StyleGAN-based image generator with an architecture that enables guidance by the classifier ("StylEx"); (iii) Automatically detect, extract, and visualize the top visual attributes that the classifier is sensitive towards. For visualization, we independently modify each of these attributes to generate counterfactual visualizations for a set of images (i.e., what the image would look like with the attribute increased or decreased); (iv) Formulate hypotheses for the underlying mechanisms, to stimulate future research. Specifically, present the discovered attributes and corresponding counterfactual visualizations to an interdisciplinary panel of experts so that hypotheses can account for social and structural determinants of health (e.g., whether the attributes correspond to known patho-physiological or socio-cultural phenomena, or could be novel discoveries).

Findings To demonstrate the broad applicability of our approach, we present results on eight prediction tasks across three medical imaging modalities—retinal fundus photographs, external eye photographs, and chest radiographs. We showcase examples where many of the automatically-learned attributes clearly capture clinically known features (e.g., types of cataract, enlarged heart), and demonstrate automatically-learned confounders that arise from factors beyond physiological mechanisms (e.g., chest X-ray underexposure is correlated with the classifier predicting abnormality, and eye makeup is correlated with the classifier predicting low hemoglobin levels). We further show that our method reveals a number of physiologically plausible, previously-unknown attributes based on the literature (e.g., differences in the fundus associated with self-reported sex, which were previously unknown).

Interpretation Our approach enables hypotheses generation via attribute visualizations and has the potential to enable researchers to better understand, improve their assessment, and extract new knowledge from AI-based models, as well as debug and design better datasets. Though not designed to infer causality, importantly, we highlight that attributes generated by our framework can capture phenomena beyond physiology or pathophysiology, reflecting the real world nature of healthcare delivery and socio-cultural factors, and hence interdisciplinary perspectives are critical in these investigations. Finally, we will release code to help researchers train their own StylEx models and analyze their predictive tasks of interest, and use the methodology presented in this paper for responsible interpretation of the revealed attributes.

*Corresponding author.
**Corresponding author.
E-mail addresses: bbabenko@google.com (B. Babenko), oranl@google.com (O. Lang).

Funding Google.

Keywords: Artificial intelligence; Medical imagery; Explainability; Interpretability; Deep learning; Generative AI

Research in context

Evidence before this study

We searched Google Scholar and PubMed for published articles and book chapters available in English related to manifestation of systemic disease and demographic information in chest radiography, external eye and fundus photography and related to interdisciplinary domain expert panels for explainable AI in health, up to May 22, 2023. Our specific search terms were "cxr", "race", "ethnicity", "bone mineral density", "eyelid margin", "meibomian glands", "conjunctiva", "cataract", "diabetes", "anemia", "retinal arteries", "retinal veins", "choroid", "smoking", "cardiovascular", "sex", and "gender". Additional search terms included "interdisciplinary", "domain expert", "health", "bias", "equity", and "social determinants". Prior studies identified well known signs of or associations with systemic disease in fundus photography and chest radiography, either by human observers or AI based models. Our previous work established that AI based models could predict systemic medical conditions from external eye photographs. Although various explainability approaches have been used to interrogate medical models, there have been no studies to date which employed a generative model to identify independent visual attributes that are most relevant for AI model classification. While prior studies highlight the need for domain expertise in explainability, there is a paucity of practical methods for implementing multidisciplinary expert review, which specifically includes social scientists and sociotechnical experts who contextualize factors related to the social determinants of health and human–computer interaction.

Added value of this study

We developed an approach to investigate medical imaging models and datasets where a StyleGAN-based model (StylEx) is first used to generate visualizations of independent attributes which influence a medical imaging-based classifier, followed by a multidisciplinary expert review of each attribute. To demonstrate this approach, we applied it to a number of classifiers across multiple image modalities (external eye photographs, fundus images, and chest radiographs) and prediction targets ranging from demographic information, such as race and sex, to systemic conditions such as elevated blood pressure. Our expert panel review consisted of technical, medical, and socio-technical subject matter experts who critically evaluated and discussed the highlighted features and demographic or medical association. This unique approach unearthed associations that would not likely be identified from review by technical or clinical experts alone. Our study demonstrates a new approach for distinguishing the most relevant, and sometimes novel, visual attributes and holistically evaluating for bias as well as directions for future research.

Implications of all the available evidence

Our work shows how a generative model can be used to identify discrete medical imaging features associated with demographic information and systemic conditions. This approach allows researchers to improve their understanding of AI based models and extract new knowledge not previously identifiable by human experts. Further, we outline a multidisciplinary approach for attribute evaluation that takes into consideration the pathophysiologic and socio-cultural factors that inform healthcare delivery and thus impact our datasets, diagnosis or labels, and model development. To enable further research using this generative model, we will release the code to allow other researchers to utilize it for their modeling tasks.

Introduction

Interest in developing Artificial Intelligence (AI) models, particularly deep learning based models for medical prediction tasks has grown considerably over the past several decades. The hope is that AI-based models can help medical providers and researchers rapidly and accurately analyze complex medical data, particularly medical imaging (e.g., radiography, endoscopy, fundus photography). Many machine learning models have performed as well as highly trained physicians,[1–3] while others have reportedly achieved higher performance than their human counterparts.[4–6] While many of these prediction tasks are based on reproducing tasks that human experts are capable of, several AI models surprisingly can identify diseases and patient characteristics far beyond what the imaging modality was known to reveal.[7–9] In some cases this is due to the model picking up on confounders,[10] selection/information bias in dataset selection[11,12] or structural, cultural, and historic biases reflected in data[13,14]; whereas in others the model is picking up on previously unknown manifestations of a condition.[15,16]

Explainability methods help provide insights into AI models[17] and the data on which these models are trained. In the medical domain, one desirable property of an explainability method is an ability to enable a panel

of experts (e.g., clinicians, statisticians, social scientists, and human factors engineers) to investigate the visual cues that the model learns to perform its task. The panel can determine whether these cues are expected given the task, or whether the cues warrant further investigation, for example, because the dataset is reflecting systematic or structural bias,[18] confounding,[19] or risk factors.[20] Such cases are of particular interest when developing a model which determines healthcare access, and might require steps to modify the model or the dataset on which the model was trained to ensure accuracy and limit bias. Alternatively, such a technique might lead to hypotheses for novel research directions that could improve our understanding of physiology and disease.

Among explainability methods in computer vision,[21] the output of most techniques[22–24] is a heatmap of per-pixel importance, for instance based on the saliency of the image features. While these explainability methods can provide information about the *spatial location* (the "where") of important features, they do not typically explain higher-level features of the pixels in the highlighted region is predictive, such as texture, shape, or size (the "what"), limiting their utility in explaining possible underlying mechanisms. Recently a new line of research[25–29] showed how generative models can be used to transform images of one class into another, i.e., creating counterfactual images for these classes. While these methods show how an image changes when the class is changed, they cannot disentangle individual fine-grained attributes.

To address these challenges, in this paper we build on a StyleGAN-based approach called StylEx,[30] which is able to generate visualizations of fine-grained attributes. A notable benefit of this StyleGAN-based approach beyond that of CycleGAN-based approaches is that attributes are generated using different axes of the style space. This separation enables the extraction of separate attributes that may otherwise have been merged into one, thus avoiding issues with some attributes being hidden by the most visually striking or familiar change. These attributes can potentially enable exploration of new avenues of scientific inquiry for AI models that have been found to identify surprising associations beyond what is identifiable by humans. In this study, we suggest a framework to use AI classifiers to learn new insights from medical data (Fig. 1). Our approach is based on 4 fundamental steps that encompass both technical and interpretation tasks: (i) Train a classifier on a given image dataset to perform the desired machine learning task. If the overall accuracy of the trained classifier is high, then the information for visual assessment is assumed to be present in those images. (ii) Train a "StylEx" model—a StyleGAN-based image generator guided by the classifier—on this dataset of images. The loss function encourages the GAN to generate images which both resembles the original input and are correctly classified by the classifier, forcing the generator to put an emphasis on the visual attributes which are most relevant for classification. (iii) Automatically extract the top attributes which affect the classifier. (iv) Generate a visualization of these attributes for subject matter experts' evaluation, discussion, and consensus. As part of this framework, an interdisciplinary panel including both data domain experts and social scientists then interpret the attributes for consistency with prior literature and reason about whether any could be novel discoveries. Importantly, the interdisciplinary nature of the panel reduced blind spots based on individual expertise.

To demonstrate this approach, we selected three different imaging modalities and 8 different prediction tasks based on recent literature, including both "positive control' predictions and predictions that clinicians are not trained to perform using these images. The tasks were: elevated glycosylated hemoglobin (HbA1c), low hemoglobin (Hgb), and cataract presence in external eye photography; systolic hypertension, smoking status, and self-identified sex in retinal fundus photography; and abnormality and race/ethnicity in chest radiography (CXR). Generated counterfactual visualizations (i.e., what that image would look like with the attribute increased or decreased) for each attribute were reviewed by an interdisciplinary panel of machine learning engineers, social and socio-technical scientists, and modality-specific clinical specialists: a comprehensive ophthalmologist for fundus and external eye photographs, and a cardiothoracic radiologist for CXR. Based on the attribute visualizations, our specialists formed hypotheses as to what signals these attributes were reflecting. This resulted in both expected findings (i.e., visual cues that one may expect to find for a given task, which served as sound positive controls) and surprising discoveries for each imaging modality. Hypotheses generated by the interdisciplinary panel suggest future areas for research needed to better understand the observed results.

Methods

In this section we present our end-to-end framework for automatic visual explanation of medical findings. Our method consists of 4 stages, 3 algorithmic and a final manual stage involving expert panel discussion (Fig. 1).

Stage 1: training a classifier

In order to verify that we can indeed extract the visual information to explain the task at hand, we first train a classifier C to predict the task label. Then we test its performance on a held out set and measure its generalization capability. We apply our method on classifiers which achieve high performance (defined for the purposes of this study as being approximately above 0.8, a threshold described qualitatively as "excellent"[31]). The motivation behind this decision is that we want to

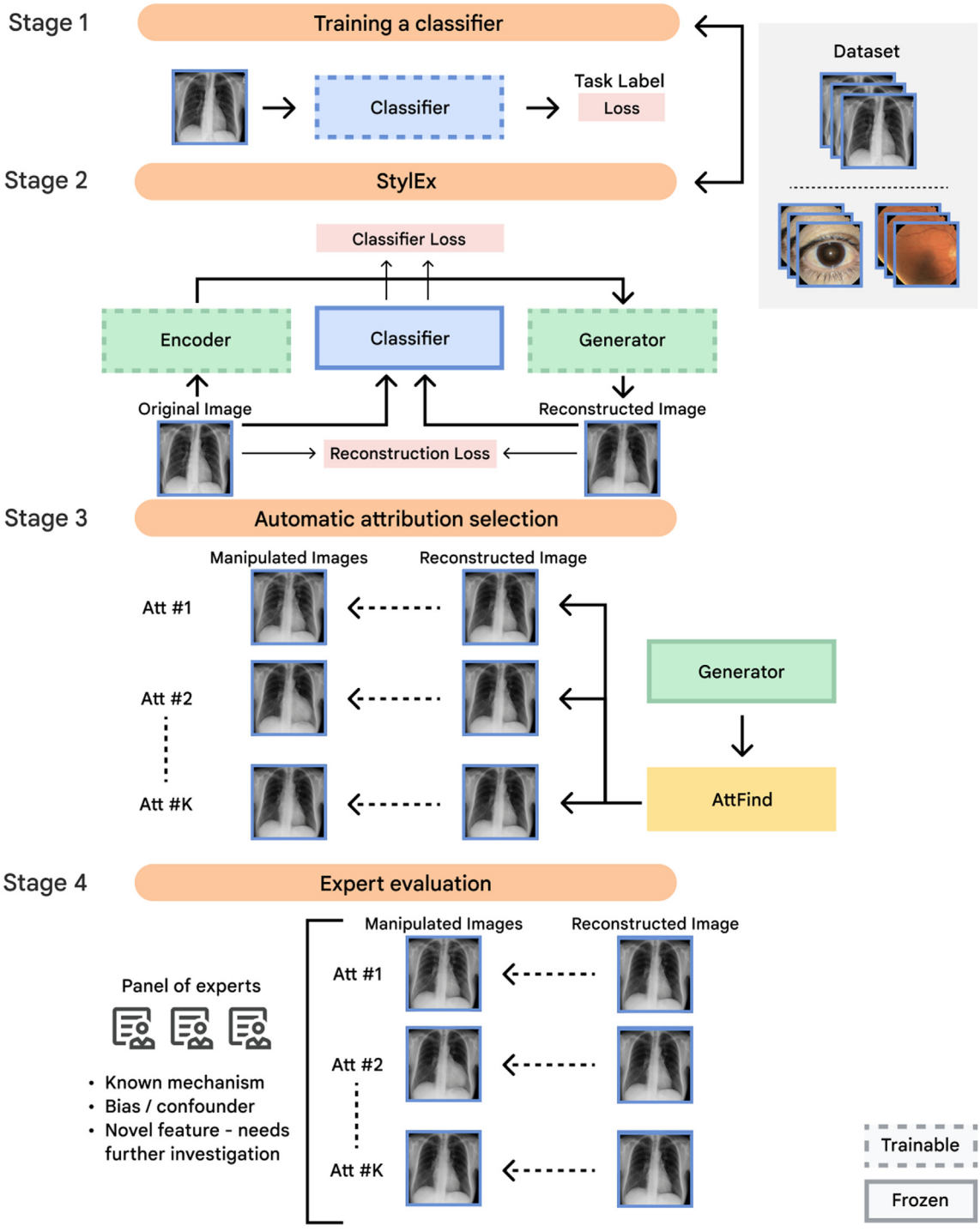


Fig. 1: Our method consists of 4 stages: (1) train a classifier on the predictive task of interest, (2) train a StylEx model on the dataset using the classifier as guidance, (3) automatically extract the top attributes which affect the classifier, and (4) generate a visualization of these attributes for experts' evaluation.

ensure that the classifier has indeed learned information relevant to the task of interest, so that it is meaningful to visualize this learned information. Further details are provided in Supplementary Note 3.

Stage 2: StylEx

We used StylEx[30] to generate counterfactual visual explanations. StylEx relies on the fact that StyleGAN2 contains a disentangled latent space called "StyleSpace",[32] which contains semantically meaningful attributes of the images. We train a StyleGAN2 model on the same dataset that is used to train the classifier, to generate realistic looking images from that domain. However, the original formulation of StyleGAN2 training does not depend on the classifier C (from Stage 1), and so it may not necessarily represent and generate subtle attributes that are important for the decision of the specific classifier C which we wish to use for explainability. Therefore, we train a StyleGAN2-like generator G to preserve the decisions of the classifier C on the training images, encouraging its StyleSpace to explicitly represent classifier-relevant attributes.

This is achieved by training the StyleGAN2 generator G with three additional changes: (i) Instead of using a mapping network from a noise vector to the latent space W, we replaced it with an encoder E, trained together with the GAN with a reconstruction-loss. This forces the generated output image to be visually similar to the input, and allows us to apply the generator on specific input images to generate counterfactuals. (ii) Together with the reconstruction loss, the classifier-output similarity loss encourages that both the generated image itself and the neuronal activations through the classification model for the generated image resemble that of the original input images. Thus, visual features important for the classifier C (such as those used to identify medical conditions) should be included in the generated image, and our expert panel's (detailed next) goals are to interpret for the presence of known and possibly novel visual features. (iii) In addition, we use a conditional version of StyleGAN2 architecture, which concatenates the classifier output of the original image to the latent vector W. This ensures that the StyleSpace vector, which is a linear mapping of this concatenated vector, uses the information from the classifier directly. To automatically find attributes which are correlated with the labels and some pre-existing set of confounding variables, we used a slightly modified method to the original StylEx work, as described in Supplementary Note 1.

Stage 3: automatic attribute selection

After training the StylEx model, we search the StyleSpace of the trained generator G for attributes that affect the output of the classifier C. To do so for a given image, we manipulate each StyleSpace coordinate of G by increasing and decreasing it by a fixed factor times the standard deviation of this feature, and measure its effect on the classification prediction (henceforth "CP" for brevity), and select the top k attributes that maximize the change in CP, in a way which resembles Individual Conditional Expectation (ICE).[33] This provides the top k image-specific attributes. We can then regenerate the input image while modifying one attribute at a time, visually displaying the meaning of each individual attribute (i.e., a counterfactual visualization). The class-specific top k attributes are obtained by repeating this process for a large number of images per class and measuring the percentage of images where CP was changed by more than some threshold (0.15), which was the same for all classifiers. To avoid outlier attributes which transformed the image out of the domain distribution, we filtered out attributes which changed the classifier prediction in opposite directions.

Stage 4: evaluation by interdisciplinary human experts

In our final stage, the interdisciplinary expert panel review allows us to critically assess the model findings and identify areas of bias as well as directions for future research. We visually display the top k attributes to a panel of human experts by generating counterfactual images where each attribute individually is increased or decreased for a set of example images (Fig. 1). For each domain, we consult with clinical, socio-technical, social science specialists in that domain, and machine learning engineers to inspect the attributes (see Supplementary Table S3). For each attribute our clinical specialists hypothesize what visual signal is captured by the attribute and the full expert panel hypothesizes possible interpretations for why the signal is useful to the classifier. The panel discusses all attributes, with specific focus being paid to attributes the panel has identified as potentially being representative of bias due to systematic or structural factors present in the dataset or confounding not accounted for in the model. Finally, the full panel reviews and identifies a list of research and validation hypotheses. These research and validation hypotheses can be tested either by comparing them to known phenomena from the literature, or via further studies to prove or disprove them. Without these kinds of interdisciplinary review and discussion, certain social or environmental factors may be assumed to be or perpetuated as biological/physiological phenomena (see Supplementary Note 5).

For interpretation and hypothesis generation, we draw from the socio-ecological theory[34] and social and structural determinants of health framework.[35] These models argue that health is not simply biologically determined, but rather it is shaped by myriad social, political and environmental factors across the lifecourse. These models are particularly useful for the interpretation of socially-constructed attributes such as race. However, much social science research has illustrated

the ways in which social, political, and environmental exposures are physically embodied and can change our biology[36] and so we apply our theoretical framework to all attributes. Additionally, research in the human factors engineering and human–computer interactions fields have demonstrated the myriad human and environmental effects, such as human error, maintenance of the technology, and environmental conditions (lighting, temperature, storage[37–39]) the accuracy and validity of the results produced by the technology. Using this information as our theoretical underpinning, prompts to the expert panel (Supplementary Note 6) are created to aid in hypothesis generation and group discussion of interpretations. As such, interpretations may consist of both biologically and socially constructed criteria. This includes physiological conditions related to the prediction task at hand, specific ways the imaging device and participants interact, representativeness of the study population, social and structural determinants of health, or a combination of them all.

Datasets and prediction tasks

We tested our method on three different input imaging modalities, and a variety of tasks. For each modality, we used datasets from prior research studies and retrained a classifier using the methods used in these works with minor modifications (more details are included in Table 1 and Supplementary Note 2). These classifiers are oftentimes trained in a multi-task fashion, such that they can predict many different targets (i.e., there is a common "backbone", and a different "head" for each prediction task). For the purposes of our work, we selected a subset of tasks that had a sufficiently high AUC to ensure that the classifier C learned a meaningfully strong signal. In all cases, we trained the StylEx model on the same training data that was used to train the classifier models. Specifically, for fundus photos, we follow the approach reported in Poplin et al.,[9] using the UK Biobank dataset; the prediction tasks in this domain include self-reported sex, systolic blood pressure (SBP)≥140, and smoking status. For external eye photos, we use the model reported in Babenko et al.,[41] which was trained on the EyePACs/LACDHS dataset; the prediction tasks in this domain include presence of cataracts, HbA1c ≥ 9, Hgb < 11. Finally, for CXR we explore two prediction tasks: CXR abnormality and race. For the former we follow the approach described in Nabulsi et al.,[42] using the IND1 and CXR-14 datasets. For the latter we trained a model similar to that described in Gichoya et al.[13] For uniformity across tasks and setups, all tasks were framed as binary classification problems.

Overall, most datasets were collected in a healthcare setting, and thus skewed older. The datasets also had unequal proportions of sex, and had racial/ethnic distributions which may not precisely represent the countries from which data were collected. While all the datasets used here have been previously described in the literature, because nuances of the populations and data collection processes affect the interpretation of the attributes, we include a brief summary of these datasets in Supplementary Note 2.

Imaging modality	External eye		Fundus photograph		Chest radiograph (CXR)			
Dataset	EyePACS/LACDHS		UKBiobank		Apollo & CXR-14		MIMIC III[40]	
Geography	Los Angeles, CA, USA		United Kingdom		India/USA		Boston	
Number of patients	49,025		62,631		228,108		47,621	
Number of visits	76,458		64,173		249,558		69,626	
Number of images	151,267		119,092		282,604		162,639	
Age (median/IQR)	57.8/12.5		59.5/13.0		48.0/21.0		63.0/26.0	
Race/ethnicity	Data available for: 42,715 Hispanic: 31,708 (64.7%) Black: 4504 (9.2%) Asian/Pacific islander: 3457 (7.1%) White: 2375 (4.8%) Other: 610 (1.2%) Native American: 61 (0.1%)		Data available for: 62,255 White: 57,344 (91.6%) Other: 2309 (3.7%) Asian/Pacific islander: 1914 (3.1%) Black: 688 (1.1%)		N/A		Data available for: 47,621 White 33,552 (70.5%) Black 8840 (18.6%) Hispanic 3315 (7.0%) Asian 1914 (4.0%)	
Self-reported sex = Male	19,267/49,014 (39.3%)		28,633/62,631 (45.7%)		140,872/228,081 (61.8%)		22,620/47,621 (47.5%)	
Task-specific statistics	**Cataract presence**	**HbA1c > 9**	**Hgb < 11**	**Sex = Male**	**Smoker**	**Systolic BP > 140**	**Abnormal**	**Race = Black**
Label								
Train counts [positive/total (%)]	2798/62,213 (4.5%)	5273/19,324 (27.3%)	2001/26,289 (7.6%)	25,032/54,771 (45.7%)	4769/55,976 (8.5%)	23,162/55,935 (41.4%)	79,444/245,065 (32.4%)	7972/42,849 (18.6%)
Tune counts [positive/total (%)]	308/14,245 (2.2%)	1188/4165 (28.5%)	451/5831 (7.7%)	3601/7860 (45.8%)	681/8021 (8.5%)	3278/8018 (40.9%)	2285/4493 (50.9%)	868/4772 (18.2%)
AUC [% (CI)]	87.3 (85.8–88.9)	70.2 (69.0–71.5)	79.1 (77.6–80.7)	93.9 (93.6–94.2)	70.8 (69.3–72.3)	77.8 (77.0–78.5)	96.9 (96.5–97.4)	96.1 (95.6–96.7)

Table 1: **Dataset characteristics.**

Ethics

Given this study was retrospective and used de-identified datasets, the need for further review was waived by the Advarra Institutional Review Board (IRB).

Role of funder

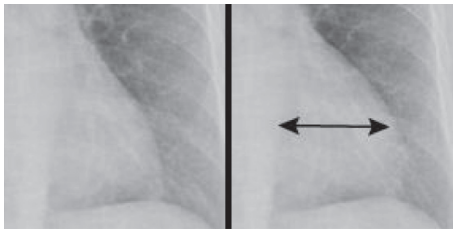
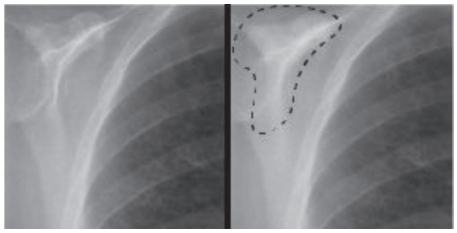
Google was involved in the design and conduct of the study; management, analysis, and interpretation of the data; preparation, review, and approval of the manuscript; and decision to submit the manuscript for publication.

Results

Across the imaging domains and prediction tasks explored in this paper, our method unveiled a number of attributes, leading to several novel areas for additional scientific inquiry according to interdisciplinary expert consensus. Tables 2, 3 and 4 lists a sample of the discovered attributes, and the full list can be found in the Supplementary Material. For each attribute, we provide a "visual explanation", by showing two generated versions of an image with increased/decreased values of the detected attribute (animated images showing these changes are attached as Supplementary Material), along with the panel's description and interpretation of the attribute, and directions for future exploration. In addition to their primary task, we trained our classifier to also predict additional variables (such as sex and age if not already being predicted), which enabled us to explore potential confounding effects (see Methods and Supp. Note 1). These results illustrate the breadth and diversity of data types and attributes for which our method is applicable and the variety of research hypotheses inspired by it. We review the highlights of these findings for each modality below, noting that while our results suggest testable hypotheses generated by the interdisciplinary panel, they do not establish causality (further discussed in the Discussion).

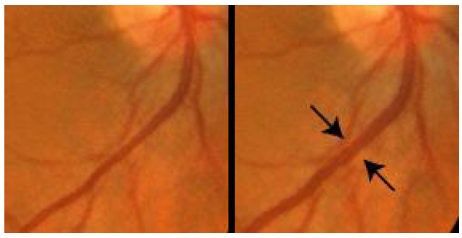
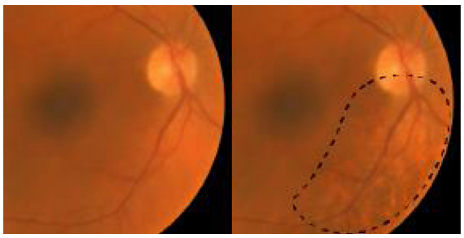
Fundus photo

For the fundus photo modality we explored three tasks: predicting smoking status, systolic blood pressure over 140 mmHg, and sex.[9] For the first two tasks, our method produced attributes related to retinal vasculature. Specifically, retinal vein dilation was correlated with a higher CP of being a smoker, whereas arteriolar narrowing was correlated with higher CP of elevated

Image Modality	CXR	
Images of increased / decreased attribute magnitude		
Prediction task	Normal / Abnormal	Race
Attribute location	Left ventricle	Bones
Human description of the attribute ("What")"	Left ventricular enlargement, demonstrated by leftward displacement of the left heart border, with increasing CP of abnormal CXR.	Decreased skeletal lucency and increased conspicuity of ribs, scapulae, humeri, or thoracic vertebral bodies on CXR with increasing CP of Black race.
Consolidated panel notes	Left ventricular enlargement often occurs in the setting of congestive heart failure, ischemic heart disease, or hypertension. This attribute is unlikely to represent a regular beating of the heart because the enlargement is focal. There are severe racial disparities in heart disease and hypertension, which should be considered in the context of the dataset demographics.	Increased bone mineral density results in bones that appear more conspicuous (relative to background) on CXR. Average bone mineral density varies among racial and ethnic groups and by age. A higher average bone mineral density in Black populations may explain the association for this model; however, we cannot conclude whether the underlying cause of this is related to biological differences or environmental exposure, nutrition, or structural artifacts that are not measured.

Please see Supplementary Tables S2a and S2b for more details.

Table 2: **A sample of attributes for the CXR domain.**

Image Modality	Fundus photography	
Images of increased / decreased attribute magnitude		
Prediction task	Smoking Status (yes/no)	Sex
Attribute location	Retinal vessels	Choroid
Human description of the attribute ("What")"	Retinal vein dilation with increasing CP of being a smoker	Increase in nasal, temporal, and inferior choroidal vasculature visibility with increasing CP of male gender
Consolidated panel notes	Retinal venular caliber is associated with cardiovascular disease, which is strongly associated with smoking. Since smoking, cardiovascular disease, and diabetes are difficult to disentangle, it would be challenging, for example, to find datasets with smokers who do not have cardiovascular disease. Even if such datasets exist, social, cultural, and environmental factors between participants would likely vary greatly, creating additional confounders.	This attribute associates greater choroidal vasculature visibility with increased probability of male sex, which is the converse of what previous research in this area has suggested. There may be differences in the dataset, such as the distribution of myopia and fundus pigmentation within male and female populations, which may drive the differences identified by the model.

Please see Supplementary Table S2c for more details.

Table 3: **A sample of attributes for the fundus domain.**

systolic blood pressure. Both of these associations have been reported in the literature.[43]

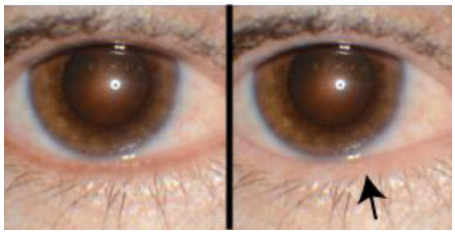
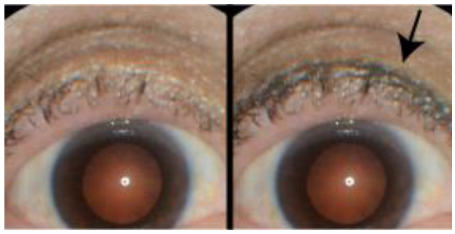
For the sex prediction task, our method found an attribute that associates greater choroidal vasculature visibility with increased CP of male sex. This is the converse of what previous research in this area has suggested.[44–47] Our panel hypothesized that dataset-specific factors, such as that related to distribution of axial length/myopia or fundus pigmentation within male and female populations, may drive the differences identified by the model, and further investigation is warranted. Further research into the association between sex, myopia,[48] and fundus pigmentation in this dataset may help understand this link.

External eye photo

In the external eye domain,[41,49] we explored three tasks: cataract presence, low Hgb (<11 g/dL), poor glycemic control (HbA1c ≥ 9%). The cataract presence task acted as a positive control since unlike the other tasks explored, cataract presence is feasible for a human to perform (and indeed is part of the reason these images were captured in this dataset). Indeed, our method found two attributes that are known to clinicians[50]: presence of cortical cataract spokes, extending from peripheral to central lens, and dimmer red reflex were both associated with higher CP of cataract presence.

For the low Hgb task, our method found an attribute showing that decreased conjunctival vessel prominence was correlated with increased CP of low hemoglobin, which our panel noted as being consistent with known biological phenomenon.[51] Another attribute appeared to show that increased eyeliner (a cosmetic used around the base of the eyelashes) thickness and density was associated with higher CP of low Hgb. Abnormally low hemoglobin levels is known to be a common condition in females of reproductive age.[52] Indeed, the CP of female sex was also correlated with this eyeliner attribute. Hence our panel hypothesized this may represent the presence of confounding picked up by the classifier that would need to be mitigated before considering the use of such a model.

Finally, for the poor glycemic control task, we report on two noteworthy attributes. The first shows that increased corneal arcus thickness from the limbus is associated with lower CP of elevated HbA1c. Generally, corneal arcus is associated with increased age,[53] and indeed we found this correlation between this attribute and CP of age above 60. The direction of the correlation with HbA1c and the attribute is therefore surprising since HbA1c tends to *increase* with age.[54] However,

Image Modality	External eye photograph	
Images of increased / decreased attribute magnitude		
Prediction task	HbA1c ≥ 9	Hgb
Attribute location	Eyelid margin	Eyelid
Human description of the attribute ("What")"	Increasing eyelid margin pallor with increasing probability of elevated HbA1c	Increasing upper eyelid makeup eyeliner thickness and density with increasing CP of low Hgb
Consolidated panel notes	The eyelid margin attribute may be depicting pathophysiologic signs of meibomian gland disease (MGD), which is associated with diabetes. Signs and symptoms of MGD or dysfunction are noted with increased frequency in individuals with diabetes. There are several other potential confouders which should be considered, such as toxic exposures to medications or environmental irritants, social and cultural practices that are related to hygiene and use of makeup or other topicals.	The eyeliner attribute identifies a known cultural association between makeup usage and low hemoglobin levels with the same demographic factors of sex and age. Additionally, the dataset was not racially and ethnically diverse, warranting questions of whether this same association would have been observed with darker skin tones or with use of different colors of eyeliner.

Please see Supplementary Tables S2d and S2e for more details.

Table 4: **A sample of attributes for the external eye domain.**

closer examination of the dataset led our panel to hypothesize that this may be due to survivorship bias—the dataset comes from a diabetic retinopathy screening program, where the sickest patients (i.e., with higher HbA1c and referrable disease) are referred out of the program, and hence older patients who are still in the program tend to be healthier. The second attribute shows that an increase in eyelid margin pallor was associated with higher CP of elevated HbA1c. Our panel hypothesized that this may be related to subtle signs of changes in meibomian glands since meibomian gland disease has been reported to be more prevalent in individuals with diabetes.[55,56] However, further investigation is required to determine the specific mechanisms involved, and these findings would not signal the use for this visual feature in isolation as a screening tool.

Chest X-ray

In the CXR domain we explored two tasks: predicting abnormality and predicting race. For predicting abnormality, StylEx highlighted three attributes, the first of which was left ventricular enlargement being correlated with higher CP of abnormality. Our panel categorized this as a known clinical phenomenon as this type of enlargement occurs in the setting of congestive heart failure, ischemic heart disease, or hypertension.[57] The second attribute showed that mediastinal widening or prominent aortic knob was associated with increasing CP of abnormality, which was also categorized as a known phenomenon since an enlarged or tortuous thoracic aorta is a common cause of mediastinal widening or prominent aortic knob in elderly individuals, particularly those with atherosclerosis.[58] The final attribute showed that more apparent over-exposure of images (i.e., darker images) were correlated with higher CP of abnormality. Our panel hypothesized that this correlation may be due to the differences between portable, or anteroposterior (AP), imagery, which tends to vary more in quality and exposure and is often taken in-patient for sicker patients, versus standard posteroanterior (PA) imagery.[59] Other hypotheses and factors to consider or warrant further investigation for this attribute are listed in Table 2.

For the task of predicting Black race from CXR, we report on three attributes. The first attribute shows that increased skeletal conspicuity is associated with increased CP of Black race. Increased skeletal conspicuity may be a radiographic indicator of increased bone mineral density, which studies have reported varies among racial/ethnic groups.[60] However, our panel noted

that due to the historical scientific racism still present in medical literature and research today[61,62] and the widely accepted understanding that race is socially, rather than biologically, constructed,[63–65] we cannot conclude these differences are related to biological differences. Differences in bone density may be the result of environmental exposures or structural artifacts[66,67] that are not measured in our dataset, suggesting further investigation is required. The next two attributes our panel categorized as being likely confounders. One showed that increased upper lung volume was associated with decreased CP of Black race. Increased upper lung volume is a sign of chronic obstructive pulmonary disease (COPD), and although some studies have suggested a lower prevalence of COPD among Black individuals,[68,69] other studies have reported decreased COPD screening in Black populations.[70] Our panel noted that there are no known racial differences in lung volume—this variation tends to be related to socio-behavioral and environmental factors.[71,72] The final attribute showed that a more superior clavicle position relative to the lung apices was correlated with decreased CP of Black race. Our panel hypothesized that this correlation could be due to associations of race distribution and AP versus PA chest radiography. A more superiorly positioned clavicle is associated with portable AP CXRs, which are usually obtained at the hospital bedside with a patient in a recumbent position. As hospitalization rates are noted to be higher in non-hispanic Black patients,[73,74] a higher frequency of AP CXR imaging was expected in Black patients, relative to other groups. As a more superior clavicle position is associated with AP imaging, a higher CP of Black race would have been expected for this feature, but was actually counter to our results. This unexpected observation could have been due to the unique characteristics of the training dataset which had a greater relative proportion of White patients to Black patients for AP images as compared to for PA images. Alternatively, thoracic kyphosis (excessive curvature of the thoracic spine often encountered in the setting of low bone mineral density) can result in more inferior, rather than more superior, clavicle position on CXR. As low bone mineral density has a weaker association with Black race than White race, one would also have expected that superior clavicle position would be associated with a higher CP of Black race–which was the opposite of what was observed. Further investigation of the dataset and model are needed to identify a complete explanation of this attribute.

Discussion

In this paper we showcase a new explainability framework incorporating generative AI and an expert consultation process to interpret medical imaging models and datasets. Our AI generates counterfactual images using GANs, and is able to discover and visualize attributes which affect the decision of a classifier model, and hence may represent an association with the underlying task. We demonstrated our method on multiple prediction tasks and imaging modalities. Importantly, our method goes beyond an understanding of what areas of an image (i.e., the "where") are responsible for a prediction (the goal of various saliency-based methods), and helps understand what change in those regions are associated with the predictions (i.e., the "what"), with an important interpretation step to understand or generate hypotheses of physiological, social or socio-technical mechanisms that link these changes to the prediction task (i.e., the "why"). This critical last step of hypothesis generation relies on the expertise and collaborative interpretation of an interdisciplinary expert panel of clinicians, statisticians, social scientists, and human factors engineers.

Some of our tasks can be performed by human experts looking for specific visual features. The fact that our method is able to rediscover such features serves as positive control experiments. For example, cardiomegaly is a known condition visible in (and defined by measurements of the heart and chest width in) CXRs, and is evident by the counterfactual for the abnormal CXR model showing an increase in the size of the cardiac silhouette. Likewise in external eye photographs, cortical spokes are visible hallmarks of one type of cataract, and thus its presence as an attribute helps sanity check that the original classification model had learned to recognize meaningful features that define the prediction tasks. In addition, our method also rediscovers known visual cues that are not directly used to determine the ground truth label, but are known to be associated with the underlying condition. For instance, smoking increases the diameter of the blood vessels,[43] which can be seen in the counterfactual attribute showing dilated vessels, but smoking status is not defined based on blood vessels' dilation status. More examples include arteriolar narrowing for elevated systolic BP and conjunctival pallor in low Hgb, both of which are known associations clinically; in both cases the attribute can be a manifestation of but do not define the respective condition. These rediscoveries provide evidence for the potential of the proposed method for scientific research.

For some of our tasks, our interdisciplinary expert panel review concluded several associations may be artifacts of dataset demographics, human interactions with technology, or the result of social and structural determinants (i.e., access to healthcare or exposure to racism and discrimination that is linked to differential hospitalization risks and utilization). For example, our method shows that low exposure increases the probability for a classifier to classify a chest X-ray as "abnormal", which could result from the patient orientation and CXR view (PA versus AP), which is in turn associated with patient mobility, inpatient versus

outpatient status, and other factors. Because these factors often depend on local patient demographics, these associations may be spurious. Another example is the attribute which shows that adding eyeliner increases the CP for low hemoglobin. This is likely a result of social and cultural association between makeup usage and low hemoglobin levels with the same demographic factors of sex and age. Another example of social or structural determinants is the prediction of race from decreased skeletal lucency and increased conspicuity of ribs, scapulae, humeri, or thoracic vertebral bodies: simply put, bone mineral density as a predictor of race. Biological and even lifestyle factors cannot fully explain this association.[75] Further exploration is warranted regarding the effect of factors such as age distribution, environmental exposures, or nutrition. Additionally, because race is socially, rather than biologically, defined and often serves as a proxy for racism[76] and there is greater genetic variation within racial groups than between them,[77] we feel it important to remind the reader that associations with race do not represent a biological phenomena. Uncovering attributes that indicate unwanted bias in the data is important in practice because such bias should be mitigated before a model is deployed. A practitioner may want to re-train the classifier with a different dataset or setup (e.g., in the case of the normal/abnormal CXR model, we could train separate models for AP and PA), or employ training methods or auxiliary data from public health datasets such as state-wide surveillance datasets, longitudinal national health surveys, and national cohort studies including factors at the individual, societal, political, and geographical levels to add context and reduce bias.[78,79]

The understanding that machine learning models can extract unwanted signals has implications for modeling and dataset design and transparency.[80] For example, researchers may desire updating model cards[81] or other transparency artifacts to ensure awareness of possible unexpected associations leveraged by the model. Moreover, this awareness could enable targeted improvements, whether by improving or modifying the training processes using custom losses or augmentations, or modifying dataset sampling during training to remove these associations.[82–84] Similarly, this knowledge could more broadly improve awareness of hidden associations that exist in both retrospective data and prospective data collection. For example, standardizing prospective data collection protocols to remove makeup, masks, jewelry, accessories, and more could help, though such efforts would need to be balanced with whether the target population during usage may not be keen on removing these. For retrospective datasets where protocols cannot be changed, more complete documentation of the data collecting environment and protocol could help researchers be more aware of potential associations.

An exciting product of our approach are attributes that might inspire hypotheses for previously unknown correlations. Such hypotheses have the potential to pave the way for novel scientific understanding, similar to the way AI techniques have been applied in non-medical domains.[25,85,86] For example, we found that increasing eyelid margin pallor was associated with increasing probability of elevated HbA1c. A possible explanation for that attribute is that it is a subtle manifestation of meibomian gland disease, which is more severe in patients with diabetes or elevated blood sugars, though the mechanism is not well understood.[55,56] Another attribute discovered by our method is that retinochoroidal pigmentation is associated with higher predicted likelihood for female sex, which adds to the evidence that sex differences have been observed in macular and peripapillary choroidal thicknesses.[46,47,87] These attribute examples are consistent with such connections, and could inspire research to better explore the underlying physiological and pathophysiological mechanisms.

It is important to note that our method was not designed to assess causality, and as such, the core assumptions of consistency, exchangeability, and positivity do not hold.[88] Prediction models are distinct from causal models in several ways, most notably in the considerations of temporality of the explanatory variables.[89] Without close attention to whether the effect truly occurred after the cause, assessing the relationship of "cause" and "effect" is impossible. However, by explicitly incorporating this understanding, interdisciplinary panel discussions helped reduce blind spots and aided in identification of both potential hypotheses (see previous paragraph) and potential confounding factors and as a result suggested opportunities for model improvement. Furthermore, it is worth noting that the workflow we present in this paper is aimed at aiding practitioners gain a better understanding of models and datasets, rather than a tool that can be employed in clinical practice (i.e., this approach as-is may not consistently aid in understanding why a model produces a particular output for a given input).

This new explainability framework contains several other notable limitations. Firstly, after attributes are extracted, their interpretation is not straightforward and requires close collaboration between the machine learning researchers and expert panel to generate hypotheses, conduct literature reviews, and even manually engineer features to validate the hypotheses. Unsurprisingly, in contrast to some of the more visually striking examples provided here, we have also found cases where the experts cannot readily interpret the counterfactuals described by our method. In other words, while the smooth animations generated by this approach render the (often subtle) change in the image visible and detectable, experts are unable to clearly describe or rationalize the change, whether from a pathophysiological perspective or otherwise. Depending

on the magnitude of the change in the latent space applied to generate the counterfactual, it is also possible for the generator to produce unrealistic-looking images. Secondly, our method requires training a StyleGAN model, which works well on structured, closed domains such as images taken using standardized protocols and backgrounds (CXR, fundus photographs, external eye photographs), but can struggle in general open-ended domains such as images with a wide variety of backgrounds, poses, and distances (e.g., clinical photographs in dermatology where the background and camera orientation and lighting can differ substantially across cases). Training StyleGAN models is also both computationally and data intensive at present, narrowing the number of projects where this approach is feasible. Additionally, this method is subject to selection bias, measurement error, and other forms of confounding unless explicitly accounted for. Along these lines, it is important to consider the role of social and structural determinants of health in shaping both the health outcomes captured in datasets and its representativeness. Datasets used in our analyses are not necessarily demographically representative of their geographic source populations and thus our findings about potential attributes and their associations are not always generalizable to each respective country in which data were collected. Furthermore, because social, political, and economic circumstances vary vastly between countries, findings from data collected in one country cannot be generalizable to another country (i.e., data from the U.S. may not be generalizable to India).

In summary, we have presented a StyleGAN-based method (StylEx) for generating counterfactuals to explain AI models, and demonstrated its utility across multiple imaging modalities and prediction tasks. StylEx was able to rediscover known associations, detect confounders, and generate new insights that can be tested in future studies. With this paper, we release sample training code to aid researchers, and look forward to more novel discoveries by the community.

Contributors

OL, DYS, HCL, YL, NH, and BB conceived and designed the study. OL, DYS and BB developed the models and analyzed their performance, with scientific guidance from MI. HCL, CRB and CRL designed and conducted the expert panel review process. OL, IT, HCL, CRB, CL and BB participated in the panel review sessions. OL, CRB and BB drafted the manuscript with feedback from all authors. CS, DRW, GSC, MI, AH, and YM provided strategic guidance, oversight and obtained funding. OL, DYS, and BB had access to the underlying data; OL, DYS, IT, HCL, CRB, CRL, CL, YL, NH and BB had access to the generated attributes. All authors read and approved the final version of the manuscript.

Data sharing statement

A subset of EyePACS data is available at https://www.kaggle.com/competitions/diabeticretinopathy-detection/data. To enquire about access to the full EyePACS dataset, researchers should contact Jorge Cuadros (jcuadros@eyepacs.com). CXR-14 is a public dataset provided by the NIH (https://nihcc.app.box.com/v/ChestXray-NIHCC). The UK Biobank data are available for approved projects (application process detailed at https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access) through the UK Biobank Access Management System (https://www.ukbiobank.ac.uk). MIMIC III and MIMIC-CXR are available via the physionet portal (https://physionet.org/content/mimiciii/1.4/and https://physionet.org/content/mimic-cxr/2.0.0/, respectively). Code including the loss function can be found at https://github.com/google/explaining-in-style.

Declaration of interests

OL, DYS, HCL, CS, DRW, GSC, AH, YM, YL, NH and BB are current or past Google employees and may own Alphabet stock. IT, CRB, and CL are paid consultants to Google. All other authors declare no competing interests.

Acknowledgements

This research was conducted using the UK Biobank Resource under application number 65275. The authors thank Dr. Sreenivasa Raju Kalidindi and his team at Apollo Radiology International for their aid with the Apollo dataset, Andrew Sellergren and Zaid Nabulsi for help with CXR modeling infrastructure, Dr. Jorge Cuadros and Dr. Lauren P. Daskivich for their help with the EyePACS/LACDHS dataset, Elvia Figueroa and the LAC DHS TDRS program staff for data collection and program support, Nikhil Kookkiri and EyePACS staff for data collection and support, and Preeti Singh for support with dataset and annotation logistics. Finally, the authors would like to thank Mat Fleck for participating the panel sessions, Avinash Varadarajan and Yossi Gandelsman for early feedback on the project, Cameron Chen, Ivor Horn, and Lily Peng for providing feedback on the manuscript, and Tiya Tiyasirichokchai for Fig. 1 design. Part of the data processing team at LAC DHS was supported by grants UL1TR001855 and UL1TR000130 from the National Center for Advancing Translational Sciences (NCATS) of the US National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, or the US Government.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2024.105075.

References

1 Swiecicki A, Li N, O'Donnell J, et al. Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. *Comput Biol Med*. 2021;133:104334.
2 Kashou AH, Ko W-Y, Attia ZI, Cohen MS, Friedman PA, Noseworthy PA. A comprehensive artificial intelligence-enabled electrocardiogram interpretation program. *Cardiovasc Digit Health J*. 2020;1:62–70.
3 Stidham RW, Liu W, Bishu S, et al. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA Netw Open*. 2019;2:e193963.
4 Ellertsson S, Loftsson H, Sigurdsson EL. Artificial intelligence in the GPs office: a retrospective study on diagnostic accuracy. *Scand J Prim Health Care*. 2021;39:448–458.
5 Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digital Medicine*. 2019;2:25. https://doi.org/10.1038/s41746-019-0099-8.
6 Qian H, Dong B, Yuan J-J, et al. Pre-consultation system based on the artificial intelligence has a better diagnostic performance than the physicians in the outpatient department of pediatrics. *Front Med*. 2021;8:695185. https://doi.org/10.3389/fmed.2021.695185.
7 Xiao W, Huang X, Wang JH, et al. Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. *Lancet Digit Health*. 2021;3:e88–e97.
8 Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. *bioRxiv*. 2019. https://doi.org/10.1101/681676.

9 Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2:158–164.
10 Duffy G, Clarke SL, Christensen M, et al. Confounders mediate AI prediction of demographics in medical imaging. *NPJ Digital Med*. 2022;5:1–6.
11 Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci*. 2014;7:342–346.
12 Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak*. 2014;14:51.
13 Gichoya J, Banerjee I, Bhimireddy A, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digital Health*. 2022;4:e406–e414.
14 Richardson R, Schultz J, Crawford K. Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. Published online Feb 13. https://papers.ssrn.com/abstract=3333423; 2019. Accessed March 5, 2023.
15 Mitani A, Traynis I, Singh P, et al. Retinal fundus photographs capture hemoglobin loss after blood donation. *medRxiv Preprint*. 2022. https://doi.org/10.1101/2021.12.30.21268488.
16 L'Imperio V, Wulczyn E, Plass M, et al. Pathologist validation of a machine learning–derived feature for colon cancer risk stratification. *JAMA Netw Open*. 2023;6:e2254891.
17 Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion*. 2021;76:89–106.
18 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447–453.
19 Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol*. 2019;155:1135–1141.
20 Berger JS, Haskell L, Ting W, et al. Evaluation of machine learning methodology for the prediction of healthcare resource utilization and healthcare costs in patients with critical limb ischemia-is preventive and personalized approach on the horizon? *EPMA J*. 2020;11:53–64.
21 Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. *J Imaging Sci Technol*. 2020;6:52. https://doi.org/10.3390/jimaging6060052.
22 Ribeiro Marco Tulio, Singh Sameer, Guestrin Carlos. 'Why should I trust you?' ACM conferences. https://dl.acm.org/doi/10.1145/2939672.2939778. Accessed November 7, 2022.
23 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. In: *Workshop at International Conference on Learning Representations*. 2014.
24 Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016:2921–2929.
25 Thiagarajan JJ, Thopalli K, Rajan D, Turaga P. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Sci Rep*. 2022;12:1–15.
26 Dravid A, Schiffers F, Gong B, Katsaggelos AK. medXGAN: visual explanations for medical classifiers through a generative latent space. In: *2022 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*. 2022. https://doi.org/10.1109/cvprw56347.2022.00331.
27 Narayanaswamy A, Venugopalan S, Webster DR, et al. Scientific discovery by generating counterfactuals using image translation. In: *Medical image Computing and computer assisted intervention – MICCAI 2020*. 2020:273–283.
28 DeGrave AJ, Cai ZR, Janizek JD, Daneshjou R, Lee S-I. Dissection of medical AI reasoning processes via physician and generative-AI collaboration. *medRxiv Preprint*. 2023. Published online May 16. https://doi.org/10.1101/2023.05.12.23289878.
29 Mertes S, Huber T, Weitz K, Heimerl A, André E. GANterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Front Artif Intell*. 2022;5:825565.
30 Lang O, Gandelsman Y, Yarom M, et al. Explaining in style: training a GAN to explain a classifier in StyleSpace. In: *2021 IEEE/CVF international conference on computer vision (ICCV)*. 2021. https://doi.org/10.1109/iccv48922.2021.00073.
31 Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied logistic regression*. John Wiley & Sons; 2013.
32 Wu Z, Lischinski D, Shechtman E. StyleSpace analysis: disentangled controls for StyleGAN image generation. In: *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 2021. https://doi.org/10.1109/cvpr46437.2021.01267.
33 Molnar C. 9.1 Individual conditional expectation (ICE). Published online Aug 21. https://christophm.github.io/interpretable-ml-book/ice.html#ice; 2023. Accessed September 21, 2023.
34 Bronfenbrenner U. Ecological Systems Theory. In: Vasta R, ed. *Six theories of child development: revised formulations and current issues*. 2000.
35 World Health Organization. *A conceptual framework for action on the social determinants of health*. 2010:76.
36 Krieger N. *Ecosocial theory of disease distribution: embodying societal & ecologic context*. 2011. Published online March 23. https://doi.org/10.1093/acprof:oso/9780195383874.003.0007.
37 Zamzam AH, Abdul Wahab AK, Azizan MM, Satapathy SC, Lai KW, Hasikin K. A systematic review of medical equipment reliability assessment in improving the quality of healthcare services. *Front Public Health*. 2021;9:753951.
38 Proctor RW, Van Zandt T. *Human factors in simple and complex systems*. 2017.
39 McEntee M, Gafoor S. *Ambient temperature variation affects radiological diagnostic performance*. 2009.
40 Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database (version 1.4). *PhysioNet*. https://doi.org/10.13026/C2XW26.
41 Babenko B, Traynis I, Chen C, et al. A deep learning model for novel systemic biomarkers in photographs of the external eye: a retrospective study. *Lancet Digital Health*. 2023;5:e257–e264.
42 Nabulsi Z, Sellergren A, Jamshy S, et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. *Sci Rep*. 2021;11:1–15.
43 Drobnjak D, Taarnhøj NC, Kessel L, Jørgensen T, Larsen M. Association between retinal vessel diameters and cigarette smoking. *Invest Ophthalmol Vis Sci*. 2011;52:2213.
44 Cheong KX, Tan CS. Sex-dependent choroidal thickness differences in healthy adults: a study based on original and synthesized data. *Curr Eye Res*. 2019;44:236.
45 Zeng J, Liu R, Zhang X-Y, et al. [Relationship between gender and posterior pole choroidal thickness in normal eyes]. *Zhonghua Yan Ke Za Zhi*. 2012;48:1093–1096.
46 Shibata H, Sawada Y, Ishikawa M, Yoshitomi T, Iwase T. Peripapillary choroidal thickness assessed by spectral-domain optical coherence tomography in normal Japanese. *Jpn J Ophthalmol*. 2021;65:666–671.
47 Yang H, Luo H, Gardiner SK, et al. Factors influencing optical coherence tomography peripapillary choroidal thickness: a multicenter study. *Invest Ophthalmol Vis Sci*. 2019;60:795–806.
48 Cumberland PM, Bountziouka V, Hammond CJ, Hysi PG, Rahi JS, UK Biobank Eye and Vision Consortium. Temporal trends in frequency, type and severity of myopia and associations with key environmental risk factors in the UK: findings from the UK Biobank Study. *PLoS One*. 2022;17:e0260993.
49 Babenko B, Mitani A, Traynis I, et al. Detection of signs of disease in external photographs of the eyes via deep learning. *Nat Biomed Eng*. 2022. Published online March 29. https://doi.org/10.1038/s41551-022-00867-5.
50 Glasspool MG. Cataract. In: Springer D, ed. *Atlas of ophthalmology*. 1982.
51 Kent AR, Elsing SH, Hebert RL. Conjunctival vasculature in the assessment of anemia. *Ophthalmology*. 2000;107:274–277. https://doi.org/10.1016/s0161-6420(99)00048-2.
52 Le CHH. The prevalence of anemia and moderate-severe anemia in the US population (NHANES 2003-2012). *PLoS One*. 2016;11:e0166635. https://doi.org/10.1371/journal.pone.0166635.
53 Hashemi H, Malekifar P, Aghamirsalim M, Yekta A, Mahboubipour H, Khabazkhoob M. Prevalence and associated factors of corneal arcus in the geriatric population; Tehran geriatric eye study. *BMC Ophthalmol*. 2022;22:354.
54 Dubowitz N, Xue W, Long Q, et al. Aging is associated with increased HbA1c levels, independently of glucose levels and insulin resistance, and also with decreased HbA1c diagnostic specificity. *Diabet Med*. 2014;31:927–935.
55 Yu T, Han X-G, Gao Y, Song A-P, Dang G-F. Morphological and cytological changes of meibomian glands in patients with type 2 diabetes mellitus. *Int J Ophthalmol*. 2019;12:1415–1419.

56 Yu T, Shi W-Y, Song A-P, Gao Y, Dang G-F, Ding G. Changes of meibomian glands in patients with type 2 diabetes mellitus. *Int J Ophthalmol*. 2016;9:1740–1744.
57 Boxt L, Abbara S. *Cardiac imaging: the requisites*. Elsevier; 2009.
58 Isselbacher EM. Thoracic and abdominal aortic aneurysms. *Circulation*. 2005;111:816–828.
59 Mothiram U, Brennan PC, Robinson J, Lewis SJ, Moran B. Retrospective evaluation of exposure index (EI) values from plain radiographs reveals important considerations for quality improvement. *Journal of Medical Radiation Sciences*. 2013;60:115–122.
60 Looker AC, Melton LJ 3rd, Harris T, Borrud L, Shepherd J, McGowan J. Age, gender, and race/ethnic differences in total body and subregional bone density. *Osteoporos Int*. 2009;20:1141–1149.
61 Fausto-Sterling A. The bare bones of race. *Soc Stud Sci*. 2008;38:657–694.
62 Smedley A, Smedley BD. Race as biology is fiction, racism as a social problem is real: anthropological and historical perspectives on the social construction of race. *Am Psychol*. 2005;60:16–26.
63 *New AMA policies recognize race as a social, not biological, construct*. American Medical Association; 2020. Published online Nov 16. https://www.ama-assn.org/press-center/press-releases/new-ama-policies-recognize-race-social-not-biological-construct. Accessed May 5, 2023.
64 Romualdi C, Balding D, Nasidze IS, et al. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res*. 2002;12:602–612.
65 Krieger N. Does racism harm health? Did child abuse exist before 1962? On explicit questions, critical science, and current controversies: an ecosocial perspective. *Am J Public Health*. 2003;93:194–199.
66 Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. *Lancet*. 2017;389:1453–1463.
67 Krieger N. Refiguring 'race': epidemiology, racialized biology, and biological expressions of race relations. *Int J Health Serv*. 2000;30:211–216.
68 Gilkes A, Ashworth M, Schofield P, et al. Does COPD risk vary by ethnicity? A retrospective cross-sectional study. *Int J Chron Obstruct Pulmon Dis*. 2016;11:739–746.
69 Martin A, Badrick E, Mathur R, Hull S. Effect of ethnicity on the prevalence, severity, and management of COPD in general practice. *Br J Gen Pract*. 2012;62:e76–e81.
70 Mamary AJ, Stewart JI, Kinney GL, et al. Race and gender disparities are evident in COPD underdiagnoses across all severities of measured airflow obstruction. *Int J Chron Obstruct Pulmon Dis*. 2018;5:177–184.
71 Braun L. Race, ethnicity and lung function: a brief history. *Can J Respir Ther*. 2015;51:99–101.
72 Van Sickle D, Magzamen S, Mullahy J. Understanding socioeconomic and racial differences in adult lung function. *Am J Respir Crit Care Med*. 2011;184:521–527.
73 Doshi RP, Aseltine RH Jr, Sabina AB, Graham GN. Racial and ethnic disparities in preventable hospitalizations for chronic disease: prevalence and risk factors. *J Racial Ethn Health Disparities*. 2017;4:1100–1106.
74 Laditka JN, Laditka SB. Race, ethnicity and hospitalization for six chronic ambulatory care sensitive conditions in the USA. *Ethn Health*. 2006;11:247–263.
75 Ettinger B, Sidney S, Cummings SR, et al. Racial differences in bone density between young adult black and white subjects persist after adjustment for anthropometric, lifestyle, and biochemical differences. *J Clin Endocrinol Metab*. 1997;82:429–434.
76 Braveman P, Parker Dominguez T. Abandon 'race.' focus on racism. *Front Public Health*. 2021;9:689462.
77 Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. *Science*. 2002;298:2381–2385.
78 Beutel A, Chen J, Zhao Z, Chi EH. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv Preprint*; 2017. Published online July 1. http://arxiv.org/abs/1707.00075.
79 Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: *Proceedings of the 2018 AAAI/ACM conference on AI, Ethics, and Society*. 2018:335–340.
80 Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J. Explainable AI: a brief survey on history, research areas, approaches and challenges. *Nat Lang Process Chinese Comput*. 2019:563–574.
81 Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. https://doi.org/10.1145/3287560.3287596.
82 Suresh H, Guttag JV. A framework for understanding sources of harm throughout the machine learning life cycle. https://doi.org/10.1145/3465416.3483305; 2021.
83 He Yuzi, Burghardt Keith, Lerman Kristina. A geometric solution to fair representations. In: *ACM Conferences*; 2020. https://dl.acm.org/doi/10.1145/3375627.3375864. Accessed September 21, 2023.
84 Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, Varshney KR. Optimized pre-processing for discrimination prevention. *Adv Neural Inf Process Syst*. 2017;30. https://proceedings.neurips.cc/paper_files/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf. Accessed September 21, 2023.
85 Guimerà R, Reichardt I, Aguilar-Mogas A, et al. A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Sci Adv*. 2020;6:eaav6971.
86 Udrescu S-M, Tegmark M. AI Feynman: a physics-inspired method for symbolic regression. *Sci Adv*. 2020;6:eaay2631.
87 Ooto S, Hangai M, Yoshimura N. Effects of sex and age on the normal retinal and choroidal structures on optical coherence tomography. *Curr Eye Res*. 2015;40:213–225.
88 Hernán MA, Robins JM. *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC; 2020.
89 Laubach ZM, Murray EJ, Hoke KL, Safran RJ, Perng W. A biologist's guide to model selection and causal inference. *Proc Biol Sci*. 2021;288:20202815.