Supplementary Information:

The effect of ImageNet

ImageNet is often used as a basis for transfer learning, specifically because of the flexibility of the patterns it already recognizes. To confirm that both weights of the beans and the beer model are indeed changing and learning new patterns beyond what ImageNet knows, it can be helpful to compare against the baseline performance of a pure ImageNet-based model. To do this, we trained a series of models to predict the confounding/latent variables (Sex, Race, etc.) and the diet variables using the original ImageNet weights, allowing only the final linear layer to change. By comparing the performance of these models against the fully trained models described in the paper, we can confirm that these results are not a side-effect of the generalizability of ImageNet-based weights but truly from new learning from fine-tuning training.

Results

The models that use the original ImageNet-based weights all show significant drops in performance when compared across the confounding and latent variables (Table S1). This is expected, but it confirms that shortcutting is not just a side effect of the original ImageNet weights. When compared to the results in Table 4, it can also be seen that the weights trained to predict beer preferences exceed the performance of raw ImageNet weights for predicting Sex and Site.

	Accuracy		Adjusted Balanced Accuracy	
	Fine-tuned	ImageNet	Fine-tuned	ImageNet
SEX	0.987	0.829	0.973	0.636
RACE	0.921	0.838	0.244	0.014
SITE	0.982	0.947	0.873	0.764
MFG	0.999	0.999	0.955	0.635
YEAR	0.644	0.546	0.338	0.184

Table S1. Performance of fine-tuned confounding/latent variable models vs models using original ImageNet weights. Fine-tuned models are initialized with ImageNet weights, but all weights are modified during training. ImageNet models only allowed for modification of the final layer weights during training.

As binary variables, refired bean and beer consumption use AUC and accuracy, which make the differences less stark (Table S2). These results show both how much ImageNet alone weights can do in making surprising predictions, but also that the last mile does come from fine-tuning training.

	Accuracy		AUC	
	Fine-tuned	ImageNet	Fine-tuned	ImageNet
Refried Beans	0.600	0.541	0.631	0.591
Beer	0.702	0.659	0.734	0.677

Table S2. Performance of fine-tuned diet models vs models using original ImageNet weights. Fine-tuned models are initialized with ImageNet weights, but all weights are modified during training. ImageNet models only allowed for modification of the final layer weights during training.

Methods

This experiment repeats steps 1-5 in the original study methods, with two exceptions. First, all Z-score normalization was done according to the ImageNet mean and standard deviation values. Second, all layers of the neural networks were frozen except for the final layer. The models from steps 1-5 in the original methods trained over 11 million parameters. Training only the final layer limits the number of trainable parameters in the new models to between 513 and 5,643 (based on the number of prediction classes for each variable).

Normalization

Image variations created from differences across X-ray machines and their setup across collection sites can create statistical fingerprints within the images. Min-max normalization and then Z-score normalization, as done in this work, limit the most obvious issues (range and scale differences). It is natural to question whether more aggressive normalization would alter these findings. While removal of these differences is an active area of research in medical imaging[1], for X-rays, the only commonly used technique beyond what was used is contrast limited adaptive histogram equalization (CLAHE)[2]. Though originally designed to enhance image contrast, its redistribution of pixel values does provide a more aggressive alteration of an image's histogram.

To examine whether applying CLAHE as a final normalization step makes a noticeable difference, we looked at how it changed model performance in predicting two latent variables: Site and Mfg. CLAHE was applied to the images at two different values. First, the mild clipping threshold parameter of 2, which is typically used on X-rays. Second, a more aggressive clipping threshold of 40 was tried for greater smoothing of the image histograms.

Results

Overall, when CLAHE was applied, the network's ability to predict the X-ray site or manufacturer saw only negligible change. The adjusted balanced accuracy scores only changed to the third decimal place (see Table S3).

	SITE			MFG		
	None	CLAHE 2	CLAHE 40	None	CLAHE 2	CLAHE 40
Accuracy	0.982	0.978	0.979	0.999	0.998	0.998
Adjusted Balanced Accuracy	0.947	0.941	0.942	0.999	0.999	0.994

Table S3. Model accuracy learning differences in predicting latent variables with and without more normalization. None is the original experiment results. CLAHE 2 is for images processed with a clipping threshold of 2. CLAHE 40 used a clipping threshold of 40.

Method

This experiment was applied to predictions of Site and Mfg. For both predictions, it repeated steps 1-4 in the original study methods three times. In the first case, exactly. In the CLAHE cases, CLAHE was applied after Min-max normalization and before the images were Z-score normalized according to the mean and standard deviation of the training data set. In one case, a CLAHE clipping threshold of 2 and a tile grid of 8x8 pixels was used. In the other, a clipping threshold of 40 and a tile grid of 8x8 pixels was used.

X-ray Manufacturer vs. Clinical Sites

	AGFA	FUJI	Swissray	GE	LS100	Siemens	Philips	Total
A	6	1	3,952	3	1		65	4,031
B		2,206			1,884			5,724
C	3,933			3,881				7,814
D	4727	1,495		406				6,628
E		1,198	401		248	324	30	2,298
Total	8,666	4,900	4,353	4,290	2,133	324	95	24,761

Table S4. Distribution of X-ray manufacturers across clinical sites. X-ray manufacturer was listed for all but 1,759 of the available X-rays.

References

1. Seoni, S. *et al.* All you need is data preparation: A systematic review of image harmonization techniques in Multi-center/device studies for medical support systems. *Comput. Methods Programs Biomed.* **250**, 108200 (2024).

2. Pizer, S. M. *et al.* Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **39**, 355–368 (1987).