Synthetic Ground Truth Counterfactuals for Comprehensive Evaluation of Causal Generative Models in Medical Imaging

Emma A. M. Stanley[1,2,3,4⋆][0000−0002−7802−6820], Vibujithan Vigneshwaran[2,3,4⋆], Erik Y. Ohara[1,2,3,4⋆], Finn G. Vamosi[5], Nils D. Forkert[2,3,4⋆⋆][0000−0003−2556−3224], and Matthias Wilms[2,3,4,6,7⋆⋆][0000−0001−8845−360X]

[1] Department of Biomedical Engineering, University of Calgary, Canada
[2] Department of Radiology, University of Calgary, Canada
[3] Hotchkiss Brain Institute, University of Calgary, Canada
[4] Alberta Children's Hospital Research Institute, University of Calgary, Canada
[5] Department of Computer Science, University of Calgary, Canada
[6] Department of Pediatrics, University of Calgary, Canada
[7] Department of Radiology, University of Michigan, United States
{emma.stanley,vibujithan.vigneshwa,erik.ohara}@ucalgary.ca

Abstract. Counterfactuals in medical imaging are synthetic representations of how an individual's medical image might appear under alternative, typically unobservable conditions, which have the potential to address data limitations and enhance interpretability. However, counterfactual images, which can be generated by causal generative models (CGMs), are inherently hypothetical—raising questions of how to properly validate that they are realistic and accurately reflect the intended modifications. A common approach for quantitatively evaluating CGM-generated counterfactuals involves using a discriminative model as a 'pseudo-oracle' to assess whether interventions on specific variables are effective. However, this method is not well-suited for in-depth error identification and analysis of CGMs. To address this limitation, we propose to leverage synthetic, 'ground truth' counterfactual datasets as a novel approach for debugging and evaluating CGMs. These synthetic datasets enable the computation of global performance metrics and precise localization of CGM failure modes. To further quantify failures, we introduce a novel metric, the *Triangulation of Effectiveness and Amplification* (TEA), which precisely quantifies the effectiveness of target variable interventions and the additional amplification of unintended effects. We test and validate our evaluation framework on two state-of-the-art CGMs where the results demonstrate the utility of synthetic datasets in identifying failure modes of CGMs, and highlight the potential of the proposed TEA metric as a robust tool for evaluation of their performance. Code and data are available at https://github.com/ucalgary-miplab/TEA.

⋆ E. Stanley, V. Vigneshwaran, E. Ohara — Contributed equally. *These authors may list their name first on their own CV.*
⋆⋆ N. Forkert, M. Wilms — Shared senior authorship.

Keywords: counterfactuals · causal generative modeling · brain MRI

1 Introduction

Counterfactual images in medical imaging are synthetic images that aim to represent how an individual's acquired image might appear under alternative, hypothetical conditions. For instance, plausible counterfactual images could be generated by asking an artificial intelligence (AI) system, "What would the brain look like if the subject were 75 years old?". The utility of counterfactual images in medicine is multifaceted. They can, for example, serve as powerful diagnostic and prognostic tools, helping medical professionals visualize and predict disease progression and treatment effects on an individual level [13,2]. In research, they can augment training datasets, addressing data scarcity and imbalance [20]. While any class of (conditional) generative model can be utilized to generate counterfactual images, *causal* generative models (CGMs) provide the most theoretically grounded way of doing so. Several causal models in medical imaging [7,18,10,17] have successfully addressed all levels of Pearl's ladder of causation [8] and are capable of generating theoretically sound counterfactuals.

However, since counterfactuals are hypothetical by definition (*i.e.*, that alternative version of the image usually does not exist in the real world), how can developers validate that their CGMs are capable of producing images that are realistic and accurately reflect the targeted changes based on the original image? Currently, when evaluating counterfactuals generated by CGMs, a common approach is to use an additional discriminative model as a 'pseudo-oracle' to assess whether a variable was *effectively* intervened on [5]. For instance, a classifier can be trained to predict whether a CGM properly removed attributes associated with Alzheimer's disease from an individual's brain magnetic resonance imaging (MRI) scan when producing a counterfactual showing what they would look like if they were healthy.

Another important consideration for CGM developers is to ensure that their models only modify the targeted aspects of the original image, without making unintended changes. For example, when generating a counterfactual for Alzheimer's disease, it is important to verify that the CGM does not unintentionally alter other attributes, such as making a male's brain MRI appear female. Prior works to quantify such unwanted modifications proposed distance-based metrics, such as *proximity* [6] and *minimality* [12,4]. Briefly described, these metrics measure the distance between the generated counterfactual and the original image, assuming that a good counterfactual should closely resemble the original image while only modifying the target variable. However, since these methods do not account for the true extent or direction of targeted changes, they often favour counterfactuals that remain largely unchanged. Again, a pseudo-oracle can also be used to evaluate whether a CGM *amplified* attributes unrelated to the intended target [19]. For example, a classifier model can help to determine if the CGM unintentionally modified the sex attribute.

Therefore, pseudo-oracles are currently one of the most practical approaches for assessing both the effectiveness and amplification of CGMs on real-world data. However, their utility for troubleshooting and debugging during development and improvement of CGMs is limited. A major drawback of this approach is shortcut learning, where the pseudo-oracle may rely on spurious correlations during training, undermining its efficacy as an evaluation technique. Furthermore, these methods reduce the whole complex problem to a single scalar global output metric: a performance value corresponding to whether the pseudo-oracle believed that an image was properly intervened on. This single, global value may not be useful for developers trying to identify *why* their CGM does not produce reliable counterfactual images and where the problem areas of those images may be.

Instead, we argue that the flexible and tractable nature of synthetic data is much better suited for in-depth error identification and analysis of CGMs. Within this context, MorphoMNIST [1] is commonly used as a first tool for validating CGMs, but has limited utility if the ultimate goal is to produce medical images, which are much more complex. Alternatively, an established tool for generating synthetic, realistic brain MRI data is the Simulated Bias in Artificial Medical Images (SimBA) framework [16,15], which was originally proposed to study the impacts of medical imaging 'biases' on neuroimaging deep learning pipelines. Crucially, SimBA enables the generation of counterfactual datasets in which a particular synthetic but realistic subject can be produced with and without precisely specified, spatially localized morphological effects. This synthetic counterfactual setup has, for example, facilitated rigorous evaluation into the impact that a specific effect within a medical image would have on discriminative model performance, explainability, and learned features [15,14].

In this work, we propose a novel use of the synthetic medical image counterfactuals generated with SimBA as a tool for the development and improvement of CGMs. We demontrate how access to these exact ground truth counterfactual images enables the computation of a quantitative global metric that informs CGM performance, as well as exact spatial localization of failure modes. To better quantify these failure modes, we also propose a new metric, *triangulation of effectiveness and amplification* (TEA), which precisely quantifies the extent to which target variables are effectively intervened on and unwanted effects are amplified or changed. We test our proposed framework using SimBA-generated neuroimaging data on two generative models capable of generating causally-grounded counterfactuals: HVAE [10] and MACAW [17].

2 Methods

2.1 Ground truth counterfactuals with SimBA

The core idea of the SimBA framework [16,15] is that medical image-specific effects (*e.g.*, morphological or intensity variations in brain MRI) can be systematically integrated into generated images to allow for exact traceability in how such effects manifest within deep learning models. A key component of SimBA is that paired counterfactual dataset scenarios are generated that enable a direct comparison of the impact of specific effects to baseline scenarios. In this work, we propose to use these ground truth counterfactual datasets as a strategy to evaluate the ability of CGMs to generate medical image counterfactuals. In this context, SimBA synthetic images are generated via augmentation of a template image (*e.g.*, a brain MRI atlas) with global subject-specific morphological variations (*'subject effects'*), localized morphological effects distinguishing classes for a downstream task (*'task effects'*), and additional specific morphology or intensity-based effects under study (*'target effects'*); where target effects are what we aim to causally intervene on. Therefore, variable intervention with a CGM aims to generate counterfactual images that only differ in the absence or presence of the target effect, while keeping the remaining image attributes (subject and task effects) constant.

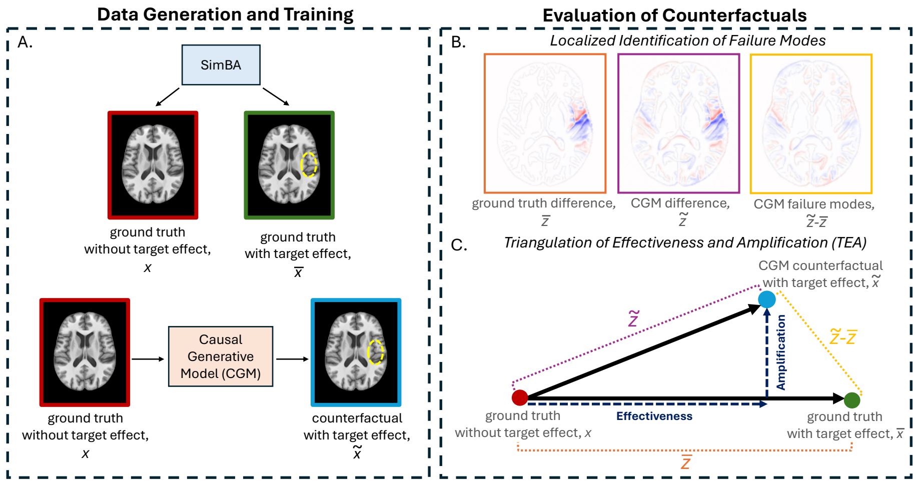
Data Generation and Training
A.
SimBA
ground truth without target effect, x
ground truth with target effect, $\bar{x}$
ground truth without target effect, x
Causal Generative Model (CGM)
counterfactual with target effect, $\tilde{x}$
Evaluation of Counterfactuals
B.
Localized Identification of Failure Modes
ground truth difference, $\bar{z}$
CGM difference, $\tilde{z}$
CGM failure modes, $\tilde{z}$-$\bar{z}$
C.
Triangulation of Effectiveness and Amplification (TEA)
CGM counterfactual with target effect, $\tilde{x}$
$\tilde{z}$
$\tilde{z}$-$\bar{z}$
Amplification
Effectiveness
ground truth without target effect, x
ground truth with target effect, $\bar{x}$
$\bar{z}$

Fig. 1. Given a synthetic image $\boldsymbol{x}$ without the target effect, the ground truth counterfactual with the target effect, $\overline{\boldsymbol{x}}$, is obtained via SimBA, and the CGM generates a counterfactual image $\widetilde{\boldsymbol{x}}$ with the target effect. In image space, the vector $\overline{\boldsymbol{z}}$ represents the ground truth intervention, while $\widetilde{\boldsymbol{z}}$ represents the CGM intervention, and their difference shows failure modes of the CGM. This setup can also apply to the removal of a target effect using a CGM.

Consider a synthetic image $\boldsymbol{x}$, which does not contain the target effect. The task of the CGM is then to generate the counterfactual image $\widetilde{\boldsymbol{x}}$ in which the target effect is added. Using SimBA, we also have access to the ground truth counterfactual which does contain the target effect, $\overline{\boldsymbol{x}}$ (Fig. 1A). In the image space, the difference vector $\overline{\boldsymbol{z}}$ between images $\boldsymbol{x}$ and $\overline{\boldsymbol{x}}$ represents the ground truth intervention on the target effect, while the vector $\widetilde{\boldsymbol{z}}$ between $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$ represents the CGM intervention on the target effect. Moreover, the difference between $\overline{\boldsymbol{z}}$ and $\widetilde{\boldsymbol{z}}$ reflects the extent to which the CGM failed to generate the intended counterfactual $\boldsymbol{x}$. We assume that all unwanted changes are orthogonal to the intended direction of change ($\overline{\boldsymbol{z}}$) in the image space. Intuitively, an ideal method would transform the image from one with the target effect to one without it (ground truth), strictly along this vector. Any deviation from this direction indicates an undesired amplification. Both $\overline{\boldsymbol{z}}$ and $\widetilde{\boldsymbol{z}}$ can be represented visually in difference maps of pixel values, or quantitatively, for example by calculating their L^2 norms. In this pixel value difference map, the non-zero regions indicate where $\widetilde{\boldsymbol{x}}$ needs to be adjusted to match the ideal ground truth counterfactual $\overline{\boldsymbol{x}}$ (Fig. 1B). Note that this setup can be applied to both addition and removal of a target effect.

2.2 Triangulation of effectiveness and amplification (TEA)

$\overline{\boldsymbol{z}}$ and $\widetilde{\boldsymbol{z}}$ provide both a global quantitative metric (L^2 norm) and a visual representation (pixel-level difference maps) of the failures of a CGM in producing an intended counterfactual. However, it is also highly useful for developers of CGMs to identify what those specific failure modes are. More precisely, to what extent did a CGM properly intervene on the target effect (effectiveness), and undesirably introduce other effects to the image (attribute amplification)? Utilizing this ground truth counterfactual setup, we propose a novel metric that represents these two quantitative properties, referred to as TEA. Here, *effectiveness* (E) quantifies how well $\widetilde{\boldsymbol{z}}$ aligns with $\overline{\boldsymbol{z}}$, where *amplification* (A) measures the orthogonal distance from $\overline{\boldsymbol{z}}$ to $\widetilde{\boldsymbol{z}}$ (see Fig. 1C):

$$E = \frac{\widetilde{\boldsymbol{z}} \cdot \overline{\boldsymbol{z}}}{\|\overline{\boldsymbol{z}}\|_2^2} \qquad\qquad A = \sqrt{\|\widetilde{\boldsymbol{z}}\|_2^2 - (E\|\overline{\boldsymbol{z}}\|_2)^2}$$

where $\cdot$ represents the dot product of the vectors.

2.3 Experimental setup

Data used in this work was generated with SimBA following the process detailed in [15]. An axial slice was extracted from each of the 3D brain MRI scans of 2,002 simulated subjects. The final size of the extracted slice was 192 × 192 to adhere to the existing CGM architectures used in this work. In these datasets, the task effect and the target effect were local morphological deformations in the left insular cortex and the right postcentral gyrus, respectively. Global subject effects were sampled from the normal distribution $\mathcal{N}(0, 1)$ using the subject effect model (as in [15]). Paired ground truth counterfactual datasets were generated: one that *did not* contain the target effect (*i.e.,* $\boldsymbol{x}$), and one that *did* contain the target effect (*i.e.*, $\overline{\boldsymbol{x}}$). Both $\boldsymbol{x}$ and $\overline{\boldsymbol{x}}$ had identical subject and task effects, and underwent the same splits of 50%/25%/25% for training, validation, and testing of the CGMs.

MACAW [17] encodes a structural causal model (SCM) into a normalizing flow, incorporating causal domain knowledge by masking connections to preserve the dependencies between parent and child nodes. After applying the masking, the model is trained using maximum likelihood estimation to learn the joint probability distribution. First, the predefined SCM, which includes independent target and task variables that influence the image, was encoded into the model and trained until convergence, and the version with the best validation loss (negative likelihood loss) was selected. The following hyperparameters achieved the best validation loss: learning rate $= 1 \times 10^{-3}$, weight decay $= 5 \times 10^{-5}$, and number of layers $= 4$. Finally, counterfactual images were generated using the best model to achieve target effects; these images are referred to as $\widetilde{\boldsymbol{x}}_{\boldsymbol{M}}$.

HVAE [10] extends classical variational autoencoders (VAEs) by introducing a group of VAEs with their respective latent variables. In this approach, each latent distribution is trained conditioned on the their causal parents by maximaxing the evidence lower bound (ELBO) on the marginal log-likelihood of the data to learn its distribution. The same SCM used for MACAW was enocoded in the model, which was trained until convergence, with the following hyperparameters achieving the best validation loss: learning rate $= 1 \times 10^{-3}$, weight decay $= 5 \times 10^{-2}$, using the originally proposed architecture [11]. Finally, counterfactual images were generated; these images are referred to as $\widetilde{\boldsymbol{x}}_{\boldsymbol{H}}$.

Pseudo-oracle evaluation [5] was used to determine effectiveness, by classifying the presence of the target effect from the counterfactual images. We used a standard discriminative model (an SFCN [9]) for this purpose and trained it to distinguish between images with and without target effects using the training split of the SimBA data. The model was trained until convergence, with the version achieving the best validation loss selected for evaluation. The optimal hyperparameters were as follows: learning rate $= 1 \times 10^{-4}$, and weight decay $= 1 \times 10^{-5}$.

3 Results

Global and local identification of failure modes in counterfactual generation. Pixel value difference maps between counterfactuals for a exemplary image from the dataset are shown in Fig. 2. It can be seen that while MACAW was able to successfully add the target effect, it also introduced changes throughout the brain (Fig. 2C). In contrast, HVAE was highly successful in introducing only the target effect, although a single pixel artifact was present (Fig. 2E).

Quantitatively, the L^2 norm of $\overline{\boldsymbol{z}}$ and $\widetilde{\boldsymbol{z}}$ provides a single number corresponding to how well a CGM performed in generating counterfactuals, with values closer to zero indicating that the CGM counterfactual is more similar to the ground truth counterfactual. These L^2 norm values for the the example shown in Fig. 2 were 203.20 and 85.71 for MACAW and HVAE, respectively, confirming the results of the visual inspection.

Quantification of effectiveness and attribute amplification. However, since L^2 norm values do not provide quantitiative information on the specific failure modes of a CGM, we also visualize the proposed TEA metric of effectiveness and amplification in Fig. 3 for all test images. In these plots, each point

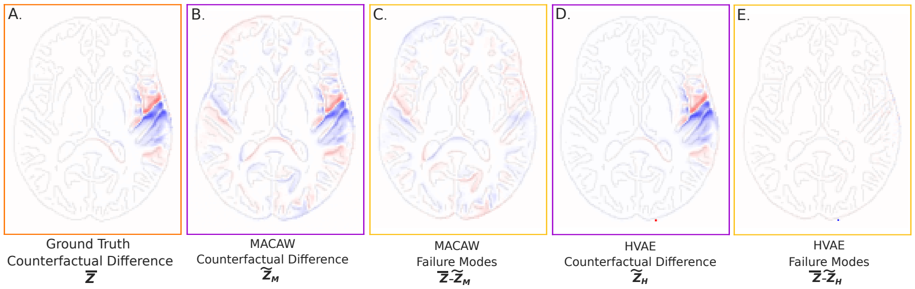
A.
B.
C.
D.
E.
Ground Truth
Counterfactual Difference
$\overline{\mathbf{Z}}$
MACAW
Counterfactual Difference
$\widetilde{\mathbf{Z}}_M$
MACAW
Failure Modes
$\overline{\mathbf{Z}}$-$\widetilde{\mathbf{Z}}_M$
HVAE
Counterfactual Difference
$\widetilde{\mathbf{Z}}_H$
HVAE
Failure Modes
$\overline{\mathbf{Z}}$-$\widetilde{\mathbf{Z}}_H$

Fig. 2. Representative pixel gray value difference maps for the ground truth (A), MACAW (B), and HVAE (D) counterfactuals for a representative sample. (C) and (E) illustrate the failure modes for MACAW and HVAE, respectively. MACAW counterfactual: E=0.895, A=2.238, pseudo-oracle logit=0.998. HVAE counterfactual: E=0.994, A=1.310, pseudo-oracle logit=0.999.

represents a counterfactual image generated with the corresponding CGM. It can be seen that while the HVAE generated counterfactuals mostly concentrated near an effectiveness value of 1.0 (perfect target effect addition), MACAW generated counterfactuals across a wider range of effectiveness values. MACAW also introduced a higher degree of amplification compared to the HVAE, which is exemplified in the failure modes shown in Fig. 2. These results indicate that HVAE performed better than MACAW at successfully introducing the target effect and was less prone to amplification of other (unwanted) effects in the image.

Fig. 3A colour-codes the TEA metrics by the absolute value of subject effect variation sampled from the distribution $\mathcal{N}(0,1)$. It can be seen that, for MACAW, subject effect variation values further away from the mean of this distribution are correlated with a higher level of amplification. This may be due to the fact that MACAW was exposed to less data from these higher degrees of subject effects, which led to problems when trying to intervene on a target variable. However, this relationship was not found for the HVAE results – conversely, it appears that for this CGM, effectiveness was impacted more by subject effect variation (as the counterfactuals with lower effectiveness values tended to have a higher degree of subject effects, regardless of the amplification value).

Comparison of TEA and pseudo-oracle. The pseudo-oracle model classified 100% of the ground truth counterfactuals, 98.0% of MACAW-generated counterfactuals, and 83.9% of HVAE-generated counterfactuals as having the target effect added at a logit threshold of 0.5. Fig. 3B displays the TEA scatterplots colour-coded by logit values of the pseudo-oracle. It can be seen that the pseudo-oracle confidently classified MACAW counterfactuals as having the target effect added, even when amplification was relatively high – only predicting that the target effect was not added (or being unsure) when both effectiveness was low and amplification was high. Interestingly, the pseudo-oracle classified HVAE

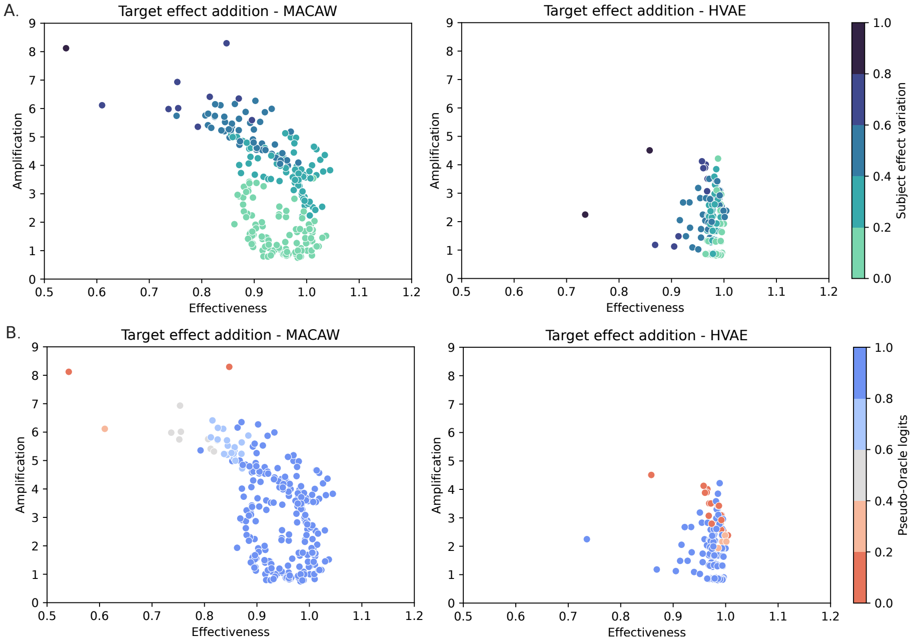

Fig. 3. Plots representing the TEA metrics of effectiveness and amplification for counterfactuals generated with MACAW (left) and HVAE (right). The TEA metrics are colour-coded by subject variation in (A) and pseudo-oracle logit value in (B).

counterfactuals as not having the target effects added for some counterfactuals where the effectiveness was at or near 1.0.

4 Discussion and Conclusion

This work demonstrated how synthetic MRI datasets with ground truth counterfactuals, combined with the newly proposed TEA metric, can successfully identify failure modes in CGMs and facilitate in-depth error analysis. As a proof-of-concept, we evaluated two causally-grounded CGMs, MACAW and HVAE, and analyzed their differences in counterfactual generation in terms of effectiveness and amplification. Importantly, this otherwise infeasible analysis was only possible due to the availability of ground truth counterfactuals through SimBA.

Pseudo-oracle evaluation is a common quantitative measure of counterfactual effectiveness, providing a single value per image to indicate CGM success in target variable intervention. However, this metric lacks specific information needed for troubleshooting and refining CGMs. For instance, in our experiments, the pseudo-oracle showed higher effectiveness for MACAW compared to HVAE. Yet, TEA plots revealed that HVAE performed better overall in effective intervention and avoiding unwanted amplification while generating counterfactuals. This highlights that pseudo-oracle evaluations may not only be uninformative, but also potentially misleading in certain scenarios. In contrast, TEA enables the disentanglement of CGM failure modes into quantitative measures of the effectiveness of target intervention and the amplification of unintended effects.

The use of SimBA datasets allows for direct comparison of counterfactual on intervention localized morphological effects, and future work should also investigate global morphology and intensity-based effects in this context. While these effects are useful for initial troubleshooting and refinement of CGM, a limitation of using this synthetic data framework is that analyses would be restricted to these effects, and may not guarantee that the model would perform the same on different imaging modalities, for instance. However, moving away from simple troubleshooting and failure mode analysis toward more realistic counterfactual validation and broader applicability, TEA could also be used in cases where ground truth counterfactuals are emulated/approximated in the real world (*e.g.,* 'traveling subjects' [3] or longitudinal data) or when using a well-validated CGM to generate pseudo 'ground truth' counterfactuals.

Here, we showed the feasibility of using SimBA and TEA to explore CGMs through a simple causal graph, which could also be extended to more complex graphs by incorporating multiple and interacting target effects, such as sex and age. In addition, while we demonstrated its utility on causally-grounded CGMs, our framework is broadly applicable to any conditional generative model.

Acknowledgments. The authors acknowledge support from the Natural Science and Engineering Research Council of Canada, the Killam Laureates, Alberta Innovates, the Parkinson Association of Alberta, the Canadian Neuroanalytics scholar program, the Canada Research Chairs Program, and the River Fund at Calgary Foundation.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Castro, D.C., Tan, J., Kainz, B., Konukoglu, E., Glocker, B.: Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. Journal of Machine Learning Research **20**(178) (2019)
2. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (2020)
3. Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S.C., Koike, S.: Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. Human Brain Mapping **42**(16), 5278–5287 (2021)
4. Melistas, T., Spyrou, N., Gkouti, N., Sanchez, P., Vlontzos, A., Panagakis, Y., Papanastasiou, G., Tsaftaris, S.A.: Benchmarking counterfactual image generation. arXiv preprint arXiv:2403.20287 (2024)

5. Monteiro, M., Ribeiro, F.D.S., Pawlowski, N., Castro, D.C., Glocker, B.: Measuring axiomatic soundness of counterfactual image models. arXiv preprint arXiv:2303.01274 (2023)
6. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 607–617 (2020)
7. Pawlowski, N., Coelho de Castro, D., Glocker, B.: Deep Structural Causal Models for Tractable Counterfactual Inference. In: Advances in Neural Information Processing Systems. vol. 33, pp. 857–869. Curran Associates, Inc. (2020)
8. Pearl, J.: The Causal Foundations of Structural Equation Modeling:. Tech. rep., Defense Technical Information Center, Fort Belvoir, VA (Feb 2012)
9. Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M.: Accurate brain age prediction with lightweight deep neural networks. Medical Image Analysis **68**, 101871 (2021)
10. Ribeiro, F.D.S., Xia, T., Monteiro, M., Pawlowski, N., Glocker, B.: High Fidelity Image Counterfactuals with Probabilistic Causal Models. In: Proceedings of the 40th International Conference on Machine Learning. pp. 7390–7425. PMLR (Jul 2023), iSSN: 2640-3498
11. Ribeiro, F.D.S., Xia, T., Monteiro, M., Pawlowski, N., Glocker, B.: High fidelity image counterfactuals with probabilistic causal models. arXiv preprint arXiv:2306.15764 (2023)
12. Sanchez, P., Tsaftaris, S.A.: Diffusion causal models for counterfactual estimation. arXiv preprint arXiv:2202.10166 (2022)
13. Sanchez, P., Voisey, J.P., Xia, T., Watson, H.I., O'Neil, A.Q., Tsaftaris, S.A.: Causal machine learning for healthcare and precision medicine. Royal Society Open Science **9**(8), 220638 (2022)
14. Stanley, E.A.M., Souza, R., Wilms, M., Forkert, N.D.: Where, why, and how is bias learned in medical image analysis models? a study of bias encoding within convolutional networks using synthetic data. eBioMedicine **111**, 105501 (Jan 2025)
15. Stanley, E.A.M., Souza, R., Winder, A.J., Gulve, V., Amador, K., Wilms, M., Forkert, N.D.: Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. Journal of the American Medical Informatics Association p. ocae165 (2024)
16. Stanley, E.A.M., Wilms, M., Forkert, N.D.: A Flexible Framework for Simulating and Evaluating Biases in Deep Learning-Based Medical Image Analysis. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 489–499. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2023)
17. Vigneshwaran, V., Ohara, E., Wilms, M., Forkert, N.: Macaw: A causal generative model for medical imaging. arXiv preprint arXiv:2412.02900 (2024)
18. Xia, K., Pan, Y., Bareinboim, E.: Neural causal models for counterfactual identification and estimation. arXiv preprint arXiv:2210.00035 (2022)
19. Xia, T., Roschewitz, M., De Sousa Ribeiro, F., Jones, C., Glocker, B.: Mitigating attribute amplification in counterfactual image generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 546–556. Springer (2024)
20. Xia, T., Sanchez, P., Qin, C., Tsaftaris, S.A.: Adversarial counterfactual augmentation: application in alzheimer's disease classification. Frontiers in radiology **2**, 1039160 (2022)