Purpose: To detect outliers from a medical imaging distribution used to train deep learning models. The proposed method can identify Computed Tomography (CT) scans that would make a clinically deployed segmentation model fail.

Methods: A cohort of 143 patients with 147 non-contrast enhanced abdominal CTs (97 training, 50 test) was used. A StyleGAN2 network, a state-of-the-art high-resolution generative model that uses a style-based generator and backpropagation to encode, was trained to reconstruct slices. Data was preprocessed with windowing, masking, and conversion to 512x512 PNG images. The network's generative quality was measured with the Fréchet Inception (FID) and Wasserstein (WD) distances. Slice reconstructions from the test CTs with a learned perceptual image patch similarity score (compares VGG network feature representations) over 0.1 were classified as out-of-distribution.

Results: Randomly generated slices had FID and WD metric values of 8.77 and 0.10, respectively. All test images on which a segmentation model failed (the model had a Dice coefficient of 0.96 on test CTs) were classified as out-of-distribution.

Conclusion: A paradigm was optimized to predict when a clinically deployed segmentation model would fail with a 100% success rate. The paradigm could be further used to create heterogenous imaging datasets and prioritize images for labeling.