Research Proposal: Novel AI for Biomedical Studies

Note: **this research proposal contains two parts:**

- **Part 1 introduces a novel perspective to address a stable prediction problem.**
- **Part 2 is a short proposal as a case study to apply the proposed variable selection method and other AI approaches to a biological question.**

Contents

Part 1: Stable Clinical Prediction Model Towards Unknown Real-World Populations

Abstract

The existing data-driven clinical prediction models cannot perform stably in the unknown real-world populations, which results in several risk factors in applications. To end this, a stable prediction perspective is introduced to analyze clinical prediction model development for the first time. More specifically, this research considers issues of variable selection and feature representation in clinical prediction models, and evaluates the advances in stable prediction compared with that of classical regularizers with respect to models' performance in unknown populations. Furthermore, R packages will be built to implement the above investigation.

Introduction

Precision medicine is a field of medicine that considers individual differences in omics data, clinical measures, environments, family history, and lifestyles to make precision healthcare strategies while developing data-driven clinical prediction models is one of the typical practices in this field (Zhang, 2015). Clinical prediction models are categorized by downstream prediction tasks, including but not limited to, prognostic models for prognostic analysis (e.g., survival analysis, scoring system), and diagnostic models for diagnosis with individualized information (e.g., disease classification, biomarker identification) (Lee, Bang, & Kim, 2016; Steyerberg et al., 2013). Still, from a machine learning perspective, a clinical prediction model is a typical regression or prediction problem. Therefore, developing a successful clinical prediction model is consistent with successful machine learning practices, including three steps: representation, optimization, and evaluation (Domingos, 2012). Representation refers to screening out the input predictors from demographic, clinical measurements, and variables in high-throughput omics data, including several to dozens of variables for pretreatment, and finally to generate biologically significant feature representation. Optimization is to maximize the prediction accuracy of the model, that is, to minimize the error of prediction during training. As for model evaluation, generalization error is evaluated through the external validation cohorts (testing dataset). Meanwhile, complementary efforts are made to perform downstream wet lab experiments validating biological significance in feature representation and to generate novel scientific insights.

Clinical practices are risk-sensitive domains, while several risk factors in clinical prediction models have limited their applications in real-world practices. A primary risk factor involves feature representation. Clinical prediction models select variables with biological significance (i.e., signatures, biomarkers) to ensure good feature representation. In practice, clinical variables selection could be supported by experts, and variable screening in high-dimensional omics variables should be implemented by statistical tools (such as regularization tools, e.g., least absolute shrinkage and selection operator; also Lasso) due to complex but unknown biological

mechanisms underlying. However, the existing model diagnosis approaches are not designed to evaluate whether biologically significant variables are selected. It also results in risk factors in subsequent model evaluation and validation. On one hand, both a training dataset and an external validation dataset (testing set) are obviously not identical to a real-world population. For a classical correlation-based prediction model, if the feature representation (selected variables) lacks biological significance, the performance towards unknown real-world populations is unstable. On the other hand, researchers conduct wet lab experiments to validate the biological significance of the selected variables, which results in another risk factor that, when failed feature representation is performed, the wet lab experiments may generate false-positive discoveries. Even worse, as many publications perform downstream bioinformatics analysis to investigate the biological significance of the selected variables, the failed representation leads to failed efforts of the analysis. In summary, selecting variables (representations) with biological significance to ensure the performance of a clinical prediction model in various real-world populations is critical to promoting real-world applications in clinical practices and biological discoveries.

To address the above risk factors, I hereby introduce a stable prediction perspective for the first time to analyze clinical prediction model development. Stable prediction is first proposed in 2018 to promote applications of machine learning prediction models in the risk-sensitive domains. Stable prediction tackles such a problem in building a prediction model, how can we ensure the model stability when applied in different testing environments with unknown distribution shifts. Note that the stability here is defined over prediction performance, rather than estimation stability. Figure 1 illustrates the relationships between stable learning, traditional machine learning (IID learning, under the assumption that the training and testing datasets are independently and identically distributed), and transfer learning (or domain adaption). The optimization goal of the latter two learning paradigms is to maximize the prediction performance (accuracy) of the model with prior knowledge of the testing dataset. (Kuang, Cui, Athey, Xiong, & Li, 2018)

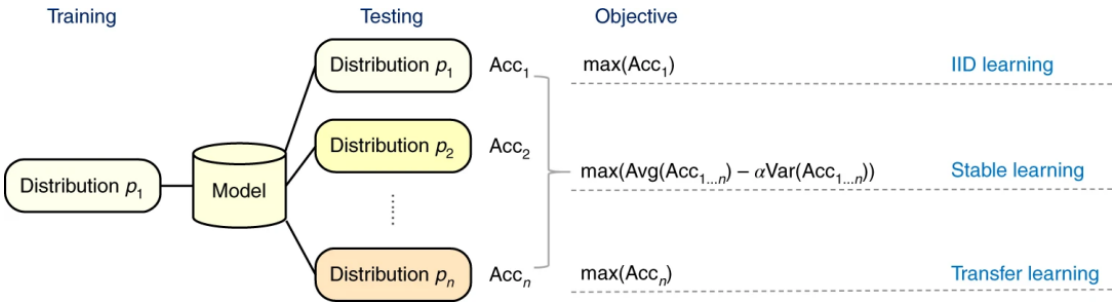

Figure 1 Stable Learning, IID Learning and Transfer Learning (Cui & Athey, 2022)

To achieve stable prediction, a causality perspective is further introduced to analyze the prediction models (Cui & Athey, 2022). Commonly used clinical prediction models make predictions based on the associations between the predictors and the target variables, thus spurious correlation among variables could lead to false predictions. From a causality perspective, ways of generating correlations are identified into three patterns: causation, confounding and data selection bias, as shown in Figure 2. Among the three patterns, only the causation reflects the intrinsic dependence among variables, that is, the biological significance we are looking for; the other two types are spurious correlations sensitive to the joint distribution of features and the data collection processes. Therefore, without identifying a causal relationship (as existing clinical prediction

models done), the prediction performance would depend heavily on how much the validation population/real-world population distribution shifts from the training distribution, leading to unstable performance under varying real-world populations.

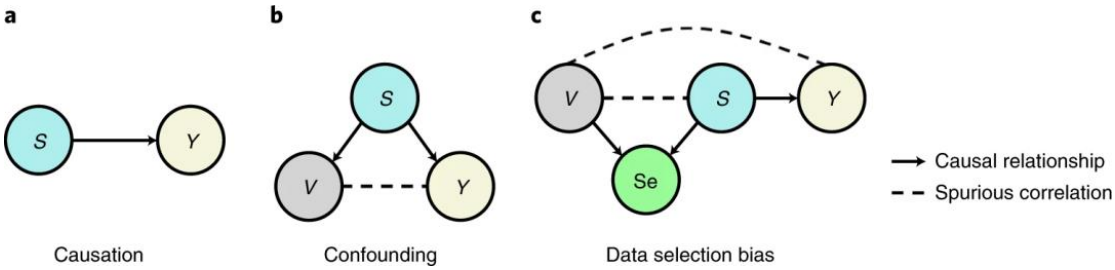

Figure 2 Three patterns of correlations generating. (Cui & Athey, 2022)

Currently, several algorithms have been proposed to achieve stable prediction with causal feature representation, but none of them is applied to a clinical prediction model. My research tries to bridge the research gap between stable prediction and its applications to clinical prediction models. I will consider variable selection and feature representation in clinical prediction model development, integrating with downstream tasks to investigate whether the stable prediction approaches will bring benefits over traditional methods (e.g., Lasso). Furthermore, I will analyze the cost of the new approaches: reducing the effective sample size. Additionally, I would like to "translate" the advances in stable prediction into R packages to support the open-source R community and R-powered biomedical studies.

Literature Review

A bunch of clinical prediction models has been proposed to address various biomedical questions. We focus on the variable selection and feature representation of clinical prediction models involving omics data, whose data is generally retrieved from several public databases, including but not limited to The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/), Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), Molecular Taxonomy of Breast Cancer International Consortium (METABRIC, https://www.cbioportal.org/datasets). To screen out the predictors, researchers perform bioinformatics analysis according to the specific biomedical questions as the first step, with the help of differentially expressed genes analysis (Zhao & Cui, 2020) or weighted gene co-expression network analysis (WGCNA) and so on, generating hundreds to dozens of selected variables (e.g., signatures or biomarkers) from tens of thousands of raw variables (e.g., genes) which may be in binary or numerical form. In general, the number of the screened variables is less than the sample size. Afterward, statistical tools are applied to the selection of these screened variables. The tools include but are not limited to Variance Inflation Factor (VIF) (Huang et al., 2021), deep autoencoder approaches (Ko, Choi, & Ahn, 2021), and regularization tools like Lasso (Sui et al., 2020). Searching the PubMed with "Search: (((prognostic) OR (diagnosis) OR (predict)) AND (LASSO)) AND (gene) Filters: from 2017 – 2022", a total of 2,784 publications are obtained, which suggests that Lasso is a commonly used approach in this topic. The selected variables are subsequently transferred into a downstream model for processing according to the specific prediction tasks. The potential model would include linear models (e.g., cox regression for survival analysis, logistics regression for diagnosis and classification) or artificial

neural network models (e.g., a backpropagation neural network, or a deep learning model).
As discussed in the previous section, the existing variables selection methods and learning models cannot guarantee the performance of a clinical prediction model in unknown real-world environments, which limits its applications in clinical practices. Stable prediction is introduced to address the concerns. To achieve stable prediction, inspired by the covariate balancing approaches in causal inference, two sample reweighting algorithms based on Variable Decorrelation have been applied to the regression tasks (Kuang et al., 2018; Kuang, Xiong, Cui, Athey, & Li, 2020). Additionally, theoretical analysis shows when collinearity exists, the prediction performance is unstable across various testing environments when even there is a large sample size (i.e., infinite), and a general data pretreatment method based on sample reweighting (Shen, Cui, Zhang, & Kunag, 2020) is proposed to reduce the collinearity effect, which can be seamlessly integrated into classical linear models for specific downstream prediction tasks. The stable prediction advances try to feature causal representations, so as to identify the invariable predictors which described the data generation mechanism, that is, the biological significance in the biomedical domain (Xu et al., 2020). Classical regularization tools like Lasso are regarded as baselines, while simulation research has revealed that regularization tools cannot address stable prediction problems since their regularizers would generally estimate larger coefficients on the unstable features (Kuang et al., 2020). In summary, it is necessary to apply the stable prediction perspective to analyze the issues of variables selection and feature representation in clinical prediction models.

Research Design and Methods

Research Objectives

Introduce stable prediction to tackle variable selection and feature representation in clinical prediction models. More specifically, in the first phase, I will evaluate the performance of a data pretreatment method by sample reweighting (Shen et al., 2020) in variable selection task integrating with downstream linear models, and further design the criteria for variable selection. In the second phase, I will investigate the availability of several other stable learning approaches which jointly optimizes a variable decorrelation regularizer and a weighted regression model (Kuang et al., 2018; Kuang et al., 2020), and further evaluate the requirements in the sample size issues. Alongside, in order to contribute to the open-source R community and to support other R-powered biomedical studies, this research will be primarily implemented via R programming.

Methods and Materials

This subsection introduces the first phase of research.

Sample Reweighted Decorrelation Operator (SRDO) method
SRDO method is designed to reduce collinearity among input variables in a design matrix in a linear model, generating a column-decorrelated counterpart by performing random resampling column-widely, which breaks down the joint distribution of variables in the raw matrix into independent marginal distributions. (Shen et al., 2020)

Simulation Study

A simulation study will be carried out to evaluate the performance of the SRDO method. The synthesis data is generated by linear models with bias terms: one is the Cox regression model for survival analysis, and another is a logistic regression model for classification. Lasso, Elastic Net, ULasso, and IILasso are employed as baselines. Absolute error is used to evaluate the performance of parameter estimation, and Area Under the ROC Curve (AUC) is used to evaluate the performance of the classifier. By customizing the parameters of bias terms, testing datasets with different distributions are generated, and the prediction performance of the models in various unknown distributions is further investigated with the help of average AUC and stability AUC (variance of average AUC generated via bootstrap method).

Issues of sample size and variable selection

In a clinical prediction model, the sample size is larger than the number of selected variables. Meanwhile, the variables are assumed to be sparse (all biological variables are sparse). Still, we have to investigate the effective sample size reduction in sample resampling to present practical criteria. Practical criteria of variable selection are also required, and if possible, the visualization of hyper-parameter can be designed (just as Lasso does in the variable selection visualization).

Case Study

If possible, design a case of a data-driven clinical prediction model, and compare whether the SRDO and Lasso methods generated the same results.

Building R packages

If possible, all my algorithm implementations and data will be bundled as open-source R packages, contributing to the R community.

Practicalities and Potential Obstacles

The primary practical consideration is that this research addresses a small statistical problem abstracted from the general clinical prediction model. Therefore, this research can contribute to the academic community independently without the support of clinical resources.

For the first phase of research, a potential obstacle is to provide an effective sample size and the criteria of variable selection, which needs to utilize statistical inference approaches. For the second phase, there will be many technical problems when I implement the Python-based deep learning model by R programming.

In sample size estimation, a smaller the effect size results in the larger the sample size. If the contribution of genome variables to complex diseases is relatively weak, we need a larger sample size to support the development of clinical prediction models. Hence, in practice, whether the method of stable prediction can bring gains may depend on the specific clinical questions, handling the number of selected variables, and the sample size.

Implications and Contributions to Knowledge

This research has three main contributors: (1) This is the first effort to analyze the performance of the clinical prediction model in the various real-world populations from the stable prediction perspective. (2) This research Introduces the causal feature representation approaches, integrating with the downstream tasks to carry out a simulation study and evaluate the stable prediction issues of clinical prediction models. In particular, this research is the first effort to apply stable learning for survival analysis. (3) The advances in stable prediction will be introduced into the R community to support R-powered biomedical studies.

Reference

Cui, P., & Athey, S. (2022). Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence, 4*(2), 110-115. doi:10.1038/s42256-022-00445-z

Domingos, P. (2012). A few useful things to know about machine learning. *55*(10 %J Commun. ACM), 78–87. doi:10.1145/2347736.2347755

Huang, C., Deng, M., Leng, D., Leung, E. L.-H., Sun, B., Zheng, P., & Zhang, X. D. (2021). MIRS: an AI scoring system for predicting the prognosis and therapy of breast cancer. *medRxiv*, 2021.2012.2016.21267775. doi:10.1101/2021.12.16.21267775

Ko, S., Choi, J., & Ahn, J. (2021). GVES: machine learning model for identification of prognostic genes with a small dataset. *Scientific Reports, 11*(1). doi:10.1038/s41598-020-79889-5

Kuang, K., Cui, P., Athey, S., Xiong, R., & Li, B. (2018). *Stable Prediction across Unknown Environments*. Paper presented at the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom. https://doi.org/10.1145/3219819.3220082

Kuang, K., Xiong, R., Cui, P., Athey, S., & Li, B. (2020). Stable Prediction with Model Misspecification and Agnostic Distribution Shift. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(04), 4485-4492. doi:10.1609/aaai.v34i04.5876

Lee, Y.-H., Bang, H., & Kim, D. J. (2016). How to Establish Clinical Prediction Models. *Endocrinology and Metabolism, 31*(1), 38. doi:10.3803/enm.2016.31.1.38

Shen, Z., Cui, P., Zhang, T., & Kunag, K. (2020). Stable Learning via Sample Reweighting. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(04), 5692-5699. doi:10.1609/aaai.v34i04.6024

Steyerberg, E. W., Moons, K. G. M., Van Der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., . . . Altman, D. G. (2013). Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLOS Medicine, 10*(2), e1001381. doi:10.1371/journal.pmed.1001381

Sui, S., An, X., Xu, C., Li, Z., Hua, Y., Huang, G., . . . Li, M. (2020). An immune cell infiltration-based immune score model predicts prognosis and chemotherapy effects in breast cancer. *Theranostics, 10*(26), 11938-11949. doi:10.7150/thno.49451

Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., & Cui, W. (2020, 2020-08-23). *Algorithmic Decision Making with Conditional Fairness.* Paper presented at the Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Zhang, X. D. (2015). Precision Medicine, Personalized Medicine, Omics and Big Data: Concepts and Relationships. *Journal of Pharmacogenomics & Pharmacoproteomics, 06*(02). doi:10.4172/2153-0645.1000e144

Zhao, X., & Cui, L. (2020). A robust six‐miRNA prognostic signature for head and neck squamous cell carcinoma. *Journal of Cellular Physiology, 235*(11), 8799-8811. doi:10.1002/jcp.29723

Part 2: Biomarkers and Drug Discoveries for Comorbid Depression and Obesity: Multi-Omics Approaches

Introduction

Both depression and obesity place heavy burdens on public health (Murray et al., 2012). Depression is a common mental disorder with persistent sadness and a lack of interest or pleasure in previously rewarding or enjoyable activities. It is estimated that more than 3.8% of the population is affected by depression ("Depression," 13 September 2021). Overweight and obesity are characterized by abnormal or excessive fat accumulation. According to WHO, more than 1.9 billion adults are overweight while over 650 million of these are obese ("Obesity and overweight," 9 June 2021). Recently, obesity is recognized as a multi-causal chronic disease that increases the risk of developing numerous comorbidities, including depression (Aleksandrova, Egea Rodrigues, Floegel, & Ahrens, 2020). Indeed, epidemiology evidence reveals a longitudinally bidirectional dose-dependent relationship between depression and obesity (Faith et al., 2011; Luppino et al., 2010). This co-occurrence within individuals brings excess obstacles to the treatment of each condition separately. To address the obstacles, researchers have investigated the shared biological mechanisms of these two conditions. Such mechanisms involve biological pathways across multiple levels, including genetics, alterations in systems involved in homeostatic adjustments (hypothalamic-pituitary-adrenal (HPA) axis, immuno-inflammatory activation, neuroendocrine regulators of energy metabolism including leptin and insulin, and microbiome), and brain circuitries integrating homeostatic and mood regulatory responses (Milaneschi, Simmons, van Rossum, & Penninx, 2019). Still, to date, there are huge challenges in integrating multi-level pathways for uncovering key molecular mechanisms and handling the heterogeneity of patients. To end this, multi-omics approaches should be introduced to study etiology and pathophysiology in this topic. Employing multi-omics approaches, potential causative changes that lead to diseases or the novel therapeutic targets can be elucidated by modeling multi-layer omics data as complex networks, and the results can be then tested in further molecular and animal studies (Aleksandrova et al., 2020; Hasin, Seldin, & Lusis, 2017; Maes et al., 2016). Such multi-omics investigation could bring substantial promise to biomarkers and drug discoveries, and help identify more homogeneous subgroups of patients in pathophysiological terms.

Research Objectives

This research aims to explore shared drivers of comorbid depression and obesity via multi-omics approaches and to identify novel biomarkers and drug targets for further validation.

Research Design

Firstly, we construct two multi-omics datasets separately on depression and obesity from available published resources and identify potential biomarkers from these two datasets through machine learning and deep learning tools. Subsequently, we will integrate the shared biomarkers to perform network analysis, pathways enrichment, and target prioritization on comorbid depression and obesity. Candidate pathways and therapeutic targets will be identified as a result, and virtual screening against the identified targets will be conducted to find hit molecules for further validation. The potential drug candidates will then be screened by deep learning-assisted molecular docking approaches.

Implications

By investigating shared mechanisms linking depression and obesity, we can discover novel therapeutic targets and therefore widen the toolbox for treating patients with comorbid depression and obesity.

Reference

Aleksandrova, K., Egea Rodrigues, C., Floegel, A., & Ahrens, W. (2020). Omics Biomarkers in Obesity: Novel Etiological Insights and Targets for Precision Prevention. *Current Obesity Reports, 9*(3), 219-230. doi:10.1007/s13679-020-00393-y

Depression. (13 September 2021). *Fact Sheets* Retrieved from https://www.who.int/news-room/fact-sheets/detail/depression

Faith, M. S., Butryn, M., Wadden, T. A., Fabricatore, A., Nguyen, A. M., & Heymsfield, S. B. (2011). Evidence for prospective associations among depression and obesity in population-based studies. *Obesity Reviews, 12*(5), e438-e453. doi:10.1111/j.1467-789x.2010.00843.x

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology, 18*(1). doi:10.1186/s13059-017-1215-1

Luppino, F. S., De Wit, L. M., Bouvy, P. F., Stijnen, T., Cuijpers, P., Penninx, B. W. J. H., & Zitman, F. G. (2010). Overweight, Obesity, and Depression. *Archives of General Psychiatry, 67*(3), 220. doi:10.1001/archgenpsychiatry.2010.2

Maes, M., Nowak, G., Caso, J. R., Leza, J. C., Song, C., Kubera, M., . . . Berk, M. (2016). Toward Omics-Based, Systems Biomedicine, and Path and Drug Discovery Methodologies for Depression-Inflammation Research. *Molecular Neurobiology, 53*(5), 2927-2935. doi:10.1007/s12035-015-9183-5

Milaneschi, Y., Simmons, W. K., van Rossum, E. F. C., & Penninx, B. W. J. H. (2019). Depression and obesity: evidence of shared biological mechanisms. *Molecular Psychiatry, 24*(1), 18-33. doi:10.1038/s41380-018-0017-5

Murray, C. J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., . . . Memish, Z. A. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet, 380*(9859), 2197-2223.

doi:10.1016/s0140-6736(12)61689-4

Obesity and overweight. (9 June 2021). *Fact Sheets.* Retrieved from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight