Research on Machine Learning for Biomedical Research

ZHAO, Kai

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
School of Biomedical Sciences

Supervised by

Prof. So Hon-cheong

The Chinese University of Hong Kong
July 2020

Thesis Assessment Committee

Professor Cheng Sze Lok Alfred (Chair)

Professor So Hon-cheong (Thesis Supervisor)

Professor Chen Yangchao (Committee Member)

Professor Sham Pak Chung (External Examiner)

Abstract of thesis entitled:
Research on Machine Learning for Biomedical Research
Submitted by ZHAO, Kai
for the degree of Doctor of Philosophy
at The Chinese University of Hong Kong in July 2020

This is the abstract in no more than 350 words in English

Acknowledgement

I would like to thank my supervisor Prof. So Hon-cheong for offering me the opportunity to work with him. He teaches me how to conduct research works and gave me generous help in my daily life. He is humorous and highly empathic. Truth to be told, it is a wonderful journey to study and work with him, and I am lucky to have the experience.

I also would like to thank my wife. It's nearly ten years since the first meet, and she has become the most important person in my life. Thanks for supporting my decision to pursue a higher degree and bearing all burden from family to free me from distractions. This is not easy to her. You are a brave girl and wonderful mother for the kids. My any achievement is impossible without you.

Thanks my father and mother for never saying NO to the decision of further my education and for taking care of my kids in the past years. I know the hardness for you to bear the burden. I highly appreciate the support.

Thanks my kids for tolerating my absence of parental responsibility for the past year. You let me be your father and share tremendous joy with me. I promise my love to you will alway be the same.

Thanks my lab mates for having some awesome years with you. You are a part of my daily life in those years. The times we spent together will be a precious memory.

Thanks everybody who helped me for their kindness!

Dedicated to those who risked their life to fight against COVID-19.

Contents

List of Figures

List of Tables

Symbols and Acronyms

In general, we denote a scalar by an italic lower case letter, a vector by a roman lower case bold letter, and a matrix by a roman upper case bold letter respectively, e.g., $a \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{M} \in \mathbb{R}^{p \times q}$, with any exceptions to be mentioned in the context case by case.

An identity matrix is written as $\mathbf{I}$. Specifically, an $n \times n$ identity matrix is written as $\mathbf{I}_n$. A zero matrix or vector is written as $\mathbf{0}$. Specifically, an $m \times n$ zero matrix is written as $\mathbf{0}_{m \times n}$.

Specialized symbols and major acronyms are defined as follows:

$p(\cdot)$	the probability density function (PDF)
$\Pr(\cdot)$	the probability value
$\mathbb{E}(\cdot)$	the expectation
$\mathbf{\Sigma}$	a covariance matrix
$\mathbf{N}(\mu, \mathbf{\Sigma})$	a normal distribution with mean μ and covariance $\mathbf{\Sigma}$
ε	a noise vector
$\mathbf{e}(\cdot)$	an error/residual function
$\mathbf{H}$	Hessian matrix
$\mathrm{tr}(\cdot)$	trace of a matrix
$\det(\cdot)$	determinant of a matrix

DNN	deep neuron network
GBM	gradient boosting machine
SVM	support vector machine
CF	causal forests
RF	risk factors
ML	machine learning
EN	elastic net
ITE	individual treatment effects
TE	treatment effects
tx	treatment
CM	cardiometabolic
GWAS	genome-wide association studies
CNV	copy number variation
SML	supervised machine learning
MCMC	Markov chain Monte Carlo
GRF	general causal forests
CF	causal forests
CV	cross validation
MSE	mean squared error
CI	confidence interval

Chapter 1

Introduction

□ **End of chapter.**

Chapter 2

Background Study

□ **End of chapter.**

Chapter 3

Drug Repurposing

3.1 Background

3.1.1 Motivation

Development of new medications is a very lengthy and costly process. While investment in research and development has been increasing, there is a lack of proportional rise in the number of drugs approved in the past two decades, especially for drugs with novel mechanisms of actions [84]. There is an urgent need for innovative approaches to improve the productivity of drug development. This is particularly true for some areas like psychiatry, for which there has been lack of therapeutic advances for some time [61, 52].

Finding new indications for existing drugs, an approach known as drug repositioning or repurposing, can serve as a useful strategy to shorten the development cycle [28]. Repurposed drugs can be brought to the market in a much shorter time-frame and at lower costs. With the exponential growth of "omics" and other biomedical data in recent years, computational drug repositioning provides a fast, cost-effective and systematic way to identify promising repositioning opportunities [28].

In this study we describe a general drug repositioning approach by predicting drug indications based on their expression profiles, with a focus on applications in psychiatry. We treat drug repositioning as a supervised learning problem and apply different state-of-the-art machine learning methods for prediction. Drugs that are not originally indicated for the disease but have high predicted probabilities serve as good candidates for repositioning. There are several advantages of this approach. Firstly, the presented approach is a general and broad framework that leverages machine learning (ML) methodologies, a field with very rapid advances in the last decade. This provides great flexibility and opportunities for further improvement in the future as virtually any supervised learning methods can be applied. Newly developed prediction algorithms can also be readily incorporated to improve the detection of useful drug candidates. In addition, the method described here is widely applicable to any chemical or drugs with expression profiles recorded, even if the drug targets or mechanisms of actions are unknown. For example, herbal medicine products may contain a mixture of ingredients with uncertain drug targets; even for many known medications (e.g. lithium [73]), their mechanisms of actions and exact targets are not completely known. If transcriptomic profiling has been performed, they can still be analyzed for therapeutic or repositioning potential under the current approach.

3.1.2 Related Works

There has been increasing interest in computational drug repositioning recently, in view of the rising cost of new drug development. Hodos et al. [50] provided a comprehensive and updated review on this topic. Similarity-based methods (e.g. ref. [40, 81, 67, 71, 77, 66]) represent one common approach, but as noted by Hodos et al., the dependence on data in the "nearby pharmacological space" might limit the ability to find medications with novel mechanisms of actions. Another related methodology is the network-based approach [68], which typically requires data on the relationship between drugs, genes and diseases as well as connections within each category (e.g. drug-drug similarities). It can integrate different sources of information but may still be constrained by the focus on a nearby pharmacological space and the choice of tuning parameters in network construction or inference is often ad hoc [50]. The present work is different in that we apply a broad framework for repositioning and we do not focus on one but many different kinds of learning methods. There is comparatively less reliance on known drug mechanisms or the “nearby pharmacological space” as we let the different algorithms “learn” the relationship between drugs, genes and disease in their own ways. We note that kernel-based ML methods such as support vector machine (SVM) are also based on some sort of similarity measures. A related work [77] have also examined SVM as a promising ML approach for drug repositioning and identified several interesting candidates. However, here our focus is different in that we considered a variety of other approaches and SVM is one of the methods which falls under the broader framework of ML for repositioning. We also employed more in-depth validation strategies, such as assessing enrichment for drugs considered in clinical trials and correlations with the level of literature support. As for other advantages of ML approaches, for high-throughput omics data, often only a subset of genes or input features are relevant, and many ML methods are able to “learn” which features to consider for repositioning. As we shall discuss later, ML approaches may also provide a new avenue to explore the mechanisms of different drug classes, by studying the variable importance of gene features.

We are particularly interested in drug repositioning for psychiatric disorders in view of the lack of novel treatments in the area. Although computational drug repositioning has attracted increased attention recently, relatively few studies focus on psychiatric disorders (except e.g. [130, 27, 89, 102]), compared to other areas like oncology. Psychiatric disorders are leading causes of disability worldwide [123], however there have been limited advances in the development of new pharmacological agents in the last two decades or so [52]. Development of new therapies is also limited by the difficulty of animal models to fully mimic human psychiatric conditions [79]. Investment by drug companies has in general been declining [52], and new approaches for drug discoveries are very much needed in this field. We will explore repositioning opportunities for schizophrenia along with depression and anxiety disorders. Here depression and anxiety disorders are analyzed together as they are highly clinically comorbid [64, 58], show significant genetic correlations [83], and share similar pharmacological treatments [8].

3.1.3 Significances

Contributions of this study are summarized below. Firstly, we presented a general workflow and approach to drug repositioning of a disease based on ML methods, leveraging drug expression profiles as predictors. While previous work [2] has also proposed the use of ML on drug transcriptome profiles for classifying drugs into groups (e.g. anti-cancer drugs, cardiovascular drugs, drugs acting on the central nervous system etc.), we focused on drug repositioning for particular diseases instead of predicting the big therapeutic groups, as disorders in the same group can have diverse treatments. Secondly, we have performed a comparison of the predictive performances of five state-of-the-art and perhaps most commonly employed ML algorithms, including deep neural networks, support vector machines, elastic net, random forest and gradient boosted trees. Thirdly, we identified new repositioning opportunities for schizophrenia and depression/anxiety disorders and validated the relevance of the repositioned drugs by showing their enrichment among drugs considered for clinical trials, as well as support by previous literature. As another means of validation, we also showed that the predicted probabilities of treatment potential are significantly and positively correlated with the level of literature support (using the number of research articles support as proxy). Finally, we explored which genes and pathways contributed the most to our predictions, hence shedding light on the molecular mechanisms underlying the actions of antipsychotics and antidepressants.

3.2 Datasets and Methods

3.2.1 Datasets

We present a general drug repositioning approach adopting a supervised learning approach. We construct prediction models in which the outcome is defined as whether the drug is a known treatment for the disease, and the predictors are expression profiles of each drug. Drugs that are not originally known to treat the disease but have high predicted probabilities are regarded as good candidates for repositioning.

Drug expression data

The expression data is downloaded from Connectivity Map (CMap), which captures transcriptomic changes when three cell lines (HL60, PC3, MCF7) were treated with a drug or chemical [64]. We downloaded raw expression data from Cmap, and performed normalization with the MAS5 algorithm [86]. Expression levels of genes represented on more than one probe sets were averaged. We employed the limma package [92] to perform analyses on differential expression between treated cell lines and controls. Analyses were performed on each combination of drug and cell line, with a total of 3478 instances. Note the expression of each drug was measured on three different cell lines. Expression measurements were available for 12436 genes. Thus, the scale of our dataset is 3478×12436. Statistical analyses were performed in R3.2.1 with the help of the R package ”longevityTools”.

Defining drug indications

Drug indications were extracted from two known resources, namely the Anatomical Therapeutic Chemical (ATC) classification system and the MEDication Indication Resource high precision subset (MEDI-HPS)[62]. We focus on schizophrenia as well as depression and anxiety disorders in this study. From the ATC classification system, two groups of drugs were extracted, consisting antipsychotics and antidepressants. On the other hand, the MEDI-HPS dataset integrates four public medication resources, including RxNorm, Side Effect Resource 2 (SIDER2) [62], Wikipedia and MedlinePlus. We used the MEDI high-precision subset (MEDI-HPS) which only include drug indications found in RxNorm or in at least 2 out of 3 other sources [128]. This subset achieves a precision of up to 92% according to Wei et al. [128]. To our knowledge, antidepressants from ATC and depression / anxiety from MEDI-HPS roughly fall into the same category, and this is also the same for antipsychosis from ATC and schizophrenia from MEDI-HPS.

3.2.2 Methods

We employed different state-of-the-art machine learning approaches including deep neural networks (DNN) [38], support vector machine (SVM) [20], random forest (RF) [13], gradient boosted machine with trees (GBM) [34] and logistic regression with elastic net regularization (EN) [135] to predict indications with binary classifiers. Our data is imbalanced as only a minority of the drugs are indicated for schizophrenia or depression/anxiety disorders. We performed both unweighted and weighted analyses in this study; in the weighted analysis, class weights are adjusted such that the minority group (drugs indicated for the disorder) will receive higher weight to achieve a balance between positive and negative instances.

In our unweighted model, DNN was implemented in the python package keras. We employed a fully connected feedforward neural network. Hyperparameters were chosen by the ”fmin” optimization algorithm from ”hyperopt”, which employs a sequential model-based optimization approach [11]. The tree-structured Parzen estimator (TPE) was used. The more sophisticated hyper-parameter search strategies provided by sequential model-based methods may produce better results than simpler approaches (e.g. grid search) when the number of hyper-parameters is large, such as in deep learning settings [11]. Fifty evaluations were run for each search of optimal hyper-parameters. Dropout and mixed $L1/L2$ penalties were employed to reduce over-fitting. The neural networks consisted of two or three layers, with number of nodes selected uniformly from the range [64, 1024]. Dropout percentage was selected uniformly from [0.25, 0.75], and $L1/L2$ penalty uniformly from [1E-5, 1E-3]. Optimizer was chosen from ”adadelta” [134], ”adam” [59] and ”rmsprop” [48], and the activation function chosen from ”relu”, ”softplus” or ”tanh”. One hundred epochs were run for each model and we extracted the model weights corresponding to the best epoch.

We also attempted ”hyperopt” for the weighted analysis, however the predictive performance was unexpectedly poor due to unclear reasons yet to be revealed. We therefore turned to hyperparameter selection with grid search, with some adjustments in the parameter ranges. A two-layer neural network was constructed with dropout and mixed $L1/L2$ to avoid over-fitting. The number of neurons in the first hidden layer was selected from $\{1000, 1500, 2000\}$, the number in the second layer from $\{500, 1000, 1500\}$, dropout rate from $\{0.4, 0.5, 0.6, 0.7, 0.8\}$, $L1/L2$ penalty from intervals [-13, -3] and [-9, -8] in log space and the number of epochs from [10, 20, 30, 50]. In order to speed up hyperparameter selection, we first chose the number of epochs, the best optimizer and activation function (following the same parameter range as described above) with other parameters fixed, and then used the best parameters chosen in the first step to find the optimal complexity of our neural networks by selecting the number of neurons in each layer, dropout and mixed $L1/L2$ penalties.

SVM, RF and GBM models were implemented in ”scikit-learn” (sklearn) in python. Hyper-parameter selection was performed by the built-in function GridSearchCV in sklearn. For SVM, we chose radial basis function as the kernel. The two hyperparameters C and γ were selected from [-5, 2] and [-6, 2] in log10-space respectively.

For RF, the number of bagged trees was set to 1000, the maximum number of features used for splitting was selected from $\{800, 1000, 1500, 2000, 3000, 5000\ \}$ and min_samples_leaf (the minimum number of samples required at a leaf node) was selected from $\{1, 3, 5, 10, 30, 50, 80\ \}$. For GBM, the number of boosting iterations was selected from a sequence of 100 to 1001 with step size 50, learning rate from {0.005, 0.01, 0.015, 0.02, 0.03, 0.05 }, subsampling proportion from {0.8, 1 }, maximum depth of each estimator from {2, 3, 5, 10} and maximum number of features from {10, 30, 50, 100, 500, 1000 }. Finally, the EN model was implemented by the R package ”glmnet” [33]. The elastic-net penalty parameter α was chosen from seq(0, 1, by=0.1), with other settings following the default.

Nested Cross-validation

We adopted nested three-fold cross validation (CV) to choose hyper-parameters and evaluate model performances. It has been observed that optimistic bias will result if one uses simple CV to compute an error estimate for a prediction algorithm that itself is tuned using CV [122]. Nested CV avoids this problem and is able to give an almost unbiased estimate of prediction accuracy [122]. The inner loop CVs were used to choose the parameters that optimized predictive performance. In each outer loop CV we made predictions on the corresponding test set using the best model trained from the inner CV loops. To achieve maximum consistency in our comparisons, we compared different methods on the same test set in each loop. Note that the test sets were not involved in model training or parameter tuning.

Predictive performance measures

The performances of the machine learning methods were evaluated in the test sets using three metrics, including log loss, area under the receiver operating characteristic curve (ROC-AUC) and area under the precision recall curve (PR-AUC). Log loss compares the predicted probabilities against the true labels. The receiver operating characteristic curve, which plots the sensitivity (i.e. recall) against (1- specificity), is a very widely used approach to evaluate predictive performances in biomedical applications. The precision-recall curve on the hand plots the precision (i.e. positive predictive value) against the sensitivity (recall). Since precision depends on the overall proportion of positive labels, the PR-AUC is also dependent on such proportions. Davies et al. [26] that the PR curve may give more informative comparisons when working with imbalanced data.

Identifying important genes and pathways

We also performed analyses to reveal the genes which contribute the most to the prediction model. For elastic net, we extracted the genes with non-zero coefficient in at least one cross-validation fold, and the resulting genes were subject to an over-representation analysis (ORA) (using hypergeometric tests) to reveal the pathways involved. For RF and GBM, feature importance was computed using built-in functions in sklearn based on Gini importance (i.e. the average decrease in node impurity). We then performed a gene-set enrichment analysis (GSEA [110]) based on the genes together with their respective importance scores (the highest score across three folds was taken). For SVM and DNN, there is a lack of widely adopted importance measures, so we focused on the rest of ML methods in this part. Pathway analyses were conducted by the web-based program WebGestalt [126]. Four pathway databases were considered in our analyses, including KEGG, PANTHER, Reactome and Wikipathways.

External validation by testing for enrichment of psychiatric drugs considered for clinical trials

We then performed additional analyses to assess if the drugs with high predicted probabilities from our machine learning models are indeed good candidates for repositioning. Briefly, we tested whether the drugs with no known indication for the disease but high predicted probabilities are more likely to be included in clinical trials.

In the first step, we filtered off drugs that are known to be indicated for the disease as derived from ATC and MEDI-HPS. This is because we are mainly interested in repositioning other drugs of unknown therapeutic potential, and that the labels of drug indication (from ATC or MEDI-HPS) have already been utilized in the ML prediction steps. Including known indications will lead to over-optimistic estimates of significance of enrichment. Next, we extracted a list of drugs that were included in clinical trials for schizophrenia as well as depression and anxiety disorders. The list was derived from clincialTrial.gov and we downloaded a pre-compiled version from.

We then tested for enrichment of those drugs listed in clinicalTrial.gov among the top repositioning results. We performed an enrichment analysis of "drug-sets", similar to a gene-set analysis approach widely used in bioinformatics [27]. We performed one-tailed t-tests to assess if the predicted probabilities (derived from machine learning models) are significantly higher for psychiatric drugs considered in clinical trials.

Searching for literature support

We manually search for literature support for the top 15 repositioning hits for each method in PubMed and Google scholar. The search strategy is given in Supplementary text. A limitation of manual search is that it is extremely time-consuming to perform such a search for all drugs. It should be noted that publication bias is likely to be present (as negative studies are less likely to be reported), although it is difficult to exactly quantify the extent of bias. As we shall discuss later, we did find literature support for a number of top repositioning candidates, but it is still possible that similar evidence may be found for drugs ranked in the middle or lower. Also, one may wish to assess whether drugs with less (or no) support by prior studies would indeed have lower predicted probabilities (similar to having 'negative controls' in an experiment).

To ensure an unbiased and more comprehensive comparison, we conducted an analysis with "automated" literature mining on all drugs in PubMed. We extracted the number of research articles supporting each drug's association with the corresponding psychiatric disorder (schizophrenia or depression/anxiety). We then examined the correlation between the number of research articles support and the predicted probabilities of treatment potential from ML models. Similar approaches for validating repositioning candidates based on literature mining have also been used in other studies [51]. As the number of articles is typically skewed and not normally distributed, we employed Spearman and Kendall correlation measures. We also compared the predicted probabilities of drugs with no article support versus those with at least one article support. We hypothesized that drugs without any literature support would have lower predicted probabilities from ML models, and vice versa. We used the Wilcoxon rank-sum test for such comparison. All tests were carried out in R 3.2.3 and the tests were one-tailed. Similar to the enrichment test discussed above, drugs that are known to be indicated for the disorder (from ATC or MEDI-HPS) were excluded from this analysis. This is because the indication label of these drugs have been used in the ML model, which may lead to over-optimistic results, and that we wish to focus on repositioning potential of other drugs of unknown significance.

3.3 Experiment Results

3.3.1 Predictive performance comparison

Unweighted analysis

The average predictive performances (averaged over three folds) of different machine learning methods are listed in Table 3.3.1. When considering log loss as the criteria of interest, SVM gave the best result overall, though EN showed the best performance in one of the four datasets. DNN and EN showed quite similar predictive performances. RF and GBM were slightly worse than other methods, but the difference was small. When ROC-AUC was considered as the performance metric, SVM and EN gave similar performances. SVM outperformed EN in the schizophrenia datasets, while EN showed better results in the other two datasets. The performance of DNN was worse than that of SVM and EN, although the differences were not large. The two tree-

Table 3.1: Average predictive performance of different ML models across four datasets in unweighted (top) and weighted analysis (bottom)

Unweighted analysis	Average Log Loss			
	MEDI-HPS DEP/ANX	ATC ATD	MEDI-HPS SCZ	ATC ATP
SVM	**0.1188**	0.0943	**0.1018**	**0.0895**
EN	0.1249	**0.0916**	0.1097	0.0954
DNN	0.124	0.0948	0.1111	0.0992
GBM	0.1293	0.1018	0.1157	0.1039
RF	0.1294	0.1002	0.1155	0.1013
	Average ROC-AUC			
SVM	**0.7141**	0.7619	**0.7705**	**0.7755**
EN	0.725	**0.779**	0.7515	0.7681
DNN	0.6952	0.7456	0.7533	0.7604
GBM	0.6536	0.6042	0.7172	0.7433
RF	0.6315	0.639	0.7036	0.7501
	Average PR-AUC			
SVM	**0.2026**	**0.1485**	**0.2973**	**0.3379**
EN	0.1372	0.1008	0.1586	0.2087
DNN	0.1447	0.0877	0.1577	0.2156
GBM	0.091	0.0417	0.1426	0.1528
RF	0.1193	0.0639	0.1677	0.1703
weighted analysis	Average Log Loss			
	MEDI-HPS DEP/ANX	ATC ATD	MEDI-HPS SCZ	ATC ATP
SVM	**0.1189**	**0.0934**	**0.1022**	**0.0898**
EN	0.5803	0.5344	0.5028	0.5112
DNN	0.1309	0.099	0.1308	0.1098
GBM	0.1281	0.1032	0.1114	0.0981
RF	0.1234	0.0943	0.106	0.0939
	Average ROC-AUC			
SVM	0.7198	0.7718	0.7731	0.7765
EN	0.661	0.7394	0.7494	**0.7997**
DNN	**0.7424**	**0.7979**	0.741	0.7576
GBM	0.7155	0.7578	0.7584	0.7794
RF	0.689	0.7355	**0.7843**	0.7801
	Average PR-AUC			
SVM	**0.2017**	**0.151**	**0.298**	**0.3361**
EN	0.0751	0.0896	0.152	0.203
DNN	0.1796	0.1107	0.2278	0.2641
GBM	0.18	0.1168	0.2697	0.278
RF	0.1771	0.1165	0.2721	0.2707

1. Values of evaluation metrics for algorithms with the best performance in each dataset (for each evaluation metrics) are marked in bold.

2. SVM: support vector machines; EN: logistic regression with elastic net regularization; DNN: deep neural networks; RF: random forest; GBM, gradient boosted machines.

3. MEDI-HPS: MEDication Indication - High Precision Subset; ATC: Anatomical Therapeutic Chemical classification.

based methods performed worse especially in the depression/anxiety datasets. We then considered PR-AUC, which is more sensitive to imbalanced data, as the measure of predictive performance. SVM was the best-performing method. EN and DNN followed with very similar performances. Consistent with other performance measures, GBM and RF did not perform as well in the depression/anxiety datasets, but the performance was more comparable for the schizophrenia datasets.

Weighted analysis

Compared with unweighted analysis, we observed improvements in predictive performance for several methods including GBM, RF and deep learning. SVM and EN performed similarly in general. Considering ROC-AUC, deep learning performed the best for depression and anxiety disorders, while RF and EN showed highest ROC-AUC for schizophrenia. SVM achieved the best PR-AUC and log-loss compared to other ML approaches.

3.3.2 Enrichment for psychiatric drugs considered clinical trials

We further tested whether the top repositioning results are enriched for drugs included in clinical trials for psychiatric disorders. As shown in Table 3.2, we observed significant enrichment of such drugs for both schizophrenia and depression/anxiety disorders across all methods in the weighted analysis. In addition, most results in the unweighted analysis were also significant. This external validation provides further support to the usefulness of our approach in identifying new repositioning opportunities.

3.3.3 Correlation of predicted probabilities with degree of literature support

We also examined Spearman and Kendall correlations between predicted probabilities from ML models and the number of PubMed articles retrieved, which serves as a proxy for the level of literature support 3.3. We focused on results from the weighted

Table 3.2: Enrichment for psychiatric drugs included in clinical trials among the repositioning hits

Dataset		unweighted analysis	weighted analysis
		P-value	P-value
SVM	MEDI-HPS DEP/ANX	**0.0014**	**0.0011**
	ATC ATD	**0.018**	**0.0039**
	MEDI-HPS SCZ	**0.0205**	**0.0264**
	ATC ATP	**0.0098**	**0.0084**
EN	MEDI-HPS DEP/ANX	**0.0022**	**0.0023**
	ATC ATD	**0.0032**	**0.0087**
	MEDI-HPS SCZ	**0.0294**	**0.0315**
	ATC ATP	**0.0104**	**0.0033**
DNN	MEDI-HPS DEP/ANX	**0.0105**	**0.0009**
	ATC ATD	0.1369	**0.0017**
	MEDI-HPS SCZ	0.0908	**0.019**
	ATC ATP	**0.0237**	**0.0021**
GBM	MEDI-HPS DEP/ANX	**0.0494**	**0.0003**
	ATC ATD	**0.0433**	**0.0002**
	MEDI-HPS SCZ	0.2283	**0.0269**
	ATC ATP	0.2482	**0.0005**
RF	MEDI-HPS DEP/ANX	0.0651	**0.0005**
	ATC ATD	0.2518	**0.0007**
	MEDI-HPS SCZ	0.1299	**0.0427**
	ATC ATP	0.5232	**0.0063**

P-values < 0.05 and with FDR less than 0.05 are in bold.

Table 3.3: Correlations between predicted probability of treatment potential with number of research articles supporting association with schizophrenia or depression/anxiety

	MEDI-HPS DEP/ANX	ATC ATD	MEDI-HPS SCZ	ATC ATP
		Spearman's rho		
SVM	**0.078**	0.049	**0.098**	0.088
EN	0.031	0.043	0.091	0.073
DNN	0.075	**0.065**	0.085	**0.11**
GBM	0.065	0.05	0.082	0.102
RF	0.066	0.051	0.052	0.055
		Kendall's tau		
SVM	**0.06**	0.037	**0.077**	0.068
EN	0.024	0.033	0.071	0.058
DNN	0.058	**0.05**	0.067	**0.086**
GBM	0.05	0.038	0.064	0.08
RF	0.052	0.04	0.041	0.044
		Spearman correlation p-value		
SVM	**2.09E-05**	4.91E-03	**6.46E-09**	1.47E-07
EN	4.89E-02	4.91E-03	5.96E-08	9.16E-06
DNN	3.45E-05	**2.64E-04**	3.37E-07	**7.53E-11**
GBM	2.91E-04	3.87E-03	9.64E-07	1.52E-09
RF	2.35E-04	3.51E-03	1.35E-03	6.86E-04
		Kendall correlation p-value		
SVM	**2.27E-05**	5.46E-03	**6.81E-09**	1.81E-07
EN	5.00E-02	1.10E-02	5.81E-08	9.05E-06
DNN	3.56E-05	**2.83E-04**	3.33E-07	**8.52E-11**
GBM	2.90E-04	4.22E-03	1.00E-06	1.62E-09
RF	2.33E-04	3.63E-03	1.37E-03	6.93E-04
		Wilcoxon rank-sum test p-value		
SVM	**1.94E-04**	2.10E-02	**1.49E-07**	3.27E-06
EN	1.09E-01	2.50E-02	3.07E-07	3.40E-05
DNN	3.85E-04	**1.30E-03**	8.63E-07	**8.76E-10**
GBM	1.43E-03	2.11E-02	1.61E-05	1.91E-08
RF	1.00E-03	1.08E-02	4.49E-03	2.98E-03

The highest correlation coefficient or lowest p-value in each analysis is marked in bold.

analysis as they have better predictive performances in general. We found significant and positive correlations for all ML methods across all four analyses. DNN was the best performing method (in terms of the correlation metric and level of significance) in two out of four tasks (ATC antipsychotics and ATC antidepressants), and was relatively close to the best ones for the other two analyses. SVM performed the best in these analyses, but its deficit when compared to DNN was proportionately large for the two ATC tasks. The results of Wilcoxon rank-sum test were generally concordant with those from correlation tests.

3.3.4 Identifying contributing genes and pathways

Supplementary Tables [1] 1-4 show the top genes as identified by variable importance measures (for RF and GBM) and regression coefficients (for EN). The enriched pathways are shown in Table IV and Supplementary tables 5-12. Since the number of genes involved is large, we only highlighted a few top enriched pathways here. Interestingly, steroid and cholesterol biosynthesis are among the most significantly enriched pathways for drugs against schizophrenia and depression/anxiety. Notably, abnormalities in the hypothalamic-pituitary-adrenal (HPA) axis have long been suggested as one of the key pathological mechanisms underlying depression [121]. The steroid (cortisol) synthesis inhibitor metyrapone has been shown to be effective for depression in a double-blind randomized controlled trial (RCT) [55] and other studies [72], although another trial failed to show any benefits [76]. Antidepressants have also been shown to regulate glucocorticoid receptor functioning in vivo. On the other hand, neuroactive steroids may be implicated in the pathophysiology of schizophrenia [98]. Cholesterol biosynthesis, including regulation by sterol regulatory element-binding protein (SREBP), was frequently top-listed in our pathway analysis. Antipsychotics and some antidepressants are associated with metabolic syndrome and weight gain, and previous in vitro and in vivo studies have shown lipogenic effects of

[1] Available at https://drive.google.com/open?id=1YDtk-uTVX5gsnZvWM7q3PRz3_x8yWrpQ

Table 3.4: Selected enriched pathways based on variable importance of genes in ML models with FDR < 0.2

Method	Name	#Gene	FDR
ATC antidepressants and MEDI-HPS depression/anxiety (weighted analysis)			
eNet-ORA_Reactome	Cholesterol biosynthesis	7	1.78E-06
eNet-ORA_Wikipathway	Sterol Regulatory Element-Binding Proteins (SREBP) signalling	8	1.73E-04
eNet-ORA_KEGG	Steroid biosynthesis - Homo sapiens (human)	5	2.38E-04
gbm_KEGG	Fat digestion and absorption - Homo sapiens (human)	34	8.95E-03
gbm_Panther	Insulin/IGF pathway-protein kinase B signaling cascade	34	1.05E-02
rf_Wikipathway	Mismatch repair	9	1.38E-01
rf_Wikipathway	ID signaling pathway	16	1.58E-01
rf_Wikipathway	Statin Pathway	25	1.61E-01
rf_Wikipathway	Photodynamic therapy-induced HIF-1 survival signaling	37	1.63E-01
eNet-ORA_Panther	TGF-beta signaling pathway	4	1.70E-01
ATC antipsychotics and MEDI-HPS schizophrenia (weighted analysis)			
rf_Wikipathway	Sterol Regulatory Element-Binding Proteins (SREBP) signalling	64	0.00E+00
eNet-ORA_Reactome	Cholesterol biosynthesis	7	9.33E-05
eNet-ORA_KEGG	Steroid biosynthesis - Homo sapiens (human)	4	2.01E-03
rf_Wikipathway	Statin Pathway	25	2.51E-02
eNet-ORA_Wikipathway	Fatty Acid Beta Oxidation	5	2.98E-02
eNet-ORA_Reactome	Asparagine N-linked glycosylation	15	4.36E-02
eNet-ORA_KEGG	Metabolic pathways - Homo sapiens (human)	37	4.87E-02
eNet-ORA_Reactome	Synthesis of UDP-N-acetyl-glucosamine	3	8.11E-02
gbm_Panther	5HT3 type receptor mediated signaling pathway	14	8.58E-02
eNet-ORA_KEGG	Citrate cycle (TCA cycle) - Homo sapiens (human)	3	8.96E-02
gbm_Reactome	G1/S-Specific Transcription	18	1.32E-01
eNet-ORA_Reactome	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	4	1.32E-01
gbm_Panther	Androgen/estrogene/progesterone biosynthesis	9	1.40E-01
rf_Wikipathway	Photodynamic therapy-induced unfolded protein response	23	1.42E-01
eNet-ORA_Reactome	COPII (Coat Protein 2) Mediated Vesicle Transport	6	1.44E-01

We aggregated pathway analysis results from 4 databases (KEGG, Reactome, Panther and Wikipathways). Pathways that were highly similar were filtered. Only results from weighted analysis are included here.

these drugs as controlled by SREBP transcription factors [91, 30]. Interestingly, some studies showed lower cholesterol may be associated with suicidality [22], depressive symptoms [87, 133], and poorer cognition in schizophrenia [60], but these findings are controversial. Whether pathways related to cholesterol synthesis may play a role in the therapeutic effects of psychotropic drugs remain a topic for further investigation. Some other pathways are also worth mentioning. For example, IGF signaling pathway was significantly enriched under antidepressants. IGF-I has been reported to improve depression and anxiety symptoms in clinical samples [113], and showed antidepressant-like effects in animal models [14, 42]. The 5-HT3 signaling pathway was also top-listed under antipsychotics. 5-HT3 has been proposed as a new drug target and improvements in negative and cognitive symptoms have been reported in clinical trials [29].

3.3.5 Top repositioning hits and literature support from previous studies

Table 3.5 show some of the selected top repositioning candidates with literature support, which will also be discussed below. More detailed tables showing the top 100 hits for each ML method in both unweighted and weighted analyses are presented in Supplementary Tables 13-16 [2]. Note that drugs that are known to be indicated for these disorders by ATC or MEDI-HPS were excluded from the lists. We noted overlap in top hits derived from different machine learning methods, but some repositioning candidates are unique to one or few ML approaches. This suggests that employing a diverse set of ML methods may be advantageous in "learning" different potential repositioning candidates. We will chiefly focus on the top 15 hits for each ML method in the exposition below.

[2] Available at https://drive.google.com/open?id=1YDtk-uTVX5gsnZvWM7q3PRz3_x8yWrpQ

Table 3.5: Some literature-supported candidates selected from top hits derived from machine learning methods (known antipsychotics and antidepressants are not included in this list)

Drug	Method	Relationship with disease
Depression/anxiety		
Cyproheptadine	SVM, RF, GBM, EN	5-HT2 receptor antagonist, improve depression in a small cross-over trial
Chlorcyclizine	DNN, RF, GBM, SVM	phenylpiperazine group to which many other antidepressants and antipsychotics belong
Pizotifen	EN	5-HT2A/2C antagonist, positive result in an RCT
TrichostatinA/ Vorinostat	DNN, EN	HDAC inhibitors may have antidepressant effects as shown in animal models
Tetrandrine	DNN, RF, GBM	CCB; antidepressant-like effects in mice; may increase 5-HT, NE and BDNF concentrations
Apigenin	GBM	Antidepressant and anxiolytic effects in animal models and in an RCT
Metformin	EN	may reduce depression risk among DM subjects
Schizophrenia		
Valproate	DNN, SVM	open RCTs reported symptom improvement when used as adjunctive treatment
Raloxifene	DNN, EN	improve SCZ symptoms in an RCT of post-menopausal women
Nordihydroguaiaretic acid	DNN, GBM, SVM	antioxidant; oxidative stress implicated in SCZ
Pioglitazone	DNN	Another drug in the same class (pioglitazone) improved SCZ symptoms in RCT
Tretinoin	DNN	Retinoid; dysfunction in retinoid signaling may be implicated in SCZ
Felodipine	GBM	CCB; CCB added to antipsychotics may be beneficial
Aspirin	SVM	NSAID; may improve SCZ symptoms as shown in RC
Genistein	GBM	Phytoestrogen; animal model shows possible anti-dopaminergic effects

As a number of top results were known antipsychotics or antidepressants (please refer to the main text for details), these were not presented in the above table. RCT, randomized controlled trial; HDAC, Histone deacetylases; CCB, calcium channel blocker; 5-HT, serotonin; NE, norepinephrine; BDNF, brain-derived neurotrophic factor; NSAID, non-steroidal anti-inflammatory drugs.

Repositioning candidates for depression/anxiety disorders

Regarding depression and anxiety disorders, many of the top results are antipsychotics, such as trifluoperazine, perphenazine, fluphenazine and thioridazine, among others. Antipsychotics have long been used for the treatment for depression [126]. In earlier studies, phenothiazines (a class of antipsychotic to which many of our top hits belong) was observed to produce similar anti-depressive effects as tricyclic antidepressants [93]. Due to the risk of extra-pyramidal side-effects, typical antipsychotics are less commonly used these days and second-generation (atypical) antipsychotics are more often prescribed. Meta-analyses have shown that atypical antipsychotics are effective as adjunctive or primary treatment for depression [93, 105]. Antipsychotics are also commonly prescribed for severe depressive episodes with psychotic symptoms.

A few other drugs on the lists are also worth mentioning. Cyproheptadine (top-listed by SVM, RF, GBM and EN) is a 5-HT2 receptor antagonist and was shown to improve depression in a small cross-over trial [41]. It was also reported that the drug reduced the neuropsychiatric side-effects of the antiviral therapy efavirenz, including depressive and anxiety symptoms [23]. Chlorcyclizine belongs to the phenylpiperazine class and numerous antidepressants and antipsychotics also belong to this class [72]. Pizotifen, listed by EN, is a 5-HT2A/2C antagonist which was shown to possess antidepressant effects in a double-blind RCT [106]. DNN and EN have identified histone deacetylase (HDAC) inhibitors including trichostatin A and vorinostat as top repositioning hits for depression/anxiety and schizophrenia. HDAC have been implicated in the pathogenesis of psychiatric disorders including depression, as reviewed by Fuchikami et al. [35]. HDAC inhibitors have been reported to produce antidepressant-like effects in animal models [49, 21], although no clinical trials on psychiatric disorders were available. Interestingly, in a recent study which employed gene-set analysis on de novo mutations to uncover repositioning opportunities, HDAC inhibitors were highlighted as candidates for schizophrenia and other neurodevelopmental disorders [103].

Another candidate was tetrandrine, a calcium channel blocker top-listed by DNN, RF, GBM and SVM. Tetrandrine demonstrated antidepressant-like effects in mice [37] in forced swimming and tail suspension tests. The drug also increased the concentration of 5-hydroytrytamine (5-HT) and norepinephrine in mice treated with reserpine or chromic mild stress, and raised the levels of brain-derived neurotrophic factor (BDNF) in the latter case [37].

Repositioning candidates for schizophrenia

With regards to repositioning results for schizophrenia, some of the hits are antidepressants, such as protriptyline, maprotiline and clomipramine, among others. Antidepressants are frequently prescribed in schizophrenia patients due to possibility of comorbid depression or obsessive-compulsive disorder [74]. In meta-analyses antidepressants were also found to improve negative symptoms of schizophrenia [99, 96]. For the antidepressants on the list, maprotiline (listed by RF, GBM, EN, SVM) has been reported to improve negative symptoms in chronic schizophrenia patients [131] as an adjunctive treatment. Other drug clomipramine (listed by DNN, GBM, EN) has been shown to ameliorate not only obsessive-compulsive but also overall schizophrenic symptoms in patients with comorbid disorders [12]. Interestingly, the mood stabilizer valproate was also listed among the top (by DNN and SVM). Valproate may improve clinical response when added to antipsychotics, although the evidence is mainly based on open RCTs [95]. The EN algorithm also “re-discovered” spiperone, an antipsychotic not listed in ATC or MEDI-HPS, as one of the top repositioning hits.

Several other drugs less well-known for psychiatric disorders are also worth mentioning. The selective estrogen receptor modulator raloxifene (listed by DNN and EN) was shown to improve schizophrenia symptom scores in double-blind RCTs of postmenopausal women [116, 117]. Another drug nordihydroguaiaretic acid (listed by DNN, GBM, SVM) has antioxidant properties [69] and may be useful in combating oxidative stress in schizophrenia [129]. Pioglitazone, top-ranked by DNN, belongs to the class of thiazolidinediones and has anti-diabetic and anti-inflammatory properties. Although this drug was withdrawn due to unexpected adverse effects on the liver, our finding suggested that other thiazolidinediones may be useful for schizophrenia. Indeed, another drug of the same class known as pioglitazone has been shown to improve negative symptoms in schizophrenia patients in a double-blind RCT [54]. Another RCT also showed improvements in depressive symptoms [101]. Tretinoin (listed by DNN) is a retinoid and retinoid dysfunction has been linked to schizophrenia [39, 65] Clinical trials with another retinoid (bexarotene) showed some benefits of the drug as an add-on agent in schizophrenia. Again retinoid signaling was implicated for schizophrenia in a recent study on drug repositioning leveraging de novo mutations [103]. Felodipine (listed by GBM) is a calcium channel blocker and GWAS on schizophrenia and bipolar disorder have revealed many genes related to calcium channels [45, 18]; a recent study also suggested concomitant use of CCB and antipsychotics may be more beneficial than antipsychotics alone [118].

Some hits from the unweighted analysis

The top repositioning candidates from unweighted analysis for each ML method are listed in Supplementary Tables 13-16. There were a number of overlaps with the candidates from the weighted analysis. Here we highlight a few prioritized drugs (that have not been mentioned above) with literature support. Aspirin (acetylsalicylic acid) is a non-steroidal anti-inflammatory agent (listed by SVM), which has been shown to improve schizophrenia symptoms in a recent meta-analysis of RCTs [104]. Genistein is a phytoestrogen and can bind to estrogen receptors [127]. An animal study showed that genistein may possess anti-dopaminergic actions [111]; interestingly, clinical studies have shown potential therapeutic benefits of estrogens on schizophrenia [104].

The EN algorithm identified metformin as one of the top repositioning hits for depression/anxiety disorders. A study in Taiwan reported that the risk of depression in diabetic patients was reduced by 60% for those given metformin with sulfonylurea [125]. Another study reported improved depressive symptoms and cognitive functions for patients with comorbid diabetes and depression [43]. Another drug apigenin, top-listed by GBM, was supported by a number of in vitro and animal studies for possible antidepressant-like and anxiolytic effects [85]. A clinical trial of oral chamomile (which was standardized to contain 1.2% of apigenin) showed benefits for anxiety and depression [3, 4].

3.4 Discussion

In this study, we have presented a repositioning approach by predicting drug indications based on expression profiles. We employed and compared five state-of-the-art ML methods to perform predictions. We also observed that the top repositioning hits are enriched for psychiatric drugs considered for clinical trials and that many hits are backed up by evidence form animal or clinical studies, supporting the validity of our approach.

Concerning the performance of different machine learning classifiers, we have employed five methods in total, and all but one (EN) are non-linear classifiers. SVM is a kernel-based learning approach that is widely used in bioinformatics. On the other hand, deep learning methods (such as DNN) that are based on the principles of representation learning [9] have witnessed rapid advances in the last decade, especially in the field of computer vision. While potentially powerful, current successful applications typically require very large datasets for training, and we suspect that the relatively modest sample size (N = 3478) of our dataset may have limited DNN to achieve the optimal predictive ability. We have used at most two hidden layers in view of the moderate sample size, and the complexity of the network may be increased with larger samples, although larger samples would lead to greater computational costs.

This study shows that deep learning can achieve reasonable performance in drug repurposing, and indeed DNN achieved the best ROC-AUC for depression/anxiety disorders in the weighted analysis. Given the rapid growth in the area, deep learning approaches might be worthy of further investigations. While logistic regression with EN is a linear classifier, it performed well overall though lagging behind SVM. The performances of the two tree-based methods (RF and GBM) were largely comparable with other methods in the weighted analysis, although they were less satisfactory without weighting. Notwithstanding the differences in predictive performances, different algorithms are based on diverse model assumptions and principles, and as shown above, methods with slightly lower predictive accuracy may still reveal useful repositioning candidates that are of different mechanisms of actions.

To the best of our knowledge, this is the first study to employ a comprehensive array of machine learning methods on drug expression profiles for drug repositioning of any particular disease; it is also the first application in psychopharmacology. This is also the first work to suggest an ML approach to explore the molecular mechanisms underlying drug actions. In a related work, Aliper et al. made use of the drug transcriptome to predict large drug classes e.g. drugs for neurological diseases, drugs for cardiovascular diseases, anti-cancer agents etc [2]. Here our focus is different and clinically more relevant in that we directly identify repositioning opportunities for a particular disorder. It should be noted that drugs for the same body system can have diverse (or sometimes even opposite) effects. For example, antipsychotics like haloperidol are used to treat schizophrenia but they also cause Parkinsonism [75]. Statins reduces low-density lipoprotein levels and coronary heart disease risk [17], but can cause weight gain and increased diabetic risk [112]. In addition, we concentrated on the study of psychiatric disorders, which was not explicitly considered in Aliper et al. [2] or other previous works. Interestingly, DNN was reported to be the best performing method in their study. However their study [2] and the present work are not directly comparable as the outcomes studied are different and the evaluation metrics also differ. F1 score was used in Aliper et al. [2] (although the choice of a cut-off probability for classification was not explicitly stated) while we used ROC-AUC, PR-AUC and log loss as performance indicators.

Here we aim to provide a proof-of-concept example showing that the application of machine learning methodologies on drug expression profiles may help to identify candidates for repositioning, particularly for psychiatric disorders. The approach is intuitive and also highly flexible. Nevertheless, there is still room for improvement. Firstly, we only consider drug indications and the drug-induced transcriptomic changes in our prediction model. This makes the method very flexible and widely applicable to any compounds or drugs for which an expression profile is available. The use of drug transcriptome evades the need of specifying targets and knowing the mechanisms of actions, and the approach may even be applicable to a mixture of chemical ingredients as may be the case for herbal medicines. However, it is possible that our methods may be further improved by incorporating other information such as drug targets and chemical structure, if such information is available. For example, dopaminergic and monoaminergic pathways have known importance in SCZ and depression treatment respectively, and incorporating such information into our ML framework may further improve prediction accuracy.

As for the prediction algorithm, in our dataset the number of positive labels is small. We tackled this problem by adjusting the weighting of positive and negative instances and indeed found improvements for several ML approaches. Other methodologies to account for imbalanced data are also possible [44], and this may be a topic for further explorations. We covered five commonly used algorithms here but this coverage is obviously not complete; further studies may benefit from the use of more advanced or recently developed learning methods. We also notice that there is an on-going effort to expand the coverage of CMap [109], and that the study with updated full data and documentations have just been released. We are planning to further explore the current framework in the expanded dataset.

It is reassuring to observe that many repositioning hits are supported by previous studies, the predicted probabilities from our model significantly correlate with the degree of literature support, and that the results are enriched for psychiatric drugs considered in clinical trials. However, we stress that further well-designed pre-clinical and clinical studies are necessary before the any results can be brought into clinical practice. The analytic validation we employed in this study aims to provide evidence for the overall validity of the presented repositioning approach. Validation of individual candidates via detailed animal and clinical studies are essential before a drug can be brought into practice; however it should also be noted that detailed experimental/clinical validation of one or two candidates is less suited to provide evidence on whether the approach as a whole works or not, since most drugs cannot be covered and chance findings are possible.

We have also made use of ML methods to explore potentially important genes and pathways that may contribute to treatment effects. Nevertheless, the results again require further experimental validations. Computational approaches for repositioning and explorations of drug mechanisms, such as ML-based methods, provide a cost-effective and systematic way to assess and prioritize drug candidates. While we believe the current approach can improve the prioritization of drug candidates, not all top-ranked drugs will be effective and we do not expect to uncover all potential new treatments. However, given the rising cost in developing a new drug (up to USD 2.6 billion [57]), if a method can reduce the failure rate by even a tiny margin, it will already result in large savings in absolute terms. Further work might involve combining the current approach with other computational and experimental methods to further improve the accuracy of repositioning. As an example, ketamine is one of the most promising new therapies for depression, but the current method did not reveal any similar drugs (ketamine is not listed in CMap). However, another computational repositioning method which compared drug and disease transcriptomes suggested several NMDA antagonists for depression [57], highlighting the potential of integrating different methods in future studies.

3.5 Conclusion

In this work we have presented and applied a machine learning to drug repositioning for schizophrenia and depression/anxiety disorders. We found the candidates were enriched for psychiatric drugs considered in clinical trials, and that numerous top hits were supported by previous studies; the degree of literature support also showed a significant correlation with predicted treatment probabilities from ML models.

It is widely acknowledged that drug development in psychiatry has become stagnant for some years, and that traditional approaches to drug discovery has not been as successful as anticipated. On the other hand, the past few years have seen an extremely rapid development in ML methods and applications; in this regard, we hope that this study will open a new avenue for drug repositioning/discovery, and stimulate further research to bridge the gap between ML and biomedical applications especially drug development. The list of repositioning candidates might also serve as a useful resource for researchers and clinicians working on schizophrenia as well as depression and anxiety disorders, which are illnesses very much in need of new therapies.

□ **End of chapter.**

Chapter 4

Drug Target Discovery

4.1 Motivation

Traditionally, drug discovery involves five steps: target identification, target validation, lead identification, lead optimization and introduction of the new drugs to the public [88]. Nevertheless, the speed of new drug development has been slower than anticipated, despite increasing investment [84]. It is estimated that the cost of developing a new drug is USD 2.6 billion [119]. One of the main reasons for the enormous cost of drug discovery is due to the high failure rate.

Success of drug development largely depends on the validity of targets. However, the majority of drugs fail to complete the development process due to lack of efficacy, and this is often due to the wrong target being pursued [97]. Traditionally, drug targets are often identified from hypothesis-driven pre-clinical models, yet preclinical models may not always translate well to clinical applications. For some diseases such as psychiatric disorders, current animal or cell models are still far from capturing the complexity of the human disorder (cite). In addition, some have hypothesized the hypothesis-driven nature of many studies may have led to ”filtering” of findings and publication bias, exacerbating the reliability and reproducibility issues of some research findings (cite). On the other hand, the recent decade has witnessed a remarkable growth in “omics” and other forms of big data. As increasing amount of biomedical data has been made available, computational methods can offer a fast, cost-effective and unbiased way to prioritize promising drug targets. Given the limitation of current approaches and the urgent need to develop therapies for diseases, addressing the problem of target identification and drug development from different angles is essential. We believe that computational and experimental approaches can complement each other to improve the efficiency and reliability of drug target finding.

In this work, we present a novel computational framework to prioritize drug targets for specific diseases. Briefly, we first fit machine learning (ML) models to predict drug indications from drug-induced expression profiles. The aim is to “learn” expression patterns that are associated with successful treatment of the studied disease. The fitted ML models were then applied to transcriptome data derived from gene perturbations (i.e. over-expression or knock-down of specific genes). We could then prioritize drug targets based on the predicted probabilities from the ML model, which reflects treatment potential. Intuitively, for example over-expression (OE) of gene X leads to an expression profile similar to that of five other drugs that are known to treat diabetes. Then an agonist targeted at X (or other drugs that activate X and related pathways) may also be useful for treating diabetes.

In this case we expect the ML model (trained on drugs but applied to gene perturbation data) would output a high predicted probability for gene X, and it can be prioritized for further studies. Let’s consider an opposite scenario in which over-expression of gene Y increases the disease risk. In this case we may observe a lower-than-expected predicted probability of ‘treatment potential’ from the ML model. Here we emphasis more on the potential of drug target discovery of our computational framework, since high quality gene expression data that meets characteristics of our approach is still limited (cite).

□ **End of chapter.**

Chapter 5

Evaluating ITE of Genetic RFs on Survival

5.1 Motivation

Traditional biomedical or clinical studies in the area of estimating treatment effect mainly focus on the average effect of risk factors (RFs) or treatment (tx) in population level. However, in the clinical environment we can easily find that the same risk factor may affect patients differently. Thus, patients may pay more attention to how a risk factor will affect them in an individual level rather than in a population level, given their clinical backgrounds and genetic characteristics. The main aim of this study is to resolve this concern by estimating the ITEs for each patient, with consideration of their unique genetic and clinical information. In this study we treat the two terms ”risk factor” and ”treatment” conceptually equivalent, since a risk factor can be considered as a ”treatment” with adverse effects. The approach for estimating the ITEs for each patient allows us to offer tailored health management to individual patients. This enables us to deliver more cost-effective prevention or treatment strategies to benefit them the most. This idea is also in the line with ”personalized medicine”, which has been advocated in recent years.

In spite of an increasing number of studies in this area, current studies in ITE are rather limited. Some critical limitations include a lack of well-established validation methods for treatment effect estimations and the contribution of key features ITE estimation, and failure in incorporating censored data. Even though genetic factors may determine heterogenous response to tx/RFs, especially to cancer treatments [31], current studies on ITE have not included genomic features. Here we proposed several methodologies to address the above limitations and applied the ITE framework to genomic data. In our approach genomic features were considered as risk factors or covariates that contribute to the heterogeneity of treatment effect.

5.2 Background

It has been well-known that different individuals response differently to the same risk factor (RF) or treatment (tx). For example, even though obesity is a risk factor for cardiometabolic (CM) diseases, there still are obese subjects who don't develop related complications [78]. The type and severity of such CM complications can also show heterogeneity among subjects [78]. Another evidence is that not all people suffering stressful life events are affected by depression, even if stressful events are risk factors for depression [132]. This fact can also be applied to other RFs or treatments. The heterogenous effect can be contributed by different genetic and/or environmental factors of subjects, and these factors affect them differently. Here we would like to investigate the different treatment effect contributed by variants or mutations instead of clinical factors, since studies have shown that same variant/mutation can have varying outcomes on different subjects [1, 82, 120].

There are dramatic advances in the omics technology and a shape rising availability of biomedical data. However, current studies in cancer still mainly focus on one clinical/genetic RF at a time, without the consideration of presence of complex interactions among the subject's genetic and/or clinical factors.

One of most crucial concerns to patients is how a RF or treatment will affect them given their genetic and clinical information. However, current researches on this issue largely focus on the average treatment effect of RFs in population rather than individualized treatment effects.

Here we built a computational framework to unravel the individualized effects of RFs/treatment so that we can estimate the treatment effect for each individual with the incorporation of his/her genetic and/or clinical background. We also developed methods to discover genetic and/or clinical features contributing the most to the estimation. We employed our approaches to cancer data to estimate treatment effects of genetic changes (e.g. changes of expression level of risk genes, mutations, CNVs etc.) and other RFs on each individual's survival.

5.3 Overview of Related Work

5.3.1 Background Methods

Here we define notations for the clarification of following presentation. Let $\mathbf{X}^{n\times m}$ denotes the covariate variable, $\mathbf{Y}^n$ denotes outcome variable, and $\mathbf{W}^n$ denotes treatment variable. Given an observation i we denote its covariates as $\mathbf{x_i}^m$, the risk factor/tx status as w_i and outcome y_i. Here we restate that since a risk factor may also be considered as a "treatment" with adverse effects, methods for ITE estimation can also be employed to RFs. Assume that the outcome $\mathbf{Y}$ satisfy

$$\mathbf{Y} = f(\mathbf{X}, \mathbf{W}) + \varepsilon, \tag{5.1}$$

Where ε follows $N(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ is the covariance matrix.

Assumption assume $\mathbf{X}, \mathbf{W}, \mathbf{Y}$ fulfill unconfoundedness assumption (randomization conditional on the covariates),

$$\left[\mathbf{Y}_i(1), \mathbf{Y}_i(0)\right] \perp\!\!\!\perp \mathbf{W}_i \,|\, \mathbf{X}_i. \tag{5.2}$$

Under the unconfoundedness assumption 5.2 the key is to estimate the expected difference, the estimation of ITE, for each individual in response between treatment and control. The ITE for subject i is formulated as

$$\tau(\mathbf{x_i}) = \mu_1(\mathbf{x_i}) - \mu_0(\mathbf{x_i}), \tag{5.3}$$

with $\mu_1(x_i)$ and $\mu_0(x_i)$ defined as

$$\begin{aligned} \mu_1(x_i) &= \mathbb{E}(\mathbf{Y} = y_i | \mathbf{X} = \mathbf{x_i}, \mathbf{W} = w_i) = f(x_i, 1) \\ \mu_0(x_i) &= \mathbb{E}(\mathbf{Y} = y_i | \mathbf{X} = \mathbf{x_i}, \mathbf{W} = 1 - w_i) = f(x_i, 0) \end{aligned} \tag{5.4}$$

respectively, where $w_i = 1$ without the loss of generality. For a given subject i, y_i and w_i are scalars, and $\mathbf{x_i}$ is a vector of length m. Traditional machine learning method cannot handle this situation since they cannot capture the difference of ITE when the outcomes for RF/tx were absent or present.

A traditional solution to measure ITE is to estimate the difference of averaged outcome between treatments and controls in pre-specified subgroups [36] or subgroup defined by learning algorithms [108, 107, 6, 32]. Su et. al. employed interaction tree to iteratively searching subgroups based on treatment effect [108, 107]. Similarly, causal trees proposed by Athey and Imbens estimate the treatment effect at the leaves of the tree [6]. However, main drawbacks of the approach are that there is no ground truth for subgroup definition, and that the impediment of iteratively searching for subgroups present obvious treatment effect and reporting only the results for subgroups with extreme treatment effects to highlight heterogeneity may be highly spurious [5, 19]. In the high dimensional setting, it's still very challenging to divide subjects into appropriate subgroups [90]. It's the same case for genomic data.

Alternatively, a feasible approach is to use any supervised machine learning (SML) methods to fit $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$ simultaneously / separately and estimate the difference by putting them together. Specifically, one may fit a single model $\mu(\mathbf{x}, w)$ or separate models for the treated and control groups, and compute the different between $\mu(\mathbf{x}, w)$ and $\mu(\mathbf{x}, 1-w)$. Studies [70, 24] utilize different counterfactual random forests algorithms to estimates treatment effects by fitting separate random forests models to treatment and control groups. Several literature, including Green and Kern [41], Hill [47], and Hill and Su [46], has employed bayesian forest-based machine learning methods to estimate heterogeneous treatment effects. These studies utilize Bayesian additive regression tree (BART) method [16], and can obtain reliable intervals for treatment effects by MCMC sampling. For lasso-like methods for causal inference [53, 114], it's difficult to capture interactions, which may be naturally present in genomic data, in high-dimensional setting, in split of its simplicity and good ability in feature selection. A limitation of these studies is lack of formal statistical inference results [124]. Some other methods, like Meta learners and deep learning based, for ITE estimation could be found in [63, 56].

Here we are interested in methods with following characteristics: automatically select important features, well capture high interactions present, and have good asymptotic properties. Thus, methods such as causal forests [124] or GRF [7] are much more preferred. Causal forests [124] have been proposed with an objective to maximize the heterogeneity of $\tau(\mathbf{x})$. Recently, Athey et al. proposed an extension of causal forests, GRF [7], inspired by the R-learner proposed in [80]. Both of the two methods [124, 7] inherits the excellent capability of random forest in capturing complex interactions.

However, there are still substantial research gaps, including relative lack of methods for result validation, evaluation of key features contributing to ITE estimation and handling censored data. Here we proposed methods to address these key issues and pioneer new applications to genomic data, which is the first of its kind.

5.3.2 Causal Forests

In this section, we will explain causal forests (CF) technically, a basis of our ITE framework. CF [124] originate from random forests [13], which are related to kernel or nearest neighborhood methods. However, random forests differ in that they determine weights received by nearby observations in a data-driven way, and this characteristics is critical in high dimensional environment or the present of high order interactions among covariates [124]. This is the same case for CF.

Here we begin with causal trees (CT) [6] since CF are made of a number of CT. In this part we follow a similar notations as appeared in [6]. A tree can be considered as a partitioning of the feature space $\mathbb{X}$, denoting as Π. A partition Π with a number of elements $\#(\Pi)$ can be written as

$$\Pi = \{l_1, l_2, \ldots, l_{\#(\Pi)}\},$$

and a union of all elements in partition is the whole feature space $\mathbb{X}$.

Let $\mathbb{P}$ denote the space of partitions, and $\mathbb{S}$ be the space of samples from a population of observations. We seek for a algorithm $\pi : \mathbb{S} \rightarrow \mathbb{P}$ that splits sample space $\mathbb{S}$ into partition Π.

Given a partition Π and sample S, the estimated conditional mean for observation $\mathbf{x}$ is

$$\hat{\mu}(\mathbf{x}; \Pi) \equiv \frac{1}{\#(i \in \mathrm{S} : \mathbf{x}_i \in l(\mathbf{x}; \Pi))} \sum_{i \in \mathrm{S}: \mathbf{x}_i \in l(\mathbf{x}; \Pi)} y_i,$$

which is an unbiased estimator for $\mu(\mathbf{x}; \Pi)$. Here $l(\mathbf{x}; \Pi)$ is the leaf to which $\mathbf{x}$ belongs.

A adjusted MSE criteria including $\mathbb{E}[\mathbf{Y}_i^2]$, a term that does not depend on the estimator, is defined as

$$\mathrm{MSE}_{\mu}(\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}, \Pi) \equiv \frac{1}{\#(\mathrm{S}^{\mathrm{te}})} \sum_{i \in \mathrm{S}^{\mathrm{te}}} \left\{ (y_i - \hat{\mu}(\mathbf{x}_i; \mathrm{S}^{\mathrm{est}}, \Pi))^2 - y_i^2 \right\}. \tag{5.5}$$

The expectation of the modified MSE is

$$\mathrm{EMSE}_{\mu}(\Pi) \equiv \mathbb{E}_{\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}} \left[\mathrm{MSE}_{\mu}(\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}, \Pi) \right].$$

The objective of CT is to maximize the criteria

$$Q^{\mathrm{H}}(\pi) \equiv -\mathbb{E}_{\mathrm{S}^{\mathrm{te}},\mathrm{S}^{\mathrm{est}},\mathrm{S}^{\mathrm{tr}}}\left[\mathrm{MSE}_{\mu}(\mathrm{S}^{\mathrm{te}},\mathrm{S}^{\mathrm{est}},\pi(\mathrm{S}^{\mathrm{tr}}))\right]. \tag{5.6}$$

This criteria shows better convergence properties of confidence intervals, compared with conventional practice that $\mathrm{S}^{\mathrm{est}}$ and S^{tr} are the same sample for both tree construction and estimation [124].

With the above setup for treatment effect estimation, a similar definition to 5.5, is defined as

$$\mathrm{MSE}_{\tau}(\mathrm{S}^{\mathrm{te}},\mathrm{S}^{\mathrm{est}},\Pi) \equiv \frac{1}{\#(\mathrm{S}^{\mathrm{te}})}\sum_{i\in\mathrm{S}^{\mathrm{te}}}\left\{(\tau_i - \hat{\tau}(\mathbf{x}_i;\mathrm{S}^{\mathrm{est}},\Pi))^2 - \tau_i^2\right\}. \tag{5.7}$$

In reality, τ_i cannot be directly observed. The estimated counterparts are defined as

$$\hat{\tau}(\mathbf{x};\mathrm{S},\Pi) \equiv \hat{\mu}(1,\mathbf{x},\mathrm{S},\Pi) - \hat{\mu}(1,\mathbf{x};\mathrm{S},\Pi), \tag{5.8}$$

where

$$\hat{\mu}(w,\mathbf{x},\mathrm{S},\Pi) \equiv \frac{1}{\#(i\in\mathrm{S}_{\mathrm{w}}:\mathbf{x}_i\in l(\mathbf{x};\Pi))}\sum_{i\in\mathrm{S}_{\mathrm{w}}:\mathbf{x}_i\in l(\mathbf{x};\Pi)} y_i^{\mathrm{obs}}.$$

With the fact that $\hat{\tau}$ is constant within each leaf and the fact that

$$\mathbb{E}_{\mathrm{S}^{\mathrm{te}}}[\tau_i | i\in\mathrm{S}^{\mathrm{te}} : i\in l(\mathbf{x},\Pi)] = \mathbb{E}_{\mathrm{S}^{\mathrm{te}}}[\hat{\tau}(\mathbf{x};\mathrm{S}^{\mathrm{te}},\Pi)],$$

A crucial estimator for the infeasible in-sample goodness-of-fit criterion is derived as

$$-\mathrm{MSE}_{\tau}(\mathrm{S}^{\mathrm{tr}},\mathrm{S}^{\mathrm{tr}},\Pi) = \frac{1}{N^{\mathrm{tr}}}\sum_{i\in\mathrm{S}^{\mathrm{tr}}}\hat{\tau}^2(\mathbf{x}_i;\mathrm{S}^{\mathrm{tr}},\Pi). \tag{5.9}$$

Then this leads to an estimator for the criterion relying only on S^{tr} and N^{est}

$$\begin{aligned} -\mathrm{E\hat{M}SE}_{\tau}(\mathrm{S}^{\mathrm{tr}}, N^{\mathrm{est}}, \Pi) \equiv & \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathrm{S}^{\mathrm{tr}}} \hat{\tau}^2(\mathbf{x}_i; \mathrm{S}^{\mathrm{tr}}, \Pi) - \\ & (\frac{1}{N^{\mathrm{tr}}} + \frac{1}{N^{\mathrm{est}}}) \sum_{l \in \Pi} \left(\frac{\mathrm{S}^2_{\mathrm{S}^{\mathrm{tr}}_{\mathrm{tx}}}(l)}{p} + \frac{\mathrm{S}^2_{\mathrm{S}^{\mathrm{tr}}_{\mathrm{ctl}}}(l)}{1-p} \right) \end{aligned} \tag{5.10}$$

where the last term in the second line of 5.10 is pooled within-leaf variance. Definitely the splits of the tree are chosen to maximize the variance of $\tau(\mathbf{x}_i)$. For more detailed derivations, refer to original publication [6].

Briefly, the objective for splitting for the adaptive version of CT, denoted CT-A, uses $-\mathrm{MSE}_{\tau}(\mathrm{S}^{\mathrm{tr}}, \mathrm{S}^{\mathrm{tr}}, \Pi)$. The same objective function also is applicable to CV version of CT-A but evaluated at the samples $\mathrm{S}^{\mathrm{tr,cv}}$ and $\mathrm{S}^{\mathrm{tr,cv}}$. The splitting objective function for the honest CTS, CT-H, is $-\mathrm{MSE}_{\tau}(\mathrm{S}^{\mathrm{tr}}, N^{\mathrm{est}}, \Pi)$. The same objective function also can be applied to a CV version of CT-H, but evaluated at the cross-validation sample $\mathrm{S}^{\mathrm{tr,cv}}$ with known $N^{\mathrm{est,cv}}$.

Procedure for causal forests with honesty and subsampling is proposed as follows: In above table, CT stands for causal tree algorithm defined above in the section.

Algorithm 1 Causal forests with honesty and subsampling

```
Require: a samples $\mathrm{S}^{n \times m}$ and pre-specified parameters, including the number of trees B,
    mtry and the sub-sampling rate s
for all $i$ such that $0 \leq i \leq$ B do
    samples for model construction $\mathrm{S}^{\mathrm{tr,est}}_{\mathrm{i}} \leftarrow$ SUBSAMPLE (S, s), with remaining as test set $\mathrm{S}^{\mathrm{te}}_{\mathrm{i}}$
    training samples $\mathrm{S}^{\mathrm{tr}}_{\mathrm{i}}, \mathrm{S}^{\mathrm{est}}_{\mathrm{i}} \leftarrow$ SUBSAMPLE ($\mathrm{S}^{\mathrm{tr,est}}_{\mathrm{i}}$)
    causal tree $\mathrm{T}_{\mathrm{i}} \leftarrow \mathrm{CT}(\mathrm{S}^{\mathrm{tr}}_{\mathrm{i}}, \mathrm{S}^{\mathrm{est}}_{\mathrm{i}})$
    make out-of-bag prediction $\hat{\tau}(\mathbf{x}^i_j)$ for $\mathbf{x}^i_j \in \mathrm{S}^{\mathrm{te}}_{\mathrm{i}}$
end for
return CTs $\mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{T}_{\mathrm{i}}, \ldots, \mathrm{T}_{\mathrm{B}}$ and $\hat{\tau}(\mathbf{x}_j)$ by averaging all available out-of-bag predictions $\tau^{(.)}_j$ for $\mathbf{x}_j \in S$.
```

5.4 ITE Framework

In order to assess the ITE of RFs on disease outcome and discover key features that contribute to estimation of the heterogeneity we proposed an analytic framework to estimate the ITE of RFs/tx, with genetic features as primary RFs and/or covariates. The term 'individualized treatment effect' (ITE) will be used regardless of an RF or treatment being considered, since we have declared that the two are conceptually equivalent entities in this study.

The estimated ITE may be formulated as 5.4 under a counterfactual outcomes framework [94]. In reality the true value of $\tau(\mathbf{x})$ cannot be directly observed, as we only have one of the two potential outcomes. In observational studies, the tx assignment may be associated with potential outcomes due to confounding variables. If the unconfoundedness assumption 5.2 satisfies then the causal ITE can still be captured. Then, in most observational studies, the study is still of significance in despite of the presence of residual confounding.

In that case we may still gain insights into TE heterogeneity at an association level, and covariates responsible for heterogeneity may still deserve further studies, and in practice a continuous variable $\mathbf{W}$ is also allowed [7]. When $\mathbf{W}$ is continuous, an average partial effect is estimated $Cov[\mathbf{Y}, \mathbf{W} \mid \mathbf{X} = \mathbf{x}]/Var[\mathbf{W} \mid \mathbf{X} = \mathbf{x}]$, which may be considered as the increase in Y given a unit increase in W, conditional on the covariates.

Our experiment studies rely on the framework of GRF, as it is a state-of-the-art approach which directly optimizes an objective function for ITE estimation, rather than fit conditional means for treatment and control observations. However, most of the methodologies and extensions presented in this study is data-driven such that it can be widely applied to any other ITE estimation models.

5.4.1 Novel Tests for the presence of heterogeneity

Unlike a SML model, an ITE model is not straightforward since the actual TE is not directly observed. There is no either empirically or theoretically well-established methods in ITE validation. We notice that the function `test_calibration` provided in the R package `grf` [115], and the idea behind it is borrowed from [15]. Thus, we would compare our methods with it in simulations in future section 5.5.

We would propose several novel statistical tests for ITE model evaluation:

Split-half correlation with multiple splits This method borrows the idea of cross validation (CV) in SML. Briefly, we proposed to split the dataset into 2 halves. Specifically, an ITE model is first fitted on the 1st half, then applied to the 2nd half to predict $\tau(\mathbf{x})$. The process can be repeated by reversing the training and testing sets. Then for each half, we have out-of-bag predictions $\tau_{\text{oob}}(\mathbf{x})$ and predictions $\tau_{\text{pred}}(\mathbf{x})$ by applying models fitted on the other half to it. We can assess the model fitting by examining the correlation between the two $\tau(\mathbf{x})$s.

The rational behind this statistical test for the presence of heterogeneity is that we expect the ITE to vary randomly around a constant in the absence of heterogenous treatment effects that can be explained by covariates, and hence the correlation between the two estimated τ values should be close to 0; otherwise, the stability and generality of ITE models can be examined by checking the replicability on an independent dataset using split-half correlation.

In order to reduce random variations that may be introduced by a single split, we performed split-half correlation test with multiple splits on the data in practice and combined the results together. Standard Simes test [100] may be an options for the combination, since it is robust to positive dependency of p-values. Its alternative hypothesis (H1) is that at least one of the hypotheses is non-null (at least one out of n splits yields a positive significant correlation), so it may be relatively loose for our problem. To increase stringency, we introduced a partial conjunction test (partial Simes test) based on the work from Benjamini et al. [10], whose H1 assumes that at least r out of the n splits yield a positive significant correlation.

The threshold r defines the stringency/level of consistency for a finding that deserves further study. In experiment we set r to be 10% of n, which can give adequate type I error control, and employed 3 correlation measures (Pearson, Spearman and Kendall) to evaluate the split-half correlation.

A new permutation framework We proposed a novel permutation statistical test to assess the presence of heterogeneity. That is, it can be used to assess whether the predicted ITE are significantly better than predictions assuming a constant TE (which is the norm in most studies).

The objective of CF is to maximize $\widehat{\text{Var}}(\tau)$, as stated in section 5.3.2. When there is no heterogeneity that can be explained by covariates, $\widehat{\text{Var}}(\tau)$ should be low and close to 0. Equivalently, in this situation $\widehat{\text{Var}}(\tau)$ is roughly equal to $\widehat{\text{Var}}(\tau)$ yielded by model fitted on arbitrary covariates. Thus, a permutation approach we proposed is based on this rational to test the significance of $\widehat{\text{Var}}(\tau)$ observed.

To model the null hypothesis we shuffled the covariates for each permutation, such that there is no heterogeneity can be explained by covariates, and then computed the $\widehat{\text{Var}}_{\text{observed}}$ for the permuted data.

If we repeat this process N times, then a probability for the null hypothesis of $\widehat{\text{Var}}(\tau)$ statistical test is defined as

$$\Pr(\text{null}_{\text{var}}) = \frac{\#(\widehat{\text{Var}}_{\text{perm}}(\hat{\tau}) \geq \widehat{\text{Var}}_{\text{observed}}(\hat{\tau})}{N}, \tag{5.11}$$

A related but clinically relevant question is: whether the model that allows ITE outperforms that predicting a constant ITE? That is, whether patients benefit more with the introduction of individualized treatments than a conventional treatments with a consideration of averaged treatment effects only. In ordinary regression problem, the goodness-of-fit of model is assessed by the mean squared error (MSE) between the expected outcome and predictions, so it's preferred to compute the mean squared error (MSE) between the true and estimated τ for model assessment. If ITE model has a lower MSE than a constant model, which assumes that the ITE is the same for every subject, then ITE model outperforms the constant one. In reality, the true $\tau(x)$ cannot be directly observed, but Nie et al. [80] proposed that the MSE between the true and estimated $\tau(\mathbf{x})$, or $\tau_{\text{risk}}(\mathbf{x})$ can be defined as

$$\begin{aligned}\hat{\tau}_{\text{risk}} &= \sum_i \Big((y_i - \hat{y}(\mathbf{x}_i) - \big(w_i - \hat{w}(\mathbf{x}_i)\big)\,\hat{\tau}(\mathbf{x}_i) \Big)^2 \\ &= \sum_i \Big(\tilde{y}_i - \tilde{w}_i \hat{\tau}(\mathbf{x}_i) \Big)^2 .\end{aligned} \tag{5.12}$$

Here $\hat{y}(\mathbf{x}_i)$ is an estimate of $\mathbb{E}(y_i|\mathbf{X} = \mathbf{x}_i)$ and $\hat{w}(\mathbf{x}_i)$ is an estimate of $\Pr(w_i = 1\,|\,\mathbf{X} = \mathbf{x}_i)$ by any SML models. For simplicity, $\tilde{y}_i$ stands for $y_i - \hat{y}(\mathbf{x}_i)$, and $\tilde{w}_i(\mathbf{x}_i)$ for $w_i - \hat{w}(\mathbf{x}_i)$. The two terms can be regarded as residualized outcome and treatment respectively. These quantities are out-of-bag estimations. We can then compute the τ_{risk} assuming an unrestricted $\hat{\tau}(\mathbf{x})$ estimated from an ITE model and a constant effect based on the **average treatment effect (ATE)** $\bar{\tau}(\mathbf{x})$. We proposed the following definition to assess the improvement in τ_{risk} due to the incorporation of ITE versus a constant treatment effect

$$\hat{\tau}_{\text{improve}} = \sum_i (\tilde{y} - \tilde{w}\tau(\hat{x_i}))^2 - \sum_i (\tilde{y} - \tilde{w}\tau(\bar{x_i}))^2. \tag{5.13}$$

To model the null distribution of $\hat{\tau}_{\text{improve}}$, we propose a permutation approach in which permuted $\hat{\tau}_{\text{improve}}$s are obtained by shuffling the covariates for a predefined number of times. The null hypothesis of above test is that there is no statistical difference between the observed $\hat{\tau}_{\text{improve}}$ and permuted $\hat{\tau}_{\text{improve}}$s. Note that the test does not require any distributional assumptions.

An adaptive permutation strategy with early stopping was adopted to reduce the computing time. Permutations will be stopped earlier if the result is unlikely to be significant in future runs. Specifically, we will calculate a 99% CI for the permutation p-value after each $k(k \ll N)$ runs. Here N is the total number permutations. If the lower CI > 0.05, the permutation will be terminated early.

5.4.2 Modeling Survival Outcomes

Time-to-event data are common in biomedical research, and standard ITE estimation methods may not work on survival data due to censoring or lost to follow-up. Usually some subjects have not experienced the event at the end of follow-up, that is, their records of survival time are unavailable (right censored), so the actual survival time for them is unknown. We proposed a flexible approach that can incorporate survival data into GRF and any other ITE models, an approach based on weighted 'mean imputation'.

Given a subject, denote its actual survival time as T_i and its censor time as c_i. In reality we observe y_i which is $\min(T_i, c_i)$ due to censoring. Let $t_{(1)} < t_{(2)} < \cdots < t_{(j)}$ be the censored survival times in ascending order, and $\hat{K}$ be the Kaplan-Meier (KP) estimator function of survival. Given $T_i > c_i$ for subject i, its log of censored survival times can be estimated by

$$\log(y_i^*) = \sum_{t(j)>T_i^c} \log t(j) \frac{\Delta \hat{T}(t(j))}{\hat{T}(T_i^c)}, \tag{5.14}$$

where $\Delta \hat{T}(t(j))$ refers to to the jump size of $\hat{K}$ at $t_{(j)}$ [25]. Under the assumption of the log-normal distribution of survival time, the imputed survival times can be included in ITE estimation.

5.5 Experiment Results

5.5.1 Simulations Studies

In design of simulations to evaluate the performance of our ITE framework and to compare the power and type I error rate of our proposed statistical tests with that of `test_calibration` provided in the R package `grf`, we adopted similar strategies in the generation of synthetic data as introduced in [90]. Here we assume the log survival time is normally distributed without the loss of generality. We considered the following six elements in simulations design:

1. *Sample size* The number of observations in the data is n, and p is the number of covariates.

2. *Distributions of covariates* Across all simulation scenarios, we drawn samples from standard normal distribution for features with odd column number, and for features with even column number we sampled from a Bernoulli distribution with $p = 1/2$. For simplicity, we denotes the distribution for the policy as D_x.

3. *Key functions* we denote propensity function for observations receiving a treatment $\pi(\cdot)$, average treatment effect $\mu(\cdot)$, and treatment effect $\tau(\cdot)$. The conditional mean effect for treatments and controls can be designed to be $\mu_1(\cdot) = \mu(\cdot) + \tau(\cdot)/2$ and $\mu_0(\cdot) = \mu(\cdot) - \tau(\cdot)/2$ respectively.

4. *Generation of survival time* Under the log-normal distribution of survival time, the survival time of observation can be generated by taking the natural exponentiation of the mean effect using $\exp(\cdot)$.

5. *Censor* We considered a censor rate of roughly 20% for all our scenarios. Given the uncensored simulated survival time, we found the cutoff r for the 80% quantile, and then generated censor time $T(\cdot)$ using the exponential distribution with rate parameter $1/r$. If the simulated $\log Y_i > T(\mathbf{x}_i)$, then $T(\mathbf{x}_i)$ should be used as outcome; otherwise $\log Y_i$ was used.

6. *Noise levels* The noise level $\sigma^2_{\log(Y)}$ was introduced in the generation of log survival time $\log Y_i$, which sampled from a normal distribution with mean $\mu(\cdot) + (w - 1/2)\tau(\cdot)$ and variance $\sigma^2_{\log(Y)}$, where $w \sim \text{Bernoulli}(\pi(\cdot))$.

Given the above predefined components, our data generation for observations i is modeled as

$$\begin{aligned}
\mathbf{x}_i &\sim D_x, \\
W_i &\sim \text{Bernoulli}\big(\pi(\mathbf{x}_i)\big), \\
\log Y_i &\sim \text{Normal}\big(\mu(\mathbf{x}_i) + (W_i - 1/2)\tau(\mathbf{x}_i), \sigma^2_{\log(Y)}\big), \\
T_i &\sim \text{Exp}(1/r), \\
c_i &= \begin{cases} 1, & \text{if } \log Y_i \le T_i \\ 0, & \text{otherwise} \end{cases}, \\
Y_i &= \exp(\min(\log Y_i, T_i)),
\end{aligned}$$

where r is a cutoff corresponding to a specific quantile of $\log Y_i$, and $\pi(\mathbf{x}_i)$ is defined as

$$\pi(\mathbf{x}_i) = \frac{\exp(\mu(\mathbf{x}_i) - \tau(\mathbf{x})/2)}{1 + \exp(\mu(\mathbf{x}_i) - \tau(\mathbf{x})/2)} \tag{5.15}$$

for observational studies; for randomized studies $\pi(\mathbf{x}_i) = 1/2$ for all $\mathbf{x}_i$, which is the same as defined in [90]. In practice, the value for 0.8 quantile was used. c_i is an indicator for censor. If there is an event ($\log Y_i \le T_i$) for subject i, then c_i is 1; otherwise 0.

We used functions with minor changes from [90] for propensity probability of receiving a treatment $\pi(\cdot)$, average treatment effect $\mu(\cdot)$, and treatment effect $\tau(\cdot)$. Within the simulation experiments, both randomized and observational studies are included, and 8 different functions of mean and treatment effects are made here to represent both univariate and multivariate, both additive and interactive, and both linear and piecewise constant relationships. They are defined as follows:

$$
\begin{aligned}
&f_1(x) = 0, f_2(x) = 5\mathbb{I}(x_1 > 1) - 5 * pnorm(-1), f_3(x) = 5x_1,\\
&f_4(x) = x_2x_4x_6 + 2x_2x_4(1-x_6) + 3x_2(1-x_4)x_6\\
&\qquad + 4x_2(1-x_4)(1-x_6) + 5(1-x_2)x_4x_6 + 6(1-x_2)x_4(1-x_6)\\
&\qquad + 7(1-x_2)(1-x_4)x_6 + 8(1-x_2)(1-x_4)(1-x_6) - 4.5,\\
&f_5(x) = x_1 + x_3 + x_5 + x_7 + x_8 + x_9,\\
&f_6(x) = 4\mathbb{I}(x_1 > 1)\mathbb{I}(x_3 > 0) + 4\mathbb{I}(x_5 > 1)\mathbb{I}(x_7 > 0) + 2x_8x_9 - 4 * pnorm(-1),\\
&f_7(x) = \frac{1}{\sqrt{2}}(x_1^3 + x_2 + x_3^3 + x_4 + x_5^3 + x_6 + x_7^3 + x_8 + x_9^3 - 7)\\
&f_8(x) = \frac{1}{2}(f_4(x) + f_5(x)),
\end{aligned}
$$

where $pnorm(x)$ is a function that calculates the cumulative distribution probability $F(x) = \Pr(X \leq x)$. Here X is with standard normal distribution.

Table 5.1: Specifications for simulation scenarios

	Scenarios							
	1,9	2,10	3,11	4,12	5,13	6,14	7,15	8,16
n	300	300	200	600	400	300	450	700
p	400	400	300	300	200	200	100	100
$\mu(x)$	$f_8(x)$	$f_5(x)$	$f_4(x)$	$f_7(x)$	$f_3(x)$	$f_1(x)$	$f_2(x)$	$f_6(x)$
$\tau(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_3(x)$	$f_8(x)$
$\sigma^2_{\log Y}$	1	1/4	1	1/4	1	1	4	4

The 8 functions listed are centered and scaled to have mean close to 0 and roughly the same variance. Table 5.1 gives the specifications for simulation scenarios, including sample size n, number of features p, functions for mean and treatment effect $\mu(\cdot)$ and $\tau(\cdot)$, and variance of noise $\sigma^2_{\log Y}$. Specifications for sample size have been adjusted to accommodate our simulated survival data, compared with those in study [90]. Scenarios of odd number are randomized experiments, with $\pi(\mathbf{x}_i) = 1/2$ for all $\mathbf{x}_i$, but scenarios of even number are observational studies, in which $\pi(\mathbf{x}_i)$ is defined by equation 5.15 for each subject $\mathbf{x}_i$. Note there is no heterogeneity in treatment effects in scenario 1 and 9.

Table 5.2: Comparison of power/type I error rate of different tests for the presence of heterogeneity

Scenarios	SHC-P	SHC-K	SHC-S	TC	$\widehat{\text{Var}}_\tau$	τ_{improve}
1*	0.002	0.002	0.002	0.002	0.058	0.054
2	0.266	0.268	0.26	0.896	0.976	0.98
3	1	1	1	1	1	1
4	0.656	0.652	0.652	0.782	0.924	0.924
5	0.548	0.548	0.546	0.752	0.926	0.934
6	0.888	0.88	0.882	0.944	0.992	0.992
7	0.862	0.844	0.85	0.886	0.956	0.974
8	0.752	0.732	0.73	0.42	0.592	0.664
9*	0.004	0.004	0.004	0.002	0.054	0.068
10	0.162	0.164	0.164	0.652	0.856	0.88
11	0.672	0.666	0.672	1	1	1
12	0.904	0.908	0.908	0.92	0.968	0.986
13	0.636	0.628	0.632	0.742	0.924	0.934
14	0.6	0.6	0.598	0.74	0.938	0.95
15	0.784	0.774	0.774	0.604	0.85	0.87
16	0.99	0.986	0.988	0.956	0.978	0.988

1. SHC-P stands for split-half correlation with pearson method, SHC-K for split-half correlation with kendall method, and SHC-S for split-half correlation with kendall method method for correlation testing.
2. TC represents `test_calibration` from the R package `GRF`.
3. $\widehat{\text{Var}}_\tau$ and τ_{improve} stand for two statistical tests for the presence of heterogeneity using them, proposed in section 5.4.1.
4. Scenario 1 and 9 are masked with star, since they are two scenarios without heterogeneity.

To explore the reliability of our developed statistical tests and compare them with the `test_calibration` (TC) from GRF, we repeated application of our ITE framework with different random seed to the above simulation scenarios 500 times and examined the fitting of our approach with the statistical tests and test_calibration for each run. We calculated the proportion of repeats with p-values ≤ 0.05 for every statistical test. Because of no heterogeneity in scenarios 1 and 9, the proportion for statistical tests in scenarios 1 and 9 are type I error rate. The proportion for other simulation scenarios is the power of statistical methods. Simulation results are shown in table 5.2.

Statistical tests with $\widehat{\text{Var}}_\tau$ and τ_{improve} and TC maintain good validity. They all have very strong power in capturing the presence of heterogeneity, and notably our methods with $\widehat{\text{Var}}_\tau$ and τ_{improve} completely dominate TC provided in package GRF cross all simulation scenarios except 1 and 9. In scenarios 1 and 9 they all show relatively low type I error rate, with roughly 0.05 for statistical tests with $\widehat{\text{Var}}_\tau$ and τ_{improve} and 0.002 for TC. Type I error rates for our methods are also within an acceptable range, and TC has lower type I error rate than our methods. However, our methods with $\widehat{\text{Var}}_\tau$ and τ_{improve} are more favored in detecting the presence of heterogeneity, even though they shows slight inflated type I error rate.

Split-half correlation (SHC) approaches with three different correlation evaluation methods show good type I error rates in scenario 1 and 9. However, their powers varies greatly cross all other scenarios. By investigating their performance in scenario 6, 7, 12, and 16, we found that they also can maintain good performance if there is strong heterogeneity present. Surprisingly, they apparently outperform the other three methods in Scenario 8. In short, we can still consider SHC approaches as good supplements to the other three methods.

5.5.2 Applications on Real Data

5.6 Conclusion

Bibliography data is put in database.bib.

□ **End of chapter.**

Chapter 6

Conclusion

Appendix A

Proof of Propositions

□ **End of chapter.**

Appendix B

Publication List

□ **End of chapter.**

Bibliography

[1] C. Aggarwal, C. W. Davis, R. Mick, J. C. Thompson, S. Ahmed, S. Jeffries, S. Bagley, P. Gabriel, T. L. Evans, J. M. Bauml, et al. Influence of tp53 mutation on survival in patients with advanced egfr-mutant non–small-cell lung cancer. *JCO precision oncology*, 2018.

[2] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, and A. Zhavoronkov. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7):2524–2530, 2016.

[3] J. D. Amsterdam, Y. Li, I. Soeller, K. Rockwell, J. J. Mao, and J. Shults. A randomized, double-blind, placebo-controlled trial of oral matricaria recutita (chamomile) extract therapy of generalized anxiety disorder. *Journal of clinical psychopharmacology*, 29(4):378, 2009.

[4] J. D. Amsterdam, J. Shults, I. Soeller, J. J. Mao, K. Rockwell, and A. B. Newberg. Chamomile (matricaria recutita) may have antidepressant activity in anxious depressed humans-an exploratory study. *Alternative therapies in health and medicine*, 18(5):44, 2012.

[5] S. F. Assmann, S. J. Pocock, L. E. Enos, and L. E. Kasten. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069, 2000.

[6] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[7] S. Athey, J. Tibshirani, S. Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[8] J. C. Ballenger. Anxiety and depression: optimizing treatments. *Primary care companion to the Journal of clinical psychiatry*, 2(3):71, 2000.

[9] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[10] Y. Benjamini and R. Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008.

[11] J. Bergstra, D. Yamins, and D. D. Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, pages 13–20. Citeseer, 2013.

[12] I. Berman, B. L. Sapers, H. H. Chang, M. F. Losonczy, J. Schmildler, and A. I. Green. Treatment of obsessive-compulsive symptoms in schizophrenic patients with clomipramine. *Journal of clinical psychopharmacology*, 15(3):206–210, 1995.

[13] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[14] J. Burgdorf, X.-l. Zhang, E. M. Colechio, N. Ghoreishi-Haack, A. Gross, R. A. Kroes, P. K. Stanton, and J. R. Moskal. Insulin-like growth factor i produces an antidepressant-like effect and elicits n-methyl-d-aspartate receptor independent long-term potentiation of synaptic transmission in medial prefrontal cortex and hippocampus. *International Journal of Neuropsychopharmacology*, 19(2):pyv101, 2016.

[15] V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018.

[16] H. A. Chipman, E. I. George, R. E. McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[17] R. Chou, T. Dana, I. Blazina, M. Daeges, and T. L. Jeanne. Statins for prevention of cardiovascular disease in adults: evidence report and systematic review for the us preventive services task force. *Jama*, 316(19):2008–2024, 2016.

[18] A. Cipriani, K. Saunders, M.-J. Attenburrow, J. Stefaniak, P. Panchal, S. Stockton, T. A. Lane, E. M. Tunbridge, J. R. Geddes, and P. J. Harrison. A systematic review of calcium channel antagonists in bipolar disorder and some considerations for their future development. *Molecular psychiatry*, 21(10):1324–1332, 2016.

[19] D. I. Cook, V. J. Gebski, and A. C. Keech. Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6):289, 2004.

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[21] H. E. Covington, I. Maze, Q. C. LaPlant, V. F. Vialou, Y. N. Ohnishi, O. Berton, D. M. Fass, W. Renthal, A. J. Rush, E. Y. Wu, et al. Antidepressant actions

of histone deacetylase inhibitors. *Journal of Neuroscience*, 29(37):11451–11460, 2009.

[22] M. da Graça Cantarelli, A. C. Tramontina, M. C. Leite, and C.-A. Goncalves. Potential neurochemical links between cholesterol and suicidal behavior. *Psychiatry research*, 220(3):745–751, 2014.

[23] F. Dabaghzadeh, P. Ghaeli, H. Khalili, A. Alimadadi, S. Jafari, S. Akhondzadeh, and Z. Khazaeipour. Cyproheptadine for prevention of neuropsychiatric adverse effects of efavirenz: a randomized clinical trial. *AIDS patient care and STDs*, 27(3):146–154, 2013.

[24] A. Dasgupta, S. Szymczak, J. H. Moore, J. E. Bailey-Wilson, and J. D. Malley. Risk estimation using probability machines. *BioData mining*, 7(1):2, 2014.

[25] S. Datta, J. Le-Rademacher, and S. Datta. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics*, 63(1):259–271, 2007.

[26] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[27] S. de Jong, L. R. Vidler, Y. Mokrab, D. A. Collier, and G. Breen. Gene-set analysis based on the pharmacological profiles of drugs to identify repurposing opportunities in schizophrenia. *Journal of psychopharmacology*, 30(8):826–830, 2016.

[28] J. T. Dudley, T. Deshpande, and A. J. Butte. Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, 12(4):303–311, 2011.

[29] B. A. Ellenbroek and E. P. Prinssen. Can 5-ht3 antagonists contribute toward the treatment of schizophrenia? *Behavioural pharmacology*, 26(1 and 2-Special Issue):33–44, 2015.

[30] J. Ferno, S. Skrede, A. O. Vik-Mo, G. Jassim, S. Le Hellard, and V. M. Steen. Lipogenic effects of psychotropic drugs: focus on the srebp system. *Front Biosci (Landmark Ed)*, 16(1):49–60, 2011.

[31] R. Fisher, L. Pusztai, and C. Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.

[32] J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[33] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[34] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[35] M. Fuchikami, S. Yamamoto, S. Morinobu, S. Okada, Y. Yamawaki, and S. Yamawaki. The potential use of histone deacetylase inhibitors in the treatment of depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 64:320–324, 2016.

[36] M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, pages 361–372, 1985.

[37] S. Gao, Y.-L. Cui, C.-Q. Yu, Q.-S. Wang, and Y. Zhang. Tetrandrine exerts antidepressant-like effects in animal models: role of brain-derived neurotrophic factor. *Behavioural brain research*, 238:79–85, 2013.

[38] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. Deep learning. vol. 1, 2016.

[39] A. B. Goodman. Three independent lines of evidence suggest retinoids as causal to schizophrenia. *Proceedings of the National Academy of Sciences*, 95(13):7240–7244, 1998.

[40] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1), 2011.

[41] D. P. Green and H. L. Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

[42] N. Grunbaum-Novak, M. Taler, I. Gil-Ad, A. Weizman, H. Cohen, and R. Weizman. Relationship between antidepressants and igf-1 system in the brain: possible role in cognition. *European Neuropsychopharmacology*, 18(6):431–438, 2008.

[43] M. Guo, J. Mi, Q.-M. Jiang, J.-M. Xu, Y.-Y. Tang, G. Tian, and B. Wang. Metformin may produce antidepressant effects through improvement of cognitive function among depressed patients with diabetes mellitus. *Clinical and experimental pharmacology and physiology*, 41(9):650–656, 2014.

[44] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[45] S. Heyes, W. S. Pratt, E. Rees, S. Dahimene, L. Ferron, M. J. Owen, and A. C. Dolphin. Genetic disruption of voltage-gated calcium channels in psychiatric and neurological disorders. *Progress in neurobiology*, 134:36–54, 2015.

[46] J. Hill and Y.-S. Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

[47] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[48] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8), 2012.

[49] T. Hobara, S. Uchida, K. Otsuki, H. Yamagata, and Y. Watanabe. Molecular mechanisms of the antidepressant actions by histone deacetylase inhibitors. *Neuroscience Research*, (68):e316, 2010.

[50] R. A. Hodos, B. A. Kidd, K. Shameer, B. P. Readhead, and J. T. Dudley. In silico methods for drug repurposing and pharmacology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(3):186–210, 2016.

[51] L. Huang, F. Li, J. Sheng, X. Xia, J. Ma, M. Zhan, and S. T. Wong. Drugcomboranker: drug combination discovery based on target network analysis. *Bioinformatics*, 30(12):i228–i236, 2014.

[52] S. E. Hyman. Psychiatric drug development: diagnosing a crisis. In *Cerebrum: the Dana forum on brain science*, volume 2013. Dana Foundation, 2013.

[53] K. Imai, M. Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

[54] N. Iranpour, A. Zandifar, M. Farokhnia, A. Goguol, H. Yekehtaz, M.-R. Khodaie-Ardakani, B. Salehi, S. Esalatmanesh, A. Zeionoddini, P. Mohammadinejad, et al. The effects of pioglitazone adjuvant therapy on negative symptoms of patients with chronic schizophrenia: a double-blind and placebo-controlled trial. *Human Psychopharmacology: Clinical and Experimental*, 31(2):103–112, 2016.

[55] H. Jahn, M. Schick, F. Kiefer, M. Kellner, A. Yassouridis, and K. Wiedemann. Metyrapone as additive treatment in major depression: a double-blind and placebo-controlled trial. *Archives of general psychiatry*, 61(12):1235–1244, 2004.

[56] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[57] N. Katalinic, R. Lai, A. Somogyi, P. B. Mitchell, P. Glue, and C. K. Loo. Ketamine as a new treatment for depression: a review of its efficacy and adverse effects. *Australian & New Zealand Journal of Psychiatry*, 47(8):710–727, 2013.

[58] R. C. Kessler, N. A. Sampson, P. Berglund, M. Gruber, A. Al-Hamzawi, L. Andrade, B. Bunting, K. Demyttenaere, S. Florescu, G. De Girolamo, et al. Anxious and non-anxious major depressive disorder in the world health organization world mental health surveys. *Epidemiology and psychiatric sciences*, 24(3):210–226, 2015.

[59] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[60] M. Krakowski and P. Czobor. Cholesterol and cognition in schizophrenia: a double-blind study of patients randomized to clozapine, olanzapine and haloperidol. *Schizophrenia research*, 130(1-3):27–33, 2011.

[61] J. H. Krystal and M. W. State. Psychiatric disorders: diagnosis to therapy. *Cell*, 157(1):201–214, 2014.

[62] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

[63] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[64] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.

[65] V. Lerner, P. J. McCaffery, and M. S. Ritsner. Targeting retinoid receptors to treat schizophrenia: rationale and progress to date. *CNS drugs*, 30(4):269–280, 2016.

[66] J. Li and Z. Lu. A new method for computational drug repositioning using drug pairwise similarity. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–4. IEEE, 2012.

[67] Z. Liu, F. Guo, J. Gu, Y. Wang, Y. Li, D. Wang, L. Lu, D. Li, and F. He. Similarity-based prediction for anatomical therapeutic chemical classification of drugs by integrating multiple data sources. *Bioinformatics*, 31(11):1788–1795, 2015.

[68] M. Lotfi Shahreza, N. Ghadiri, S. R. Mousavi, J. Varshosaz, and J. R. Green. A review of network-based approaches to drug repositioning. *Briefings in bioinformatics*, 19(5):878–892, 2018.

[69] J.-M. Lü, J. Nurko, S. M. Weakley, J. Jiang, P. Kougias, P. H. Lin, Q. Yao, and C. Chen. Molecular mechanisms and clinical applications of nordihydroguaiaretic acid (ndga) and its derivatives: an update. *Medical science monitor: international medical journal of experimental and clinical research*, 16(5):RA93, 2010.

[70] M. Lu, S. Sadiq, D. J. Feaster, and H. Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

[71] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F.-X. Wu, and Y. Pan. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, 32(17):2664–2671, 2016.

[72] R. d. C. Maia, R. Tesch, and C. A. M. Fraga. Phenylpiperazine derivatives: a patent review (2006–present). *Expert opinion on therapeutic patents*, 22(10):1169–1178, 2012.

[73] G. S. Malhi, M. Tanious, P. Das, C. M. Coulston, and M. Berk. Potential mechanisms of action of lithium in bipolar disorder. *CNS drugs*, 27(2):135–153, 2013.

[74] Y.-M. Mao and M.-D. Zhang. Augmentation with antidepressants in schizophrenia treatment: benefit or risk. *Neuropsychiatric disease and treatment*, 11:701, 2015.

[75] S. Markey, J. Johannessen, C. Chiueh, R. Burns, and M. Herkenham. Intraneuronal generation of a pyridinium metabolite may cause drug-induced parkinsonism. *Nature*, 311(5985):464–467, 1984.

[76] R. H. McAllister-Williams, I. M. Anderson, A. Finkelmeyer, P. Gallagher, H. C. Grunze, P. M. Haddad, T. Hughes, A. J. Lloyd, C. Mamasoula, E. McColl, et al. Antidepressant augmentation with metyrapone for treatment-resistant depression (the add study): a double-blind, randomised, placebo-controlled trial. *The Lancet Psychiatry*, 3(2):117–127, 2016.

[77] F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. D'Amato, and D. Greco. Drug repositioning: a machine-learning approach through data integration. *Journal of cheminformatics*, 5(1):30, 2013.

[78] I. J. Neeland, P. Poirier, and J.-P. Després. Cardiovascular and metabolic heterogeneity of obesity: clinical challenges and implications for management. *Circulation*, 137(13):1391–1406, 2018.

[79] E. J. Nestler and S. E. Hyman. Animal models of neuropsychiatric disorders. *Nature neuroscience*, 13(10):1161, 2010.

[80] X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

[81] M. Oh, J. Ahn, and Y. Yoon. A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions. *PloS one*, 9(10), 2014.

[82] W. K. O'Neal and M. R. Knowles. Cystic fibrosis disease modifiers: complex genetics defines the phenotypic diversity in a monogenic disease. *Annual review of genomics and human genetics*, 19:201–222, 2018.

[83] T. Otowa, K. Hek, M. Lee, E. M. Byrne, S. S. Mirza, M. G. Nivard, T. Bigdeli, S. H. Aggen, D. Adkins, A. Wolen, et al. Meta-analysis of genome-wide association studies of anxiety disorders. *Molecular psychiatry*, 21(10):1391–1399, 2016.

[84] F. Pammolli, L. Magazzini, and M. Riccaboni. The productivity crisis in pharmaceutical r&d. *Nature reviews Drug discovery*, 10(6):428–438, 2011.

[85] L. Pathak, Y. Agrawal, and A. Dhir. Natural polyphenols in the management of major depression. *Expert opinion on investigational drugs*, 22(7):863–880, 2013.

[86] S. D. Pepper, E. K. Saunders, L. E. Edwards, C. L. Wilson, and C. J. Miller. The utility of mas5 expression summary and detection call algorithms. *BMC bioinformatics*, 8(1):273, 2007.

[87] J. E. Persons, J. G. Robinson, W. H. Coryell, M. E. Payne, and J. G. Fiedorowicz. Longitudinal study of low serum ldl and depressive symptom onset in postmenopause. *The Journal of clinical psychiatry*, 77(2):212, 2016.

[88] Y.-P. Phoebe Chen and F. Chen. Identifying targets for drug discovery using bioinformatics. *Expert Opinion on Therapeutic Targets*, 12(4):383–389, 2008.

[89] T. R. Powell, T. Murphy, S. H. Lee, J. Price, S. Thuret, and G. Breen. Transcriptomic profiling of human hippocampal progenitor cells treated with antidepressants and its application in drug repositioning. *Journal of Psychopharmacology*, 31(3):338–345, 2017.

[90] S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani. Some methods for heterogeneous treatment effect estimation in high-dimensions. *arXiv preprint arXiv:1707.00102*, 2017.

[91] M. B. Raeder, J. Fernø, A. O. Vik-Mo, and V. M. Steen. Srebp activation by antipsychotic-and antidepressant-drugs in cultured human liver cells: relevance for metabolic side-effects? *Molecular and cellular biochemistry*, 289(1-2):167–173, 2006.

[92] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[93] M. M. Robertson and M. Trimble. Major tranquillisers used as antidepressants: a review. *Journal of affective disorders*, 4(3):173–193, 1982.

[94] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[95] C. Schwarz, A. Volz, C. Li, and S. Leucht. Valproate for schizophrenia. *Cochrane Database of Systematic Reviews*, (3), 2008.

[96] A. A. Sepehry, S. Potvin, R. Élie, and E. Stip. Selective serotonin reuptake inhibitor (ssri) add-on therapy for the negative symptoms of schizophrenia: a meta-analysis., 2007.

[97] H.-P. Shih, X. Zhang, and A. M. Aronov. Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. *Nature Reviews Drug Discovery*, 17(1):19, 2018.

[98] Y. Shulman and P. G. Tibbo. Neuroactive steroids in schizophrenia. *The Canadian Journal of Psychiatry*, 50(11):695–702, 2005.

[99] P. D. Sigalas, H. Garg, S. Watson, R. H. McAllister-Williams, and I. N. Ferrier. Metyrapone in treatment-resistant depression. *Therapeutic advances in psychopharmacology*, 2(4):139–149, 2012.

[100] R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[101] R. C. Smith, H. Jin, C. Li, N. Bark, A. Shekhar, S. Dwivedi, C. Mortiere, J. Lohr, Q. Hu, and J. M. Davis. Effects of pioglitazone on metabolic abnormalities, psychopathology, and cognitive function in schizophrenic patients treated with antipsychotic medication: a randomized double-blind study. *Schizophrenia research*, 143(1):18–24, 2013.

[102] H.-C. So, C. K.-L. Chau, W.-T. Chiu, K.-S. Ho, C.-P. Lo, S. H.-Y. Yim, and P.-C. Sham. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nature neuroscience*, 20(10):1342, 2017.

[103] H.-C. So and Y.-H. Wong. Implications of de novo mutations in guiding drug discovery: A study of four neuropsychiatric disorders. *Journal of psychiatric research*, 110:83–92, 2019.

[104] I. E. Sommer, R. van Westrhenen, M. J. Begemann, L. D. de Witte, S. Leucht, and R. S. Kahn. Efficacy of anti-inflammatory agents to improve symptoms in patients with schizophrenia: an update. *Schizophrenia bulletin*, 40(1):181–191, 2014.

[105] G. I. Spielmans, M. I. Berman, E. Linardatos, N. Z. Rosenlicht, A. Perry, and A. C. Tsai. Adjunctive atypical antipsychotic treatment for major depressive disorder: a meta-analysis of depression, quality of life, and safety outcomes. *Focus*, 14(2):244–265, 2016.

[106] J. E. Standal. Pizotifen as an antidepressant. *Acta Psychiatrica Scandinavica*, 56(4):276–279, 1977.

[107] X. Su, K. Meneses, P. McNees, and W. O. Johnson. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):457–474, 2011.

[108] X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

[109] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[110] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[111] P. Suresh and A. B. Raju. Antidopaminergic effects of leucine and genistein on shizophrenic rat models. *Neurosciences*, 18(3):235–241, 2013.

[112] D. I. Swerdlow, D. Preiss, K. B. Kuchenbaecker, M. V. Holmes, J. E. Engmann, T. Shah, R. Sofat, S. Stender, P. C. Johnson, R. A. Scott, et al. Hmg-coenzyme a reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *The Lancet*, 385(9965):351–361, 2015.

[113] J. Thompson, G. Butterfield, U. Gylfadottir, J. Yesavage, R. Marcus, R. Hintz, A. Pearman, and A. Hoffman. Effects of human growth hormone, insulin-like growth factor i, and diet and exercise on body composition of obese postmenopausal women. *The Journal of Clinical Endocrinology & Metabolism*, 83(5):1477–1484, 1998.

[114] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

[115] J. Tibshirani, S. Athey, S. Wager, R. Friedberg, L. Miner, M. Wright, M. J. Tibshirani, L. Rcpp, R. I. DiceKriging, and G. SystemRequirements. Package 'grf'. 2018.

[116] J. Usall, E. Huerta-Ramos, R. Iniesta, J. Cobo, S. Araya, M. Roca, A. Serrano-Blanco, F. Teba, and S. Ochoa. Raloxifene as an adjunctive treatment for postmenopausal women with schizophrenia: a double-blind, randomized, placebo-controlled trial. *Journal of Clinical Psychiatry*, 72(11):1552, 2011.

[117] J. Usall, E. Huerta-Ramos, J. Labad, J. Cobo, C. Núñez, M. Creus, G. G. Parés, D. Cuadras, J. Franco, E. Miquel, et al. Raloxifene as an adjunctive treatment for postmenopausal women with schizophrenia: a 24-week double-blind, randomized, parallel, placebo-controlled trial. *Schizophrenia bulletin*, 42(2):309–317, 2016.

[118] P. D. Van and K. L. Thomas. Concomitant calcium channel blocker and antipsychotic therapy in patients with schizophrenia: Efficacy analysis of the catie-sz phase 1 data. *Annals of clinical psychiatry: official journal of the American Academy of Clinical Psychiatrists*, 30(1):6–16, 2018.

[119] R. Van den Oever, B. Hepp, B. Debbaut, and I. Simon. Socio-economic impact of chronic venous insufficiency: an underestimated public health problem. *International angiology*, 17(3):161, 1998.

[120] P. A. VanderLaan, D. Rangachari, S. M. Mockus, V. Spotlow, H. V. Reddi, J. Malcolm, M. S. Huberman, L. J. Joseph, S. S. Kobayashi, and D. B. Costa. Mutations in tp53, pik3ca, pten and other genes in egfr mutated lung cancers: Correlation with clinical outcomes. *Lung Cancer*, 106:17–21, 2017.

[121] F. P. Varghese and E. S. Brown. The hypothalamic-pituitary-adrenal axis in major depressive disorder: a brief primer for primary care physicians. *Primary care companion to the Journal of clinical psychiatry*, 3(4):151, 2001.

[122] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91, 2006.

[123] D. Vigo, G. Thornicroft, and R. Atun. Estimating the true global burden of mental illness. *The Lancet Psychiatry*, 3(2):171–178, 2016.

[124] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[125] M. L. Wahlqvist, M.-S. Lee, S.-Y. Chuang, C.-C. Hsu, H.-N. Tsai, S.-H. Yu, and H.-Y. Chang. Increased risk of affective disorders in type 2 diabetes is minimized by sulfonylurea and metformin combination: a population-based cohort study. *BMC medicine*, 10(1):150, 2012.

[126] J. Wang, S. Vasaikar, Z. Shi, M. Greer, and B. Zhang. Webgestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research*, 45(W1):W130–W137, 2017.

[127] T. T. Wang, N. Sathyamoorthy, and J. M. Phang. Molecular effects of genistein on estrogen receptor mediated pathways. *Carcinogenesis*, 17(2):271–275, 1996.

[128] W.-Q. Wei, R. M. Cronin, H. Xu, T. A. Lasko, L. Bastarache, and J. C. Denny. Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association*, 20(5):954–961, 2013.

[129] J. Q. Wu, T. R. Kosten, and X. Y. Zhang. Free radicals, antioxidant defense systems, and schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 46:200–206, 2013.

[130] R. Xu and Q. Wang. Phenopredict: a disease phenome-wide drug repositioning approach towards schizophrenia drug discovery. *Journal of biomedical informatics*, 56:348–355, 2015.

[131] S. Yamagami and K. Soejima. Effect of maprotiline combined with conventional neuroleptics against negative symptoms of chronic schizophrenia. *Drugs under experimental and clinical research*, 15(4):171–176, 1989.

[132] L. Yang, Y. Zhao, Y. Wang, L. Liu, X. Zhang, B. Li, and R. Cui. The effects of psychological stress on depression. *Current neuropharmacology*, 13(4):494–504, 2015.

[133] H. You, W. Lu, S. Zhao, Z. Hu, and J. Zhang. The relationship between statins and depression: a review of the literature. *Expert opinion on pharmacotherapy*, 14(11):1467–1476, 2013.

[134] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[135] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.