Deep Structural Causal Modelling of the Clinical and Radiological Phenotype of Alzheimer's Disease

Anonymous Author(s)
Affiliation
Address
email

Abstract

Alzheimer's disease (AD) has a poorly understood aetiology. Patients often have different rates and patterns of brain atrophy, and present at different stages along the natural history of their condition. This means that establishing the relationships between disease-related variables, and subsequently linking the clinical and radiological phenotypes of AD is difficult. Investigating this link is important because it could ultimately allow for a better understanding of the disease process, and this could enable tasks such as treatment effect estimates, disease progression modelling, and better precision medicine for AD patients. We extend a class of deep structural causal models (DSCMs) to the clinical and radiological phenotype of AD, and propose an aetiological model of relevant patient demographics, imaging and clinical biomarkers, and cognitive assessment/educational scores based on specific current hypotheses in the medical literature. The trained DSCM produces biologically plausible counterfactuals relating to the specified disease covariates, and reproduces ground-truth longitudinal changes in magnetic resonance images of AD. Such a model could enable the assessment of the effects of intervening on variables outside a randomized controlled trial setting. In addition, by being explicit about how causal relationships are encoded, the framework provides a principled approach to define and assess hypotheses of the aetiology of AD. Code to replicate the experiment can be found at: Counterfactual AD.

1 Introduction

Deep learning (DL) has demonstrated a wide range of utilities in medical healthcare, from producing accurate prognostic systems for novel diseases [1] to elucidating the previously unknown structures of complex monomeric proteins [2]. However, DL systems can still under-perform human clinicians in specific domains, for example in the task of producing differential diagnoses [3]. One factor that may contribute to the relative under-performance of current models in specific predictive tasks is that a majority of these systems rely on associative inference. Indeed, associative inference is the first rung in the hierarchy of possible inference schemes [4]. Counterfactual inference sits at the final rung, and allows for causal explanations to be used in modelling the data. In producing disease differentials, it has been argued that diagnosis is fundamentally a counterfactual process, and defining models to carefully reflect this leads to more accurate systems [3]. Additionally, DL is susceptible to learning spurious correlations [5], is sensitive to changes in the input distribution [6], and can amplify biases due to both inductive model-centric biases, and potentially biased datasets [7–9]. However, by explicitly modelling causal relationships between variables of interest, we might define more robust, transparent, and fair DL models [5].

Precision medicine requires asking questions of a causal nature, such as 'does this therapy treat these symptoms?' or 'what would this brain scan look like if the patient had a higher amyloid protein

Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

load?'. Such questions aim to determine the outcomes of interventions on variables of interest [10]. Whilst establishing causal relationships between variables normally requires a randomized controlled trial (RCT) setting, the tools of causal inference allow us to specify biologically plausible models and investigate such causal questions using observational data alone [11, 12].

Pawlowski et al. [5] introduced a modular framework for using structural causal models (SCMs) to learn the functional causal dependencies between variables of interest using DL elements. They used amortized inference to make counterfactual inference tractable for high-dimensional problems, and applied their framework to healthy brain magnetic resonance (MR) scans. This work extends their model to the radiological and clinical phenotype of Alzheimer's disease (AD). The principle aim is to enable disease progression modelling (in the imaging space) of a complex neurodegenerative condition which still respects the functional causal relationships between disease variables of interest. Such a model could not only support clinical inference, for example in tasks such as treatment effect estimation, but could also improve our general understanding of how the clinical phenotype of AD relates to its presentation on MR imaging. This is of particular importance because AD has a poorly understood aetiology (causal explanation), and patients have heterogeneous rates and patterns of brain atrophy, and also often present along different stages in the natural history of their condition. Finally, we propose a partial validation method based on brain segmentations with which model counterfactuals can be assessed.

2 Deep structural causal models

A SCM $\mathfrak{S} := (\mathbf{S}, p(\boldsymbol{\epsilon}))$ is defined as a collection $\mathbf{S} = (f_1, \ldots, f_k)$ of mechanisms $x_k := f_k(\epsilon_k; \mathsf{pa}(x_k))$, where $\mathsf{pa}(x_k)$ are the parents (direct causes) of x_k, and a joint $p(\boldsymbol{\epsilon}) = \prod_{k=1}^{K} p(\epsilon_k)$, which represents unaccounted sources of variation as independent exogenous noise variables [5]. The SCM $\mathfrak{S}$ satisfies the Markov condition [13], in that every node in the network is independent of its non-descendants given its parents. The joint distribution on the observed variables therefore factorises as $p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k|\mathsf{pa}(x_k))$. Each conditional $p(x_k|\mathsf{pa}(x_k))$ is determined by its mechanism f_k and the corresponding noise distribution $p(\epsilon_k)$ [5, 12]. Counterfactual queries can be performed in three steps referred to as abduction, action, and prediction [5, 14]: 1) In abduction, we wish to infer the 'state of the world' that is compatible with the observations $\boldsymbol{x}$, $p_{\mathfrak{S}}(\boldsymbol{\epsilon}|\boldsymbol{x})$; 2) next, we perform an action of interest, such as $\mathrm{do}(x_{I,1} := \mathsf{x}_{I,1})$, resulting in a modified SCM $\tilde{\mathfrak{S}} = \mathfrak{S}_{\mathrm{do}(x_{I,1}:=\mathsf{x}_{I,1})} := (\tilde{\mathbf{S}}, p_{\mathfrak{S}}(\boldsymbol{\epsilon}|\boldsymbol{x}))$; 3) we now infer a quantity of interest according to the distribution corresponding to the modified SCM $p_{\tilde{\mathfrak{S}}}(\boldsymbol{x})$.

For N observed variables, mechanisms $\{f_i\}_{i=1}^{N}$ must be invertible so that $\{\epsilon_i\}_{i=1}^{N}$ can computed (as per the abduction step). For scalar variables, conditional normalising flows can be used to learn bijective mappings between exogenous noise and the observed variables. Such mappings operate in the space of the data, and would therefore be computationally costly for modelling MR images. To circumvent this, Pawlowski et al. [5] decompose the image mechanism f_k into invertible h_k and non-invertible g_k functions. The noise is correspondingly decomposed as $e_k = (u_k, z_k)$, with $p(e_k) = p(u_k)p(z_k)$. Specifically, the non-invertible noise term z_k is computed by the recognition model of a conditional variational autoencoder (CVAE). The class of SCMs which use deep learning elements as functional approximations to be estimated from the data are known as deep structural causal models (DSCMs). See [5] for additional details.

3 A causal model of Alzheimer's disease

Having given an overview of DSCMs, we propose a SCM for AD (Figure 1). The following is a discussion of the medical literature encoded by the SCM. AD is a progressive, chronic neurodegenerative disease characterised by non-reversible global impairment of the function of the cerebrum [15]. The natural history is that of a deteriorating course over approximately a decade, and is marked by impairment of daily activities, memory loss, and neurobehavioural abnormalities. Such abnormalities include but are not limited to personality change, psychological disturbance, reduced executive function, loss of occupational or social functioning, and motor and speech deficits [15]. On gross examination, patients with AD have reduced brain weights of up to 200g less than average, and this could be greater in more severe disease [16]. As the most common form of dementia, accounting for up to 70% of all dementia cases [17], AD has a prevalence of approximately 30% in people over 80

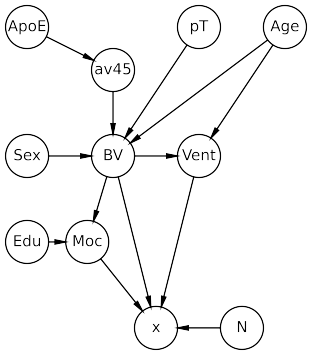

Figure 1: A structural causal model (SCM) for MR images in Alzheimer's disease (AD). The directed edges capture the causal relations between variables. ApoE = APOE4 allele, av45 = Amyloid level, as measured by the PET-derived marker AV45, pT = hyperphosphorylated tau, BV = Brain volume, Vent = ventricular volume, Edu = educational score, Moc = Montreal Cognitive Assessment score, x = axial MR image slice, N = slice number.

years [18]. Between 1990 and 2016, the number of people living with diagnosed dementia increased from 20.2 to 43.8 million worldwide [17], and the World Health Organisation (WHO) projects this to increase to 152 million by 2050 [19], underscoring the increasing importance of this condition, and the necessity with which its aetiology and natural history should be better understood.

Whilst the complete aetiology of AD has not been completely described, there are thought to be inter- and intra-neural pathways leading to its development. With respect to the former, it has been demonstrated that through reduced clearance and/or excess production, the brains of patients with AD have an increased level of beta-amyloid (Aβ) peptides [20]. Oligomers formed by these peptides are deposited as diffuse plaques, activating a number of immunological mechanisms including the complement cascade system, microglial recruitment, and cytokine formation. The resulting inflammation produces neuritic plaques, which cause neural damage and eventual cell death [21]. A gene of particular interest is the apolipoprotein E (APOE) gene, which encodes the ApoE protein. The gene has three major alleles, $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$ [22]. The ApoE-$\epsilon 4$ protein influences clearance and deposition of neurotoxic Aβ, and has been linked to sporadic AD [23].

Amyloid can be measured in the cerebrospinal fluid (CSF) or blood, and in fact the use of blood amyloid measurements remains contentious, with groups finding both increased [24–27] and decreased levels [24, 28–32] to be associated with AD risk. In view of this, we consider the use of Florbetapir F 18 (otherwise known as ^{18}FAV45 and henceforth referred to as AV45), as a surrogate marker of amyloid plaque load in-vivo. In particular, this positron emission tomography (PET)-derived imaging biomarker demonstrates three desirable qualities which make it amenable for use in the AD SCM: 1) AV45 labels Aβ plaques in anatomically appropriate areas of the brain in patients with pathologically confirmed AD [33]; 2) the marker binds to at-risk areas in patients with AD, but with minimal cortical binding in healthy controls [34]; 3) there is a high correspondence between Aβ plaques and AV45 binding in postmortem series [35]. As expected, cortical amyloid burden as assessed by AV45 is highly correlated with APOE4 carrier status [36] ('APOE4 $\rightarrow$ AV45').

With respect to intraneural aetiology, patients with AD develop aggregations of abnormally phosphorylated tau (P-tau) protein, which forms dystrophic neurites and neurofibrillary tangles. Tau helps to stabilise microtubules within cells, and the formation of neurites can cause direct neural damage [37] ('pTau $\rightarrow$ brain vol.'). Other relevant risk factors include age, sex, and educational attainment/cognitive scores. The incidence of AD doubles for each 5-year period after the age of 65 [38], and less than secondary school-level education is associated with an increased risk of a number of dementias, including AD [39–41]. The Montreal cognitive assessment (MOCA) score has recently been shown to be effective at detecting AD ('brain vol. $\rightarrow$ MOCA'), and is more sensitive to detecting mild cognitive impairment in at-risk patients than other cognitive scores such as the MMSE [42]. Baseline MOCA scores are affected by education ('education $\rightarrow$ MOCA'). Finally, AD is more common in women than in men [17, 43].

4 Experimental setup

All data was acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, specifically the ADNI2 protocol. There were 268 participants, with 123 (45.9%) females and median age of 73.05 years. There were 61 cognitively normal participants, 160 with mild cognitive impairment, and 47 with a diagnosis of AD. Structural T1 weighted MRI scans were accrued for all participants. Each scan was linked to the relevant demographic, disease biomarker, and cognitive assessment variables for a given participant. Data entries were sporadically missing (assumed missing at random; MAR), and were imputed using multiple imputations with chained equations (MICE) [44]. The earliest Inversion Recovery Spoiled Gradient echo sequence (SAG IR-SPGR) MRI was retrieved for each participant, skull-stripped using the HD-BET brain extraction tool [45], and bias field-corrected with the N4 software [46]. The MNI ICBM152 brain atlas was loaded using the NiLearn package [47], and images were resampled to the same size as the atlas ($197\times233\times189$) with linear interpolation. Images were then rigidly registered to the MNI template using ANTs [48]. The intensity values of the 2D axial slices were normalised by rescaling the minimum and maximum values of each slice to $[0, 255]$. The middle 10 axial slices of each MRI were then saved as PNG files for training the DSCM. During training, the images were uniformly dequantized by addition of random noise [49]. To prevent overfitting, the images were randomly cropped from their original size to 192×192, and are then downsampled to 64×64 during training. Images were centre-cropped during counterfactual image inference.

With regards to the DSCM, the mechanisms $f_i \in \mathbf{S}$ are represented by (conditional) rational spline normalizing flows [5, 10] with the exception of the images, sex, slice number, and APOE status variables. The images are modelled using a CVAE architecture, where the encoder and decoder functions consist of 5 levels of 3 modules of $(\text{LeakyReLU}(0.1), \text{BN}_\theta, \text{Conv}_\theta)$, where $\text{LeakyReLU}(\phi)$ is a leaky ReLU with an angle of negative slope parameter ϕ, BN is a batch normalisation layer, and Conv is a convolutional layer. We learn the binary probability of the sex variable by sampling from a Bernoulli distribution (female = 1, male = 0). APOE4 status and slice number are sampled from uniform distributions (APOE4 status in $\{0, 1, 2\}$, and minimum to maximum number of slices, respectively), as per Reinhold et al. [10]. The normalising flows had Gaussian base distributions, with the scale and location parameters set to the logarithm of the variance and mean of the training set, respectively. For conditional flows, the hypernetworks predicting the transformation parameters were multi-layer perceptrons with two hidden layers. The number of nodes in the hidden layers are (8, 16) for an input dimension of ≤ 2, and (16, 24) for the brain volume network, which has an input dimension of 4 (sex, AV45, P-tau, and age). The parameters for the image CVAE and scalar flows were optimised for the evidence lower bound (ELBO; estimated using 4 MC samples) using the Adam optimiser [50] with learning rates of 10^{-5} and 5×10^{-3}, respectively, for 300 epochs. For image reconstruction and counterfactuals, 32 MC samples were used. At inference, all learned mechanisms $f_i \in \mathbf{S}$ were fixed, and the single world intervention graph (SWIG) formalism was used to produce counterfactuals [51]. The model is trained on a single NVIDIA RTX 3090 GPU.

5 Results

Qualitative investigation We test the hypothesis that single atomic interventions in the SCM describing the relationships between non-imaging variables of interest produce changes in images that reflect biologically plausible associations. The results of selected single intervention counterfactuals on a single participant are illustrated in Figure 2. Altering biological sex has little morphological effect (column 1), and this is reflected in the difference map. This is in fact the expected finding in a dataset with a high AD signal, as it was found that there were no statistically significant sex-based differences in cortical volumes in AD patients [52]. Decreasing age broadly increases intensity values in the cortical regions, whereas increasing age has the opposite effect with visible cortical degeneration produced in the counterfactual image (columns 2 and 3, respectively). It should be noted that increasing the age value not only demonstrates cortical neurodegenration in the counterfactual image, but the ventricles are also expanded, which is the expected neuro-radiological result of increasing age [53]. Directly intervening on whole brain and ventricular volumes (columns 3 - 6) leads to expected morphological effects in both instances. Increases in these volumes leads to pixel-wise intensity changes that reflect increases the size of the whole brain and/or ventricular space. Meanwhile, setting these volumes to lower values produces the opposite effect, with contractions of the whole brain and/or ventricular space.

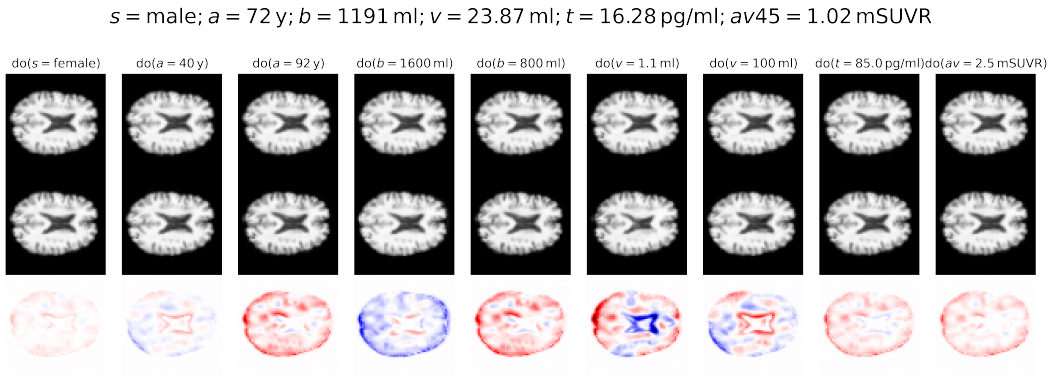

Figure 2: Single intervention counterfactual queries for a single participant using a deep structural causal model (DSCM) of MR images in Alzheimer's disease (AD). The top row shows the original image, the middle row illustrates the counterfactuals, and the third row is a pixel-wise difference map between the original image and the counterfactual. Red indicates a pixel-wise reduction in intensity, whereas blue indicates an increase intensity in the counterfactual image relative to the original image. s = sex, a = age, b = brain volume, v = ventricular volume, t = hyperphosphorylated tau, av = AV45.

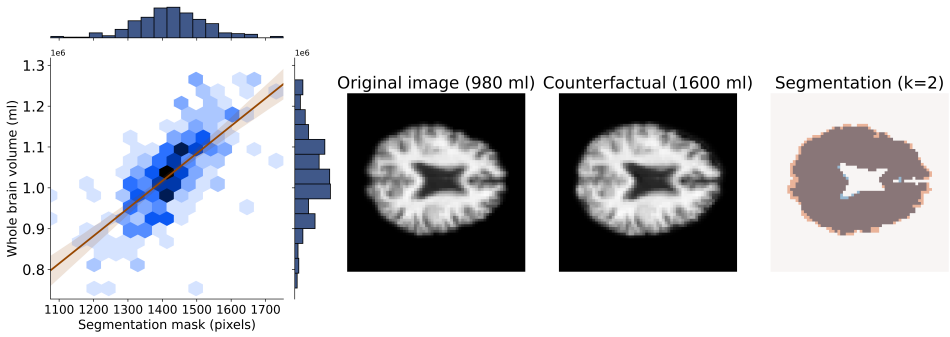

Figure 3: **Left:** Brain segmentation masks (2D) correlate to whole brain volumes calculated from 3D MR images. **Right:** An example with a brain volume of 980 ml, counterfactually increased to 1600 ml. The original segmentation mask has 1284 pixels which represent the brain, and the counterfactual has an appropriately larger mask of 1386 pixels. The segmentation masks are overlaid and the larger (counterfactual) mask can be seen to subsume the original mask.

With respect to disease biomarkers, increasing the P-tau or AV45 level (to 85 pg/ml and 2.5 mSUVR, respectively) leads in both instances to a reduction in pixel intensity values reflecting degenerative change (columns 8 and 9). It can be seen that there is slightly greater reduction in intensity values around the ventricle for higher AV45 levels than higher P-tau levels. More examples of selected counterfactual interventions can be seen in Appendix A.1. Counterfactuals in alternative anatomical slices and multiple-intervention counterfactuals can additionally be seen in Appendix A.2.

Quantitative evaluation To quantitatively assess the counterfactual images, we first segment the brains in each of the 2D slices using K-means clustering of the pixel intensity values. We then consider the correlation between brain segmentation mask sizes and whole brain volume, where the latter metric is calculated from application of the FSL neuroimaging toolkit to the entire 3D MR image [54]. Figure 3 illustrates a positive correlation between the two metrics with a Pearson's correlation coefficient of $r = 0.71, p < 0.001$. Segmentation masks in 2D therefore act as a reasonable proxy of whole brain volume. We perform an atomic intervention directly on brain volume and assess whether the segmentation mask changes in the expected way. Figure 3 illustrates an example of this.

We consider three related approaches to the quantitative assessment of counterfactuals. First, we intervene on a variable which can be directly accrued from the 2D image slices (in this case, intervening on brain volume and directly measuring the change in segmentation mask size). Second,

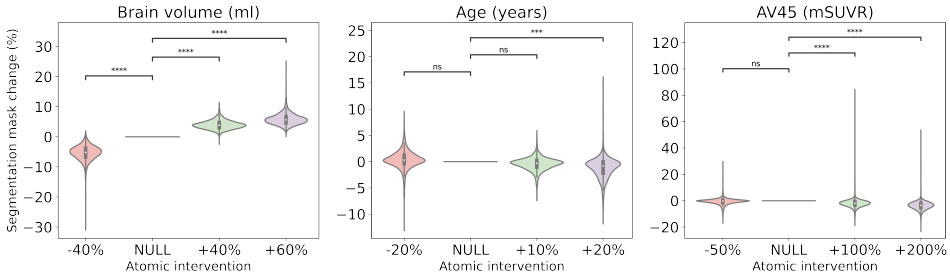

Figure 4: For each 2D brain slice, an intervention is performed for each of the brain volume, age, and AV45 variables. The interventions are set as percentages of the original measurement of the image so that counterfactual measurements are expected to change in the same direction (e.g. setting brain volume values to +40% for each participant should increase segmentation mask size across all slices). Statistical annotations represent Welch's t-tests with a Bonferonni correction. ***: $p \leq 10^{-3}$, ****: $p \leq 10^{-4}$, ns: No statistical significance.

we intervene on a related variable where we have a known hypothesis of causal influence and assess how segmentation masks change (we intervene on age, and measure brain volume, with an expectation that aging leads to neurodegeneration and therefore smaller brain volumes). Finally, we assess the influence of intervening on a disease biomarker of AD (we increase AV45, which is an in-vivo measure of amyloid load, and expect that greater AV45 levels are related to more severe disease and therefore smaller brain volumes). Welch's ANOVA demonstrates a significant difference in all three groups of interventions $F(\text{age}) = 11.36, p < 0.001, F(\text{brain volume}) = 731.04, p < 0.001, F(\text{AV45}) = 38.59, p < 0.001$, respectively. Welch's t-tests with Bonferroni corrections show statistically significant differences in the expected directions in all three groups of interventions. Figure 4 illustrates these results. Finally, the DSCM is capable of approximating ground-truth longitudinal changes, as can be seen in Appendix A.3.

6 Discussion

This work extends the DSCM to the radiological and clinical phenotype of AD by proposing a biologically plausible SCM of the condition. To the best of our knowledge, this is the first DSCM which captures the aetiological process of neurodegenration in AD. The model is capable of producing appropriate single- and multiple-intervention counterfactual images based on the SCM. By learning the causal dependencies between the model variables, we can perform queries based on participant demographics, relevant imaging and clinical biomarkers, and cognitive assessment scores.

An important limitation of this work is that the current setup assumes that there are no unobserved confounders. However, the DSCM is lacking a number of potentially important variables to properly control for confounding. For example, it has been observed that patients with psoriatic arthritis are more likely to develop AD [55]. Therefore, accounting for autoinflammatory disease could produce a more robust model with more reliable counterfactuals. The interventions presented in this work lead to broadly appropriate counterfactuals. For example, increasing brain volume produces a counterfactual with a larger brain segmentation mask. However, increasing brain volume by 40% does not necessarily increase mask size by 40%. This represents another limitation which is likely due to using 2D slices to produce counterfactuals which use statistics that are computed from 3D volumes, which is a sub-optimal setup [10]. Extending the DSCM to 3D volumes therefore represents a natural avenue of future research.

Despite these limitations, the DSCM provides a framework for principled counterfactual inference. The model is capable of producing biologically plausible counterfactual images which directly relate to the proposed aetiological process of AD in MR images. The model achieves this with a relatively limited subset of the ADNI2 cohort. The counterfactuals are produced on a per-participant basis, which essentially offers a framework for explaining the data, since we can analyse the changes resulting from manipulating any given set of variables. There are a wide variety of potential use-cases for these counterfactuals, such as developing a better understanding of the aetiological process of a complex neurodegerative disease such as AD, and disease progression modelling.

References

[1] Abdulaal A, Patel A, Charani E, Denny S, Mughal N, Moore L. Prognostic modeling of COVID-19 using artificial intelligence in the United Kingdom: Model development and validation. Journal of Medical Internet Research. 2020 8;22(8):e20259.

[2] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873. 2021 7;596(7873):583-9.

[3] Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nature Communications 2020 11:1. 2020 8;11(1):1-9.

[4] Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:180104016. 2018.

[5] Pawlowski N, Castro DC, Glocker B. Deep Structural Causal Models for Tractable Counterfactual Inference. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020); 2020. p. 857-69.

[6] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. 2013 12.

[7] Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings. 2017 7:2979-89.

[8] Hüllermeier E, Fober T, Mernberger M. Inductive Bias. Encyclopedia of Systems Biology. 2013:1018-8.

[9] Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:180601261. 2018.

[10] Reinhold JC, Carass A, Prince JL. A Structural Causal Model for MR Images of Multiple Sclerosis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021;12905 LNCS:782-92.

[11] Morgan SL, Winship C. Counterfactuals and causal inference: Methods and principles for social research. Counterfactuals and Causal Inference: Methods and Principles for Social Research. 2007 1:1-319.

[12] Peters J, Janzing D, Schölkopf B. Elements of Causal Inference. Foundations and Learning Algorithms. 2017:288.

[13] Geiger D, Pearl J. On the Logic of Causal Models. Machine Intelligence and Pattern Recognition. 1990 1;9(C):3-14.

[14] Pearl J. Causality: Models, reasoning, and inference, second edition. Causality: Models, Reasoning, and Inference, Second Edition. 2011 1:1-464.

[15] Bird TD. Alzheimer disease overview. GeneReviews®[Internet]. 2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1161/.

[16] Double K, Halliday G, Krill J, Harasty J, Cullen K, Brooks W, et al. Topography of brain atrophy during normal aging and Alzheimer's disease. Neurobiology of aging. 1996;17(4):513-21.

[17] Nichols E, Szoeke CEI, Vollset SE, Abbasi N, Abd-Allah F, Abdela J, et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology. 2019 1;18(1):88-106.

[18] Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. American Journal of Public Health. 1998;88(9):1337-42.

[19] World Health Organization. Dementia;. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia.

[20] O'Brien RJ, Wong PC. Amyloid precursor protein processing and Alzheimer's disease. Annual review of neuroscience. 2011 7;34:185-204.

[21] Joe E, Ringman JM. Cognitive symptoms of Alzheimer's disease: clinical management and prevention. BMJ (Clinical research ed). 2019 12;367.

[22] Christensen KD, Roberts JS, Uhlmann WR, Green RC. Changes to perceptions of the pros and cons of genetic susceptibility testing after APOE genotyping for Alzheimer disease risk. Genetics in Medicine. 2011;13(5):409-14.

[23] Marcus C, Mena E, Subramaniam RM. Brain PET in the Diagnosis of Alzheimer's Disease. Clinical nuclear medicine. 2014 10;39(10):e413.

[24] Yang Y, Giau VV, An SSA, Kim S. Plasma Oligomeric Beta Amyloid in Alzheimer's Disease with History of Agent Orange Exposure. Dementia and Neurocognitive Disorders. 2018;17(2):41.

[25] Wang MJ, Yi S, Han JY, Park SY, Jang JW, Chun IK, et al. Oligomeric forms of amyloid-β protein in plasma as a potential blood-based biomarker for Alzheimer's disease. Alzheimer's Research and Therapy. 2017 12;9(1):1-10.

[26] An SSA, Lee BS, Yu JS, Lim K, Kim GJ, Lee R, et al. Dynamic changes of oligomeric amyloid β levels in plasma induced by spiked synthetic Aβ42. Alzheimer's Research and Therapy. 2017 10;9(1):1-10.

[27] Youn YC, Kang S, Suh J, Park YH, Kang MJ, Pyun JM, et al. Blood amyloid-β oligomerization associated with neurodegeneration of Alzheimer's disease. Alzheimer's Research and Therapy. 2019 5;11(1):1-8.

[28] Graff-Radford NR, Crook JE, Lucas J, Boeve BF, Knopman DS, Ivnik RJ, et al. Association of low plasma Abeta42/Abeta40 ratios with increased imminent risk for mild cognitive impairment and Alzheimer disease. Archives of neurology. 2007 3;64(3):354-62.

[29] Pesaresi M, Lovati C, Bertora P, Mailland E, Galimberti D, Scarpini E, et al. Plasma levels of beta-amyloid (1-42) in Alzheimer's disease and mild cognitive impairment. Neurobiology of aging. 2006 6;27(6):904-5.

[30] van Oijen M, Hofman A, Soares HD, Koudstaal PJ, Breteler MM. Plasma Abeta(1-40) and Abeta(1-42) and the risk of dementia: a prospective case-cohort study. The Lancet Neurology. 2006 8;5(8):655-60.

[31] Zhou L, Chan KH, Chu LW, Kwan JSC, Song YQ, Chen LH, et al. Plasma amyloid-β oligomers level is a biomarker for Alzheimer's disease diagnosis. Biochemical and biophysical research communications. 2012 7;423(4):697-702.

[32] Bagyinszky E, Kang MJ, Van Giau V, Shim KH, Pyun JM, Suh J, et al. Novel amyloid precursor protein mutation, Val669Leu ("Seoul APP"), in a Korean patient with early-onset Alzheimer's disease. Neurobiology of Aging. 2019 12;84:1-236.

[33] Choi SR, Golding G, Zhuang Z, Zhang W, Lim N, Hefti F, et al. Preclinical Properties of 18F-AV-45: A PET Agent for Aβ Plaques in the Brain. Journal of Nuclear Medicine. 2009 11;50(11):1887-94.

[34] Wong DF, Rosenberg PB, Zhou Y, Kumar A, Raymont V, Ravert HT, et al. In Vivo Imaging of Amyloid Deposition in Alzheimer Disease Using the Radioligand 18F-AV-45 (Flobetapir F 18). Journal of Nuclear Medicine. 2010 6;51(6):913-20.

[35] Clark CM, Schneider JA, Bedell BJ, Beach TG, Bilker WB, Mintun MA, et al. Use of Florbetapir-PET for Imaging β-Amyloid Pathology. JAMA. 2011 1;305(3):275-83.

[36] Johnson KA, Sperling RA, Gidicsin CM, Carmasin JS, Maye JE, Coleman RE, et al. Florbetapir (F18-AV-45) PET to assess amyloid burden in Alzheimer's disease dementia, mild cognitive impairment, and normal aging. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2013 10;9(0):S72.

[37] Reddy PH, Oliver DMA. Amyloid Beta and Phosphorylated Tau-Induced Defective Autophagy and Mitophagy in Alzheimer's Disease. Cells. 2019 5;8(5).

[38] Bachman DL, Wolf PA, Linn RT, Knoefel JE, Cobb JL, Belanger AJ, et al. Incidence of dementia and probable Alzheimer's disease in a general population: the Framingham Study. Neurology. 1993;43(3 Pt 1):515-9.

[39] Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet. 2020 8;396(10248):413-46.

[40] Gottesman RF, Albert MS, Alonso A, Coker LH, Coresh J, Davis SM, et al. Associations Between Midlife Vascular Risk Factors and 25-Year Incident Dementia in the Atherosclerosis Risk in Communities (ARIC) Cohort. JAMA Neurology. 2017 10;74(10):1246.

[41] Larsson SC, Traylor M, Malik R, Dichgans M, Burgess S, Markus HS. Modifiable pathways in Alzheimer's disease: Mendelian randomisation analysis. The BMJ. 2017 12;359:j5375.

[42] Pinto TCC, Machado L, Bulgacov TM, Rodrigues-Júnior AL, Costa MLG, Ximenes RCC, et al. Is the Montreal Cognitive Assessment (MoCA) screening superior to the Mini-Mental State Examination (MMSE) in the detection of mild cognitive impairment (MCI) and Alzheimer's Disease (AD) in the elderly? International psychogeriatrics. 2019 4;31(4):491-504.

[43] Niu H, Álvarez-Álvarez I, Guillén-Grima F, Aguinaga-Ontoso I. Prevalence and incidence of Alzheimer's disease in Europe: A meta-analysis. Neurología (English Edition). 2017 10;32(8):523-32.

[44] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011;30(4):377-99.

[45] Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. Human Brain Mapping. 2019 12;40(17):4952-64.

[46] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE transactions on medical imaging. 2010 6;29(6):1310-20.

[47] nilearn/nilearn: Machine learning for NeuroImaging in Python;. Available from: https://github.com/nilearn/nilearn.

[48] Avants BB, Tustison N, Song G, et al. Advanced normalization tools (ANTS). Insight j. 2009;2(365):1-35.

[49] Theis L, Van Den Oord A, Bethge M. A note on the evaluation of generative models. 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings. 2015 11.

[50] Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR; 2015. p. 0.

[51] Richardson TS, Robins JM. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Center for the Statistics and the Social Sciences, University of Washington Series Working Paper. 2013;128(30):2013.

[52] Sangha O, Ma D, Popuri K, Stocks J, Wang L, Beg MF. Structural volume and cortical thickness differences between males and females in cognitively normal, cognitively impaired and Alzheimer's dementia population. Neurobiology of aging. 2021 10;106:1-11.

[53] Dinsdale NK, Bluemke E, Smith SM, Arya Z, Vidaurre D, Jenkinson M, et al. Learning patterns of the ageing brain in MRI using deep convolutional networks. NeuroImage. 2021 1;224:117401.

[54] Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JLR, Griffanti L, Douaud G, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage. 2018 2;166:400-24.

[55] Kim M, Park HE, Lee SH, Han K, Lee JH. Increased risk of Alzheimer's disease in patients with psoriasis: a nationwide population-based cohort study. Scientific reports. 2020 12;10(1).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 5 and 6.
   (b) Did you describe the limitations of your work? [Yes] Please see paragraph 2 of 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Whilst associative deep learning is susceptible to inductive model-centric biases and potentially biased datasets, graphical modelling helps to clarify modelling assumptions, and could subsequently produce more transparent and fairer models, circumventing these biases to an extent. A brief restatement of this is given in Section 1.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see the hyperlink at the end of the abstract.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Complete hyperparameter details are given in the code - please see end of the abstract. For the model setup in this work, please see Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Whilst significance testing is reported for the quantitative analysis of counterfactuals in Section 5, the same model was chosen to produce all counterfactual images in this work.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Please see paragraph 3 of section 1.
   (b) Did you mention the license of the assets? [Yes] The asset licenses (MIT license) for the original DSCM is given by the authors of the paper, and is repeated again in this work - please see code which is linked at the end of the abstract.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please see end of abstract.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to investigate neurodegeneration in Alzheimer's disease (AD). Informed, written consent is accrued from all participants. Additional details can be found on the ADNI website at: ADNI background.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Please note all ADNI data is anonymized at point of serving on the data repository (the LONI Image and Data Archive). Additional details can be found here: ADNI data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

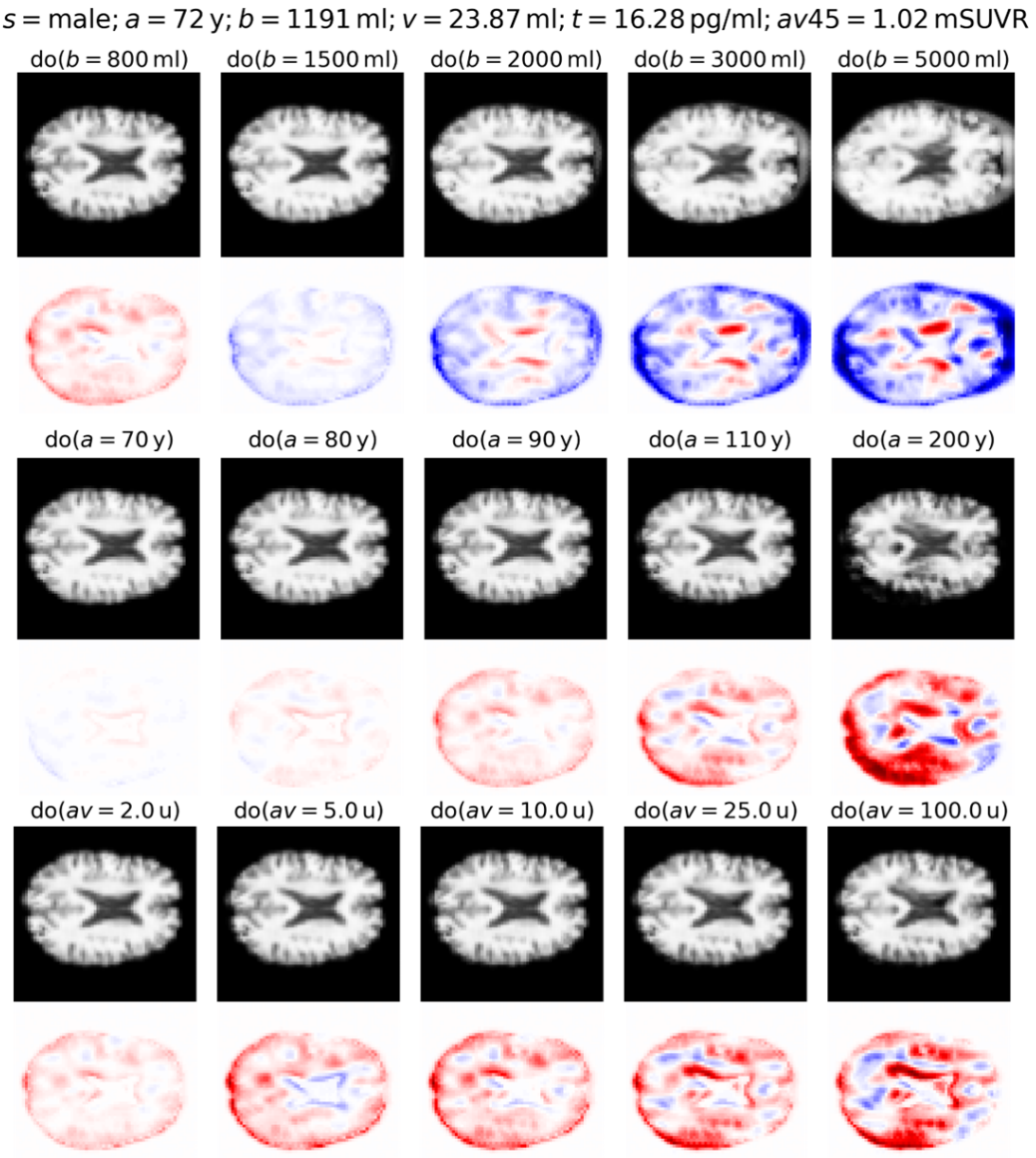

Figure 5: Single intervention counterfactual queries for a single participant using a deep structural causal model (DSCM) of MR images in Alzheimer's disease (AD). Out-of-distribution counterfactuals are investigated for three selected variables. Pixel-wise intensity-difference maps are shown following each row of counterfactual images. Red indicates a pixel-wise reduction in intensity, whereas blue indicates an increase in pixel-wise intensity in the counterfactual image relative to the original image. b = brain volume, a = age, av = AV45.

A Appendix

A.1 Out-of-distribution counterfactual queries

Figure 5 illustrates multiple counterfactual queries for selected demographic and imaging biomarkers, namely age, brain volume, and AV45. Note that out-of-distribution counterfactual queries are investigated, for example by setting the brain volume to 5000ml, or the age to 200 years. Increasing brain volume leads to large increases in cortical tissue, whilst increasing age leads to dramatic neurodegenrative change. As AV45 increases, there is increased expansion of the ventricular space with associated cortical degeneration, particularly of the temporoparietal cortex.

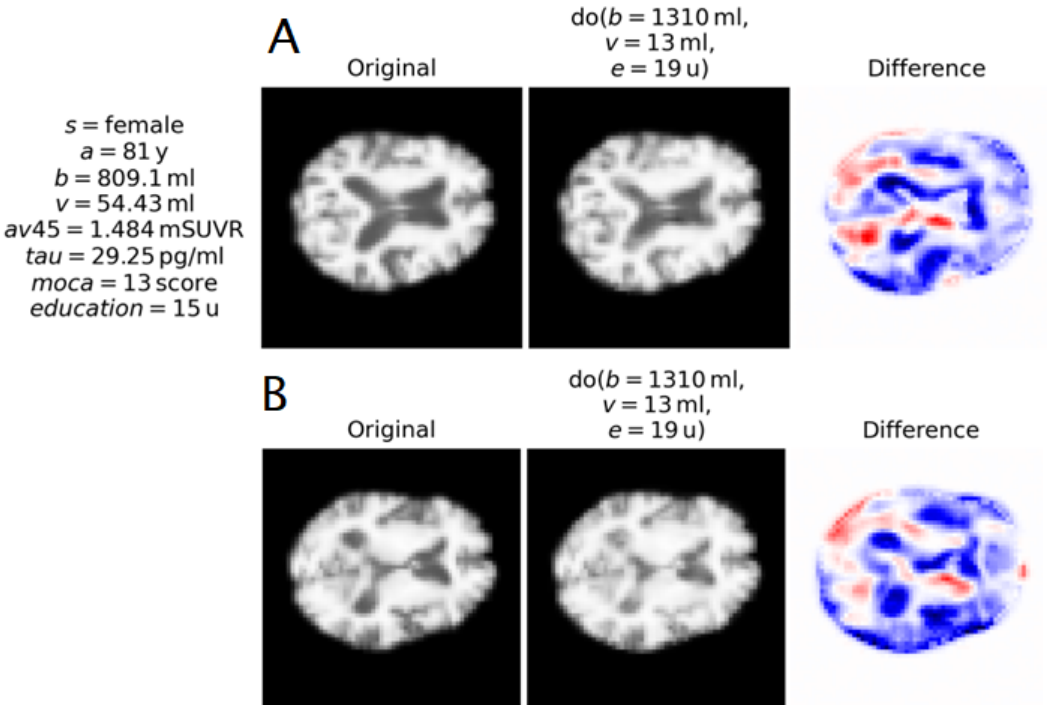

Figure 6: Multiple intervention counterfactual query for a single participant using a deep structural causal model (DSCM) of MR images in Alzheimer's disease (AD). The leftmost images are the original images, the centre images are the counterfactuals produced by the model, and the rightmost images are pixel-wise intensity-difference maps. Red indicates a pixel-wise reduction in intensity, whereas blue indicates an increase in pixel-wise intensity in the counterfactual image relative to the original image. Having set brain volume and educational attainment to a larger value, and ventricular volume to a lower one, it can be seen that the cortical regions have broadly increased intensity values in the counterfactual images, and that the ventricular space is contracted. Panel A illustrates an approximately medial axial slice, whereas panel B is a more caudal (inferior) slice. s = sex, a = age, v = ventricular volume, av45 = AV45, tau = hyperphosphorylated tau, moca = Montreal cognitive assessment score, education/e = educational attainment score; u is a unit-measure of educational attainment.

A.2 Alternative anatomy and multiple-intervention counterfactuals

Following the single intervention counterfactuals experiment, we test the hypothesis that multiple atomic interventions (i.e., constant reassignments of multiple nodes in the SCM simultaneously) produce changes in images that reflect biologically plausible associations.

The DSCM allows for multiple model variables to be set to an intervention state, producing a multiple intervention counterfactual. For example, we could produce a counterfactual query of the nature 'what would this MR image look like if the participant had a larger brain volume, a smaller ventricular volume, and a higher educational attainment score?'. Figure 6A illustrates the output of such a query for a different ADNI participant than in the single intervention case above. Here, setting the brain volume to a larger value, whilst reducing the volume of the ventricular space and increasing participant educational attainment produces clear expansion in the cortex. Results are particularly noticeable in the temporal regions, where atrophic changes appear to be restored in the counterfactual image. There is dramatic reduction in the ventricular space, as can be seen in the counterfactual and as reflected in the intensity-difference map.

As in the single intervention counterfactuals case, alternative anatomical counterfactuals can be produced for multiple intervention queries. Figure 6B illustrates the same counterfactual query as in Figure 6A in a more caudal anatomical slice.

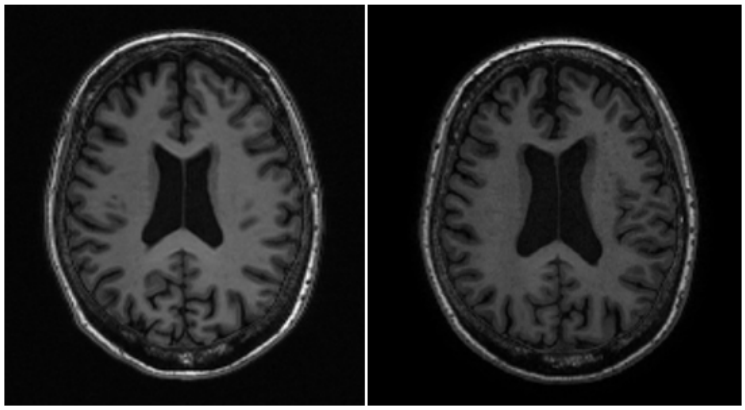

Figure 7: Longitudinal axial MR slices for an ADNI participant. The images are accrued six years apart and have been matched to be as anatomically equivalent as possible. It can be seen that the ventricular space has asymmetrically enlarged in this participant over the follow-up period.

A.3 Counterfactual image validation through longitudinal case-study analysis

Whilst validating counterfactuals in real datasets is often severely constrained (or impossible as true counterfactuals may never be observed), we may consider approximate counterfactuals in the form of longitudinal data as a framework for partial validation of the model. If the DSCM has learned an (approximately) correct set of functional mechanisms and generative procedure, then the model counterfactuals should approximately match trends in longitudinal data.

A.3.1 Participant characteristics

To assess DSCM counterfactuals we consider a participant case study for a 72 year old female with a brain volume of 980.60 ml, a ventricular volume of 30.19 ml, a P-tau level of 25.11 pg/ml and AV45 of 1.52 mSUVR. The participant has a MOCA score of 23 at baseline, and an educational score of 16. The participant then goes on to have a follow-up MRI scan six years later, and has all relevant clinical variables recorded during their visit. On their follow-up visit, the participant now has a brain volume of 960.19 ml, a ventricular volume of 38.97 ml, a P-tau level of 29.25 pg/ml, and an AV45 of 1.42 mSUVR. Their educational score is unchanged. Despite broadly worsening disease and imaging biomarkers (with the exception of AV45), the participant has a higher MOCA score of 24 on their follow-up visit.

A.3.2 Ground-truth changes and counterfactual comparison

Figure 7 illustrates two unprocessed axial slices from the baseline scan and the follow-up scan. The slices have been matched to be as anatomically equivalent as possible. It can be seen that the most evident macroscopic change is an enlarged ventricular space between the scans.

Performing atomic interventions of the form $\text{do}(x_{I,N} = \mathsf{x}_{I,N})$ where $x_{I,N}$ is an intervention variable graphically has the effect of disconnecting the intervention variable $x_{I,N}$ from its parents and setting it to a fixed state. This has implications for how we are able to produce counterfactuals given the current SCM of AD. For example, if we intervene on brain volume and ventricular volume simultaneously, then the effects of age, P-tau, and AV45 are negated (i.e., additionally intervening on these variables will have no effect). There are therefore a number of ways to produce counterfactuals to test the model. In this instance, we consider two approaches to produce counterfactuals. The first is to intervene on maximally distal continuous ancestors (i.e. nodes with no parents with the exception of AV45, because participant genotype is fixed), and the second is to intervene on proximal child nodes. Figure 8 illustrates the nodes being intervened on in both cases. The two approaches and their relevant counterfactuals along with the ground-truth change are shown in Figure 9.

The difference map between the baseline image and the follow-up image is noisy due to imperfect registration of the follow-up to the baseline. This leads to some mismatch between the sulci and gyri in the 2D plane. However, the map nonetheless faithfully captures the principle morphological change

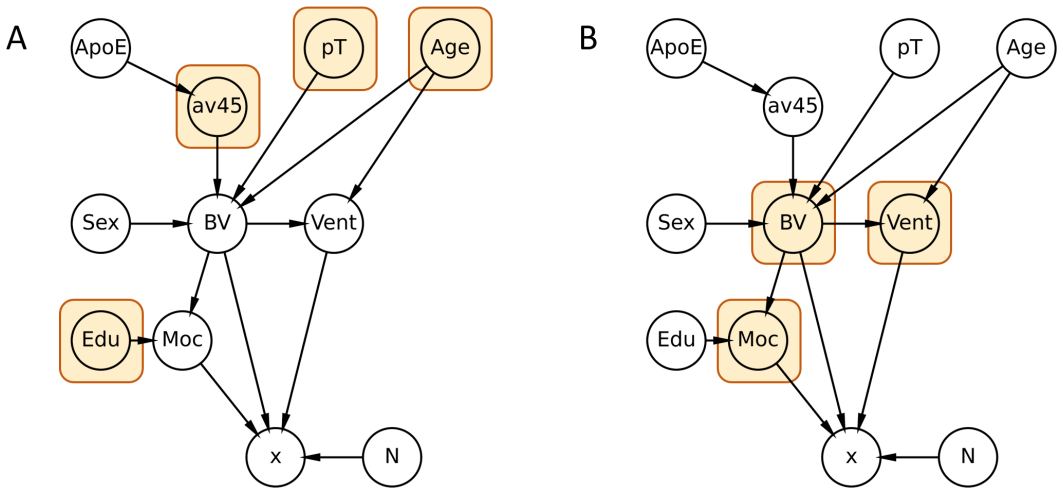

Figure 8: Alternative approaches to producing validation counterfactuals. The nodes being intervened upon are highlighted in both instances. In panel A, distal continuous ancestors are intervened upon, and in panel B, intervention states are set for proximal causes.

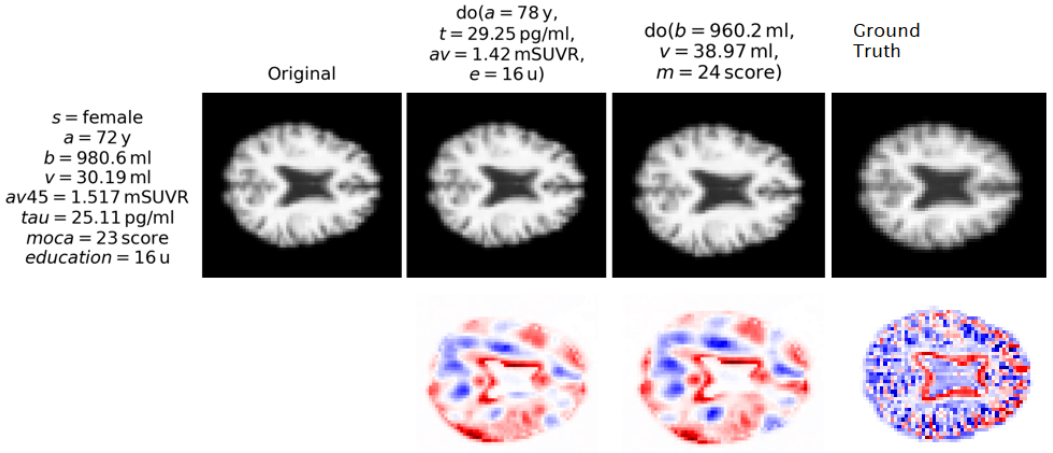

Figure 9: Assessment of counterfactuals produced by a deep structural causal model (DSCM) for MR images of Alzheimer's disease (AD). Counterfactual images relating to two sets of interventions on the causal graph are shown alongside the ground truth follow-up image on the top row. The second row shows pixel-wise intensity-difference maps between the counterfactuals/ground truth and the original (baseline) image. Red indicates a pixel-wise reduction in intensity, whereas blue indicates an increase in pixel-wise intensity in the counterfactual image relative to the original image.

(as illustrated in unprocessed axial slices from the original MRI scans in Figure 7) of an enlarged ventricular space. With the exception of the measured AV45 level, which decreased slightly in this participant over time, the remaining metrics are expected to be associated with increased atrophic change. Indeed, the counterfactuals demonstrate some atrophic change around the superior temporal gyrus and in a region just anterior to the precuneus. This provides evidence that the DSCM has learnt broadly appropriate mechanisms $f_i \in \mathbf{S}$, as it produces plausible counterfactuals in both the single- and multi-intervention settings. The counterfactual images are also capable of appropriately approximating ground-truth longitudinal changes.