Principal Investigator

First Name: Yang
Last Name: LI
Degree: Ph.D. in statistics
Primary Affiliation: School of Statistics, Renmin University Of China
E-mail: 1533899859@qq.com
Phone number: 13810064660
Address:

City: Beijing
State or Province: Beijing
Zip or Postal Code: 100089
Country: China

General Information

Key Personnel (in addition to PI):
First Name: Haoyu
Last name: Yang
Degree: bachelor
Primary Affiliation: School of Statistics, Renmin University Of China
SCOPUS ID:

Are external grants or funds being used to support this research?: No external grants or funds are being used to support this research.
How did you learn about the YODA Project?: Scientific Publication

Conflict of Interest

https://yoda.yale.edu/system/files/yoda_project_coi_yangli.pdf
https://yoda.yale.edu/system/files/yoda_project_coi_haoyuyang.pdf

Certification

Certification: All information is complete; I (PI) am responsible for the research; data will not be used to support litigious/commercial aims.
Data Use Agreement Training: As the Principal Investigator of this study, I certify that I have completed the YODA Project Data Use Agreement Training
Associated Trials:

1. NCT00968812 - 28431754DIA3009 - A Randomized, Double-Blind, 3-Arm Parallel-Group, 2-Year (104-Week), Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of JNJ-28431754 Compared With Glimepiride in the Treatment of Subjects With Type 2 Diabetes Mellitus Not Optimally Controlled on Metformin Monotherapy

What type of data are you looking for?: Individual Participant-Level Data, which includes Full CSR and all supporting documentation

Research Proposal

Project Title

A new randomization procedure: pairwise sequential randomization (PSR), properties and applications in both causal inference and clinical trials.

Narrative Summary:

In causal inference and clinical trials, balancing important covariates is often one of the most important concerns for both efficient and credible comparison. However, chance imbalance still exists. To address this issue, we propose a new procedure, called pairwise sequential randomization (PSR). With a large number of covariates or units, PSR shows substantial advantages over the traditional methods in terms of the covariate balance, estimation accuracy, and computational time, making it an ideal technique in the era of big data. The estimated treatment effect under PSR attains its minimum variance asymptotically. Also PSR is widely applicable in both causal inference and clinical trials.

Scientific Abstract:

Background
As we know, randomization is the foundation for the treatment effect evaluation. In the framework of causal inference although RR works well in the case of a few covariates, it is incapable of scaling up to address massive amounts of data. In clinical trials, properties for balancing continuous covariates were not well investigated in literature.
Objective
Our objective is to propose a new approach to generate a more balanced allocation and thus to improve the subsequent analysis for both causal inference and clinical trails settings.
Study Design
We allocate units adaptively and sequentially by assigning one randomly chosen pair of units at a time. For each pair of units, using their covariate information and the existing level of imbalance of the previously allocated units, adjust the probability with which the pair is allocated to treatment groups to avoid imbalance.
Participants
We will utilize information about all participants(1452) obtained from the CANTATA-SU trial, NCT00968812.
Main Outcome Measure
The main outcome measure(yi) of this research is the change of HbA1C. As reported in the statistical case report, the study were to evaluate the effect of canagliflozin compared with glimepiride on the change of HbA1C.
Statistical Analysis
We will allocate these units our proposed method, and further simulate the outcome variable using a linear regression model which was fitted to the original data. Our analysis will be focused on the average treatment effect under PSR compared with other different methods.

Brief Project Background and Statement of Project Significance:

Traditional randomization methods often generate unsatisfactory configurations with unbalanced prognostic covariates. The advantages of balanced covariates are at least threefold. First, covariate balance improves the efficiency of estimation for the treatment effect. Second, it increases the interpretability of the estimated treatment effect by making the units in the treatment groups more comparable, thereby enhancing the credibility of the analysis. Third, it makes the analysis more robust against model misspecification. Consequently, covariate imbalance can significantly undermine the validity of subsequent analysis. In the absence of covariate balance, various problems must be addressed before a valid conclusion can be drawn.
In causal inference and clinical studies, if a significant imbalance exists, any inferences regarding the treatment effect will be inaccurate, and any claims about the treatment effect will need to rely on unveriable assumptions. Researchers must assess the balance in the covariate distribution before estimating the causal effect. In addition, these adjustments often rely on at least a nearly correct model, which can be difficult to test.
More recently, covariate balance has attracted growing interest in the field of crowdsourced-internet experimentation. Researchers increasingly recruit workers from online labor markets into their experiments. Because of the nature of the recruiting process, a large number of workers with many covariates, typically are enrolled in such studies, which consequently pose challenges for traditional randomization methods.
In the framework of causal inference, Morgan and Rubin have proposed rerandomization (RR). They propose to

repeatedly randomize the units into treatment groups using complete randomization (CR), until certain the balance criterion is satisfied. They has also assumed fixed equal numbers of units in two treatment groups and demonstrated various desirable properties under rerandomization. Although rerandomization works well in the case of a few covariates, it is incapable of scaling up to address massive amounts of data. For example, as the number of covariates increases, the probability of acceptance, pa = P(M < a), of each complete randomization decreases drastically, causing the rerandomization procedure to remain in loop for a long time. To compromise the computational burden, one can increase a, which unavoidably leads poorer covariate imbalance.
In clinical trials, to balance important covariates, most existing methods are only for discrete covariates. Discretizing continuous covariates is often less efficient and changes the nature of the covariates. A variety of methods for balancing continuous covariates have been proposed in the literature: the methods based on ranks; based on p-value; based on empirical cumulative distribution; based on kernel density, etc. However, the performance of those procedures was usually evaluated by simulation studies, their theoretical properties are not well investigated in literature. Also these methods are usually applicable for only a few covariates.

Specific Aims of the Project:

In this research, we try to propose a new approach-pairwise sequential randomization (PSR), to generate a more balanced treatment allocation and thus to improve the subsequent analysis for both causal inference and clinical trails settings. The properties of the PSR procedure are illustrated both theoretically and numerically.
The advantages of the proposed method are: (i) For cases with a large number of covariates or a large number of units, the proposed method exhibits superior performance, with more balanced randomization and less computational time; (ii) The PSR procedure attains the optimal covariate balance, in the sense that the estimated treatment effect under the proposed method attains its minimum variance asymptotically; and (iii) The proposed procedure is designed for directly randomizing units with both continuous and discrete covariates. Therefore the PSR procedure is widely applicable for balancing many important covariates in comparative studies.

What is the purpose of the analysis being proposed? Please select all that apply.
Confirm or validate previously conducted research on treatment effectiveness
Research on clinical trial methods
Research on comparison group

Research Methods

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study:

We apply to use the real clinical trial data named CANTATA-SU, NCT00968812, to illustrate our proposed method. In order to ensure sufficient observations, we don’t anticipate excluding any participants. We are able to produce a much more balanced allocation units through our method, in that case we will check whether our method can improve the accuracy of efficacy estimation and analysis. Moreover, as this data set contains multiple treatment effects, we are able to extend our method to multiple treatment groups, compare the results under different randomization methods and improve the accuracy of causal inference.
In this way, the proposed algorithm can be easily adopted in clinical trial studies where patients are sequentially enrolled and the treatment is conducted after the individual enrollment. It is important to note that the proposed method is designed for directly randomizing units with continuous covariates. Also the PSR procedure works well for large p and n, while the other methods only work for small p. Moreover through simulation studies, we can show that the above scenarios yield similar results in terms of covariate balance especially when sample size is large.

Main Outcome Measure and how it will be categorized/defined for your study:

In this research, we have introduced a new randomization procedure for balancing the covariates to improve the estimation accuracy for causal inference and clinical trials. Our main outcome yi is the change of HbA1C. As reported in the SCR, the study were to evaluate the effect of canagliflozin compared with glimepiride on the change of HbA1C. Similar to the analysis of this report, we will compare the durability of hemoglobin A1C (HbA1C)-lowering efficacy in each canagliflozin group with the glimepiride group from Week 26 to Week 104. Thus, we will compare the causal inference estimation through our randomization method compared with traditional methods.
As mentioned above, using the covariate information and the existing level of imbalance of the previously allocated units, we adjust the probability with which the pair is allocated to treatment groups to avoid incidental covariate

imbalance. In this way, we are able to produce a much more balanced allocation of units and improve the accuracy for causal inference and clinical trials.

Main Predictor/Independent Variable and how it will be categorized/defined for your study:

Through the analysis of this real clinical trial, we found there are many predictors that influence our main outcome variable(HbA1C). In order to have a more complete and comprehensive analysis, the predictor variables we wish to obtain include but are not limited to the following: FPG; body weight; SBP; DBP; fasting plasma lipids, including LDL-C, HDL-C, non-HDL-C, total cholesterol, ratio of LDL-C to HDL-C, and triglycerides; HOMA2-%B, insulin/proinsulin (and ratio), and waist circumference and BMI. On the one hand, this predictor variables will be treated as covariates for each observation when we allocate these units. On the other hand, under each patient allocation scheme, we will further simulate the outcome variable using a linear regression model which was fitted to the original data(our main predictor variables and the main outcome) . To closely mimic the original data, we will fit the linear regression to the original data with the original patient allocation, store the residuals and the coefficient estimates. Using the simulated outcome variable, we can estimate the average treatment effect under the proposed method, rerandomization, and complete randomization.

Other Variables of Interest that will be used in your analysis and how they will be categorized/defined for your study:

We have listed the desired predictor variables in the previous section. If there are any other continuous variables in this clinical trial that have impact on our outcome variable, please provide them to us.

Statistical Analysis Plan:

We apply to use the data named CANTATA-SU, illustrate our proposed method using the real clinical trial data: A Randomized, Double-Blind, 3-Arm Parallel-Group, 2-Year 104-Week), Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of JNJ-28431754 100 mg and JNJ-28431754 300 mg Compared With Glimepiride in the Treatment of Subjects With Type 2 Diabetes Mellitus Not Optimally Controlled on Metformin Monotherapy. In total, there are 1452 patients and we will select some important continuous covariates into our study. We do need this large sample size to guarantee the theoretical properties of PSR. The outcome variable yi, durability of hemoglobin A1C (HbA1C)-lowering efficacy in each canagliflozin group with the glimepiride group was recorded to study the treatment effect.
In the original study, the patients were randomly assigned to the treatment or control groups. We can calculate corresponding Mahalanobis distance between different treatment groups and take this distance as a measure of imbalance. That is,
(1)We first arrange all n units randomly into a sequence.
(2)Assign the first three units into three treatment groups randomly.
(3)Suppose that 3i units have been assigned to treatment groups, for the (3i+1)-th and (3i+2)-th units: (3a)if the (3i+1)-th unit is assigned to treatment 1, the (3i+2)-th unit to treatment 2 and the (3i+3)-th unit to treatment 3, then we can calculate the “potential” Mahalanobis distance between the updated treatment groups with (3i+3) units, (3b)similarly, we can calculate the other “potential” Mahalanobis distance in all of the possible assignments.
(4)Assign the (3i+1)-th, (3i+2)-th and (3i+3)-th unit to treatment groups with a certain probability according to the “potential” Mahalanobis distance.
(5)Repeat the last two steps until all units are assigned. If there is one or two units left, assign them to treatments with equal probabilities.
Through the proposed method, we can get a more balanced assignment.
To compare, we repeatedly assigned these patients to treatment groups using the proposed method, complete randomization, and rerandomization. And then plot the corresponding Mahalanobis distances. In order to compare different sample sizes, we will replicate the data many times to generate a larger data set.
For each randomization method, we will further simulate the outcome variable according to the real clinical data. Using the simulated outcome variable, we can obtain the average treatment effect under the proposed method. We can imagine that the proposed method exhibits the best performance compared with other methods especially under large sample size. It will yield the largest PRIV and the lowest variance. For rerandomization, a smaller threshold results in better performance; however, this comes at the cost of a longer computational time and a lower acceptance probability. Note that Because of the finite sample size, the optimal PRIV cannot be achieved. We can see that if we increase the sample size, the PRIV of the proposed method is greatly improved and is close to optimal, whereas that of the rerandomization method does not improve at all. The gain from the proposed method is quite substantial.

In addition, as the number of covariates increases, it is more efficient to balance only the most important covariates; therefore, we will attempt to select the important covariates to balance in our proposed framework. We consider that the proposed method may also be applied to balance important covariates in the field of crowdsourced-internet experimentation.
Software Used:
RStudio
Project Timeline:

We are desperate for this request to be approved. We estimate that it will take a total of four months from the start of the project to the completion of the report. If we are allowed to use this clinical trial data, we will begin our research within half a month unless there are some special circumstances. We will concentrate our efforts on the research and expect to have preliminary results within a month. As we know in order to achieve the same good effect as pairwise sequential randomization, the rerandomization experiment needs many iterations. As we don't know the performance of the secure data sharing platform, If it takes more time for the rerandomization experiment, we will to extend the time to calculate the data to one and a half months. We will analyze the results obtained under the pairwise sequential randomization, rerandomization and complete rerandomization, it may cost half a month. According to this results and analysis, we will complete the manuscript in half a month and first submit for publication. Finally we will send results report back to the YODA Project in a month.

Dissemination Plan:

In this research, we combine pairwise sequential randomization (PSR) method with the clinical trial data. The main objective is to generate a more balanced treatment allocation and thus to improve the subsequent analysis for both causal inference and clinical trails settings. For cases with a large number of covariates or a large number of units, the proposed method exhibits superior performance, with more balanced randomization and less computational time. The PSR procedure attains the optimal covariate balance, in the sense that the estimated treatment effect under the proposed method attains its minimum variance asymptotically. And the PSR procedure can be widely applicable for balancing many important covariates in comparative studies.
As we know in the causal inference, if there is an imbalance in the covariates, it will affect the accuracy of the treatment effect. Through this research we recommend that the experimenters use (PSR) for randomization. As mentioned before, the proposed method has many excellent theoretical properties and the post-trial analysis will also be straightforward. Any aspect of this work be publicly presented, suitable journals of this research may be Biometrika and Biometrics.

Bibliography:

Antognini, A. B. and Zagoraiou, M. (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. Biometrika, 98(3):519-535.
Begg, C. B. and Iglewicz, B. (1980). A treatment allocation procedure for sequential clinical trials. Biometrics, 36(1):81-90.
Bruhn, M. and McKenzie, D. (2008). In pursuit of balance: Randomization in practice in development field experiments. World Bank Policy Research Working Papers.
Chandler, D. and Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. Journal of Economic Behavior & Organization, 90:123-133.
Ciolino, J., Zhao, W., Martin, R., and Palesch, Y. (2011). Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization. Contemporary Clinical Trials, 32(2):250-259.
Cochran, W. G. (1965). The planning of observational studies of human populations. Journal of the Royal Statistical Society, Series A, 128(2):234-266.
Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: a review. Sankhya, A, 35(4):417-446.
Cox, D. R. (1982). Randomization and concomitant variables in the design of experiments. Statistics and Probability: Essays in Honor of C. R. Rao, pages 197-202. North-Holland, Amsterdam. MR0659470.
Fisher, R. A. (1926). The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain, 33:503-513.
Frane, J. W. (1998). A method of biased coin randomization, its implementation, and its validation. Drug Information Journal, 32:423-432.
Freedman, D. A. (2008). On regression adjustments to experimental data. Advances in Applied Mathematics, 40(2):180-193.
Hoehler, F. K. (1987). Balancing allocation of subjects in biomedical research: a minimization strategy based on

ranks. Computers and Biomedical Research, 20:209-213.
Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. Experimental Economics, 14(3):399-425.
Hu, F., Hu, Y., Ma, Z., and Rosenberger, W. F. (2014). Adaptive randomization for balancing over covariates. Wiley Interdisciplinary Reviews: Computational Statistics, 6(4):288-303.
Hu, Y. and Hu, F. (2012). Asymptotic properties of covariate-adaptive randomization. Annals of Statistics, 40(3):1794-1815.
Imbens, G. W. and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press.
Kapelner, A. and Krieger, A. (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. Biometrics, 70(2):378-388.
Lin, Y. and Su, Z. (2012). Balancing continuous and categorical baseline covariates in sequential clinical trials using the area between empirical cumulative distribution functions. Statistics in Medicine, 31:1961-1971.
Lock, K. F. (2011). Rerandomization to improve covariate balance in randomized experiments. Ph.D. Thesis, Harvard University,.
Ma, W., Hu, F., and Zhang, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. Journal of the American Statistical Association, 110(510):669-680.
Ma, Z. and Hu, F. (2013). Balancing continuous covariates based on kernel densities. Contemporary Clinical Trials, 34(2):262-269.
Meyn, S. P. and Tweedie, R. L. (2009). Markov Chains and Stochastic Stability. Cambridge University Press, 2nd edition edition.
Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. Annals of Statistics, 40(2):1263-1282.
Morgan, K. L. and Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. Journal of the American Statistical Association, 110(512):1412-1421.
Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics, 31(1):103-115.
Rubin, D. B. (2008). For objective causal inference, design trumps analysis. Annals of Applied Statistics, 2(3):808-840.
Stigsby, B. and Taves, D. R. (2010). Rank-minimization for balanced assignment of subjects in clinical trials. Contemporary Clinical Trials, 31(2):147-150.
Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. Clinical Pharmacology & Therapeutics, 15(5):443-453.
Wei, L. J. (1978). An application of an urn model to the design of sequential controlled clinical trials. Journal of the American Statistical Association, 73:559-563.