

Spring 5-2020

## PERSONALIZED ESTIMATION AND CAUSAL INFERENCE VIA DEEP LEARNING ALGORITHMS

YUANYUAN LIU

*UTHealth School of Public Health*

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/uthsph\\_dissertsopen](https://digitalcommons.library.tmc.edu/uthsph_dissertsopen)



Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

---

### Recommended Citation

LIU, YUANYUAN, "PERSONALIZED ESTIMATION AND CAUSAL INFERENCE VIA DEEP LEARNING ALGORITHMS" (2020). *UT School of Public Health Dissertations (Open Access)*. 129.  
[https://digitalcommons.library.tmc.edu/uthsph\\_dissertsopen/129](https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/129)

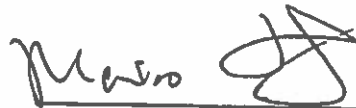
This is brought to you for free and open access by the School of Public Health at DigitalCommons@TMC. It has been accepted for inclusion in UT School of Public Health Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact [digitalcommons@library.tmc.edu](mailto:digitalcommons@library.tmc.edu).

PERSONALIZED ESTIMATION AND CAUSAL INFERENCE VIA DEEP LEARNING  
ALGORITHMS

by

YUANYUAN LIU, M.S., B.S.

APPROVED:



MOMIAO XIONG, PH.D.



GOO JUN, PH.D.



SUJA S. RAJAN, M.H.A., M.S., PH.D.



DEAN, THE UNIVERSITY OF TEXAS  
SCHOOL OF PUBLIC HEALTH

Copyright  
by  
Yuanyuan Liu, B.S., M.S., Ph.D.  
2020

## DEDICATION

To Yumei Guo, Yi Liu, Songyuan Liu

PERSONALIZED ESTIMATION AND CAUSAL INFERENCE VIA DEEP LEARNING  
ALGORITHMS

by

YUANYUAN LIU  
M.S., Harvard University, 2013  
B.S., Fudan University, 2011

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS  
SCHOOL OF PUBLIC HEALTH  
Houston, Texas  
May, 2020

## ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Momiao Xiong for the continuous support of my Ph.D study and research, for his patience, motivation, and knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Suja S. Rajan, Dr. Goo Jun, and Dr. Kai Zhang, for their insightful comments and encouragement, but also for the critiques which motivated me to explore my research topics from various perspectives.

My deep appreciation also goes to UTHHealth Innovation for Cancer Prevention Research Training Program Pre-doctoral Fellowship (Cancer Prevention and Research Institute of Texas grant # RP160015) and all my CPRIT mentors: Dr. Roberta Ness, Dr. Patricia Mullen, Dr. David Loose, Dr. Sahiti Myneni, and Dr. Trevor Cohen. I am very grateful for their support and tough love which has always encouraged me to think out of the box and has enabled me to become an innovation enthusiast.

I thank my fellow labmates for the inspiring discussions, for the encouragement when I was frustrated, and for the joy we shared over the past few years.

Last but not the least, I would like to say a heartfelt thank you to my parents and my husband, who has always been there for me and believe in me during this challenging journey. I could not have achieved any of this without their encouragement and love.

# PERSONALIZED ESTIMATION AND CAUSAL INFERENCE VIA DEEP LEARNING ALGORITHMS

Yuanyuan Liu, MS, PhD  
The University of Texas  
School of Public Health, 2020

Dissertation Chair: Momiao Xiong, PhD

Personalized approaches have shown great potential to transform modern medicine. As challenging as it may sound, we are making tremendous progress with the help of data sciences and machine learning. Two fundamental tasks in data sciences are prediction and inference. In this dissertation, I proposed to address these two tasks using deep learning approaches in the setting of personalized medicine. First, I developed a novel framework to estimate individualized treatment effects (ITE), which quantified the variation in response to the same treatment among patients with heterogeneous profiles. The ITE estimation has the potential to replace the one-size-fits-all average treatment effects (ATE) commonly used in clinical practice and provides more accurate patient-specific treatment guidance. Second, I developed a statistical test to determine pairwise causation between two sets of continuous variables. Despite of the massive data available, the primary methods to determine causation clinically are through randomized controlled experiments and animal studies, which are highly inefficient or sometimes even infeasible. With this new statistical test, we were able to draw causal conclusions from observational data instead of experimental data alone, which

was beneficial in terms of understanding underlying disease mechanisms. Statistical simulation was conducted to demonstrate the validity and accuracy of the proposed methods. Last but not least, I applied the developed methods on real-life datasets to demonstrate their usage. The TCGA lung cancer dataset was used to estimate ITE for patients with complex covariate structure. I also performed an end-to-end causal discovery for Alzheimer's disease using the medical images from the ADNI dataset. The results indicate deep learning based approaches offer great flexibility and deep insights for biomedical data, which will help us bridge the gap in precision medicine.



## TABLE OF CONTENTS

List of Tables .....	i
List of Figures .....	ii
List of Appendices .....	iii
Background .....	1
1. Literature Review.....	1
1.1 Neural Networks .....	1
1.2 Convolutional Neural Networks .....	3
1.3 Auto-Encoders .....	5
1.4 Generative Adversarial Network .....	7
1.5 Causal Inference.....	8
2. Public Health Significance.....	19
3. Specific Aims.....	21
Methods.....	22
1. Individualized Treatment Effect (ITE) Estimation .....	22
1.1 Problem Formulation .....	22
1.2 Model Description .....	23
1.3 Optimization .....	26
1.4 Simulation Experiments.....	27
2. Adversarial Causal Test (ACT) .....	29
2.1 Problem Formulation .....	29
2.2 Model Description .....	29
2.3 Optimization .....	31
2.4 Simulation Experiments.....	31
3. Application Studies.....	34
3.1 ITE Estimation for TCGA .....	34
3.2 Causal Discovery for ADNI.....	38
4. Human Subjects, Animal Subjects, or Safety Considerations .....	48
Results.....	48
1. Simulation Experiments for ITE Estimation.....	48
2. Simulation Experiments for ACT .....	50
3. Application Studies.....	51
3.1 ITE Estimation for TCGA .....	51
3.2 Causal Discovery for Alzheimer's disease .....	55
Discussion .....	59
1. ITE-Related Genes.....	59
2. Causal Discovery for AD.....	61
Conclusion .....	63

References.....	88
Appendices.....	112

## LIST OF TABLES

Table 1. 3D filters in five convolutional layers. ....	76
Table 2. Comparisons of different methods for ITE using data generated in scenario I .....	76
Table 3. Comparisons of different methods for ITE using data generated in scenario II.....	76
Table 4. Type I errors when there is no association or causation .....	77
Table 5. Type I errors when there is association but no causation .....	77
Table 6. Causation power for ACTs at varying sample sizes using the quadratic transformation. ....	77
Table 7. Causation power for ACTs at varying sample sizes using the cubic transformation. ....	77
Table 8. Power for ACTs to correctly identifying causation direction at varying sample sizes using the quadratic transformation. ....	78
Table 9. Power for ACTs to correctly identifying causation direction at varying sample sizes using the quadratic transformation. ....	78
Table 10. Comparison of different methods in simulation study.....	79
Table 11. Comparison of different methods with feature selection.....	79
Table 12. Plug-in validation of different methods for TCGA lung cancer dataset.....	80
Table 13. Chemotherapy (treatment) vs. Combinational therapy (control).....	81
Table 14. Radiation (treatment) vs. Combinational therapy (control).....	81
Table 15. Radiation (treatment) vs. chemotherapy (control).....	81
Table 16. Counts for Observed/Recommended Treatment on entire dataset .....	82
Table 17. Feature selection using lasso* and R-square using selected features .....	82
Table 18. Selected significant correlations between latent factors and genes .....	83
Table 19. AD prediction accuracy on five-fold cross validation. ....	86
Table 20. Average sensitivity and specificity over five-fold cross validation.....	86
Table 21. Causations between DTIs image ROIs and AD disease status .....	87

## LIST OF FIGURES

Figure 1. An example of neural networks.....	67
Figure 2. The structure of a conditional generative adversarial network (CGAN). ....	67
Figure 3. DAG for a randomized controlled trial .....	68
Figure 4. Causal DAGs for Observational Studies. ....	68
Figure 5. Architecture of CFGAIN.....	69
Figure 6. Procedure for conducting the proposed adversarial causal test (ACT). ....	69
Figure 7. Comparisons of ITE results from linear regression and CFGAIN .....	70
Figure 8. Comparisons of ITE results for different methods .....	71
Figure 9. Power for ACTs at varying sample sizes. Cause X was randomly sampled from normal distribution and effect Y was obtained through the non-linear transformation of X plus the random noise N: $Y = f(X) + N$ . Quadratic and cubic functions were used for the non-linear mapping. ....	72
Figure 10. Comparisons of various methods for ITE estimation without feature selection. ....	73
Figure 11. Comparisons of various methods for ITE estimation with feature selection. ....	73
Figure 12. Visualization of the brain regions with relative importance values at the baseline, 6 months, 12 months and 24 months. The deeper red color of the brain region, the more important for AD prediction.....	74
Figure 13. Three brain regions showed causation to AD.....	75

## LIST OF APPENDICES

Table S1. P-values of 43 genes that showed causation or association or both with the frontal, left temporal lobe. ....	112
Table S2. P-values of 46 genes that showed causation or association or both with the right temporal lobe region.....	118

## BACKGROUND

### 1. Literature Review

Over the past few decades, we have witnessed the rise of precision medicine, which was defined as ‘an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment and lifestyle for each person’ by the National Institutes of Health (NIH). Two major tasks in statistics that are relevant to precision medicine are prediction and inference. Prediction focuses on forecasting future events based on observed ones, whereas inference emphasizes understanding the data generating process or causes of observed events. When it comes to precision medicine, we not only want to predict individualized risks/effects, but also hope to understand the underlying causes of diseases and the diverse manifestations observed on individuals. Thanks to the increasing availability of data in this digital era and rapid development in deep learning approaches, we have been moving closer to the goal of individualized risk prediction and conducting causal discoveries in settings that are not limited to randomized clinical trials.

#### *1.1 Neural Networks*

Deep learning is a subfield of machine learning, which centers around neural networks. As the name suggests, neural networks were motivated by the neural system of human beings that is in charge of processing signals and transmitting information to our brains (Figure 1). As popular as it may sound nowadays, the concept of neural networks dates back to 1940s. In 1943, Walter Pitts and Warren McCulloch designed the ‘thresholded logic unit’, which was a step function that worked in a similar way as the human neuron did by employing a threshold (McCulloch & Pitts, 1943). Alan Turing first described the ‘Turing

test', which listed criteria to determine if machine could be considered as intelligent in his paper published in 1950 (Machinery, 1950). However, the perceptron model proposed by Frank Rosenblatt in 1958 was considered as the building blocks of neural networks (Rosenblatt, 1958). The Rosenblatt perceptron was a binary single neuron model to classify two linearly separable classes. It output 1 if the weighted average of input plus bias was positive, and output -1 otherwise. Despite of the impressive performance of perceptron, Marvin Minsky and Seymour Papert pointed out in 1969 that the perceptron was theoretically impossible to learn non-linear functions (Minsky & Papert, 2017), which marked the start of the first artificial intelligence (AI) winter.

It was not until the development of backpropagation in 1986 that the scientific community started to feel amazed again by the potential of neural networks with stacked perceptrons and activation functions to allow for non-linearity. The backpropagation learning algorithm worked by taking the derivatives of the network's loss functions and back-propagate the errors to adjust the weights and biases in the network (Rumelhart, Hinton, & Williams, 1986). In 1991, Kurt Hornik proved the universal approximation capabilities of multilayer feedforward networks (Hornik, 1991). Despite of a few early successes, neural networks failed to scale to larger problems, mainly due limited computing power, which brought another setback for AI from late 1980s to early 1990s.

Thanks to the increasing computing power and innovative designs such as deep belief networks (Hinton, 2009), neural networks with multiple hidden layers began to attract more and more attention, and the name 'deep learning' was introduced as a rebranding of neural networks in 2006. In recent years, the use of graphics processing units (GPUs) and various

regularization techniques such as dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) and batch normalization (Ioffe & Szegedy, 2015) have tremendously stabilized and speeded up the training of neural networks with more advanced designs including the convolution neural networks (CNN), recurrent neural networks (RNN) and residual neural networks (ResNet). In addition, the application of deep learning algorithms in medicine has achieved great success, producing results comparable or surpassing the traditional machine learning methods (Cheng et al., 2016; Esteva et al., 2017).

## ***1.2 Convolutional Neural Networks***

Convolutional neural networks (CNNs) was one of the most widely used deep learning techniques in the field of computer vision. This concept was first introduced by Kunihiko Fukushima in 1980 as neocognitron, where he presented the early forms of convolutional layers and down-sampling layers (Fukushima, 1980; Schmidhuber, 2015). However, Fukushima updated weights using WTA-based unsupervised learning rules (Fukushima, 2013) instead of backpropagation and the down-sampling technique he used were spatial averaging, in contrast to max pooling, which are more popular nowadays. In 1990, Yann LeCun published the LeNet, which achieved 1% error rate on zip code digits with minimal preprocessing (LeCun et al., 1990). Later he published a series paper improving the design of CNNs and expanding the applications (LeCun, Bottou, Bengio, & Haffner, 1998; LeCun, Haffner, Bottou, & Bengio, 1999). The ImageNet project, which was first presented in 2009, was considered a milestone in the development of CNNs. ImageNet has been a growing database which holds millions of images with annotations obtained from crowdsourcing and this massive amount of data enabled CNNs with much deeper



architecture. Krizhevsky et al. came up with the AlexNet that brought the error rate down by more than 10% in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where they showed that the depth of the network was crucial for its top performance (Krizhevsky, Sutskever, & Hinton, 2012). CNNs have since evolved rapidly, with applications in various fields such as automatic driving, facial recognition and medical diagnosis.

The great potential of CNNs has drawn a lot of attention from the medical imaging community, resulting in an unprecedented shift in the ways to extract information. For example, Esteva et al. presented a CNN classification of skin cancer, which had performance comparable to dermatologists (Esteva et al., 2017). A team from Google relied on CNNs to learn features from retinal fundus images to predict age, gender and major cardiac events with satisfactory results (Gulshan et al., 2016). Kermany et al. established CNNs with transfer learning to screen patients with common treatable blinding retinal diseases and demonstrated performance at the level of human experts, among many others (Kermany et al., 2018).

Despite the great success of CNNs in medical imaging, they have been criticized as being ‘black box’ methods, which were non-transparent and difficult to interpret to medical professions and patients, greatly limiting the usage of such techniques. A lot of research efforts have been put into creating interpretable CNNs. The most straightforward method would be looking at the layer activations by plotting the activation maps for each filter (Stanford, 2019). Another method would be visualizing the weights/filters of the first few layers in a CNN, where relatively smooth patterns should be observed if the network

converges well (Stanford, 2019). Google has also made their contribution to interpretable CNNs by proposing using optimization and regularization for feature visualization (Olah, Mordvintsev, & Schubert, 2017). In addition, spatial embedding methods such as t-SNE can be used to embed high-dimensional vector from the last layer before classification into two-dimensional vector for each image and all the embedded images can be visualized in a grid. Similarities in embedded images indicate that CNNs ‘see’ the images as close ones (Stanford, 2019). Perhaps the most intuitive method would be the occlusion heat maps (Zintgraf, Cohen, Adel, & Welling, 2017). These are created by iterating over regions of the image, removing the regions one at a time and assessing how the prediction probabilities change for the class of interest. Variations of this method has been applied to images of everyday objects, but much less commonly in medical research involved CNNs.

### ***1.3 Auto-Encoders***

Auto-encoder is one of the most commonly known unsupervised-learning techniques (Kingma & Welling, 2013). It consists of two part, one encoder and one decoder. The encoder takes in the data  $x$ , passes through any hidden layers, and compresses it into a latent-space representation  $h$ . The encoder performs the transformation  $h = f(x)$ . The decoder treats  $h$  as input, passes through any hidden layers, and reconstruct the output in the input space. Mathematically, the decoder performs the transformation  $y = g(h)$ . The cost function can defined as the mean squared error between  $x$  and  $y$ . An auto-encoder is optimized when  $x$  and  $y$  are as close as possible and thus the training is carried out by minimizing the mean squared error loss function.

There have been many variations of auto-encoders. One of the most popular types is the variational autoencoder (VAE) (Kingma & Welling, 2013). Contrast to the classic auto-encoder where a single scalar value is output by the encoder for each encoded dimension, VAE provides a probabilistic way to learn the data representation in the latent space. The encoder of VAE (also called the recognition model for VAE) describes a probability distribution for each encoded dimension, whereas the decoder (also called the generative model) tries to reconstruct the input based on a continuous and more smoothed latent space representation (Jordan, 2018). In order to accommodate different types of data and tasks, researchers have also been using convolutional auto-encoders for images, recurrent or long short term memory (LSTM) auto-encoders for text mining and signal enhancement, and regularized auto-encoders for tasks such as classification and denoising.

The concept of auto-encoder was first proposed around 1987 as a method for unsupervised pre-training of neural networks by DH Ballard (Ballard, 1987), they have been mostly used to perform data denoising and dimension reduction/feature learning over the past few years. Vincent et al. first proposed denoising autoencoders (DAE) to extract and compose robust features (Vincent, Larochelle, Bengio, & Manzagol, 2008), which has been further extended and applied to image denoising, language generation and signal enhancement (Freitag & Roy, 2018; Gondara, 2016; P. Xiong et al., 2016). Auto-encoders have also achieved a great success in automatic feature engineering. They are not only capable of learning linear transformations like the principal component analysis does, but also good at capturing non-linear transformations, with the help of non-linear activation functions. For example, Dong et al. used auto-encoder regularized network to perform

driving style representation learning (Dong, Yuan, Yang, Li, & Zhang, 2017), which outperformed other state-of-art methods; LSTM autoencoders achieved a great success in learning video representations, which was proven to be very helpful when the learned representations were later used in other supervised learning problems (Srivastava, Mansimov, & Salakhudinov, 2015); Auto-encoders have also been applied to speech waveforms to learn meaningful latent representation of speech, generating results comparable to the top entries in the ZeroSpeech 2017 unsupervised acoustic unit discovery task (Chorowski, Weiss, Bengio, & van den Oord, 2019).

#### ***1.4 Generative Adversarial Network***

Generative adversarial network (GAN) was first introduced by Ian Goodfellow in 2014 (Goodfellow et al., 2014). GAN is an innovative framework which consists of two neural networks, a generator and a discriminator, that are trained simultaneously in an adversarial fashion. The generator is aimed to generate data that comes from the same distribution as the training data, whereas the discriminator learns to distinguish whether a sample comes from the training data or the generator. This competition improves both the generator and the discriminator, resulting in a stage where the generated data is indistinguishable from the real training data. GAN framework does not need any assumptions about the underlying distributions or rely on Markov chains for inference, which makes it extremely valuable for many applications. Another remarkable feature is that once GAN learns about the underlying distribution of the training data, it is able to generate data that has not been seen in the training set.

Shortly after the publication of the vanilla version of GAN, Mehdi Mirza and Simon Osindero introduced the conditional GAN (CGAN, Figure 2.), where the framework was able to learn distributions conditioned on another variable (Mirza & Osindero, 2014). Radford et al. replaced the simple neural networks with CNNs in both discriminators and generators and proposed the deep convolutional generative adversarial networks (DCGANs), which have been widely applied in computer vision (Radford, Metz, & Chintala, 2015). Despite the successful stories of GANs, they were notoriously hard to train due to problems like non-convergence, mode collapse, diminished gradients and sensitivity to hyper-parameters. Arjovsky et al. introduced the Wasserstein GAN (WGAN) as a remedy to these problems by using the Wasserstein distance in the optimization (Arjovsky, Chintala, & Bottou, 2017); Gulrajani et al. proposed to add gradient penalty to the loss function of WGAN, further improving the performance of GANs (Gulrajani, Ahmed, Arjovsky, Dumoulin, & Courville, 2017). Meanwhile, GANs have been considered as one of the major approaches in unsupervised learning and many variations have been published such as Coupled GAN (M.-Y. Liu & Tuzel, 2016), Auxiliary Classifier GAN (Odena, Olah, & Shlens, 2017), DiscoGAN (Kim, Cha, Kim, Lee, & Kim, 2017) and Boundary Seeking GAN (Hjelm et al., 2017).

## ***1.5 Causal Inference***

### **1.5.1 Causation vs. Association**

Causal inference has been a fundamental task in scientific research as well as philosophical studies. Conceptually speaking, causal inference aims to determine if there

exists a causal connections between variables based on their occurrences in response to the change of other variables. The first task in causal inference is to tell the difference between causation and association. As human beings, we grasp these concept well without realizing it. For example, we know that carrying an umbrella kept us dry in the rain from the past experience. Had we not carried an umbrella, we would have become wet in the rain. Therefore, we conclude that carrying an umbrella is a cause for staying dry in the rain. In contrast, there is a positive correlation between ice cream sales and the number of sunglasses sold, but we would never say increasing ice cream sales is a cause for more sunglasses sold.

As J. Pearl has pointed out, the basic distinction between causation and association is coping with change, or to put it in another way, coping with interventions (Pearl, 2009). Causation is different from association by taking a step further to understand the dynamics under changing conditions/different interventions, whereas association only emphasizes making inference under static conditions (Pearl, 2009).

To define a causal effect on a population level, there are two sets of mathematical notations that encode the same meaning. The first set of notations was introduced by Judea Pearl, called ‘do- Calculus’, where  $E[Y|do(T = 0)] \neq E[Y|do(T = 1)]$  indicates causation between T and Y, assuming there is no confounding.  $do(T = t)$  means an intervention to artificially force the variable T to take the value t. This is different from the conditional probability, which was based on static conditions. For example, let  $p(Y|T)$  stands for the distribution of temperature Y when thermometer displays a reading of T. We could observe pairs of Y and T at random times. Knowing T will give me some information about Y.

However,  $P(Y|do(T = t))$  means that we manually force the reading of the thermometer to  $t$ , but this will not affect the actual temperature to change. In this example,  $P(Y|do(T = t))$  does not depend on the value of  $t$  here, and it should be the same as  $P(Y)$ .

Another framework would be counterfactuals proposed by Rubins (Hernán MA, 2019). It stemmed directly from the setting of randomized controlled trials. In a randomized trials, a participant will only be randomized to a single treatment group and the observed outcome is called the factual outcome. A counterfactual outcome is defined as what would have happen had the participant been randomized to a different treatment group. The average causal effect in population under this framework is defined as  $E[Y^{t=1}] \neq E[Y^{t=0}]$ , if the treatment  $T$  is binary.

In the field of artificial intelligence, causation is generally reflected in the generating mechanism for the observed data. In contrast to association, causation is less of interest when the purpose is forecasting/prediction and there is a large amount of data. Given enough amount of data, deep learning algorithms are capable to capture rich information that is predictive of the outcome, without the need to worry about the underlying mechanism. For example, gender and age may have large predictive power for certain types of cancer, but we would not reach the conclusion that they are causes for the disease. However, when the amount of data is limited, which is common in the medical setting, understanding the data generating process is crucial and can help us get further insights about disease etiology.

### **1.5.2 Directed Acyclic Graphs**

The most commonly used way to describe the causal connections between variables is through directed acyclic graphs (DAG). ‘Directed’ implies the connections in the graph have directions and ‘acyclic’ means that there are no cycles in the graph: you can never go back to the same variable where you start by following the directions of the arrows. The variables here have another name called ‘nodes’ and the arrows are referred as ‘directed edges’ in a DAG. Despite the fact that DAG have implications in many applications such as structural equation models (SEM) and Bayesian networks, I will use causal DAG and DAG interchangeably in this dissertation. Causal DAG requires that the common causes of any pair of variables are also in the graph (Hernán MA, 2019). Causal DAG is not only useful to illustrate the causal assumptions for observed data, but also encodes the associations between variables. In fact, causation always leads to association, but the reverse is not true.

Figure 3 is a DAG for randomized controlled trials (RCTs), which is considered as the gold standard to establish causation. For example, a pharmaceutical company is interested in determine the effect of a new drug T on the survival time for cancer patients. They randomize the patients to the drug and a type of placebo and compare the survival time for the two groups of patients. If the patients randomized to the new drug have significantly longer average survival time compared to patients in the placebo group, they may conclude that the new drug causes the longer survival and thus it is effective.

Figure 4 demonstrate a common DAG used for observational studies. For example, X denotes social economic status; T stands for insurance plans; Y is health care expenditure. Social economic status determines which type of insurance is a person will purchase, and insurance type has a causal effect on health care expenditure, but social economic status also



has direct causal effect on a person's health care expenditure. In this scenario, social economic status can be considered as a confounder for the causations between insurance and healthcare expenditure. If we assume there is no residual confounding, which means there is no variable other than social economic status that is the common cause of insurance and health care expenditure, we will be able to infer the direct causal effect of T on Y if and only if we block the backdoor path from T to Y. In statistics, we achieve this blocking by adjusting or controlling X in causal models. This is illustrated in a DAG by putting a box around X. After blocking the backdoor path from T to Y, we can safely erase the node X and the edges coming from X. In this way, we are able to get the same DAG in Figure 2. This type of analysis is called observational causal inference, which aims to determine causations from observational data when RCTs are impractical or unethical. For example, when we are interested in determine if a certain gene cause a type of cancer, it is impossible for us to randomize genes to different group of people; if we would like to study the effect of smoking on lung cancer, it is unethical to randomly assign participants to a smoking group where they are required to smoke heavily.

### **1.5.3 Average Treatment Effect**

Observational causal inference emphasizes to draw causal conclusions using observational data. This can be considered as a type of transfer learning, where we learn from the observational data and transfer the knowledge into a RCT setting as illustrated in Figure 2. Causal effects can be defined on two levels - the population level and the individual level, corresponding to average treatment effect (ATE) and individualized treatment effect (ITE).

Classical work focused more on estimating ATE. One of the most commonly used methods to estimate ATE from observational studies is propensity score method (Austin, 2011). Propensity score is the probability of getting the treatment for each individual given the covariate profile. It is a balancing score, which is designed as a way to block the backdoor path from the treatment to the outcome through confounders (i.e. the treatment and the confounders are independent conditional on propensity scores), so that observational studies can mimic the design of RCT, assuming the propensity score is correctly estimated. After obtaining the propensity scores, covariates should distribute identically within the same levels of propensity scores. There are various ways to incorporate propensity score in the estimation of ATE. The first way would be to treat propensity score as a covariate in the regression models (Elze et al., 2017). Depending on the type of outcomes, regression models are built with treatment and propensity score as two covariates in the model. Transformations or higher order terms for propensity scores may be needed depending the nature of causation. In contrast to this method, the other three methods separate propensity score from the analysis and consider it more as a study design feature (Austin, 2011). For example, study sample could be matched based on their propensity score, and the analysis could be carried out as a matched study (Dehejia & Wahba, 2002; Rosenbaum & Rubin, 1985). Another method would be stratification by propensity score, where the entire sample is divided into multiple strata based on propensity scores. Regression models are built within each stratus and later pooled together to get an overall ATE (Lunceford & Davidian, 2004; Rosenbaum & Rubin, 1984). In addition, Rosenbaum proposed the inverse probability of treatment weighting (IPTW) using propensity score. Each subject is assigned a weight defined by  $w_i =$

$\frac{T_i}{S_i} + \frac{1-T_i}{1-S_i}$ , where  $T_i$  stands for treatment assignment (0 or 1) for subject  $i$ , and  $S$  indicates the propensity score for that subject. By applying IPTW, we create a pseudo-population where ATE for this population can be estimated directly using regression models (Austin & Stuart, 2015; Rosenbaum, 1987). Structural equation models and Bayesian networks can also be used to estimate the average treatment effect, but they require the causal Markov assumption and the causal faithfulness assumption to hold for valid inference (Spirtes & Zhang, 2016).

#### **1.5.4 Individualized Treatment Effect**

Given the increasing amount of data and the heterogeneous effects of treatments in medical setting (Angrist, 2004; Kravitz, Duan, & Braslow, 2004), individualized treatment effect (ITE) has received more and more attention. ITE estimation is most straightforward in the framework of counterfactuals, where the difference between the expected factual outcome and the expected counterfactual is considered as the treatment effect for a subject, assuming a binary treatment. Propensity score based methods can be adapted to estimate ITE by predicting the counterfactuals using the regression models adjusting for propensity score (Dehejia & Wahba, 2002). However, they rely on parametric assumptions and thus have limited usage. Porter et al., presented the targeted maximum likelihood estimators (TMLEs) (Porter, Gruber, Van Der Laan, & Sekhon, 2011), which were double robust semi-parametric estimators, among many others (Kang & Schafer, 2007).

Machine-learning based approaches offer much flexibility in terms of estimating ITE. However, the lack of counterfactuals for ITE estimation poses a difficulty for applying any supervised learning technique directly. The majority work in this area used tree-based

methods to estimate the counterfactuals first and then made inferences about heterogeneous treatment effect on subgroup level or ITE on individual level. For example, Jennifer Hill discussed the potential of Bayesian additive regression trees (BART), which naturally learned heterogeneous treatment effects without pre-specification (Hill, 2011). Athey and Imben introduced an algorithm based on regression trees, where they first used a subsample to construct a partition of the population, and then estimated the causal effects within each partition to identify the heterogeneity (Athey & Imbens, 2016). Lu et al. compared several tree-based methods for estimating ITE and concluded that counterfactual synthetic random forest achieved best performance by constructing separate forests for each treatment (Lu, Sadiq, Feaster, & Ishwaran, 2018). Wager and Athey developed a nonparametric causal forest under the counterfactual framework, which extended previous algorithms into both the classification and regression scenarios (Wager & Athey, 2018). In addition, they proved the consistency and asymptotic properties for their estimators.

Given the increasing popularity of deep learning algorithms, several group of researchers has started to tackle the problem of ITE estimation using neural networks. Johansson et al. borrowed ideas from learning representations and domain adaption, proposing to first learn a representation of the features/confounders and then the function that maps the treatment and the learned representation to the outcome domain. Besides ensuring low-error prediction of the factuals and counterfactuals, they additionally emphasized to balance the distribution of features of the treatment populations through enforcing similarity between learned representations for treatment arms (Johansson, Shalit, & Sontag, 2016). Shalit et al. adopted the same idea by proposing a network structure that had a shared

representation layer before splitting into two arms to estimate the outcomes for two treatment respectively. They incorporated the integral probability metric (IPM) measure of distance to ensure the similarity between two treatment groups over the shared representation layer (Shalit, Johansson, & Sontag, 2017). Alaa et al. formulized ITE estimation as a multi-task learning problem and unified propensity score approach and deep neural networks, which they named as the deep counterfactual networks with propensity-dropout (DCN-PD) (A. M. Alaa, Weisz, & van der Schaar, 2017). This network consisted of two sub-networks, one multitask network with a few shared layers before splitting into two feed-forward arms for the potential outcomes, and a feed-forward network to learn propensity scores. Propensity scores were used to determine the dropout probability. The dropout probability was higher for subject whose propensity score was around 0.5, indicating ambiguous treatment assignment. Alaa and van der Schaar adopted a nonparametric Bayesian framework in which they model the potential outcomes as the outputs of a function in a vector-valued reproducing kernel Hilbert space (vvRKHS)(A. M. Alaa & van der Schaar, 2017). This multitask Gaussian process for ITE inferences enable the estimation of credible intervals, which came very handy in terms of precision medicine. Recently, Yoon et al. introduced the novel framework, termed generative adversarial nets for individualized treatment effects (GANITE) (Yoon, Jordon, & van der Schaar, 2018b), where they implemented two separate GAN blocks for generating potential outcomes and estimating ITE. This was the first attempt to incorporate GAN into the area of ITE estimation, which served as a strong motivation for this dissertation work.

### 1.5.5 Causal Discovery

All the previous discussion for causal inference focuses on determining the treatment effect assuming the causal DAG is known, which is barely the case in reality. Therefore, another problem that has attracted much attention is to determine the causal structure, termed ‘causal discovery’ (Spirtes & Zhang, 2016). Given the need to determine causal structure from observational data, several groups of researchers have proposed constraint-based search for causal structures (Spirtes & Zhang, 2016). These type of algorithms rely on checking conditional independence in the population and background knowledge, which output graphical objects representing a Markov equivalence class (Spirtes & Zhang, 2016). Thus, this type of algorithms do not output unique patterns and could not check all possible DAGs (Spirtes & Glymour, 1991; Spirtes et al., 2000). Over the past few years, the general-purpose Boolean Satisfiability Solver (SAT) was proposed as a constrained optimization technique used for causal discovery, but such algorithm does not scale well when the number of variables involved in a causal structure is large due to the difficulty in optimization (Hyttinen, Hoyer, Eberhardt, & Jarvisalo, 2013; Spirtes & Zhang, 2016; Triantafillou & Tsamardinos, 2015).

Another fundamental problem that directly ties to causal discovery is how to determine the causation direction between two variables based on observational data (Spirtes & Zhang, 2016). The key to this problem is to capture the information asymmetry in the data. The first class of methods is based on SEMs involving  $X$  and  $Y$ . Such SEMs can be formulized using the following mathematical equations:

$$Y = f(X, \varepsilon_1; \theta_1)$$

$$X = g(Y, \varepsilon_2; \theta_2)$$

Here  $\varepsilon_1$  and  $\varepsilon_2$  are error terms independent of  $X$  and  $Y$  respectively.  $f$  and  $g$  denote the functions describing how one variable is generated from another, which belong to a class of properly constrained class.  $\theta_1$  and  $\theta_2$  are parameters involved in  $f$  and  $g$ . The above two SEMs are fit using the data and the direction in which the estimated error term is independent of the hypothetical cause is considered the plausible direction based on the observed data (Spirtes & Zhang, 2016). However, this type of SEM based methods suffer from the identifiability problem, which depends on if the causal asymmetry can be reflected in the observed data. For example, when  $f$  and  $g$  are linear functions and the error terms is normally distributed, the causal direction is undefinable. Therefore, constraints are needed for the mapping functions  $f$  and  $g$  such that the independence condition holds only for one direction when it comes to SEM-based methods (Spirtes & Zhang, 2016).

When SEMs are too restrictive, nonlinear additive noise model (ANM) can be used to introduce the nonlinear generating mechanism (Hoyer, Janzing, Mooij, Peters, & Schölkopf, 2009; Peters, Janzing, & Schölkopf, 2010). Mathematically, ANM is a class of model satisfying the following formulation:

$$Y = f(X; \theta_1) + \varepsilon_1$$

The ANM is a special case of post-nonlinear causal models (PNL) (K. Zhang & Hyvärinen, 2009a, 2009b), which applies the non-linear transformation twice:

$$Y = f_2(f_1(X; \theta_1) + \varepsilon_1)$$

Determination of plausible causal direction for ANMs and PNLs also relies on fitting the models in both the forward and backward fashions and test for independence between the estimated error terms and hypothetical cause.

Despite the great potential of GANs and the remarkable similarity between the formulation of GAN's generator and SEMs, this deep learning approach has not been introduced into the field of causal discovery until recently. Lopez-Paz and Oquab introduced the classifier two-sample test (C2ST) to determine if two samples are drawn from the same distribution, which could be used as a metric to monitor the convergence of GANs (Lopez-Paz & Oquab, 2016). They discussed the novel application of this test for causal discovery. However, their results were more of a proof-of-concept realization without formally defining a test statistic and assessing the power and errors.

## **2. Public Health Significance**

Patient heterogeneity has caused many uncertainties in medical research and clinical practice. For example, individuals tend to have various environmental exposure, genomic profiles and health status, leading to differential risks for diseases; although same type of targeted therapy is administrated, people with same clinical conditions may respond systematically different due to their own unique immune background (Reuben et al., 2017). Therefore, personalized medicine has been proposed as a remedy to the problem of heterogeneity, which uses patients' specific information to predict risks, plan treatments or make a prognosis (<https://www.cancer.gov/publications/dictionaries/cancer-terms/def/personalized-medicine>).



However, the path to personalized medicine has not been straightforward (Hamburg & Collins, 2010). Technology advancement enables the collection of massive data for each individual such as genomic profile or biomedical imaging, but how to tease out useful information and guide clinical practice remains a major problem. As challenging as it may sound, we are making tremendous progress with the help of deep learning algorithms.

My dissertation will help to address the problems arising in personalized medicine from three different aspects. First, I will develop a novel framework to estimate individualized treatment effects (ITE), which quantifies the variations in response to the same treatment among patients with heterogeneous profiles. The ITE estimation will have the potential to replace the one-size-fits-all average treatment effects (ATE) commonly used in clinical practice and provide patient-specific treatment guidance. Second, I will develop a statistical test to determine pairwise causation between two sets of continuous variables. Despite of the massive data available, the primary methods to determine causation clinically are still through randomized controlled experiments and animal studies, which are highly inefficient or sometimes even infeasible. With this new statistical test, we will be able to draw causal conclusions from observational data instead of experimental data alone, which will be beneficial in terms of understanding underlying disease mechanisms. Last but not least, I will apply the previously developed methods on real-life datasets to demonstrate their usage, which hopefully will help to empower the clinicians and the patients to make informed treatment plans.

### 3. Specific Aims

Personalized approaches have shown great potentials to transform modern medicine. As challenging as it may sound, we are making tremendous progress with the help of data sciences and machine learning. Two fundamental tasks in data sciences are prediction and inference. In this dissertation, I proposed to address these two tasks using deep learning approaches in the setting of personalized medicine. Aim I focused on the prediction part, which was aimed to develop a method to estimate individualized treatment effects in contrast to population-based effect estimates. Aim II emphasized on the inference part, involving determining the pairwise causations from observational data when randomized trials are not feasible. Aim III demonstrated how to incorporate previously developed methods into biomedical analysis by leveraging various types of data including clinical factors, gene expression and images that characterize patients' individualized profiles. In addition, I also addressed the 'black-box' criticism around deep learning application by creating an interpretable diagnosis system based on images for Alzheimer's disease.

**Aim 1: To develop a generative adversarial network (GAN)-based approach to estimate individualized treatment effects.** In contrary to the one-size-fits-all population-level effect sizes, GAN was used to infer the unseen, individual counterfactuals based on factual outcomes, which later was used to obtain the personalized treatment effects. This approach was compared with other state-of-art methods through simulations.

**Aim 2: To develop a novel statistical test combined with GANs to determine pairwise causation between two sets of continuous variables.** Given the strong resemblance among the assumptions of GAN, the additive noise models (ANMs) and the

structural equation models (SEMs) for causal inferences, a novel framework was established by replacing the conventional non-linear functions in ANMs with neural networks. Three different scenarios, when there is neither association nor causation, when there is association only, and when both association and causation exist were explored.

**Aim 3: To demonstrate how to incorporate previously developed methods into biomedical analysis by leveraging various types of data including clinical factors, gene expression and images.** Patients with lung cancer from The Cancer Genome Atlas Program (TCGA) with high-dimensional complex features were included in the first analysis. I quantified the variations of days to death in response to the cancer treatment among patients with heterogeneous clinical and genetic profiles. My second analysis provided an end-to-end framework for Alzheimer’s disease from automatic diagnosis of medical images by artificial intelligence to results interpretation including feature selection and causal discovery.

## METHODS

### 1. Individualized Treatment Effect (ITE) Estimation

#### 1.1 Problem Formulation

The problem of individualized treatment effect (ITE) is best defined under the counterfactual framework introduced by Hernan and Rubin (Hernán MA, 2019). Let capital letters denote random variables and small letters denote the realization of the random variables.  $X$  stands for feature vectors.  $T$  stands for an action/treatment/intervention. In the case of binary treatment,  $T$  takes the value of 0 or 1. Let  $Y$  be the outcome. Mathematically, the problem of ITE can be defined as

$$\tau(x) = E(Y_{T=1}|x) - E(Y_{T=0}|x) = E(Y_{T=1} - Y_{T=0}|x) \quad (1)$$

This is the expected difference between the outcome when  $T$  takes the value of 1 and the outcome when  $T$  is 0 for an individual whose feature vector equals to  $x$  (Shalit et al., 2017). The key here is that only one outcome, either  $Y_{T=1}$  or  $Y_{T=0}$ , can be observed in reality, depending which treatment the patient actually receives. Therefore, the outcome that is observed is called the factual outcome, whereas the outcome that is not observed is defined as the counterfactual outcome.

Consistency, exchangeability and positivity are the necessary assumptions for valid ITE estimation using observational data to mimic a randomized controlled trial (Hernán MA, 2019). Consistency refers to a consistent definition of treatment; the values of treatment under comparison correspond to well-defined interventions that, in turn, correspond to the versions of treatment in the data (Hernán MA, 2019). Exchangeability is perhaps the most important assumption for observation causal inference. Exchangeability indicates conditional independence  $Y_{T=t} \perp\!\!\!\perp T|X$ , which is equivalent to no residual confounding in epidemiology. This is a very strong assumption because you can never guarantee the data capture every feature that is related to both the treatment and the outcome. Positivity requires that the conditional probability of receiving every value of the treatment is greater than 0 ( $0 < P(T = t|x) < 1$ ). Exchangeability and positivity are also called strong ignorability in some literature (A. M. Alaa et al., 2017; Shalit et al., 2017).

## ***1.2 Model Description***

Yoon et al. proposed to use the generative adversarial networks for data imputation, which they termed ‘GAIN’ (Yoon, Jordon, & van der Schaar, 2018a). In comparison to the

standard conditional GAN, GAIN’s generator was trained to fill in the missing values, whereas the discriminator was aimed to distinguish the observed data and the imputed data. In addition to the discriminator and the generator, the authors added a ‘hint’ mechanism where they provided the discriminator with some extra information depending on the missing patterns. They showed that their algorithms outperformed the state-of-art methods on many datasets.

ITE estimation is essentially a data imputation problem. Counterfactuals will need to be estimated/imputed before taking the difference of outcomes to get ITE. Therefore, I proposed to address the problem of ITE estimation by building on and adapting the structure of GAIN, to create a framework called Counter-Factual Generative Adversarial Imputation Networks (CFGAIN). Specifically, CFGAIN made two main changes to the structure of GAIN. First, instead of merging  $X$  and  $Y$  together as the target data vector that needed to be filled with imputed values, I borrowed the structure of conditional GANs and treat  $X$  as conditions to both the discriminator and generator. Second, in the case of high-dimensional covariates with complex structure that affect ITE distribution (e.g. gene expression profiles), an encoding block was added to perform automatic data reduction, which took  $X$  as the input and extracted relevant features  $V$  from the hidden space. The features were used as conditions to both the generator and discriminator instead of directly using  $X$ . Figure 5 illustrates the design of the proposed network CFGAIN in the case of a binary treatment.

We start with the input dataset to CFGAIN as  $D = \{\mathbf{t}_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ . Here  $\mathbf{t}_i$  denotes the treatment vector for subject  $i$ , which is in a  $m$ -dimensional space, corresponding to  $m$  types of treatment. The  $n^{\text{th}}$  element of  $\mathbf{t}_i$  (i.e.  $\mathbf{t}_i^n$ ) takes the value 1 if subject  $i$  receives the  $n^{\text{th}}$

treatment. Otherwise,  $\mathbf{t}_i^n = 0$ .  $\mathbf{y}_i$  is the observed outcome vector with the length of  $m$ . The  $n^{\text{th}}$  element of  $\mathbf{y}_i$  (i.e.  $\mathbf{y}_i^n$ ) takes the value of the observed factual outcome if subject  $i$  receives the  $n^{\text{th}}$  treatment. Otherwise,  $\mathbf{y}_i^n$  is replaced with a random noise sampled from the uniform distribution  $[0,1)$ .

$\mathbf{x}_i$  is the set of realization values of the covariates for subject  $i$ . If  $\{\mathbf{x}_i\}_{i=1}^n$  is high-dimensional with complex structure such as medical images, it will first go through a feature extractor, consisting of a few fully-connected layers or convolutional layers to tease out useful features  $\{\mathbf{v}_i\}_{i=1}^n$  for the generator  $G$  and discriminator  $D$ .

The generator  $G$  takes  $\{\mathbf{t}_i, \mathbf{v}_i, \mathbf{y}_i\}_{i=1}^n$  as input and outputs  $\{\hat{\mathbf{y}}_i\}_{i=1}^n$ , which are the generated outcome vectors, with both the factual and counterfactual outcomes replaced by generated values. I will use  $\{\tilde{\mathbf{y}}_i\}_{i=1}^n$  to denote the imputed outcome vectors, where only the counterfactual outcomes are replaced with the values generated by  $G$ .

The hint mechanism  $H$  is self-defined. Depending on the value of  $\mathbf{t}_i$ ,  $\mathbf{h}_i$  can be sampled from the distribution  $H|T=\mathbf{t}_i$ . This hint mechanism provides the discriminator  $D$  with some information about the treatment assignment, which has been shown to be necessary for  $G$  to reproduce a unique optimal distribution with respect to  $D$ .

The discriminator  $D$  takes  $\{\mathbf{h}_i, \mathbf{v}_i, \tilde{\mathbf{y}}_i\}_{i=1}^n$  as input and outputs vectors of size  $m$ , with each element indicating the probability of the corresponding element in  $\tilde{\mathbf{y}}_i$  being a factual outcome.

### 1.3 Optimization

Similar to standard GAN, the optimization is conducted in an adversarial fashion. The discriminator D is trained to maximize the probability to correctly identify factual outcomes, whereas the discriminator G is trained to minimize such probability. The loss functions for the D and G can be written as

$$L_D = -E[T^t \log(D(H, V, \tilde{Y})) + (1 - T)^t \log(1 - D(H, V, \tilde{Y}))] \quad (2)$$

$$L_G = E\left[T^t \log(D(H, V, \tilde{Y})) + (1 - T)^t \log(1 - D(H, V, \tilde{Y}))\right] + \alpha L_{MSE} \quad (3)$$

The extra  $L_{MSE}$  term respects the fact that the corresponding elements in the generated outcome vectors  $\{\hat{\mathbf{y}}_i\}_{i=1}^n$  should be as close as the factual outcomes. Let  $J_i$  be the number of factual outcomes for subject  $i$ :

$$L_{MSE} = \sum_{i=1}^n \sum_{j=1}^{J_i} (y_j^f - \hat{y}_j^f)^2 \quad (4)$$

The vanilla version of GAN is notorious for its difficulty in training. GAN is defined as a minmax game which requires to update the parameters of D and G in the same training process. G may stop learning if D is too well trained. It is difficult to find ideal sets of hyper-parameters to balance between D and G. In addition, vanilla GAN also suffers from non-convergence, mode collapse, diminished gradient and lack of quantitative measures indicating the distance between the generated distribution and the true distribution (Hui, 2018). Therefore, I used WGAN to replace GAN whenever it was too difficult to find good hyper-parameters.

WGAN proposes a new cost function, which relies on the Wasserstein distance or earth-mover distance (EM distance) (Arjovsky et al., 2017). EM distance is the smallest

distance needed to convert one distribution into another. The loss functions for WGAN can be written as

$$L_D = -E[T^t D(H, V, \tilde{Y}) + (1 - T)^t (1 - D(H, V, \tilde{Y}))] \quad (5)$$

$$L_G = E \left[ T^t D(H, V, \tilde{Y}) + (1 - T)^t (1 - D(H, V, \tilde{Y})) \right] + \alpha L_{MSE} \quad (6)$$

Here the loss of D is an estimation of the Wasserstein distance, which is a quantitative measure related to the convergence of WGAN. In addition, WGAN applies a clipping to the weights of D to ensure Lipschitz constraint (Arjovsky et al., 2017).

$$w \leftarrow clip(w, -0.01, 0.01) \quad (7)$$

As proposed by the original study, RMSProp is used as the optimizer for WGAN training.

#### 1.4 Simulation Experiments

Statistical simulation was conducted to evaluate the performance of the proposed algorithm, in comparison with linear regression, propensity score method, k nearest neighbor regression and random forest, when the treatment is binary. Synthetic data was generated for two scenarios. Scenario I corresponds to the simple case when the covariate that affects ITE has only a single dimension:

- 1) Draw the covariate vector  $\mathbf{X}$  from the standard normal distribution for 10,000 individuals;
- 2) Define a non-linearity function  $f(x) = \frac{1}{2 + \exp(-20(x - \frac{1}{3}))}$ ;
- 3) Draw  $\mathbf{y}_0 = 0.3 + \mathbf{n}_0$ , where  $\mathbf{n}_0$  is the randomly sample noise vector from normal distribution  $N(0, SD = 0.1)$ ;
- 4) Draw  $\mathbf{y}_1 = 0.3 + f(\mathbf{X}) + \mathbf{n}_1$ , where  $\mathbf{n}_1$  is the randomly sample noise vector from normal distribution  $N(0, SD = 0.1)$ ;



- 5) Treatment  $t_i$  is drawn from a bernoulli distribution with  $p = 0.5$  for each subject.

Scenario II corresponds to the case when the ITE distribution is affected by some high-dimensional feature with complex structure:

- 1) Generate  $X_i$  as an image of 28x28 pixels, containing a circle with random radius  $R_i$  and random center  $Q_i$  (R. Chen & Liu, 2018).
  - a) First determine the radius  $R_i$  by sample integers randomly from  $[0,14]$ .
  - b) To determine the origin  $Q_i$ , draw the coordinates by randomly sample from  $[0,28]$  for both coordinates.
  - c) For pixel values inside the circle, assign a value of 180. Otherwise, let the pixel values be 0.
- 2) Draw  $\mathbf{y}_0^i$  from the standard normal distributions for 10,000 subjects.
- 3) Draw  $\mathbf{y}_1^i$  from  $N(R_i, SD = 1)$  for all 10,000 subjects.
- 4) Treatment  $t_i$  is drawn from a Bernoulli distribution with  $p = 0.5$  for each subject.

The evaluation metric for ITE was the mean squared error (MSE). Average treatment effect (ATE), average treatment effect on the treated (ATT) and average treatment effect on the control (ATC) will also be reported as a sanity check. Below are the mathematical definitions for the evaluation metrics:

$$MSE = \frac{1}{N} \sum_{i=1}^n (y_i^{cf} - \hat{y}_i^{cf})^2 \quad (8)$$

## 2. Adversarial Causal Test (ACT)

### 2.1 Problem Formulation

As discussed before, causal discovery is a crucial problem that need to be addressed before any valid treatment effect estimation. Causal discovery focuses on determining the causal structure among several variables. The most fundamental and simple problem is to determine if two variables  $X$  and  $Y$  have causal relation and if so, what is the direction for that causal relation. Previous methods tend to emphasize the latter part of the problem while assuming the first part is true (Hoyer et al., 2009; Peters et al., 2010). Therefore, I propose an end-to-end algorithm to test for causations and determine the direction for the causation, which is an extension to the work by Lopez-Paz and Qquab on causal discovery using conditional GANs (Lopez-Paz & Oquab, 2016). Due to the identifiability issues associated with discrete variables (Spirtes & Zhang, 2016), I will focus on the causation between two continuous sets of variables.

### 2.2 Model Description

#### 2.2.1 Classifier Two-Sample Test (C2ST)

The classifier two-sample test was introduced by Lopez-Paz and Qquab.(Lopez-Paz & Oquab, 2016) Assuming there are two samples,  $S_P \sim P^n$  and  $S_Q \sim Q^m$ . The null hypothesis of C2ST is  $P = Q$ , whereas the alternative hypothesis is  $P \neq Q$ . C2ST aims to determine if we could reject the null hypothesis through observed samples:

$$S_P := \{x_1, \dots, x_n\} \sim P^n(X) \text{ and } S_Q := \{y_1, \dots, y_n\} \sim Q^n(Y)$$

The test will be conducted in the following steps:

- 1) Construct the dataset  $D = \{(x_i, 0)\}_{i=1}^n \cup \{(y_i, 0)\}_{i=1}^n := \{(z_i, 0)\}_{i=1}^{2n}$ ;
- 2) Shuffle the dataset  $D$  first and then split the data into a training set  $D_{train}$  and a test set  $D_{test}$ ;
- 3) Train a binary classifier on  $D_{train}$  ;
- 4) Test the classier on the test set and get the classification accuracy  $\hat{t}$ , which is the test statistic following  $N(\frac{1}{2}, \frac{1}{4n_{test}})$  under the null using the central limit theorem:

$$\hat{t} = \frac{1}{n_{test}} \sum_{(z_i, l_i) \in D_{test}} I[f(z_i) > \frac{1}{2}] = l_i \quad (9)$$

### 2.2.2 Adversarial Causal Test (ACT)

Figure 6 illustrates the procedure for conducting the ACTs. Specifically, the following steps should be taken:

- 1) Train a CGAN from X to Y  $D_{x \rightarrow y} = \{(x_i, g_y(x_i, z_i))\}_{i=1}^n$
- 2) Apply the C2ST test on  $\{(x_i, g_y(x_i, z_i))\}_{i=1}^n$  and  $\{(x_i, y_i)\}_{i=1}^n$  to get the classification

$$\text{accuracy } \widehat{t_{x \rightarrow y}} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} w_i$$

- 3) Train a CGAN from X to Y  $D_{y \rightarrow x} = \{(g_x(y_i, z_i), y_i)\}_{i=1}^n$
- 4) Apply the C2ST test on  $\{(g_x(y_i, z_i), y_i)\}_{i=1}^n$  and  $\{(x_i, y_i)\}_{i=1}^n$  to get the classification

$$\text{accuracy } \widehat{t_{y \rightarrow x}} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} v_i$$

- 5) Calculate the ACT test statistic  $T = \frac{(\widehat{t_{x \rightarrow y}} - \widehat{t_{y \rightarrow x}})^2}{\sigma^2}$ , where  $\sigma^2 =$

$$2 \left( \frac{p(1-p)}{n_{test}} - \frac{\sum_{i=1}^{n_{test}} (w_i - \widehat{t_{x \rightarrow y}})(v_i - \widehat{t_{y \rightarrow x}})}{n_{test} - 1} \right), \quad p = \frac{1}{2} \text{ and } T \text{ follows a Chi-square distribution with}$$

1 degree of freedom under the null hypothesis of no causation.

- 6) If the null hypothesis is rejected, the causal direction could be determined by comparing  $\widehat{t_{x \rightarrow y}}$  and  $\widehat{t_{y \rightarrow x}}$  with 0.5. The one closer to 0.5 corresponds to the correct direction.

### 2.3 Optimization

Similar to section 1.4, the training process of CGAN involves  $G$  and  $D$  playing the following two-player minmax game with mini batch size of  $M$ :

$$Loss_D = -V(D, G) + L_{GP} \quad (10)$$

$$Loss_G = V(D, G) + L_{fm} \quad (11)$$

Besides the standard gradient penalty for the discriminator, an extra feature mapping term will be added to the loss function of the generator to avoid the over training of the discriminator, which takes in the output from the last hidden layer and calculates the mean square error between the real data  $O_i(\mathbf{x}_i|\mathbf{y}_i)$  and the fake data  $O_i(G(\mathbf{z}_i|\mathbf{y}_i))$  generated by the generator for a mini batch of size  $M$ :

$$V(D, G) = E[D(X|Y)] - E[D(G(Z|Y))] \quad (12)$$

$$L_{GP} = \lambda E[\left(\|\nabla_{\bar{\mathbf{x}}} D(\bar{\mathbf{X}})\|_2 - 1\right)^2] \quad (13)$$

$$L_{fm} = \frac{1}{M} \sum_{i=1}^M (O_i(\mathbf{x}_i|\mathbf{y}_i) - O_i(G(\mathbf{z}_i|\mathbf{y}_i)))^2 \quad (14)$$

The algorithm updates the discriminator and the generator iteratively using the stochastic optimizer RMSprop until convergence (Tieleman & Hinton, 2012).

### 2.4 Simulation Experiments

Statistical stimulation was conducted using synthetic data to evaluate the type I error and power of the proposed test. Three different scenarios were explored. Type I error is defined the probability of significant results at a pre-specified significance level when the

null hypothesis of no causation should not be rejected. Power is defined as the probability of successfully rejecting the null hypothesis of no causation when there is causation. I also reported the power of correctly identifying the causation direction.

Scenario I corresponds to the case where there is no association and no causation.

Data will be generated in the following steps:

- 1) Sample X from the standard normal distribution, which is a vector with length 10,000;
- 2) Sample Y from the standard normal distribution, which is a vector with length 10,000;
- 3) Test the independence between X and Y using the Hilbert Schmidt independence criterion (dHSIC);(Gretton et al., 2008; Pfister, Bühlmann, Schölkopf, & Peters, 2018)
- 4) If the p-value from dHSIC test is significant under the pre-specified significance level 0.05, discard X and Y; otherwise, randomly sample realizations from X and Y as a dataset, at sample size of 500, 1000, 2000;
- 5) Repeat the above procedure until accumulate 1000 datasets, each containing X and Y;
- 6) Repeat the above procedure for a significance level at 0.01.

Scenario II corresponds to the case where X and Y are associated but no causation.

Data was generated in the following steps:

- 1) Sample X from the standard normal distribution, which is a vector with length 10,000;

- 2) Sample Y from the standard normal distribution, which is a vector with length 10,000;
- 3) Test the independence between X and Y using the Hilbert Schmidt independence criterion (dHSIC) (Gretton et al., 2008; Pfister et al., 2018);
- 4) If the p-value from dHSIC test is significant under the pre-specified significance level 0.05, randomly sample realizations from X and Y as a dataset, at sample size of 500, 1000, 2000; otherwise, discard X and Y.
- 5) Repeat the above procedure until accumulate 1000 datasets, each containing X and Y;
- 6) Repeat the above procedure for a significance level at 0.01.

Scenario III corresponds to the case where X and Y have causal relations. Data was generated in the following steps:

- 1) Randomly sample X from the standard normal distribution with a sample size of 100,000;
- 2) Generate the random noise N of the same dimension as X from the normal distribution with mean=0 and SD=0.1;
- 3) Generate Y by non-linear mappings  $f$  from X and N:  $y_i = f(x_i) + n_i$
- 4) Randomly sample X and Y to create 100 datasets at sample size of 200, 500, 1000, 2000, 5000, so there are a total of 500 datasets for each nonlinear function  $f$ .

### 3. Application Studies

#### 3.1 ITE Estimation for TCGA

##### 3.1.1 Simulation using Gene Expression for Pancreatic Cancer

Contrast to previous two simulation settings where all data was purely generated, I borrowed the gene expression profiles from TCGA pancreatic cancer study as covariates, and only synthesize data reflecting treatment effects:

- 1) Randomly select 1000 genes from whole gene expression profile and calculate the mean expression levels for the 1000 genes  $\mu_0^i$ . Draw  $y_0^i$  from  $\exp(\lambda = 1) + \mu_0^i/10$ .
- 2) Randomly select another 1000 genes from the whole expression profile. Input the selected gene values into an autoencoder, which learns the hidden representations of the genes  $L$ .
- 3) Input  $L$  to a neural network with 1 hidden layer of 256 neurons and the sigmoid activation to generate ITEs. The neural network is initialized with random weights and bias but does not go through any training process.
- 4) Draw  $y_1^i$  from  $y_1^i = y_0^i + ITE_i + normal(0,0.05)$ .

##### 3.1.2 ITE Estimation for Lung Cancer

Patient heterogeneity has caused many uncertainties in medical research and clinical practice. In this project, we applied CFGAIN to quantify variation in days to death, in response to the same treatments among patients with heterogeneous genetic profiles. Based the estimation, best treatment strategy could be recommended to future patients, taking

individual genetic profile into consideration. The ITE estimation will have the potential to replace the one-size-fits-all average treatment effects (ATE) commonly used in clinical practice and provide patient-specific treatment guidance.

Lung cancer has been the leading cause of cancer-related deaths in the US (Barta, Powell, & Wisnivesky, 2019). Although both the incidence rates have declined in the US, 5-year survival rate for lung cancer is around 20% and among the lowest for all cancer ("Cancer Facts & Figures 2019 - American Cancer Society," 2019). Lung cancer can be grouped into two major types depending on the aetiology: small-cell lung cancer (SCLC), and non-small-cell lung cancer (NSCLC), accounting for 80% of lung cancers (R., 2013). Only 16% of lung cancer cases diagnosed are at localized stages ("Cancer Facts & Figures 2019 - American Cancer Society," 2019). Advanced-stage lung cancer is usually treated with chemotherapy, radiation, a combination of both or immunotherapy.

Differential response to treatment for lung cancer has been reported extensively in literature, which can be attributed to heterogeneous genetic profiles. For example, it has been shown that mutations in HER1/EGFR are associated with positive response of targeted therapy and poor prognosis and resistance to chemotherapy (Cooke, Reeves, Lannigan, & Stanton, 2001). A few studies identified CCND1 overexpression as a marker associated with poor prognosis (Esposito et al., 2005; Ikehara et al., 2003; Jin et al., 2001). In addition, a few groups reported that some mRNAs and microRNAs (miRNAs) were significantly associated with prognosis in NSCLC patients (Hu et al., 2010; Raponi et al., 2009). The fact that genetic markers are important for the differential prognosis makes lung cancer a good candidate for us to apply CFGAIN.



TCGA (<https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>) is a large-scale cancer genomics program, supported by the National Cancer Institute (NCI) and the National Human Genome Research Institute. It was initiated in 2006 and included extensive genomic, epigenetic, transcriptomic and proteomic data for over 33 cancer types. The TCGA lung cancer data used in this study included samples from three projects: TCGA-LUSC, TCGA-LUAD and TCGA-MESO. After merging clinical information, treatment strategies and days to death, a sample of 186 subjects with complete data was used for this analysis. Their expression profiles of 60,483 genes were extracted to be used as covariates to estimate ITE. Among the study subjects, 80 subjects had chemotherapy only and 39 subjects had radiation therapy only, whereas 67 subjects had combinational therapy of both chemotherapy and radiation. Demographic and clinical factors that were also used for ITE estimation included age at diagnosis, gender, race, ethnicity, diagnosis, tumor stage, tumor morphology, tumor origin and which project the subject belonged to.

Over 60,000 gene expression values were encoded into a vector of 128 latent factors. To increase the efficiency of information processing by CFGAIN, gene expression values were arranged into a 2X2 matrix in size of 236x238 with padding 0s in the end. Convolutional filters were used in the auto-encoder. To ensure fair comparisons, the 128 latent factors were used to summarize information from gene expression profiles for all methods. Results from linear regression, propensity score regression, K nearest neighbors and random forest were compared with results from CFGAIN. After getting the complete outcome vectors, ITE was calculated for each subject. The treatment therapy corresponding

to the longest survival days was chosen as the recommended treatment. Since ITE was defined as the difference between two treatment arms, so three pairwise comparison was conducted to calculate ATE, ATT and ATC.

One fundamental difficulty for applying machine-learning based ITE estimation to real data is the lack of counterfactual outcomes to validate the results. Researchers have been using simulated data, where they manually define the data generating process so that they know both the factual outcomes and the counterfactuals, to demonstrate the superiority of their methods (Yoon et al., 2018b; W. Zhang, Le, Liu, Zhou, & Li, 2017). However, simulated data are generally not so reflective of real-life datasets. Since match-learning based estimation are data-driven approaches, it has been established that each dataset has its own suitable method. Therefore, it is important to apply validation measures to compare the performance of various methods.

Plug-in validation was applied to the TCGA lung cancer dataset, which obtained a plug-in precision of estimating heterogeneous effects (PEHE) estimate (Rolling & Yang, 2014; Schuler, Jung, Tibshirani, Hastie, & Shah, 2017). PEHE quantified a model's ability to capture the heterogeneous causal effect of a treatment among individuals (Hill, 2011). An empirical measure of PEHE is defined as  $\frac{1}{n} \sum_{i=1}^n (\hat{T}(X_i) - (Y_i^{(1)} - Y_i^{(0)}))^2$ , where  $T(x) = E_\theta[Y^{(1)} - Y^{(0)} | X = x]$  and  $\theta$  is the parameter space, with nuisance parameters  $\{\mu_0, \mu_1, \dots\}$ . For each individual, the counterfactuals, either  $Y_i^{(1)}$  or  $Y_i^{(0)}$  is unknown, it is impossible to calculate the empirical PEHE. The plug-in model  $\tilde{\theta} = \{\tilde{\mu}_0, \tilde{\mu}_1, \dots\}$  can be estimated from the observed data by fitting supervised regression models to the subsets of data  $\{(X_i, Y_i) | T_i = 0\}$

and  $\{(X_i, Y_i) | T_i = 1\}$ . Let us define  $\tilde{T}(x) = \widetilde{\mu}_1(x) - \widetilde{\mu}_0(x)$ . The plug-in validation PEHE estimator can be estimated as  $\frac{1}{n} \sum_i \left( \hat{T}(X_i) - \tilde{T}(X_i) \right)^2$ . The supervised regression models I used here was gradient boosting regression (i.e. XGBoost Regression) (T. Chen, He, Benesty, Khotilovich, & Tang, 2015).

To further understand which latent factors and genes played important roles for ITE, lasso regression with 5-fold cross validation was performed by regressing ITE estimates against latent factors. After important features were selected, multiple linear regression was performed on ITE vs. selected features and R<sup>2</sup> was reported. To map latent factors to genes, Spearman correlation was calculated for each of the genes and each of the latent factors selected from Lasso regression.

### ***3.2 Causal Discovery for ADNI***

Alzheimer's disease (AD) that causes progressive brain atrophy and memory loss is a progressive, irreversible degenerative disease of the brain and is the most common neurodegenerative disease in the world (Xiaonan Liu et al., 2018; Xin Liu et al., 2018; Struyfs et al., 2015; Zhuang, Zheng, Gu, Shen, & Ji, 2017). An estimation of 5.4 million Americans and more than 30 million people in the world are affected. It is estimated that these numbers will be tripled by 2050. AD is the sixth leading cause of death in the USA (Association, 2016; Leandrou, Petroudi, Kyriacou, Reyes-Aldasoro, & Pattichis, 2018)

Diagnosis and prediction of AD via clinical and psychometric assessment is challenging (Leandrou et al., 2018). It is difficult for AD patients to be early and accurately identified through clinical dementia rating and cognitive tests. A final diagnosis of AD can

only be confirmed by histological examination at postmortem biopsy. Although histological examination of the brain for the living patients is infeasible, individually varying brain structure, function and the pathological effects can be measured by images. Therefore, imaging plays an important role in improving diagnosis and prediction of AD. According to the recommendation by the National Institute of Neurologic, Communicative Disorders and Stroke–AD and Related Disorders Association (NINCDS-ADRDA) Work Group, the clinical classification of AD should explore the image markers: magnetic resonance imaging (MRI), diffusion tensor imaging (DTI), positron emission tomography (PET), amyloid-PET, tau-PET and abnormal neuronal cerebrospinal fluid (CSF) markers (tau and/or A $\beta$ ) (Dubois et al., 2007; Leandrou et al., 2018).

As the size of the imaging datasets increases, manual analysis of imaging data is tedious and time consuming. Computer-aided diagnosis (CAD) of AD that combines computational models and analytical tools for high dimensional imaging data analysis is emerging as one of major tools for diagnosis and prediction of AD (Dimitriadis, Liparas, & Initiative, 2018; Leandrou et al., 2018). The widely used machine learning methods in CAD include discriminant analysis (DA), logistic regression (LR), random forest, neural networks, and support vector machine (SVM) (Dimitriadis et al., 2018; Leandrou et al., 2018; Lorenzi et al., 2017; Sarica, Cerasa, & Quattrone, 2017). Deep learning, a rapidly resurged subfield of machine learning outperforms many classical ML approaches and is emerging as a major analytic platform in machine learning (Esteva et al., 2019). Deep learning with massive amount of computational power has achieved a great success in driverless cars, speech recognition and imaging analysis (Waldrop, 2019), and demonstrated great potential for

diagnosis and prediction power in tuberculosis (Heo et al., 2019), cancer (Esteva et al., 2017; Ghatwary, Zolgharni, & Ye, 2019; Haenssle et al., 2018; Ladefoged et al., 2019), diabetic retinopathy (Gulshan et al., 2016), chronic kidney disease (Ravizza et al., 2019), AD (Ding et al., 2019; Hosseini-Asl, Keynton, & El-Baz, 2016; Ju, Hu, & Li, 2017; Payan & Montana, 2015; Sarraf & Tofghi, 2016; Spasov et al., 2019; Wada et al., 2019), and conversion from mild cognitive impairment (MCI) to AD (Choi, Jin, & Initiative, 2018; Spasov et al., 2019). There is a growing interest in application of deep learning to healthcare and medicine.

Despite its great success in computer vision, natural language processing, control, decision-making, diagnosis and early detection of complex diseases, deep learning is also well known as a ‘black box’ due to its low interpretability to humans and still has a serious opacity problem (Waldrop, 2019). Overcoming the limitation of the lack of transparency and interpretation remains a great challenge for deep learning (Dubois et al., 2007). In this project, I developed a novel framework that integrates deep learning and causal inference for image classification. The new framework consists of two stages: (1) convolutional neural networks to classify AD status based on DTI and occlusion map to find image regions that are most distinctive for disease status and (2) the state-of-the-art causal inference tools to determine if the selected image regions are causal for AD.

Brain anatomy, structural connectivity and physical connection between brain regions that are characterized through water molecular diffusing within white matter tracts can be measured by DTI. The imaging signals provide intermediate endophenotypes. Genetic variants will influence brain microstructure, function and disease development.

Understanding the role that genetics has in imaging and disease variation, is a key to getting

inside into the causal chain of complex diseases (Bycroft et al., 2018; Elliott et al., 2018; Jahanshad et al., 2013). Therefore, to further cover the genetic bases of brain structures and function, and mechanism of AD, joint analysis of the genetic, brain images and AD was carried out. I assessed both association and causal relationships among genetic variants, brain regions and AD.

### **3.2.1 Materials**

The DTI images used in this study are downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI), sizes of each image was  $91 \times 109 \times 91$ . ADNI is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biomedical biomarkers for the early detection and tracking of Alzheimer's Disease (About ADNI. Retrieved January 12th, 2019, from: <http://adni.loni.usc.edu/about/#core-container>). DTI images were recorded for every participant from different time-points in which they joined in the research. In this study DTI images of 151 individuals from normal controls (NC) (100 images) and AD (51 images) groups were chosen from 4 different diagnostic time-points: baseline, 6 months, 12 months, and 24 months.

### **3.2.2 Image Preprocessing**

To make sure that all the images for this analysis are comparable, we registered all the DTI image data for every subject every time point to the common template which could be downloaded from the McConnell Brain Imaging Center (<http://www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009>). We utilized a

strategy of combination of linear and nonlinear registration algorithm to map each individual DTI data to the common template. During the linear image registration procedure, we first mapped the image data to the common template to make sure all the image are within the standard brain region by using FLIRT (FMRIB's Linear Image Registration Tool) from FSL (FMRIB software library) image analysis suite (<http://www.fmrib.ox.ac.uk/fsl/>) . Then we further applied non-linear registration algorithm which was implemented in RNiftyReg to map the image details within the standard brain. The linear image registration process helps us restrain each individual DTI image to a standard template and the nonlinear image registration helps us to make sure the registered image maintain the structures details as the original data.

To overcome the small sample size limitation of medical images, image augmentation techniques were used (Aderghal, Boissenin, Benois-Pineau, Catheline, & Afdel, 2017). The first technique we applied was Gaussian filters to blur the image to mimic the possible variations in the original images. Filter size of  $3 \times 3$ ,  $5 \times 5$  and  $7 \times 7$  were used with spread parameters of 0.7, 0.7 and 0.6 respectively. The second augmentation technique we used was translation, where we shifted the images by  $\pm 1$  pixel in each dimension. This imitates the possible variations in registration process where the images were aligned with the template. Finally yet importantly, the images were flipped horizontally because some regions of the brain (eg. the hippocampus) is symmetrical to enlarge our sample size. To balance the data, we randomly duplicated some images from the under-sampled category. Data augmentation and class balancing produced over 20 times more data than the original dataset.

### 3.2.3 Genetic Data Preprocessing

We performed QC in both individual level and SNP level QC in the plink binary format. For the individual level QC, the following steps were applied to the data.

- 1) Individuals with discordant gender information were removed from the data.
- 2) Individuals with missing rate >10% were removed from the data.
- 3) Individuals with heterozygosity rate more than 3 standard deviation from the mean were excluded from the data.
- 4) Individuals with  $IBD > 0.185$  were excluded from the data.

After the individual level QC was conducted, the following steps for SNP level QC were further applied to the data.

- 1) SNPs with missing genotype rate > 10% were excluded from the data being analyzed.
- 2) SNPs with P - value for HWE test <  $1E-6$  were excluded from the data.
- 3) SNPs without polymorphism were removed from the data.

Then pre-imputation QC tool from MaCarthy groups was further applied to check the data against 1000G reference data. The imputation of the genetic data were conduct under the SHAPEIT+ IMPUTE2 framework in the internal computational clusters. The 1000G reference data was used as the reference panel for imputation. After the imputation, the SNP level QC steps were applied again to the data to produce the final genetic data for analysis. Finally, a total of 1,589,061 common SNPs in 36480 genes genotyped in 151 individuals were included in analysis.



### 3.2.4 Architecture of Convolutional Neural Networks

We first developed a three-dimensional convolutional neural network (3D-CNN) by modifying the classic AlexNet structure to classify AD and NC in this project (Krizhevsky et al., 2012). The structure has 5 convolutional layers and 3 fully connected layers. Batch normalization and dropout with a rate of 0.5 were used after the first two convolutional layers to speed up training and deal with overfitting. The first two fully connected layers each had 128 neurons, which was different from the 2048 neurons in the original AlexNet. The input to this network was 3D DTI images (109x91x91), whereas the output is a vector of predicted probabilities for AD and NC. Five-fold cross-validation was used to develop the models.

3D whole brain images with 109×91×91 size were input into this CNN. Convolution of an image with different filters can perform operations that capture various types of features and directional information of DTI images, and preserve tract of DTI and the relationship between pixels. 3D convolutional neural networks (3D-CNN) with 5 convolutional layers and 3 fully connected layers were used for AD prediction. 3 dimensional (3D) filter was applied to the dataset and the filter moves in 3-direction (x, y, z) to calculate the low level feature representations. Specifically, 3D filters were arranged as in Table 1.

Due to computational limitations, the two dimensional convolutional neural network (2D-CNN) model VGG (Visual Geometry Group) that won the first and the second places in the localization and classification tracks, respectively, in the ImageNet Challenge 2014 was also explored and modified for image classification and prediction (Simonyan & Zisserman, 2014). To prevent over fitting and improve the image region recognition ability of the

networks, Global Average Pooling (GAP) layer was used as a structure regularizer and localizer in the model to identify the complete extent of the object and exactly which regions of an image are being used for classification (B. Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016).

The model was trained in the Texas Advanced Computing Center (TACC) Maverick2 with NVIDIA GTX 1080 Ti GPUs.

### 3.2.5 Deep Feature Selection for DTI Images

Prediction difference analysis for visualizing the response of CNN to a specific input was used to select features for DTI image classification (Zintgraf et al., 2017). Specifically, prediction difference analysis is to estimate the importance of input pixels by calculating the effect of removing information from the imaging on the class prediction precision (Zeiler & Fergus, 2014).

A sliding window (patch) of  $3 \times 3 \times 3$  was applied through each image. The imaging signals contained in the sliding window was taken as a feature. Each one  $3 \times 3 \times 3$  patch was replaced by randomly sampled values from multivariate normal distributions. The resulting new image where the imaging feature (information) was removed was input into a previously trained CNN model to obtain probability  $p_1$  for predicting AD. Let  $p_0$  be the probability of predicting AD using the original images (without removing the feature (information)). The relative importance of the feature was evaluated by (Zintgraf et al., 2017)

$$d = \log \left( \frac{\frac{p_0}{1-p_0}}{\frac{p_1}{1-p_1}} \right) \quad (15)$$

The sliding window moved across the entire image and a relevance matrix  $W$  of the same size as the whole image was generated, which reflected the relevance importance of all image pixels. A positive value indicated the pixel contributed evidence for the classification of AD, whereas a negative value showed that the pixel contributed against the classification of AD. For details, please see (Zintgraf et al., 2017).

### 3.2.6 Causal Discovery

Three-dimensional functional principal component scores (FPCs) were used to summarize the imaging signal information of the brain region (M. Xiong, 2018). Similarly, one dimensional FPCs can be used to summarize genetic information in the gene. ACTs were used to discover causal relationships between the brain neuroimaging region and AD, and causal relationships between the brain neuroimaging region and gene as well (Goodfellow et al., 2014; Lopez-Paz & Oquab, 2016). Specifically, consider two variables  $X$  and  $Y$ , which can be binary disease status, or continuous FPCs summarizing imaging signals in the brain region or genetic variation in the gene. If  $X$  causes  $Y$ , denoted by  $X \rightarrow Y$ , then we have

$$Y = f_Y(X, N_Y),$$

where  $f_Y$  is a nonlinear function and realized by conditional GAN where a neural network is used to approximate the nonlinear function  $f_Y(X, N_Y)$ , and  $N_Y$  is a noise random variable and is independent of cause  $X$ . Similarly, if  $Y$  causes  $X$  ( $Y \rightarrow X$ ), then we have

$$X = f_X(Y, N_X),$$

where  $f_X$  is a nonlinear function, and  $N_X$  is a noise random variable and is independent of cause  $Y$ . Assume that  $n$  subjects are sampled.

We define dataset  $D_w = \{u_i, v_i, i = 1, \dots, n\}$ . We assign label 0 to dataset  $D_u = \{u_i, i = 1, \dots, n\}$  and 1 to dataset  $D_v = \{v_i, i = 1, \dots, n\}$ . Let  $P$  be the distribution of  $u_i, i = 1, \dots, n$  and  $Q$  be the distribution of  $v_i, i = 1, \dots, n$ . We use the K nearest neighbor (KNN) as a binary classifier to classify two datasets and define the test statistic  $t$  as the classification accuracy to test the null hypothesis of equal distributions of two datasets  $P = Q$ . Let  $z$  be a random variable.

The procedures for causal discovery using ACTs are summarized as follows (Lopez-Paz & Oquab, 2016).

1. Use a CGAN from  $X \rightarrow Y$  to generate the dataset  $D_{X \rightarrow Y} = \{(x_i, \hat{y}_i = f_Y(x_i, z_i)), i = 1, \dots, n\}$ .
2. Use a CGAN from  $Y \rightarrow X$  to generate the dataset  $D_{Y \rightarrow X} = \{(\hat{x}_i = f_X(y_i, z_i), y_i), i = 1, \dots, n\}$ .
3. Divide the total samples into training samples and test samples.
4. Classify two datasets :  $D_u = D_y = \{y_i, i = 1, \dots, n\}$  versus  $D_v = D_{X \rightarrow Y} = \{\hat{y}_i, i = 1, \dots, n\}$  and calculate the two-sample statistic  $\hat{t}_{X \rightarrow Y}$ .
5. Classify two datasets :  $D_u = D_x = \{x_i, i = 1, \dots, n\}$  versus  $D_v = D_{Y \rightarrow X} = \{\hat{x}_i, i = 1, \dots, n\}$  and calculate the two-sample statistic  $\hat{t}_{Y \rightarrow X}$ .
6. Calculate the test statistic  $T = \hat{t}_{X \rightarrow Y} - \hat{t}_{Y \rightarrow X}$ . Under the null hypothesis of no causal relationship or test inconclusive , the statistic  $T$  is asymptotically distributed as

$N(0, \sigma^2)$  , where  $\sigma^2 = \frac{0.5}{n_{test}} - 2cov(\hat{t}_{X \rightarrow Y}, \hat{t}_{Y \rightarrow X})$  ,  $n_{test}$  is the number of subjects in the test set.

#### 4. Human Subjects, Animal Subjects, or Safety Considerations

The datasets I will use include the Alzheimer’s Disease Neuroimaging Initiative (ADNI, [adni.loni.usc.edu](http://adni.loni.usc.edu)) and The Cancer Genome Atlas Program (TCGA, <https://portal.gdc.cancer.gov/>). Both of these datasets are de-identified and available to the public. My request to use the datasets have been approved by my dissertation advisor Dr. Momiao Xiong.

## RESULTS

### 1. Simulation Experiments for ITE Estimation

Statistical simulations were conducted to evaluate the performance of the proposed algorithm, in comparison with linear regression, propensity score method, k nearest neighbor regression and random forest, when the treatment is binary. Synthetic data was generated for three scenarios. Scenario I corresponds to the simple case when the covariate that affects ITE has only a single dimension. Scenario II corresponds to the case when the ITE distribution is affected by some high-dimensional feature with complex structure.

Table 2 listed the quantitative results comparing CFGAIN and other state-of-art methods for scenario I. Based on the results, random forest and the proposed CFGAIN had the best performance in terms of PEHE and MSE, which were much smaller than the estimates from ordinary linear regression (OLS), despite the latter was the standard analytical method used in such case of a randomized control trial with a continuous outcome. In

addition, the true and estimated ITEs were plotted against the covariate  $x$  (Figure 7). It could be clearly visualized that true ITEs vary according to values of  $x$  and CFGAIN was able to capture the variation, whereas the ordinary linear regression could only estimate the ATE without acknowledging any of the causation between  $x$  and ITE.

Table 3 listed the quantitative comparison between the proposed method and other state-of-art methods for scenario II. In this comparison, fully connected layers and convolutional layers were used for encoding purpose, given the fact that the high-dimensional covariate  $X$  were generated. The results clearly showed that the proposed method outperformed other methods, which had the closest estimated PEHE to the ground truth and the smallest MSE. In addition, due to the graphical nature of the high dimensional covariate  $X$ , encoding with convolutional layers performed better in comparison with decoding with fully connected layers. Despite the fact that random forest outperformed CFGAIN in the previous simulation setting where the covariate  $x$  was a simple one-dimensional vector, CFGAIN has been proven to be more accurate when the true ITE is affected by complex and high dimensional covariates.

To better visualize the results for scenario II, estimated ITE was plotted against true ITE to check the consistency of the two (Figure 8). If the estimated ITEs were close to the ground truth, all data points should roughly follow along the 45 degree reference line without any curvature. It could be clearly shown that AEGAIN with convolutional encoding had the best performance whereas other methods all displayed some deviation from the 45 degree line.

## 2. Simulation Experiments for ACT

Statistical stimulation was conducted using synthetic data to evaluate the type I error and power of the proposed test. Three different scenarios were explored. Type I error is defined the probability of significant results at a pre-specified significance level when the null hypothesis of no causation should not be rejected. Power is defined as the probability of successfully rejecting the null hypothesis of no causation when there is causation. Power of correctly identifying the causation direction was also calculated. Scenario I corresponds to the case where there is no association and no causation. Scenario II corresponds to the case where X and Y are associated but no causation. Scenario III corresponds to the case where X and Y have causal relations.

Table 4 listed type I errors calculated from scenario I – no association and causation. Type I errors were fluctuated around nominal levels. When the nominal level was set to 0.05, the larger the sample size, the closer to 0.05 for the type I error. Similarly, Table 5 listed type I errors when there is association but no causation. Type I errors were also fluctuated around nominal levels. When the nominal level was set to 0.01, the larger the sample size, the closer to 0.01 for the type I error.

Power was calculated for varying sample sizes (Figure 9). As expected, the power of ACT increased from 0.67 to 0.90 as sample size increased from 200 to 5000 (Table 6). However, the range of power clearly depended on the type of nonlinear transformation used. Power was greatly reduced when the complexity of non-liner transformation was changed from quadratic function ( $Y = X^2 + N$ ) to cubic function ( $Y = 4X^3 + N$ ) (Table 7). In addition, it was generally more difficult to detect the correct direction of causation even after

establishing a significant causation between two variables, as proved by the lower power for direction identification (Table 8 and 9).

### 3. Application Studies

#### 3.1 ITE Estimation for TCGA

##### 3.1.1 Simulation Results

Previous CFGAIN simulation applied to a more general case of high dimensional covariate, but did not specifically address gene expression data, so I conducted another simulation study by borrowing gene expression profiles from TCGA pancreatic cancer studies. In this simulation, I randomly selected 1000 genes from around 60,000 genes available in the TCGA datasets, and assumed that these 1000 genes affected true ITE distribution through some nonlinear function:

$$ITE_i = f(1000 \text{ selected genes})$$

The outcome here was continuous and the baseline levels were sampled from exponential distributions. The method section listed details about the data generation process.

I first treated 1000 selected genes as input into all methods to compare the accuracy of estimation. Table 10 listed all the quantitative measures. ATE, ATT and ATC were included as sanity checks to confirm if the method at least estimated the averages accurately. While most of methods were able to give a close estimate of the average effects, it was worth noting that K nearest neighbor did poorly in this task, which was very likely due to the fact that Euclidean distance used in KNN could not accommodate the complexity of the non-linear effect of genes. MSE was used as the index to assess the accuracy of ITE estimation.



It was clear that CFGAIN outperformed all other methods by a large amount, indicating it was able to give the most accurate estimation of ITEs. Figure 10 plotted the true and estimated ITE distributions against one of the genes whose value had the largest variance. It can be clearly visualized that CFGAIN's estimated ITEs distribution was much more close to the ground truth, whereas other methods' estimation was more scattered.

Previous analysis did not involve any feature selection/engineering. In most of the cases, we could not narrow down to only a few genes that affect the ITEs, so feature selection is necessary before any treatment effect estimation. To select the most relevant features for the outcome, linear regressions were performed for each of genes, with the factual outcome as the dependent variable and treatment assignment as the other independent variable:  $y = \alpha + \beta_0 * treatment + \beta_1 * gene + \epsilon$ ,  $\epsilon \sim N(0, \sigma^2)$ . P-values for  $\beta_1$  were sorted and top 1000 genes with the lowest p-values were selected as important features to be included into multiple, linear regression, propensity score, KNN and random forest to estimate ITEs. In contrast to this, given the complexity and high dimension of gene expression data, convolutional auto-encoder was used in conjunction with CFGAIN to perform automatic feature engineering and ITE estimation. Table 11 listed the quantitative results. Only CFGAIN and random forest were able to estimate the average effects relatively accurately, whereas linear regression, propensity score and KNN's estimation was biased. In addition, CFGAIN once again had the smallest MSE of 0.6970, ten times smaller than linear regression, which gave the largest MSE due to its non-flexibility. We could visualize the same results in Figure 11. GAIN's estimated ITE distribution was more close to the ground truth, in comparison with other methods, whose values were more scattered. We did not see a

complete overlap between the ground truth and the estimation because of estimation error and the random noise introduced in the data generation process.

### 3.1.2 ITE Estimation for Lung Cancer

CFGAIN was applied to estimate ITE in days to death from different treatment strategies on lung cancer patients. TCGA (<https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>) is a large-scale cancer genomics program, supported by the National Cancer Institute (NCI) and the National Human Genome Research Institute. It was initiated in 2006 and included extensive genomic, epigenetic, transcriptomic and proteomic data for over 33 cancer types. The TCGA lung cancer data used in this study included samples from three projects: TCGA-LUSC, TCGA-LUAD and TCGA-MESO. After merging clinical information, treatment strategies and days to death, a sample of 186 subjects with complete data was used for this analysis. Their expression profiles of 60,483 genes were extracted to be used as covariates to estimate ITE. Among the study subjects, 80 subjects had chemotherapy only and 39 subjects had radiation therapy only, whereas 67 subjects had combinational therapy of both chemotherapy and radiation. Demographic and clinical factors that were also used for ITE estimation included age at diagnosis, gender, race, ethnicity, diagnosis, tumor stage, tumor morphology, tumor origin and which project the subject belonged to.

Over 60,000 gene expression values were encoded into a vector of 128 latent factors. To increase the efficiency of information processing by CFGAIN, gene expression values were arranged into a 2X2 matrix in size of 236x238 with padding 0s in the end. Then

convolutional filters were used in the auto-encoder part of CFGAIN. To ensure fair comparisons, the 128 latent factors were used to summarize information from gene expression profiles for all methods.

Plug-in validation PEHE was presented in Table 12. Since ITE was defined as the difference between two treatment arms, so three pairwise comparison was conducted to calculate ATE, ATT and ATC. It was worth noting that CFAIN was one of the top 2 best methods for all three pairwise comparisons. For the comparison between chemo-therapy vs. combinational therapy, random forest had the lowest PEHE, indicating its best accuracy to quantify the ITEs, whereas CFGAIN came to the second place. However, for the other two comparisons, CFGAIN provided the smallest PEHE, which made it most suitable for this two comparisons.

Table 13-15 listed the results of applying CFGAIN on TCGA lung cancer data. As stated previously, random forest was chosen as the analytical method for the comparison between chemotherapy and combination, people receiving chemotherapy were estimated to live 0.14 standard deviation longer than people receiving combinational therapy on average (Table 13). Similarly, for the comparison between radiation and combination therapy, people receiving radiation were estimated to live longer than people receiving combination therapy on average (Table 14). Last but not least, the results from CFGAIN indicated that people receiving chemotherapy would live longer than people receiving radiation on average, controlling for clinical factors (Table 15).

After getting the complete outcome vectors, ITE was calculated for each subject. The treatment therapy corresponding to the longest survival days was chosen as the recommended

treatment (Table 16). KNN, random forest and linear recommended more patients to radiation therapy, which was different from the observed distribution for the treatment strategies, indicating possible misplacement of patients to suboptimal treatment strategies. On contrary, CFGAIN recommended more patients to chemotherapy, which was the closest to the therapies that they actually received clinically.

To further understand which latent factors and genes played important roles for ITE, lasso regression with 5-fold cross validation was performed by regressing ITE estimates against latent factors. After important features were selected, multiple linear regression was performed on ITE vs. selected features and  $R^2$  was reported (Table 17). For all three pairs, selected latent factors explained around 40% of variance in ITE, indicating the predictive power of gene expression profiles. In addition, after clinical covariates were added,  $R^2$  was increased for all three pairwise comparisons, especially for radiation therapy in comparison with the other two treatments.

To further understand the relations between latent factors and genes, Spearman correlation was calculated for each of the genes and each of the latent factors selected from Lasso regression. Genes that were significantly associated with latent factors were presented in Table 18. Several EGFR-related genes were related to several latent factors. EGFR-AS1 was related to factor 101 (p-value=0.0076). ELDR was significantly associated with factor 51 (p-value=0.0063), factor 99 (p-value=0.0182) and factor 87 (p-value=0.0126).

### ***3.2 Causal Discovery for Alzheimer's disease***

#### **3.2.1 AD classification and prediction**

The network with 3D filters was used for classification and prediction of AD using 3D whole brain DTI images at 4 different time points: baseline, 6 months, 12 months and 24 months. We consider two classes: AD and NC. AD prediction accuracy using 3D-CNN was listed in Table 16 and its sensitivity and specificity in Table 17 where the left and right in the brackets represented sensitivity and specificity, respectively. Tables 19 and 20 demonstrated that the prediction accuracy, sensitivity and specificity of the model using the training dataset at baseline to predict AD in the test datasets at baseline, 6 months, 12 months and 24 months were 0.8675 (0.6873, 0.9600), 0.8452 (0.6364, 0.9600), 0.8335 (0.7295, 0.8995), 0.7463 (0.6294, 0.8853), respectively. In other cases, we can observe similar results. The AUC using the training data at baseline, 6 months, 12 months and 24 months for prediction of AD in the test datasets at the same time points, was 0.8571, 0.8291, 0.8583, and 0.7756, respectively. The low sensitivity of prediction of AD may be due to small and imbalanced sample size (51 AD and 100 controls). Much higher proportion of non-AD controls) decreased sensitivity, while increased specificity. Deep 3D-CNN that has a large number of parameters to be estimated requires large sample sizes. Although we used data augmentation methods to increase sample sizes, augmentation methods still did not provide large and reliable sample sizes. Large sample sizes is an important issue for increasing the prediction of accuracy.

### **3.2.2 Region selection and interpretation**

Relative importance of value  $d$  was sorted. Image areas whose relative importance value were in the top 10<sup>th</sup> percentile were considered as features that contributed substantially to the prediction of AD. We identified 23 important brain regions that contributed

substantially to AD prediction. The results were shown in Figure 12 where each sub-figure had  $91 \times 109$  pixel sizes where the darker the red color is, the more important the brain region is to the prediction accuracy. The brain regions with red color included the temporal lobe (the left temporal lobe, medial, and right temporal lobe), ventricles and enlarged ventricle, occipital lobe and prefrontal area. To further interpret the image analysis results and increase their transparency, we tested the causal relationships between DTIs image ROIs and AD disease at baseline, 6 months, 12 months and 24 months using CGAN-based statistics. After Bonferroni correction, P-value  $< 0.0022$  was threshold to declare significance. The number of identified brain regions that showed significant causation to AD at baseline, 6 months, 12 months and 24 months were 1, 1, 2, and 4, respectively. Table 21 listed ROIs where p-values for testing causation between the ROI and AD were less than 0.05. Three remarkable features emerged from these results. First, as time passed, AD progressed from mild (early stage), via moderate (middle stage), to severe (late stage) which resulted in atrophy of more and more brain regions. Therefore, we observed the increased number of significant causal brain regions with AD as the study time of AD increased from the baseline to 24 months. Second, in general, as AD progressed, the significance of causation between the brain region and AD increased (p-values for testing causation decreased). Third, the brain region ROI 18 (the ventricles and enlarged ventricle) (Figure 13) showed significant causation to AD at all four time points (baseline, 6 months, 12 months and 24 months). The brain regions ROI 14 (the left temporal lobe) (Figure 13) showed significant causation at 12 months and 24 months after Bonferroni correction. The literature reports that these regions are related to AD. The left temporal lobe is involved in language

and AD (Cretin, Di Bitonto, Blanc, & Magnin, 2015; Flick et al., 2018; Trimmel et al., 2018), the right temporal lobe atrophy is involved in severe impairment in emotion recognition (Everhart, Watson, Bickel, & Stephenson, 2015) and causes frontotemporal dementia (Gliebus, 2014), with the brain ventricles often affected AD (Ferrarini et al., 2006). Ventricle enlargement is a useful structural biomarker for the diagnosis of AD (Anandh, Sujatha, & Ramakrishnan, 2014).

### **3.2.3 Genetic studies of two brain regions**

To uncover genetic architecture of brain regions, in addition to genetic imaging association analysis, we conducted genetic imaging causal analysis using ACT where imaging signals within the brain region and SNPs within the gene were summarized by two dimensional functional principle scores and classical functional principle scores, respectively (Lopez-Paz & Oquab, 2016). The total number of candidate genes being tested was 61. After Bonferroni correction, p-value for declaring significance of both causation and association was 0.00082. We presented the results of P-values  $< 0.05$  in causal analysis and association analysis of genetic variation in 61 candidate genes with two brain regions: left temporal lobe and frontal and temporal left lobe, and right temporal lobe as seen in Tables S1 and S2, respectively, where 61 genes were obtained from genome-wide causation studies of AD in the manuscript (Lin et al., unpublished). In Tables S1 and S2, the P-values in green bold denoted significant causation or association after Bonferroni corrections. The majority of genes that had causal or association relationships with brain neuroimaging phenotypes were identified at all time points (baseline, 6 months, 12 months and 24 months). We also

observed that these identified genes had causal or association relationships with both the left temporal lobe and right temporal lobe regions. The identified genes *CD33*, *COBL* and *APP* that had causal relationships with brain neuroimaging regions were confirmed multiple times in the literature (Bradshaw et al., 2013; C.-C. Huang et al., 2019; C.-Y. Huang et al., 2019; Kovacs, Burchett, & Sheafor, 2018; Mez et al., 2017; Van Giau et al., 2018). It was also reported that gene *FGF4* was involved in neurodevelopmental disorders (Grillo et al., 2014), *FRMD6* was implicated in AD (Hong et al., 2012), *Dock9* played an important role in regulation of morphological changes in hippocampal neurons (Kuramoto, Negishi, & Katoh, 2009), *H3F3B* was associated with a broad schizophrenia phenotype (Manley et al., 2018), *SCYL1* was involved in cerebellar atrophy (Lenz et al., 2018), *AKAP5* played a significant role in the regulation of sympathetic nerve activities (Han et al., 2016), and *PIGC* was involved in epilepsy and intellectual disability (Edvardson et al., 2017).

## DISCUSSION

### 1. ITE-Related Genes

The genes identified in this dissertation that were associated with ITEs were also confirmed in literature as important cancer genes or prognostic markers. Recent studies found that *TRIM32* not only had an impact on chemo-resistance to breast cancer cells through NF- $\kappa$ B signaling (Zhao et al., 2018), but also negatively regulated tumor suppressor gene *p53* to contribute to tumorigenesis (J. Liu et al., 2014). *HOXB5* was shown negatively affected cell proliferation, migration and invasion in non-small cell lung cancer (B. Zhang, Li, & Zhang, 2018), but acted as an oncogenic driver in breast cancer and head and neck



carcinoma (J.-Y. Lee et al., 2015; K. Lee et al., 2019). AMIGO2 upregulation promoted metastatic tumor cells to attach to liver endothelial cells and has been considered as a novel target for cancer (Fontanals-Cirera et al., 2017; Kanda et al., 2017). Researchers have also found that overexpression of IGSF10 genes was significantly associated with good prognosis for lung cancer patients, indicating a possible prognostic marker for treatment evaluation (Ling et al., 2020). In addition to protein-coding genes, the expression level of long noncoding RNA (lncRNAs) was found to be associated with heterogeneous prognosis in this study, which has also been consistent with previous findings (Sun et al., 2019). The epidermal growth factor receptor antisense RNA 1 (EGFR-AS1) was found to be associated with a poor prognosis in patients with non small cell lung cancer and renal cancer, which could be attributed to overexpression of EGFR-AS1 induced chemotherapy resistance (Tan et al., 2017; Xu, Tu, Zhao, Xie, & Tang, 2019). Moreover, I also identified that AL092794.1, an antisense to KRAS, was significant associated with ITE distribution. KRAS mutation was established as the major driver of lung cancer cell growth and has been playing a very important role in genetic counselling and treatment selection for patients (Eberhard et al., 2005; Riely, Marks, & Pao, 2009).

The greatest challenge for machine-learning based methods to be applicable in real datasets for ITE estimations lies in the inaccessibility of counterfactuals. In this project, I used the plug-in validation, which synthesized complete outcome vectors based on observed data. However, plug-in PEHE estimates can only reveal their true comparative performance when the plugged-in distribution and the true distribution are close enough, i.e.  $\|\tilde{T} - T\|_{\theta}^2 \approx$

0 (A. Alaa & Van Der Schaar, 2019). This assumption is difficult to test because we do not have access to the ground truth. Alaa et al. recently proposed to use the influence statistics to unplug from the plugged-in distribution to adjust the plug-in PEHE estimates (A. Alaa & Van Der Schaar, 2019). Unfortunately they have not released enough details of their methods for me to apply their method. Further research will be conducted to resolve the problem of model validation without access to counterfactuals.

## **2. Causal Discovery for AD**

In this dissertation, a general artificial intelligence (AI) platform for prediction of AD using DTI images was presented. Non-transparency could be a major challenge of deep learning for medical image analysis. To meet this challenge, we introduced three approaches to medical image interpretation: feature selection and visualization, causal analysis of neuroimaging region and genetic-imaging analysis. Feature selection and visualization methods selected and visualized brain regions as a potential pathology of AD. Further ACT tests discovered potential causal relationships between the brain neuroimaging regions and candidate genes for AD. We observed the increased number of significant causal brain regions with AD when AD progressed. In general, as AD progressed, the significance of causation between the brain region and AD increased (p-values decreased). We observed the ventricles and enlarged ventricle, left and right temporal lobes had strong causal relationships with AD. Temporal lobes including the hippocampus are crucial in AD development at the early stages, whereas the ventricles and enlarged ventricle are useful structural biomarker for the diagnosis of AD. Joint causal analysis of genetic and images of left and right temporal

regions using ACTs mapped *CD33*, *COBL*, *FRMD6*, *APP* and other genes to the left and right temporal brain regions.

Many findings in the dissertation could be confirmed in the literature. For example, both prediction analysis using deep learning and causal analysis using ACT identified brain temporal lobe region that was involved in AD. Temporal lobe includes the hippocampus and its surrounding regions. It is well-known that the temporal lobe consists of structures that are vital for long-term memory. There are numerous reports that temporal lobe including left, medial and right temporal lobe are involved in AD pathology (Delgado-González, Florensa-Vila, Mansilla-Legorburo, Insausti, & Artacho-Pérula, 2017; Kakeda & Korogi, 2010; Li & Chen, 2015; Menéndez-González, de Celis Alonso, Salas-Pacheco, & Arias-Carrión, 2015; Wolk et al., 2017). DTI discovered the functional and structural connectivity between the medial temporal lobes (MTL) and posteromedial cortex (PMC) (Buckner, Andrews-Hanna, & Schacter, 2008; Pasquini et al., 2019). The MTL includes the hippocampal formation and other cortices. These regions underlie memory processing through interplay with neocortical areas from the PMC. AD-related pathological changes such as tau accumulation, amyloid- $\beta$  deposition often affect the PMC and MTL regions. The functional and structural disconnections between the MTL and PMC cause the development and progression of AD.

The literature confirmed the identified pathological paths from genetic variants to AD via brain regions: *CD33*→Medial temporal and hippocampus (Wang et al., 2019) → AD (Pasquini et al., 2019), and *CD33*→ AD (Miles et al., 2019); *APP* → medial and lateral temporal lobe (C.-Y. Huang et al., 2019) → AD (Buckner et al., 2008), and *APP* → AD (Z.-d.

Zhou et al., 2011); SCYL1  $\rightarrow$  Cerebellar Atrophy (Schmidt et al., 2015)  $\rightarrow$  AD (Gallo et al., 2017), and SCYL1  $\rightarrow$  neurodegenerative disease (Schmidt et al., 2007). These provided indirect evidences of identified biomarkers for unravelling mechanism of AD.

## CONCLUSION

In this dissertation, I addressed the causal inference problem through deep learning approaches. A framework based on the generative adversarial networks was developed to estimate the heterogeneous treatment effect using the well-established counterfactual reasoning for causal inference (Hernán MA, 2019). In addition to this experimental design to determine causation and its effect, I further explored the case where only retrospective observational data was available to perform causal discovery between two sets of variables. A novel causal test was proposed, which was inspired by the additive noise models (ANMs) for causal discovery but with relaxed assumptions and great flexibility. A few applications of developed methods to real datasets were also presented to showcase their usage, as well as to stimulate further discussions regarding the great challenges we are facing in developing robust deep learning platforms.

Given the popularity of machine learning, which are associational learning systems (Pearl, 2019), it seems less obvious how causal inference could potentially improve artificial intelligence. However, modern machine learning systems are facing several major obstacles (Pearl, 2019). The first obstacle is the lack of robustness or reliability. Current systems often fail miserably for new instances which they have not been trained for. They are optimized for the current dataset, but it is quite difficult for the systems to transfer the knowledge to a new

dataset. Another great difficulty is opacity or the lack of explainability. Current systems have mostly been considered as ‘black-box’ algorithms. This non-transparency reduces people’s trust towards such systems and thus limits their usage in practice. Last but not least, current systems could not recognize cause-effect relations, which is considered as a necessary component of human-level intelligence. Causal inference is able to provide some remedies to all these difficulties, which helps to build stronger AI systems.

Just as causal inference has the potential to revolutionize machine learning, machine learning is capable to empower modern causal inference. With the advances in graphical models and structural equations, structural causal models (SCMs) have been established, which serves as the ‘inference engine’ to solve sophisticated queries of interest (Pearl, 2019). SCMs engine accepts assumptions (in the form of DAGs), data and queries of interest and outputs an estimand (i.e. a procedure of estimation), an estimate (i.e. a consistent estimate of the estimand) and fit indices which measure the compatibility of data and the model assumptions. Many techniques have been operated under this framework such as propensity score methods and tree-based methods (Austin, 2011; Lu et al., 2018). However, as data complexity and volume increases, these techniques have their own limitations and deep learning approaches are likely to outperform them, just as what have been shown in this dissertation.

Deep learning approaches have a few remarkable strengths. One key strength is the capability to handle high-dimensional complex data without manual feature engineering. For example, I used auto-encoders to conduct automatic feature engineering and dimension reduction, which reduced the whole gene expression profile to hundreds of latent factors. In

addition, deep learning approaches have been adaptable to various form of data.

Convolutional neural networks have gained huge success in image recognition, a task that I also explored using DTI images to classify Alzheimer's cases in this dissertation. Recurrent neural networks are specially designed for sequential data to accommodate the temporal dependence. Moreover, deep learning approaches exhibits more flexibility in terms of approximating non-linear functions than standard approaches. One typical example to demonstrate this is the generative adversarial networks used in this dissertation. In contrast to the additive noise models for causal discovery which assumes  $y = f(x) + noise$ , GAN essentially tries to learn the non-linear mapping of  $y = f(x, noise)$ , which is more generous than the additive assumptions. Last but not least, deep learning approaches are generally more suitable to unsupervised learning. In this dissertation, one of the greatest obstacle was the lack of counterfactuals, which made ITE estimation an unsupervised problem. Although I was able to apply a few machine learning techniques to get the estimation, they were essentially supervised learning. Nevertheless, CFGAIN was based on the generative adversarial networks, which were aimed to learn the distribution of the input data in an unsupervised fashion. Therefore, CFGAIN is more respectful of the unsupervised nature of this problem, in comparison of other methods.

Despite the advantages of deep learning, we need to pay special attention to its limitations. First, deep learning approaches generally require large amount of data so that they can learn from many examples and capture as much the association signals as possible. However, human beings can perform object recognition with much less training using causal reasoning. This further highlights the importance of incorporating causal inference into deep

learning systems. Second, deep learning systems have always been criticized for non-transparency. In this dissertation, I tried to offset this by proposing a novel statistical test for causal discovery, which could be conducted in order to understand a complex deep learning system through causal analysis. The causal discovery analysis for Alzheimer's disease serves as an attempt to decode a deep learning system.

In conclusion, causal inference, in conjunction with deep learning approaches, is crucial to exploit full potential of artificial intelligence. Further research is needed to explore additional methods for heterogeneous effect estimation, causal discovery, missing data and confounding issues, as well as how to incorporate causal inference into complex machine learning systems in an effective and efficient way.

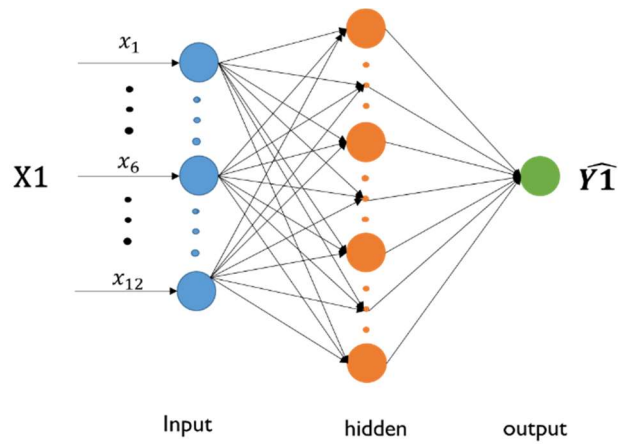


Figure 1. An example of neural networks

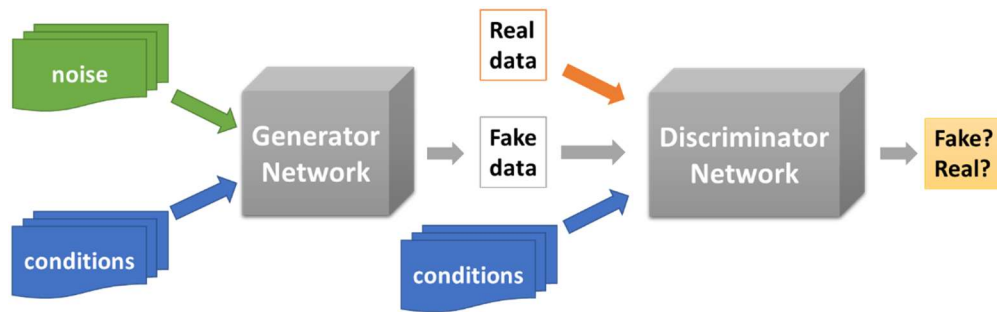


Figure 2. The structure of a conditional generative adversarial network (CGAN).



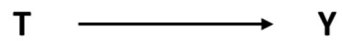


Figure 3. DAG for a randomized controlled trial

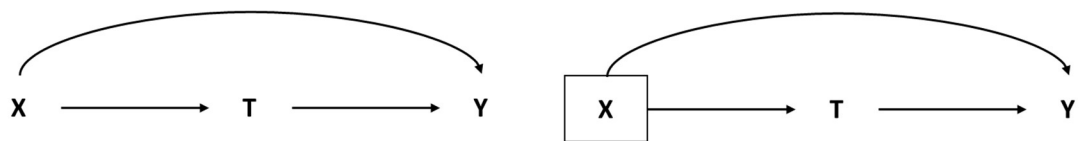


Figure 4. Causal DAGs for Observational Studies.

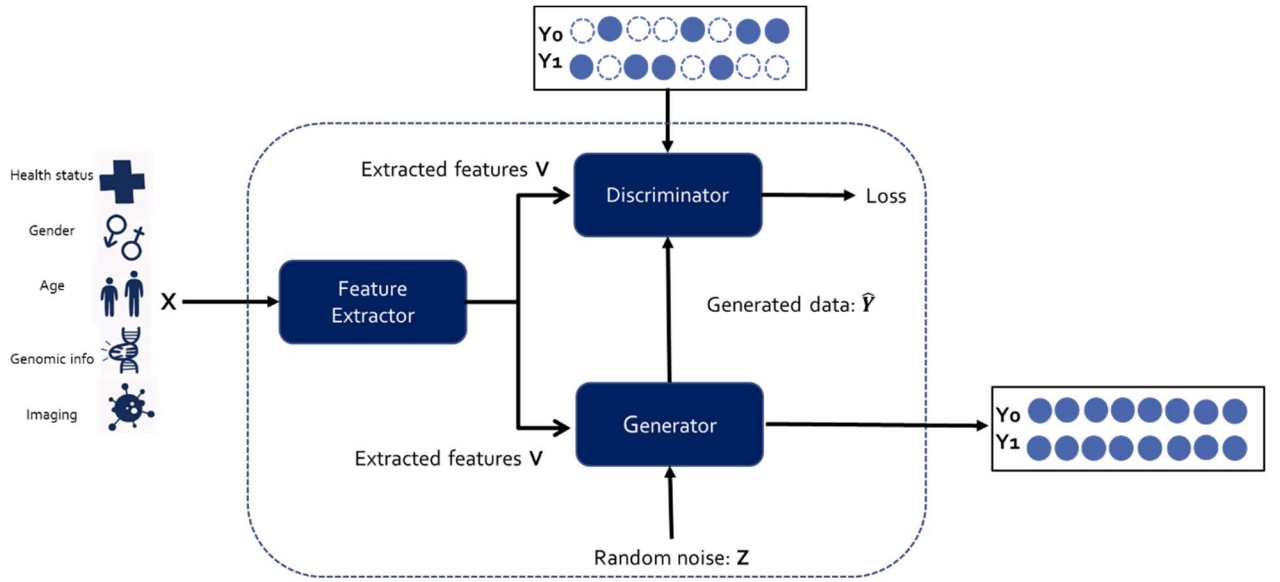


Figure 5. Architecture of CFGAIN.

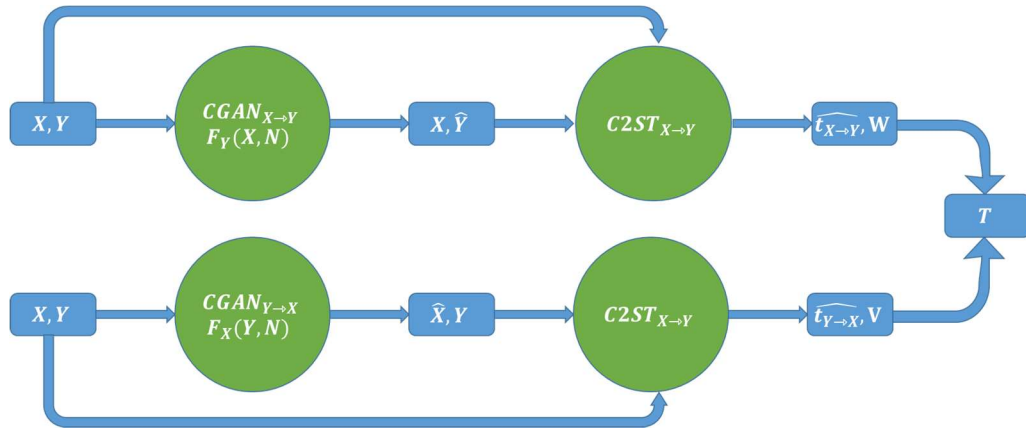


Figure 6. Procedure for conducting the proposed adversarial causal test (ACT).

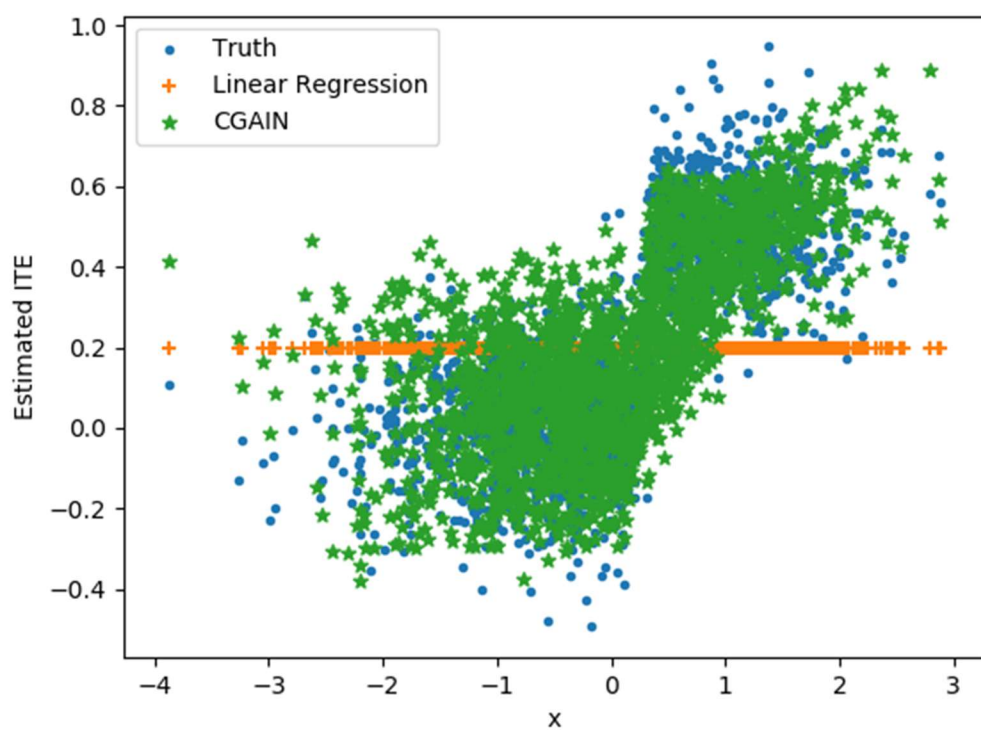


Figure 7. Comparisons of ITE results from linear regression and CFGAIN

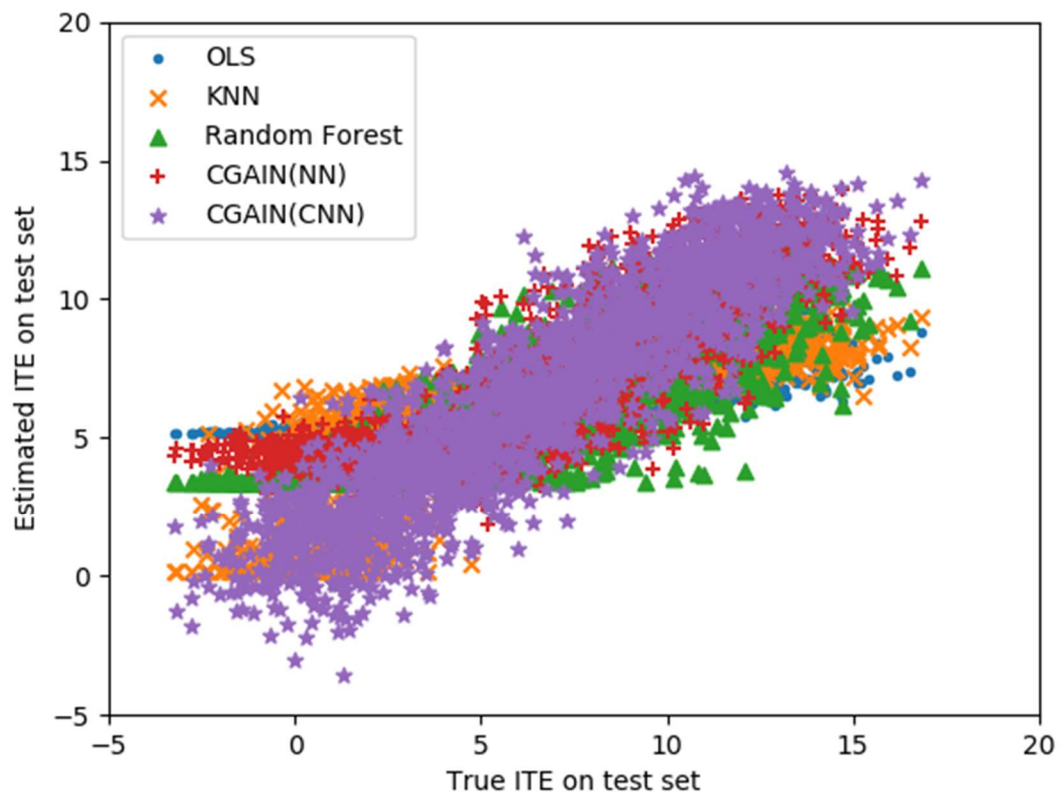


Figure 8. Comparisons of ITE results for different methods

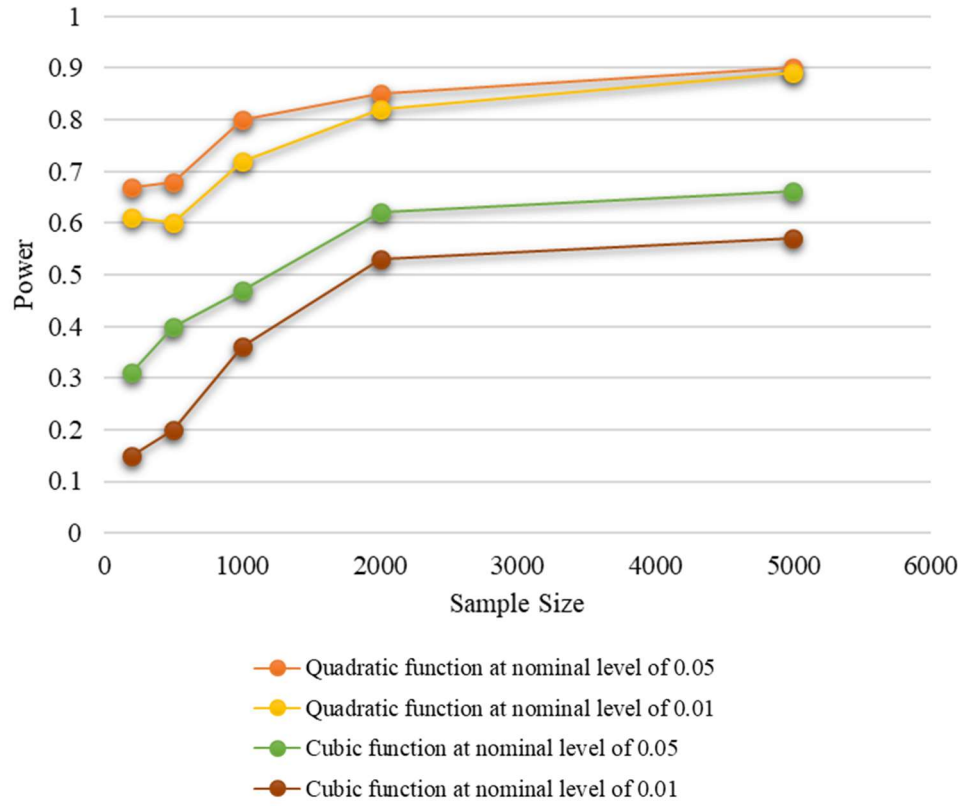


Figure 9. Power for ACTs at varying sample sizes. Cause  $X$  was randomly sampled from normal distribution and effect  $Y$  was obtained through the non-linear transformation of  $X$  plus the random noise  $N$ :  $Y = f(X) + N$ . Quadratic and cubic functions were used for the non-linear mapping.

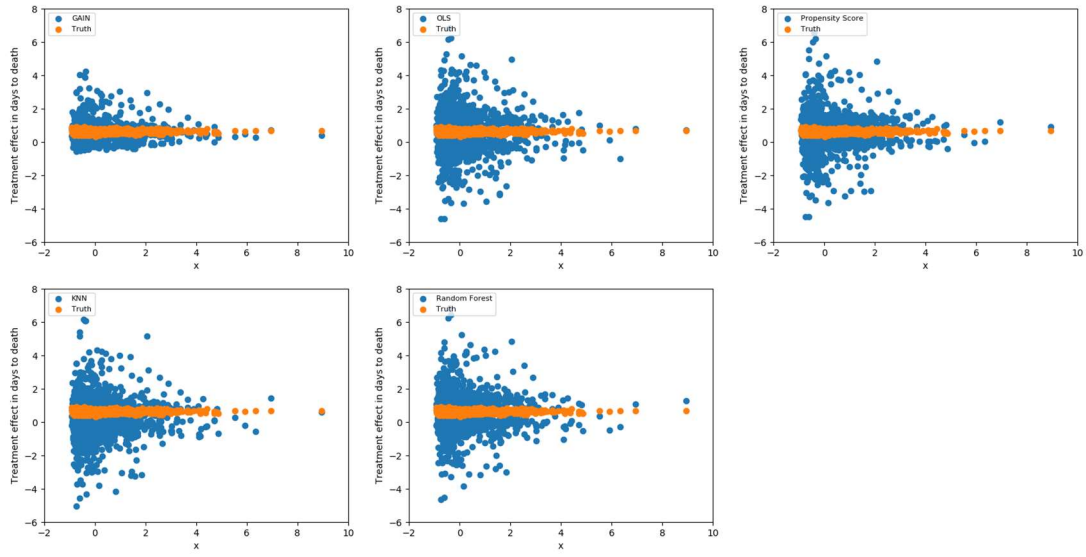


Figure 10. Comparisons of various methods for ITE estimation without feature selection.

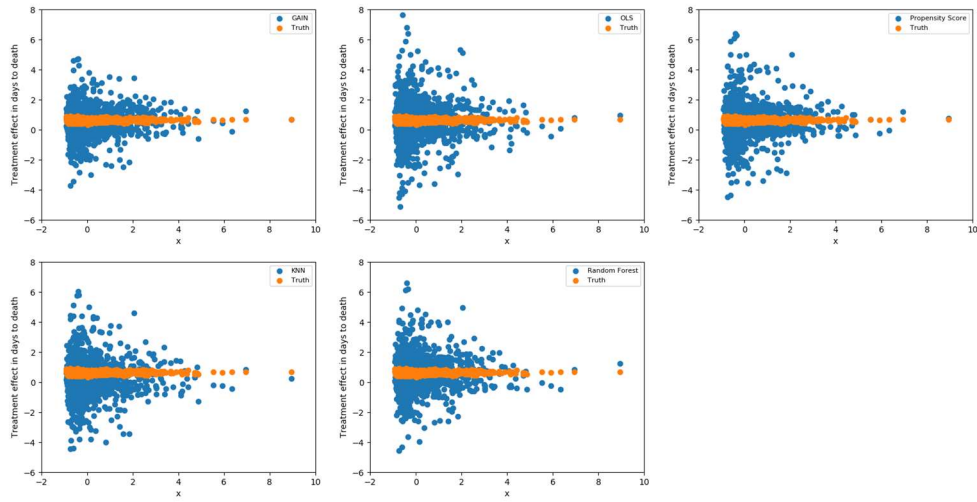


Figure 11. Comparisons of various methods for ITE estimation with feature selection.

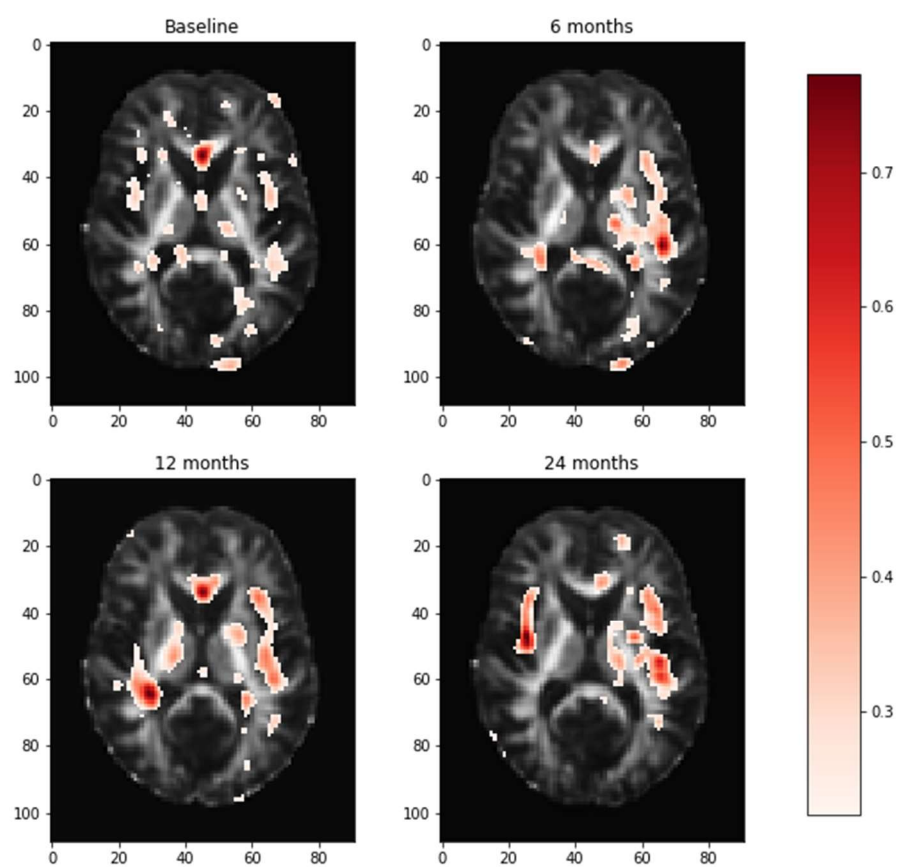
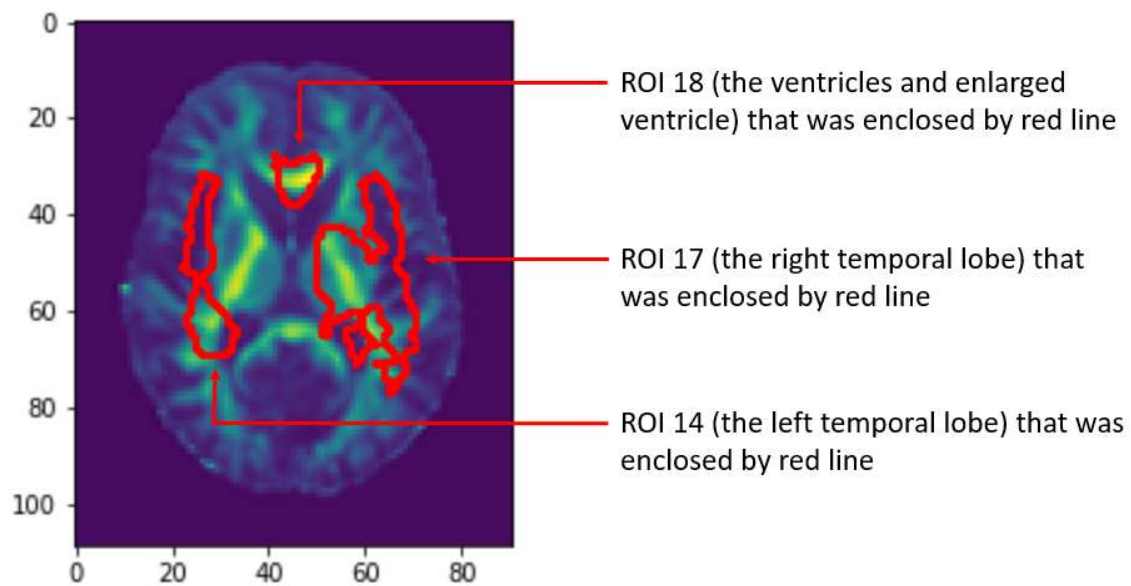


Figure 12. Visualization of the brain regions with relative importance values at the baseline, 6 months, 12 months and 24 months. The deeper red color of the brain region, the more important for AD prediction.



*Figure 13. Three brain regions showed causation to AD.*



Table 1. 3D filters in five convolutional layers.

Filter Size	Stride in (X,Y) direction	Stride in Z direction
11X11X11	4	4
5X5X5	1	1
3X3X3	1	1
3X3X3	1	1
3X3X3	1	1

Table 2. Comparisons of different methods for ITE using data generated in scenario I

Methods	ATE	ATT	ATC	PEHE	MSE
Ground Truth	0.1940	0.2020	0.1863	0.1290	0.0
Conditional GAIN	0.1898	0.1406	0.2367	0.1165	0.0295
OLS (t only)	0.2008	0.2008	0.2008	0.0549	0.0748
OLS (t and x)	0.1990	0.1990	0.1990	0.0549	0.0557
Propensity Score	0.1990	0.1990	0.1990	0.0549	0.0557
KNN	0.1943	0.1987	0.1900	0.1119	0.0212
Random Forest	0.1936	0.1972	0.1902	0.1104	0.0199

Table 3. Comparisons of different methods for ITE using data generated in scenario II

Methods	ATE	ATT	ATC	PEHE	MSE
Ground Truth	6.5385	6.6713	6.4142	2.0856	0.0
Conditional GAIN (NN)	7.0666	7.1156	7.0207	3.6481	5.5315
Conditional GAIN (CNN)	6.5531	6.7098	6.4064	1.4079	3.1188
OLS (t and x)	6.5089	6.5556	6.4651	9.6378	8.8863
Propensity Score	6.2813	6.3252	6.2402	10.0558	9.3316
KNN	6.2090	6.1913	6.2255	5.7457	6.1478
Random Forest	6.3046	6.4193	6.1972	4.0134	6.0422

Table 4. Type I errors when there is no association or causation

Sample Size Nominal Level	500	1000	2000
0.01	0.008	0.011	0.014
0.05	0.036	0.045	0.051

Table 5. Type I errors when there is association but no causation

Sample Size Nominal Level	500	1000	2000
0.01	0.006	0.010	0.011
0.05	0.046	0.057	0.045

Table 6. Causation power for ACTs at varying sample sizes using the quadratic transformation.

Causation Power		Number of samples				
		200	500	1000	2000	5000
Significance level	0.05	0.67	0.68	0.80	0.85	0.90
	0.01	0.61	0.60	0.72	0.82	0.89

Table 7. Causation power for ACTs at varying sample sizes using the cubic transformation.

Causation Power		Number of samples				
		200	500	1000	2000	5000
Significance level	0.05	0.31	0.40	0.47	0.62	0.66
	0.01	0.15	0.20	0.36	0.53	0.57

*Table 8. Power for ACTs to correctly identifying causation direction at varying sample sizes using the quadratic transformation.*

Correct Direction Power		Number of samples				
		200	500	1000	2000	5000
Significance level	0.05	0.66	0.63	0.66	0.67	0.72
	0.01	0.60	0.58	0.63	0.65	0.71

*Table 9. Power for ACTs to correctly identifying causation direction at varying sample sizes using the quadratic transformation.*

Correct Direction Power		Number of samples				
		200	500	1000	2000	5000
Significance level	0.05	0.31	0.34	0.44	0.52	0.52
	0.01	0.15	0.20	0.34	0.45	0.48

Table 10. Comparison of different methods in simulation study

Methods	ATE	ATT	ATC	MSE
Ground Truth	0.6295	0.6304	0.6285	0
CFGAIN	0.6103	0.5919	0.6292	0.2463
Linear Regression	0.6169	0.6304	0.6032	4.4341
Propensity Score	0.6059	0.6466	0.5648	1.0260
KNN	0.2804	0.3270	0.2335	1.2665
Random Forest	0.6096	0.5942	0.6251	1.1007

Table 11. Comparison of different methods with feature selection

Methods	ATE	ATT	ATC	MSE
Ground Truth	0.6295	0.6304	0.6285	0
CFGAIN	0.6133	0.6284	0.5980	0.6970
Linear Regression	0.6325	0.7588	0.5052	6.0497
Propensity Score	0.6081	0.6749	0.5407	1.0425
KNN	0.3094	0.4401	0.1776	1.2012
Random Forest	0.6306	0.6392	0.6220	1.0742

Table 12. Plug-in validation of different methods for TCGA lung cancer dataset

	Chemo vs. Combo	Radiation vs. Combo	Radiation vs. Chemo
Linear Regression	1.9838	1.4893	2.0630
Propensity Score	2.0843	1.4044	1.8632
KNN	1.9502	1.4865	2.2115
Random Forest	<b>1.6913</b>	1.5066	2.0298
<b>Auto- encoder+GAIN</b>	1.8408	<b>1.2825</b>	<b>1.8538</b>

Table 13. Chemotherapy (treatment) vs. Combinational therapy (control)

	ATE	ATT	ATC
Linear Regression	0.0694	0.2668	-0.1164
Propensity Score	-0.0153	0.0310	-0.0501
KNN	0.1137	0.3439	-0.0845
Random Forests	0.1368	0.3615	-0.0955
Auto-encoder+GAIN	0.1900	0.3988	0.0196

\*ATT denotes the average ITE effect for people in the chemotherapy group.

\*\* ATC denotes the average ITE effect for people in the combinational therapy group.

Table 14. Radiation (treatment) vs. Combinational therapy (control)

	ATE	ATT	ATC
Linear Regression	-0.0185	0.1158	-0.1164
Propensity Score	0.0854	-0.0284	0.1879
KNN	-0.0009	0.1435	-0.0845
Random Forest	-0.1450	-0.3695	-0.1142
<b>Auto-encoder+GAIN</b>	<b>0.0532</b>	<b>0.0328</b>	<b>-0.2528</b>

\*ATT denotes the average ITE effect for people in the radiation group.

\*\* ATC denotes the average ITE effect for people in the combinational therapy group.

Table 15. Radiation (treatment) vs. chemotherapy (control)

	ATE	ATT	ATC
Linear Regression	-0.0879	0.1158	-0.2668
Propensity Score	0.1007	0.0186	0.0217
KNN	-0.1146	0.1435	-0.3439
Random Forest	-0.2818	-0.4631	-0.4213
<b>Auto-encoder+GAIN</b>	<b>-0.1368</b>	<b>-0.0376</b>	<b>-0.0678</b>

\*ATT denotes the average ITE effect for people in the radiation group.

\*\* ATC denotes the average ITE effect for people in the combinational therapy group.

Table 16. Counts for Observed/Recommended Treatment on entire dataset

	Chemotherapy	Radiation	Combinational therapy
Observed	80	39	67
Linear Regression	20	108	58
Propensity Score	53	53	80
KNN	39	70	77
Random Forest	50	91	45
<b>CFGAIN</b>	<b>70</b>	<b>55</b>	<b>61</b>

Table 17. Feature selection using lasso\* and R-square using selected features

Pairwise comparison	R-squared with selected latent factors for gene expression	R-squared with selected latent factors for gene expression + clinical factors**
Chemotherapy vs. Combination	0.4128	0.4713
Radiation vs. Combination	0.4261	0.6559
Radiation vs. Chemotherapy	0.3781	0.5335

\*Five-fold cross validation was used with Lasso to determine the best alpha.

\*\*Clinical factors refer to age at diagnosis, gender, race, ethnicity, diagnosis, tumor stage, tumor morphology, tumor origin and which project the subject belonged to.

Table 18. Selected significant correlations between latent factors and genes

Gene names	Gene descriptions	Correlated latent factors	P-values
TRIM32	tripartite motif containing 32	20	$3.8265 \times 10^{-5}$
		98	$4.3738 \times 10^{-8}$
HOXB5	homeobox B5 [Source:HGNC Symbol;Acc:HGNC:5116]	86	$2.4997 \times 10^{-4}$
AMIGO2	adhesion molecule with Ig like domain 2	125	$1.4361 \times 10^{-4}$
IGSF10	immunoglobulin superfamily member 10	17	$4.3097 \times 10^{-4}$
WHAMM	WASP homolog associated with actin, golgi membranes and microtubules	119	$5.8337 \times 10^{-7}$
EBLN1	endogenous Bornavirus like nucleoprotein 1	5	$9.7148 \times 10^{-4}$
OR52B2	olfactory receptor family 52 subfamily B member 2	123	$2.3129 \times 10^{-4}$
AC241585.1	EGFR-coamplified and overexpressed protein (ECOP) pseudogene	20	0.0020
		36	0.0030
		84	$4.4538 \times 10^{-5}$
		87	0.0006
		125	0.0083
EGFR-AS1	EGFR antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:40207]	101	0.0076
ELDR	EGFR long non-coding downstream RNA [Source:HGNC Symbol;Acc:HGNC:49511]	51	0.0063
		99	0.0182
		87	0.0126
TP53TG3GP	TP53 target 3 family member G, pseudogene [Source:HGNC Symbol;Acc:HGNC:51818]	11	0.0057
		75	0.0280
		78	0.0072
		97	0.0112
		107	0.0046
		123	0.0087
CDKN2A-DT	CDKN2A divergent transcript [Source:HGNC Symbol;Acc:HGNC:23831]	20	$8.8294 \times 10^{-4}$
		16	$1.0053 \times 10^{-3}$
		36	0.0049



		100	0.0191
AL092794.1	antisense to KRAS	83	$4.8675 \times 10^{-2}$
AL021407.3	PERP, TP53 apoptosis effector (PERP) pseudogene	42	0.0115
TPRKBP1	TP53RK binding protein pseudogene 1 [Source:HGNC Symbol;Acc:HGNC:44943]	122	0.0062
CDKN2AIPNLP1	CDKN2A interacting protein N-terminal like pseudogene 1 [Source:HGNC Symbol;Acc:HGNC:39854]	112	0.0085
SMAD1-AS1	SMAD1 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:49379]	3	$6.7220 \times 10^{-4}$
		7	0.0070
		8	0.0013
		16	0.0054
		28	0.0341
		43	0.0269
		54	0.0045
		60	0.0136
		75	0.0067S
		84	$3.0237 \times 10^{-4}$
		105	0.0011
		107	0.0087
		112	0.0080
		119	$5.7954 \times 10^{-4}$
		123	0.0062
SMAD5-AS1	SMAD5 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:30586]	79	$6.8549 \times 10^{-4}$
		101	0.0066
AC125611.2	FGFR1 oncogene partner 2 (FGFR1OP2) pseudogene	108	0.0025
		119	$4.8342 \times 10^{-4}$
AC126471.1	p53 and DNA-damage regulated 1 (PDRG1) pseudogene	24	0.0045
		28	0.0347
		75	0.0295
		86	0.0148
		92	0.0030
AC068044.1	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse) (MDM2) pseudogene	122	0.0145

AC008126.1	Mdm2 p53 binding protein homolog (mouse) (MDM2) pseudogene	67	0.0205
NOP53-AS1	NOP53 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:51587]	13	0.0013
		37	$6.8694 \times 10^{-4}$
		51	0.0023
		58	0.0011
		63	$2.8379 \times 10^{-6}$
		64	$7.1582 \times 10^{-5}$
		68	$7.8089 \times 10^{-4}$
H3F3AP6	H3 histone, family 3A, pseudogene 6 [Source:HGNC Symbol;Acc:HGNC:42982]	106	0.0118
AC355297.1	H3 histone, family 3A (H3F3A) pseudogene	36	0.0101
H3F3AP4	H3 histone, family 3A, pseudogene 4 [Source:HGNC Symbol;Acc:HGNC:42980]	46	0.0189
		65	0.0048
H3F3AP4	H3 histone family member 3A pseudogene 2 [Source:HGNC Symbol;Acc:HGNC:19823]	39	0.0054
		110	0.0017
AL109618.3	fibroblast growth factor receptor 3 (FGFR3) pseudogene	61	0.0017
AL627095.1	pseudogene similar to part of fibroblast growth factor receptor 3 (achondroplasia, thanatophoric dwarfism) (FGFR3)	9	0.0066
PTENP1	phosphatase and tensin homolog pseudogene 1 [Source:HGNC Symbol;Acc:HGNC:9589]	9	0.0031
		108	0.0030
AC006600.1	phosphatase and tensin homologphosphatase and tensin homologphosphatase and tensin homolog (PTEN) pseudogene	34	0.0106

Table 19. AD prediction accuracy on five-fold cross validation.

Model Development Time Point	Prediction Time Point			
	Baseline	6 months	12 months	24 months
Baseline	0.8675	0.9123	0.8864	0.7967
6 months		0.8452	0.8963	0.7791
12 months			0.8335	0.7813
24 months				0.7643

Table 20. Average sensitivity and specificity over five-fold cross validation

Model Development Time Point	Prediction Time Point			
	Baseline	6 months	12 months	24 months
Baseline	(0.6873, 0.9600)	(0.8073, 0.9700)	(0.7524, 0.9717)	(0.6465, 0.9313)
6 months		(0.6364, 0.9600)	(0.7778, 0.9717)	(0.5977, 0.9417)
12 months			(0.7295, 0.8995)	(0.6674, 0.8833)
24 months				(0.6294, 0.8853)

Table 21. Causations between DTIs image ROIs and AD disease status

Time Point	ROI Index	P-value
Baseline	2	0.0463
	18	0.0005
6 months	8	0.0182
	14	0.0108
	17	0.0155
	18	0.0010
12 months	6	0.0117
	14	0.0018
	17	0.0107
	18	<0.00005
24 months	0	0.0245
	3	0.0133
	5	0.0092
	7	0.0063
	8	0.0030
	9	0.0007
	11	0.0084
	12	0.0002
	13	0.0082
	14	<0.00005
	15	0.0098
	17	0.0239
	18	<0.00005
	19	0.0210
	21	0.0363
	22	0.0166

ROI 14 corresponds to the left temporal lobe and 17 corresponds to the right temporal lobe. ROI 18 corresponds to ventricles and enlarged ventricle indicates atrophy of cerebral nerve tissue, which is typical in AD patients.

## REFERENCES

- Aderghal, K., Boissenin, M., Benois-Pineau, J., Catheline, G., & Afdel, K. (2017). *Classification of sMRI for AD Diagnosis with Convolutional Neuronal Networks: A Pilot 2-D+  $\epsilon$  Study on ADNI*. Paper presented at the International Conference on Multimedia Modeling.
- Alaa, A., & Van Der Schaar, M. (2019). *Validating causal inference models via influence functions*. Paper presented at the International Conference on Machine Learning.
- Alaa, A. M., & van der Schaar, M. (2017). *Bayesian inference of individualized treatment effects using multi-task gaussian processes*. Paper presented at the Advances in Neural Information Processing Systems.
- Alaa, A. M., Weisz, M., & van der Schaar, M. (2017). Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*.
- Anandh, K., Sujatha, C., & Ramakrishnan, S. (2014). Segmentation of ventricles in Alzheimer mr images using anisotropic diffusion filtering and level set method. *Biomedical sciences instrumentation*, 50, 307-313.
- Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The economic journal*, 114(494), C52-C83.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.
- Association, A. s. (2016). 2016 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4), 459-509.

- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353-7360.
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3), 399-424.
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28), 3661-3679.
- Ballard, D. H. (1987). *Modular Learning in Neural Networks*. Paper presented at the AAAI.
- Barta, J. A., Powell, C. A., & Wisnivesky, J. P. (2019). Global epidemiology of lung cancer. *Annals of global health*, 85(1).
- Bradshaw, E. M., Chibnik, L. B., Keenan, B. T., Ottoboni, L., Raj, T., Tang, A., . . . Von Korff, A. (2013). CD33 Alzheimer's disease locus: altered monocyte function and amyloid biology. *Nature neuroscience*, 16(7), 848-850.
- Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *nature*, 562(7726), 203.
- Cancer Facts & Figures 2019 - American Cancer Society. (2019).

- Chen, R., & Liu, H. (2018). Heterogeneous Treatment Effect Estimation through Deep Learning. *arXiv preprint arXiv:1810.11010*.
- Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1-4.
- Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., . . . Chen, C.-M. (2016). Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Scientific reports*, 6, 24454.
- Choi, H., Jin, K. H., & Initiative, A. s. D. N. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural brain research*, 344, 103-109.
- Chorowski, J., Weiss, R. J., Bengio, S., & van den Oord, A. (2019). Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12), 2041-2053.
- Cooke, T., Reeves, J., Lannigan, A., & Stanton, P. (2001). The value of the human epidermal growth factor receptor-2 (HER2) as a prognostic marker. *European Journal of Cancer*, 37, 3-10.
- Cretin, B., Di Bitonto, L., Blanc, F., & Magnin, E. (2015). Left temporal lobe epilepsy revealing left posterior cortical atrophy due to Alzheimer's disease. *Journal of Alzheimer's Disease*, 45(2), 521-526.
- Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1), 151-161.

- Delgado-González, J.-C., Florensa-Vila, J., Mansilla-Legorburo, F., Insausti, R., & Artacho-Pérula, E. (2017). Magnetic Resonance Imaging and Anatomical Correlation of Human Temporal Lobe Landmarks, in 3D Euclidean Space: A Study of Control and Alzheimer's Disease Subjects. *Journal of Alzheimer's Disease*, 57(2), 461-473.
- Dimitriadis, S. I., Liparas, D., & Initiative, A. s. D. N. (2018). How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database. *Neural regeneration research*, 13(6), 962.
- Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., . . . Mari Aparici, C. (2019). A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. *Radiology*, 290(2), 456-464.
- Dong, W., Yuan, T., Yang, K., Li, C., & Zhang, S. (2017). Autoencoder regularized network for driving style representation learning. *arXiv preprint arXiv:1701.01272*.
- Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., . . . Jicha, G. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology*, 6(8), 734-746.
- Eberhard, D. A., Johnson, B. E., Amler, L. C., Goddard, A. D., Heldens, S. L., Herbst, R. S., . . . Johnson, D. H. (2005). Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non–small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *Journal of clinical oncology*, 23(25), 5900-5909.



- Edvardson, S., Murakami, Y., Nguyen, T. T. M., Shahrour, M., St-Denis, A., Shaag, A., . . . Abu-Libdeh, B. (2017). Mutations in the phosphatidylinositol glycan C (PIGC) gene are associated with epilepsy and intellectual disability. *Journal of medical genetics*, 54(3), 196-201.
- Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., . . . Smith, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *nature*, 562(7726), 210-216.
- Elze, M. C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., . . . Pocock, S. J. (2017). Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology*, 69(3), 345-357.
- Esposito, V., Baldi, A., De Luca, A., Tonini, G., Vincenzi, B., Santini, D., . . . Baldi, F. (2005). Cell cycle related proteins as prognostic parameters in radically resected non-small cell lung cancer. *Journal of clinical pathology*, 58(7), 734-739.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639), 115.
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., . . . Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1), 24.
- Everhart, D. E., Watson, E. M., Bickel, K. L., & Stephenson, A. J. (2015). Right temporal lobe atrophy: a case that initially presented as excessive piety. *The Clinical Neuropsychologist*, 29(7), 1053-1067.

- Ferrarini, L., Palm, W. M., Olofsen, H., van Buchem, M. A., Reiber, J. H., & Admiraal-Behloul, F. (2006). Shape differences of the brain ventricles in Alzheimer's disease. *NeuroImage*, 32(3), 1060-1069.
- Flick, G., Oseki, Y., Kaczmarek, A. R., Al Kaabi, M., Marantz, A., & Pylkkänen, L. (2018). Building words and phrases in the left temporal lobe. *Cortex*, 106, 213-236.
- Fontanals-Cirera, B., Hasson, D., Vardabasso, C., Di Micco, R., Agrawal, P., Chowdhury, A., . . . Wu, P. (2017). Harnessing BET inhibitor sensitivity reveals AMIGO2 as a melanoma survival gene. *Molecular cell*, 68(4), 731-744. e739.
- Freitag, M., & Roy, S. (2018). Unsupervised natural language generation with denoising autoencoders. *arXiv preprint arXiv:1804.07899*.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4), 193-202.
- Fukushima, K. (2013). Training multi-layered neural network neocognitron. *Neural networks*, 40, 18-31.
- Gallo, M., Frangipane, F., Cupidi, C., De Bartolo, M., Turone, S., Ferrari, C., . . . Colao, R. (2017). The novel PSEN1 M84V mutation associated to frontal dysexecutive syndrome, spastic paraparesis, and cerebellar atrophy in a dominant Alzheimer's disease family. *Neurobiology of aging*, 56, 213. e217-213. e212.
- Ghatwary, N., Zolgharni, M., & Ye, X. (2019). Early esophageal adenocarcinoma detection using deep learning methods. *International journal of computer assisted radiology and surgery*, 14(4), 611-621.

- Gliebus, G. (2014). A case report of anxiety disorder preceding frontotemporal dementia with asymmetric right temporal lobe atrophy. *SAGE open medical case reports*, 2, 2050313X13519977.
- Gondara, L. (2016). *Medical image denoising using convolutional denoising autoencoders*. Paper presented at the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW).
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). *Generative adversarial nets*. Paper presented at the Advances in neural information processing systems.
- Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). *A kernel statistical test of independence*. Paper presented at the Advances in neural information processing systems.
- Grillo, L., Greco, D., Pettinato, R., Avola, E., Potenza, N., Castiglia, L., . . . Luciano, D. (2014). Increased FGF3 and FGF4 gene dosage is a risk factor for craniosynostosis. *Gene*, 534(2), 435-439.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). *Improved training of wasserstein gans*. Paper presented at the Advances in Neural Information Processing Systems.
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., . . . Cuadros, J. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22), 2402-2410.

- Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., . . . Enk, A. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8), 1836-1842.
- Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, 363(4), 301-304.
- Han, C., Tomita, H., Ohba, T., Nishizaki, K., Ogata, Y., Matsuzaki, Y., . . . Imaizumi, T. (2016). Modified sympathetic nerve regulation in AKAP5-null mice. *Biochemical and biophysical research communications*, 469(4), 897-902.
- Heo, S.-J., Kim, Y., Yun, S., Lim, S.-S., Kim, J., Nam, C.-M., . . . Yoon, J.-H. (2019). Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual workers' health examination data. *International journal of environmental research and public health*, 16(2), 250.
- Hernán MA, R. J. (2019). *Causal Inference*: Boca Raton: Chapman & Hall/CRC, forthcoming.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217-240.
- Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, 4(5), 5947.
- Hjelm, R. D., Jacob, A. P., Che, T., Trischler, A., Cho, K., & Bengio, Y. (2017). Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*.

- Hong, M. G., Reynolds, C. A., Feldman, A. L., Kallin, M., Lambert, J. C., Amouyel, P., . . . Prince, J. A. (2012). Genome-wide and gene-based association implicates FRMD6 in alzheimer disease. *Human mutation*, 33(3), 521-529.
- Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2), 251-257.
- Hosseini-Asl, E., Keynton, R., & El-Baz, A. (2016). *Alzheimer's disease diagnostics by adaptation of 3D convolutional network*. Paper presented at the 2016 IEEE International Conference on Image Processing (ICIP).
- Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). *Nonlinear causal discovery with additive noise models*. Paper presented at the Advances in neural information processing systems.
- <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/personalized-medicine>.  
from <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/personalized-medicine>
- Hu, Z., Chen, X., Zhao, Y., Tian, T., Jin, G., Shu, Y., . . . Zhang, C. (2010). Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer. *J Clin Oncol*, 28(10), 1721-1726.
- Huang, C.-C., Hsiao, I.-T., Huang, C.-Y., Weng, Y.-C., Huang, K.-L., Liu, C.-H., . . . Lin, K.-J. (2019). Tau PET with 18F-THK-5351 Taiwan Patients with Familial Alzheimer's Disease with the APP p. D678H Mutation. *Frontiers in neurology*, 10, 503.

- Huang, C.-Y., Hsiao, T., Lin, K.-J., Huang, K.-L., Fung, H.-C., Liu, C.-H., . . . Yen, T.-C. (2019). Amyloid PET pattern with dementia and amyloid angiopathy in Taiwan familial AD with D678H APP mutation. *Journal of the neurological sciences*, 398, 107-116.
- Hui, J. (2018). GAN — Why it is so hard to train Generative Adversarial Networks! Retrieved from [https://medium.com/@jonathan\\_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b](https://medium.com/@jonathan_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b)
- Hytinen, A., Hoyer, P. O., Eberhardt, F., & Jarvisalo, M. (2013). Discovering cyclic causal models with latent variables: A general SAT-based procedure. *arXiv preprint arXiv:1309.6836*.
- Ikehara, M., Oshita, F., Ito, H., Ohgane, N., Suzuki, R., Saito, H., . . . Kameda, Y. (2003). Expression of cyclin D1 but not of cyclin E is an indicator of poor prognosis in small adenocarcinomas of the lung. *Oncology reports*, 10(1), 137-139.
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.
- Jahanshad, N., Kochunov, P. V., Sprooten, E., Mandl, R. C., Nichols, T. E., Almasy, L., . . . de Zubicaray, G. I. (2013). Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA–DTI working group. *NeuroImage*, 81, 455-469.
- Jin, M., Inoue, S., Umemura, T., Moriya, J., Arakawa, M., Nagashima, K., & Kato, H. (2001). Cyclin D1, p16 and retinoblastoma gene product expression as a predictor for

- prognosis in non-small cell lung cancer at stages I and II. *Lung cancer*, 34(2), 207-218.
- Johansson, F., Shalit, U., & Sontag, D. (2016). *Learning representations for counterfactual inference*. Paper presented at the International conference on machine learning.
- Jordan, J. (2018). Variational autoencoders.
- Ju, R., Hu, C., & Li, Q. (2017). Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(1), 244-257.
- Kakeda, S., & Korogi, Y. (2010). The efficacy of a voxel-based morphometry on the analysis of imaging in schizophrenia, temporal lobe epilepsy, and Alzheimer's disease/mild cognitive impairment: a review. *Neuroradiology*, 52(8), 711-721.
- Kanda, Y., Osaki, M., Onuma, K., Sonoda, A., Kobayashi, M., Hamada, J., . . . Okada, F. (2017). Amigo2-upregulation in tumour cells facilitates their attachment to liver endothelial cells resulting in liver metastases. *Scientific reports*, 7(1), 1-13.
- Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4), 523-539.
- Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., . . . Yan, F. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122-1131. e1129.
- Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*.

- Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kovacs, M. D., Burchett, P. F., & Sheafor, D. H. (2018). App review: management guide for incidental findings on CT and MRI. *Journal of digital imaging*, 31(2), 154-158.
- Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4), 661-687.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. Paper presented at the Advances in neural information processing systems.
- Kuramoto, K., Negishi, M., & Katoh, H. (2009). Regulation of dendrite growth by the Cdc42 activator Zizimin1/Dock9 in hippocampal neurons. *Journal of neuroscience research*, 87(8), 1794-1805.
- Ladefoged, C. N., Marner, L., Hindsholm, A., Law, I., Højgaard, L., & Andersen, F. L. (2019). Deep learning based attenuation correction of PET/MRI in pediatric brain tumor patients: evaluation in a clinical setting. *Frontiers in neuroscience*, 12, 1005.
- Leandrou, S., Petroudi, S., Kyriacou, P. A., Reyes-Aldasoro, C. C., & Pattichis, C. S. (2018). Quantitative MRI Brain Studies in Mild Cognitive Impairment and Alzheimer's Disease: A Methodological Review. *IEEE reviews in biomedical engineering*, 11, 97-111.



- LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). *Handwritten digit recognition with a back-propagation network*. Paper presented at the Advances in neural information processing systems.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- LeCun, Y., Haffner, P., Bottou, L., & Bengio, Y. (1999). Object recognition with gradient-based learning *Shape, contour and grouping in computer vision* (pp. 319-345): Springer.
- Lee, J.-Y., Hur, H., Yun, H. J., Kim, Y., Yang, S., Kim, S. I., & Kim, M. H. (2015). HOXB5 promotes the proliferation and invasion of breast cancer cells. *International journal of biological sciences*, 11(6), 701.
- Lee, K., Chang, J. W., Oh, C., Liu, L., Jung, S.-N., Won, H.-R., . . . Koo, B. S. (2019). HOXB5 acts as an oncogenic driver in head and neck squamous cell carcinoma via EGFR/Akt/Wnt/ $\beta$ -catenin signaling axis. *European Journal of Surgical Oncology*.
- Lenz, D., McClean, P., Kansu, A., Bonnen, P. E., Ranucci, G., Thiel, C., . . . Dimitrov, B. (2018). SCYL1 variants cause a syndrome with low  $\gamma$ -glutamyl-transferase cholestasis, acute liver failure, and neurodegeneration (CALFAN). *Genetics in Medicine*, 20(10), 1255-1265.
- Li, B.-Y., & Chen, S.-D. (2015). Potential similarities in temporal lobe epilepsy and Alzheimer's Disease: from clinic to pathology. *American Journal of Alzheimer's Disease & Other Dementias®*, 30(8), 723-728.

- Ling, B., Liao, X., Huang, Y., Liang, L., Jiang, Y., Pang, Y., & Qi, G. (2020). Identification of prognostic markers of lung cancer through bioinformatics analysis and in vitro experiments. *International Journal of Oncology*, 56(1), 193-205.
- Liu, J., Zhang, C., Wang, X., Ly, P., Belyi, V., Xu-Monette, Z., . . . Feng, Z. (2014). E3 ubiquitin ligase TRIM32 negatively regulates tumor suppressor p53 to promote tumorigenesis. *Cell Death & Differentiation*, 21(11), 1792-1804.
- Liu, M.-Y., & Tuzel, O. (2016). *Coupled generative adversarial networks*. Paper presented at the Advances in neural information processing systems.
- Liu, X., Chen, K., Wu, T., Weidman, D., Lure, F., & Li, J. (2018). Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Translational Research*, 194, 56-67.
- Liu, X., Hou, D., Lin, F., Luo, J., Xie, J., Wang, Y., & Tian, Y. (2018). The role of neurovascular unit damage in the occurrence and development of Alzheimer's disease. *Reviews in the neurosciences*.
- Lopez-Paz, D., & Oquab, M. (2016). Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.
- Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., Ourselin, S., & Initiative, A. s. D. N. (2017). Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. *NeuroImage*.

- Lu, M., Sadiq, S., Feaster, D. J., & Ishwaran, H. (2018). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1), 209-219.
- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19), 2937-2960.
- Machinery, C. (1950). Computing machinery and intelligence-AM Turing. *Mind*, 59(236), 433.
- Manley, W., Moreau, M. P., Azaro, M., Siecinski, S. K., Davis, G., Buyske, S., . . . Brzustowicz, L. (2018). Validation of a microRNA target site polymorphism in H3F3B that is potentially associated with a broad schizophrenia phenotype. *PloS one*, 13(3).
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115-133.
- Menéndez-González, M., de Celis Alonso, B., Salas-Pacheco, J., & Arias-Carrión, O. (2015). Structural neuroimaging of the medial temporal lobe in Alzheimer's Disease clinical trials. *Journal of Alzheimer's Disease*, 48(3), 581-589.
- Mez, J., Chung, J., Jun, G., Kriegel, J., Bourlas, A. P., Sherva, R., . . . Buxbaum, J. D. (2017). Two novel loci, COBL and SLC10A2, for Alzheimer's disease in African Americans. *Alzheimer's & Dementia*, 13(2), 119-129.

- Miles, L. A., Hermans, S. J., Crespi, G. A., Gooi, J. H., Doughty, L., Nero, T. L., . . .
- Oehlrich, D. (2019). Small molecule binding to Alzheimer risk factor CD33 promotes A $\beta$  phagocytosis. *iScience*, 19, 110-118.
- Minsky, M., & Papert, S. A. (2017). *Perceptrons: An introduction to computational geometry*: MIT press.
- Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Odena, A., Olah, C., & Shlens, J. (2017). *Conditional image synthesis with auxiliary classifier gans*. Paper presented at the Proceedings of the 34th International Conference on Machine Learning-Volume 70.
- Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, 2(11), e7.
- Pasquini, L., Rahmani, F., Maleki-Balajoo, S., La Joie, R., Zarei, M., Sorg, C., . . .
- Tahmasian, M. (2019). Medial Temporal Lobe Disconnection and Hyperexcitability Across Alzheimer's Disease Stages. *Journal of Alzheimer's disease reports*, 3(1), 103-112.
- Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506*.
- Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3, 96-146.
- Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54-60.

- Peters, J., Janzing, D., & Schölkopf, B. (2010). *Identifying cause and effect on discrete data using additive noise models*. Paper presented at the Proceedings of the thirteenth international conference on artificial intelligence and statistics.
- Pfister, N., Bühlmann, P., Schölkopf, B., & Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 5-31.
- Porter, K. E., Gruber, S., Van Der Laan, M. J., & Sekhon, J. S. (2011). The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, 7(1), 1-34.
- R., J. K. S. A. K. A. J. (2013). Differential response to targeted therapy in non-small cell lung carcinoma patients harbouring epidermal growth factor receptor mutations—a demand for diagnostic procedure optimization—a critical review. *OA Molecular Oncology*, 1(1), 2.
- Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Raponi, M., Dossey, L., Jatkoe, T., Wu, X., Chen, G., Fan, H., & Beer, D. G. (2009). MicroRNA classifiers for predicting prognosis of squamous cell lung cancer. *Cancer research*, 69(14), 5776-5783.
- Ravizza, S., Huschto, T., Adamov, A., Böhm, L., Büsser, A., Flöther, F. F., . . . Robertson, D. H. (2019). Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nature medicine*, 25(1), 57-59.

- Reuben, A., Spencer, C. N., Prieto, P. A., Gopalakrishnan, V., Reddy, S. M., Miller, J. P., . . . Song, X. (2017). Genomic and immune heterogeneity are associated with differential responses to therapy in melanoma. *NPJ genomic medicine*, 2(1), 10.
- Riely, G. J., Marks, J., & Pao, W. (2009). KRAS mutations in non-small cell lung cancer. *Proceedings of the American Thoracic Society*, 6(2), 201-205.
- Rolling, C. A., & Yang, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4), 749-769.
- Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American statistical Association*, 82(398), 387-394.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387), 516-524.
- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533.
- Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Frontiers in aging neuroscience*, 9, 329.

- Sarraf, S., & Tofighi, G. (2016). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *BioRxiv*, 070441.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.
- Schmidt, W. M., Kraus, C., Höger, H., Hochmeister, S., Oberndorfer, F., Branka, M., . . . Macedo-Souza, L. I. (2007). Mutation in the Scyl1 gene encoding amino-terminal kinase-like protein causes a recessive form of spinocerebellar neurodegeneration. *EMBO reports*, 8(7), 691-697.
- Schmidt, W. M., Rutledge, S. L., Schüle, R., Mayerhofer, B., Züchner, S., Boltshauser, E., & Bittner, R. E. (2015). Disruptive SCYL1 mutations underlie a syndrome characterized by recurrent episodes of liver failure, peripheral neuropathy, cerebellar atrophy, and ataxia. *The American Journal of Human Genetics*, 97(6), 855-861.
- Schuler, A., Jung, K., Tibshirani, R., Hastie, T., & Shah, N. (2017). Synth-validation: Selecting the best causal inference method for a given dataset. *arXiv preprint arXiv:1711.00083*.
- Shalit, U., Johansson, F. D., & Sontag, D. (2017). *Estimating individual treatment effect: generalization bounds and algorithms*. Paper presented at the Proceedings of the 34th International Conference on Machine Learning-Volume 70.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

- Spasov, S., Passamonti, L., Duggento, A., Liò, P., Toschi, N., & Initiative, A. s. D. N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage*, 189, 276-287.
- Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1), 62-72.
- Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*: MIT press.
- Spirtes, P., & Zhang, K. (2016). *Causal discovery and inference: concepts and recent methodological advances*. Paper presented at the Applied informatics.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929-1958.
- Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). *Unsupervised learning of video representations using lstms*. Paper presented at the International conference on machine learning.
- Stanford. (2019). CS231n: Convolutional Neural Networks for Visual Recognition., from <http://cs231n.github.io/understanding-cnn/>
- Struyfs, H., Van Hecke, W., Veraart, J., Sijbers, J., Slaets, S., De Belder, M., . . . Robberecht, C. (2015). Diffusion kurtosis imaging: a possible MRI biomarker for AD diagnosis? *Journal of Alzheimer's Disease*, 48(4), 937-948.



- Sun, R., Wang, R., Chang, S., Li, K., Sun, R., Wang, M., & Li, Z. (2019). Long non-coding RNA in drug resistance of non-small cell lung cancer: A Mini Review. *Frontiers in Pharmacology*, 10.
- Tan, D. S., Chong, F. T., Leong, H. S., Toh, S. Y., Lau, D. P., Kwang, X. L., . . . Chang, M. M. (2017). Long noncoding RNA EGFR-AS1 mediates epidermal growth factor receptor addiction and modulates treatment response in squamous cell carcinoma. *Nature medicine*, 23(10), 1167.
- Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 26-31.
- Triantafillou, S., & Tsamardinos, I. (2015). Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16, 2147-2205.
- Trimmel, K., van Graan, A. L., Caciagli, L., Haag, A., Koepp, M. J., Thompson, P. J., & Duncan, J. S. (2018). Left temporal lobe language network connectivity in temporal lobe epilepsy. *Brain*, 141(8), 2406-2418.
- Van Giau, V., Senanarong, V., Bagyinszky, E., Limwongse, C., An, S. S. A., & Kim, S. (2018). Identification of a novel mutation in APP gene in a Thai subject with early-onset Alzheimer's disease. *Neuropsychiatric disease and treatment*, 14, 3015.
- Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). *Extracting and composing robust features with denoising autoencoders*. Paper presented at the Proceedings of the 25th international conference on Machine learning.

- Wada, A., Tsuruta, K., Irie, R., Kamagata, K., Maekawa, T., Fujita, S., . . . Nakanishi, A. (2019). Differentiating Alzheimer's disease from dementia with Lewy bodies using a deep learning technique based on structural brain connectivity. *Magnetic Resonance in Medical Sciences*, 18(3), 219.
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American statistical Association*, 113(523), 1228-1242.
- Waldrop, M. M. (2019). News Feature: What are the limits of deep learning? *Proceedings of the National Academy of Sciences*, 116(4), 1074-1077.
- Wang, Y.-J., Wan, Y., Wang, H.-F., Tan, C.-C., Li, J.-Q., Yu, J.-T., . . . Initiative, A. s. D. N. (2019). Effects of CD33 variants on neuroimaging biomarkers in non-demented elders. *Journal of Alzheimer's Disease*, 68(2), 757-766.
- Wolk, D. A., Das, S. R., Mueller, S. G., Weiner, M. W., Yushkevich, P. A., & Initiative, A. s. D. N. (2017). Medial temporal lobe subregional morphometry using high resolution MRI in Alzheimer's disease. *Neurobiology of aging*, 49, 204-213.
- Xiong, M. (2018). *Big Data in Omics and Imaging: Integrated Analysis and Causal Inference*: Chapman and Hall/CRC.
- Xiong, P., Wang, H., Liu, M., Zhou, S., Hou, Z., & Liu, X. (2016). ECG signal enhancement based on improved denoising auto-encoder. *Engineering Applications of Artificial Intelligence*, 52, 194-202.
- Xu, Y.-H., Tu, J.-R., Zhao, T.-T., Xie, S.-G., & Tang, S.-B. (2019). Overexpression of lncRNA EGFR-AS1 is associated with a poor prognosis and promotes chemotherapy

- resistance in non-small cell lung cancer. *International Journal of Oncology*, 54(1), 295-305.
- Yoon, J., Jordon, J., & van der Schaar, M. (2018a). Gain: Missing data imputation using generative adversarial nets. *arXiv preprint arXiv:1806.02920*.
- Yoon, J., Jordon, J., & van der Schaar, M. (2018b). GANITE: Estimation of individualized treatment effects using generative adversarial nets.
- Zeiler, M. D., & Fergus, R. (2014). *Visualizing and understanding convolutional networks*. Paper presented at the European conference on computer vision.
- Zhang, B., Li, N., & Zhang, H. (2018). Knockdown of homeobox B5 (HOXB5) inhibits cell proliferation, migration, and invasion in non-small cell lung cancer cells through inactivation of the Wnt/ $\beta$ -catenin pathway. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 26(1), 37-44.
- Zhang, K., & Hyvärinen, A. (2009a). *Causality discovery with additive disturbances: An information-theoretical perspective*. Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.
- Zhang, K., & Hyvärinen, A. (2009b). *On the identifiability of the post-nonlinear causal model*. Paper presented at the Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence.
- Zhang, W., Le, T. D., Liu, L., Zhou, Z.-H., & Li, J. (2017). Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15), 2372-2378.

- Zhao, T.-T., Jin, F., Li, J.-G., Xu, Y.-Y., Dong, H.-T., Liu, Q., . . . Yin, S.-C. (2018). TRIM32 promotes proliferation and confers chemoresistance to breast cancer cells through activation of the NF- $\kappa$ B pathway. *Journal of Cancer*, 9(8), 1349.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). *Learning deep features for discriminative localization*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Zhou, Z.-d., Chan, C. H.-s., Ma, Q.-h., Xu, X.-h., Xiao, Z.-c., & Tan, E.-K. (2011). The roles of amyloid precursor protein (APP) in neurogenesis: Implications to pathogenesis and therapy of Alzheimer disease. *Cell adhesion & migration*, 5(4), 280-292.
- Zhuang, Q.-S., Zheng, H., Gu, X.-D., Shen, L., & Ji, H.-F. (2017). Detecting the genetic link between Alzheimer's disease and obesity using bioinformatics analysis of GWAS data. *Oncotarget*, 8(34), 55915.
- Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*.

## APPENDICES

Table S1. P-values of 43 genes that showed causation or association or both with the frontal, left temporal lobe.

Baseline				
Gene	P-Value			
	Causal		Association	
	Temporal_L	Frontal & Temp_L	Temporal_L	Frontal & Temp_L
FGF4	6.8E-05	1.6E-04		
FOLR2	4.9E-03	5.7E-03		
RNU6.1272P	3.9E-03	2.1E-03	1.9E-02	4.4E-02
AL021394.1		3.6E-02		
RP11.953B20.2	6.6E-03	5.1E-03		
ANKZF1	5.0E-04	8.4E-04		
RNU6.883P	1.0E-03	1.5E-04		
APP		2.2E-02		
CD33	7.2E-04			
FRMD6	3.8E-05	2.9E-04		4.6E-02
LIPE.AS1	3.7E-03	1.2E-04		
RP11.749H17.2	2.7E-02			
RP11.749H17.1				
DOCK9.AS2				
H3F3B	5.1E-04	4.6E-03		
DDX52	3.0E-02			
SCYL1	1.8E-02			
RPL35P9	6.5E-03	5.9E-04		
RN7SKP298	2.2E-02			4.1E-02
RP11.394D2.1		5.3E-03		9.6E-03
PRM1	1.1E-03	2.5E-05		
AP000769.8	6.9E-04	2.8E-04		
RN7SL146P	7.5E-05	1.8E-02		
CTD.3222D19.7	2.9E-04			
BTBD7P2	1.3E-02	1.6E-02		
RABGGTA	1.2E-02	7.5E-03		
SPATA21	9.4E-04			
GS1.410F4.5	1.5E-02	4.6E-02	8.4E-03	

RP11.732M18.2	2.5E-03	3.0E-02	3.3E-04	
CARS				
PKDCC			1.7E-02	3.0E-02
ZBTB1			1.1E-03	
AKAP5	2.9E-02		6.9E-03	
RP11.15F12.1	8.2E-03	1.3E-03		
FAM162B	1.8E-04	1.2E-03		
YTHDC2	2.2E-03	4.7E-03	2.5E-03	2.8E-05
LINC00926			4.7E-03	4.8E-03
CCDC179	1.0E-04	6.9E-04		
RP11.84C13.2			2.4E-02	8.5E-03
COBL	2.6E-05	2.6E-04		
RP11.443B20.1			1.3E-03	3.8E-02
NBR1			3.0E-02	
PIGC			3.0E-02	

6 Months				
Gene	P-Value			
	Causal		Association	
	Temporal_L	Frontal & Temp_L	Temporal_L	Frontal & Temp_L
FGF4	4.5E-05	1.0E-04		
FOLR2	1.0E-02	6.8E-03		
RNU6.1272P	1.8E-04	1.3E-03	3.1E-02	
AL021394.1	3.1E-02			
RP11.953B20.2	6.6E-03	1.2E-02		
ANKZF1	4.6E-04	8.4E-04		
RNU6.883P	4.1E-04	1.7E-04		
APP		1.8E-02		
CD33	3.0E-04	4.9E-02		
FRMD6	4.8E-05	3.2E-04		3.8E-02
LIPE.AS1	5.5E-04	1.1E-03		
RP11.749H17.2		3.3E-02		
RP11.749H17.1		3.1E-02		
DOCK9.AS2	8.1E-03	5.7E-03		
H3F3B	5.1E-04	6.7E-03		
DDX52	1.8E-02	4.8E-02		
SCYL1	8.9E-03	4.4E-02		
RPL35P9	1.1E-03	1.8E-04		

RN7SKP298				
RP11.394D2.1		5.2E-03		1.8E-02
PRM1	6.7E-05	8.6E-06		
AP000769.8	1.0E-03	6.9E-04		
RN7SL146P	1.6E-04	1.7E-03		
CTD.3222D19.7	2.4E-03	1.6E-02		
BTBD7P2	1.2E-02	4.6E-02		
RABGGTA	7.4E-03	4.7E-03		
SPATA21	8.9E-04	3.3E-02		
GS1.410F4.5	8.6E-03	3.4E-02	2.9E-02	
RP11.732M18.2	2.5E-03	3.0E-02	3.1E-04	
CARS	2.5E-02			
PKDCC			9.0E-03	3.4E-02
ZBTB1			1.4E-03	
AKAP5			1.2E-02	
RP11.15F12.1		5.1E-04		
FAM162B	5.5E-04	4.2E-03		
YTHDC2	3.7E-05	1.9E-03	1.6E-03	2.1E-05
LINC00926			5.6E-03	5.4E-03
CCDC179	4.0E-05	3.8E-04		
RP11.84C13.2			9.7E-03	8.3E-03
COBL	7.3E-05	1.0E-04		
RP11.443B20.1			4.1E-03	5.0E-02
NBR1			1.5E-02	
PIGC		3.8E-02	4.0E-02	4.1E-02

12 Months				
Gene	P-Value			
	Causal		Association	
	Temporal_L	Frontal & Temp_L	Temporal_L	Frontal & Temp_L
FGF4	2.5E-05	1.0E-04		
FOLR2	6.4E-03	8.5E-03		
RNU6.1272P	1.8E-04	1.3E-03	3.5E-02	
AL021394.1	3.8E-02			
RP11.953B20.2	3.6E-03	1.7E-02		
ANKZF1	6.6E-04	1.4E-03		
RNU6.883P	4.1E-04	2.7E-04		

APP		4.1E-03		
CD33	5.5E-04	4.9E-02		
FRMD6	1.3E-04	3.2E-04		3.9E-02
LIPE.AS1	5.5E-04	1.1E-03		
RP11.749H17.2		4.9E-02		
RP11.749H17.1	3.1E-02	1.6E-02		
DOCK9.AS2	1.0E-03	5.7E-03		
H3F3B	5.1E-04	6.7E-03		
DDX52	2.8E-02			
SCYL1	4.2E-03	3.6E-02		
RPL35P9	3.0E-03	3.2E-04		
RN7SKP298				
RP11.394D2.1		2.5E-03		1.8E-02
PRM1	6.7E-05	8.6E-06		
AP000769.8	6.9E-04	6.9E-04		
RN7SL146P	3.0E-04	1.7E-03		
CTD.3222D19.7	2.4E-03	2.3E-02		
BTBD7P2	1.2E-02	4.6E-02		
RABGGTA	7.1E-03	2.9E-03		
SPATA21	1.6E-03			
GS1.410F4.5	8.6E-03	3.4E-02	3.4E-02	
RP11.732M18.2	8.2E-03	3.0E-02	3.7E-04	
CARS				
PKDCC			8.0E-03	3.5E-02
ZBTB1			1.2E-03	
AKAP5			1.1E-02	
RP11.15F12.1	3.0E-02	1.3E-03		
FAM162B	7.2E-04	6.4E-03		
YTHDC2	5.4E-05	6.9E-04	1.2E-03	2.1E-05
LINC00926			5.2E-03	4.9E-03
CCDC179	2.7E-05	3.8E-04		
RP11.84C13.2			8.4E-03	8.7E-03
COBL	1.8E-05	6.4E-04		
RP11.443B20.1			4.7E-03	4.9E-02
NBR1			1.3E-02	
PIGC			4.4E-02	4.4E-02

24 Months



Gene	P-Value			
	Causal		Association	
	Temporal_L	Frontal & Temp_L	Temporal_L	Frontal & Temp_L
FGF4	5.7E-05	1.0E-04		
FOLR2	2.6E-02	6.8E-03		
RNU6.1272P	3.7E-05	1.3E-03		
AL021394.1				
RP11.953B20.2	2.3E-03	1.7E-02		
ANKZF1	6.6E-04	8.4E-04		
RNU6.883P	2.0E-04	4.3E-04		
APP		4.1E-03		
CD33	2.0E-02	4.9E-02		
FRMD6	7.9E-04	5.1E-04		3.9E-02
LIPE.AS1	2.0E-04	5.7E-04		
RP11.749H17.2		3.4E-02		
RP11.749H17.1		3.1E-02		
DOCK9.AS2	7.4E-03	5.7E-03		
H3F3B	7.2E-04	4.6E-03		
DDX52	2.2E-02	3.3E-02		
SCYL1	5.9E-03	3.6E-02		
RPL35P9	6.0E-03	3.2E-04		
RN7SKP298	2.1E-02			
RP11.394D2.1		3.6E-03		1.8E-02
PRM1	6.6E-05	8.6E-06		
AP000769.8	3.7E-04	6.9E-04		
RN7SL146P	1.5E-03	2.1E-03		
CTD.3222D19.7	9.2E-04	2.6E-02		
BTBD7P2	1.0E-02	4.6E-02		
RABGGTA	7.2E-03	2.9E-03	4.8E-02	
SPATA21	2.0E-03			
GS1.410F4.5	1.3E-02	2.3E-02		
RP11.732M18.2	7.7E-04	1.8E-02	8.4E-04	
CARS		4.3E-02		
PKDCC			4.6E-03	3.5E-02
ZBTB1			7.7E-04	
AKAP5			1.0E-02	
RP11.15F12.1		1.3E-03		
FAM162B	1.3E-03	1.1E-02		
YTHDC2	2.0E-05	3.0E-03	4.0E-04	2.0E-05

LINC00926			3.8E-03	5.3E-03
CCDC179	8.9E-04	3.8E-04		
RP11.84C13.2			6.7E-03	8.6E-03
COBL	5.0E-05	6.4E-04		
RP11.443B20.1			7.8E-03	
NBR1			6.1E-03	
PIGC				4.4E-02

Table S2. P-values of 46 genes that showed causation or association or both with the right temporal lobe region.

	Baseline		6 Months		
Gene	P-Value		Gene	P-Value	
	Causation	Association		Causation	Association
FGF4	<b>5.3E-05</b>		FGF4	<b>2.9E-05</b>	
FOLR2	2.3E-03		FOLR2	7.8E-03	
RNU6.1272P	<b>1.6E-04</b>	1.2E-02	RNU6.1272P	<b>8.9E-05</b>	1.4E-02
AL021394.1			AL021394.1		
RP11.953B20.2	5.1E-03		RP11.953B20.2	6.1E-03	
ANKZF1	<b>1.4E-04</b>		ANKZF1	<b>2.1E-04</b>	
RNU6.883P	9.2E-04		RNU6.883P	<b>1.5E-04</b>	
CD33	2.1E-02		CD33	4.3E-03	
FRMD6	<b>2.3E-05</b>		FRMD6	<b>5.8E-06</b>	
LIPE.AS1	<b>2.0E-04</b>		LIPE.AS1	<b>2.0E-04</b>	
AKAP9	3.4E-02		AKAP9		
RP11.749H17.2			RP11.749H17.2	1.7E-02	
RP11.749H17.1	1.2E-02		RP11.749H17.1	3.5E-02	
DOCK9.AS2			DOCK9.AS2	2.9E-02	
H3F3B	1.5E-02		H3F3B	2.4E-03	
DDX52	1.9E-02		DDX52		
SCYL1	2.0E-02		SCYL1	1.3E-02	
RPL35P9	2.7E-03		RPL35P9	2.9E-02	
PRM1	<b>1.5E-05</b>		PRM1	7.1E-05	
AP000769.8	<b>2.2E-04</b>		AP000769.8	1.4E-02	
RN7SL146P	1.7E-03		RN7SL146P	3.1E-03	
CTD.3222D19.7	4.1E-03		CTD.3222D19.7	2.2E-02	
BTBD7P2	2.8E-02		BTBD7P2	3.3E-03	
RABGGTA	1.7E-03		RABGGTA	1.1E-03	3.1E-02
SPATA21	3.0E-02		SPATA21	8.6E-03	
GS1.410F4.5	3.0E-02	7.1E-03	GS1.410F4.5	3.9E-02	1.4E-02
RP11.732M18.2	1.2E-03	3.2E-03	RP11.732M18.2	7.5E-03	3.2E-03
SAA2		4.0E-02	SAA2		
PKDCC		2.1E-02	PKDCC		1.1E-02
ZBTB1		5.1E-03	ZBTB1	4.2E-02	5.8E-03

SLC9B2	5.4E-03		SLC9B2	7.1E-03	
AKAP5	3.8E-02	1.3E-02	AKAP5		1.9E-02
CTD.2176I21.2			CTD.2176I21.2	3.9E-02	
RP11.15F12.1	1.0E-02		RP11.15F12.1	3.1E-03	
AC079896.1		3.4E-02	AC079896.1		4.1E-02
FAM162B	6.4E-04	3.5E-02	FAM162B	1.6E-03	
YTHDC2	2.1E-03	3.6E-04	YTHDC2	1.3E-03	1.4E-04
LINC00926		4.8E-04	LINC00926		3.7E-04
CCDC179	1.0E-04		CCDC179	1.8E-05	
RP11.84C13.2		3.2E-03	RP11.84C13.2		9.8E-04
HARS	4.0E-02		HARS		
COBL	6.0E-05		COBL	2.2E-05	
RP11.443B20.1		7.0E-04	RP11.443B20.1		1.4E-03
NBR1	3.4E-02	5.7E-03	NBR1		1.2E-03
PIGC		3.0E-02	PIGC		2.3E-02

12 Months			24 Months		
Gene	P-Value		Gene	P-Value	
	Causation	Association		Causation	Association
FGF4	2.9E-05		FGF4	1.6E-05	
FOLR2	2.9E-02		FOLR2	7.8E-03	
RNU6.1272P	8.9E-05	1.4E-02	RNU6.1272P	8.9E-05	1.3E-02
AL021394.1			AL021394.1	3.2E-02	
RP11.953B20.2	6.1E-03		RP11.953B20.2	6.1E-03	
ANKZF1	1.2E-04		ANKZF1	1.2E-04	
RNU6.883P	1.5E-04		RNU6.883P	2.3E-04	
CD33	4.3E-03		CD33	4.3E-03	
FRMD6	5.8E-06		FRMD6	8.8E-06	
LIPE.AS1	2.0E-04		LIPE.AS1	2.0E-04	
AKAP9			AKAP9		
RP11.749H17.2	1.7E-02		RP11.749H17.2	1.7E-02	
RP11.749H17.1			RP11.749H17.1	1.9E-02	
DOCK9.AS2	1.4E-02		DOCK9.AS2	6.5E-03	
H3F3B	6.1E-03		H3F3B	9.2E-03	
DDX52	3.8E-02		DDX52		

SCYL1	3.5E-03		SCYL1	3.2E-03	
RPL35P9	1.9E-02		RPL35P9	1.7E-02	
PRM1	4.1E-05		PRM1	4.1E-05	
AP000769.8	9.3E-03		AP000769.8	1.4E-02	
RN7SL146P	3.1E-03		RN7SL146P	3.1E-03	
CTD.3222D19.7	2.2E-02		CTD.3222D19.7	1.4E-02	
BTBD7P2	3.3E-03		BTBD7P2	5.5E-03	
RABGGTA	1.1E-03	3.0E-02	RABGGTA	6.3E-04	3.0E-02
SPATA21	3.3E-02		SPATA21	3.3E-02	
GS1.410F4.5		1.4E-02	GS1.410F4.5	3.9E-02	1.4E-02
RP11.732M18.2	7.5E-03	3.5E-03	RP11.732M18.2	7.5E-03	3.2E-03
SAA2			SAA2		
PKDCC		1.1E-02	PKDCC		1.1E-02
ZBTB1	4.2E-02	5.9E-03	ZBTB1	4.2E-02	5.9E-03
SLC9B2	1.3E-03		SLC9B2	7.1E-03	
AKAP5		1.9E-02	AKAP5		2.0E-02
CTD.2176I21.2	3.9E-02		CTD.2176I21.2		
RP11.15F12.1	8.0E-03		RP11.15F12.1	1.0E-02	
AC079896.1		4.0E-02	AC079896.1		4.2E-02
FAM162B	1.6E-03		FAM162B	1.6E-03	
YTHDC2	1.3E-03	1.5E-04	YTHDC2	1.3E-03	1.6E-04
LINC00926		3.7E-04	LINC00926		3.8E-04
CCDC179	1.8E-05		CCDC179	1.8E-05	
RP11.84C13.2		9.9E-04	RP11.84C13.2		1.0E-03
HARS			HARS		
COBL	2.2E-05		COBL	1.3E-04	
RP11.443B20.1		1.4E-03	RP11.443B20.1		1.6E-03
NBR1	3.2E-02	1.2E-03	NBR1	4.7E-02	1.1E-03
PIGC		2.3E-02	PIGC		2.4E-02