UtilitasMathematica

ISSN 0315-3681 Volume 122 (2), 2025

Selection of Comparable Subjects from Different Treatment Groups When Randomization is Not Feasible

Shyam Bihari Tiwari[1], Abhijeet Pandey[1], Ashwini Mathur[2], Asha Kamath[3]

[1]Novartis healthcare private limited (NKC), Hyderabad, India; [2]Onesto Consulting, Dublin, Ireland, [3]Prassna School of Public Health, MAHE, Manipal, Karnataka, India,

Abstract

Objective: The aim of this research is to find solutions for selecting comparable subjects in non-randomized situations across different treatments.

Method: A series of analysis were performed to find subjects with similar profiles. The Propensity Score (PS) was estimated using logistic regression with potential covariates and a Cartesian product of PS across the subjects and treatments was created. The absolute difference in PS was derived and the pair with the minimum difference in PS was obtained, finally dose effect was estimated from selected paired.

Results: There were 23 subjects in the low dose group and 186 in the high dose group under treatment to assess the effectiveness of the ADAS-cog score in terms of change from Baseline to Week 24. The unequal proportion of subjects between the two dose groups has raised questions about balance and bias. An optimum matching technique was used to find profiles similar to the 23 subjects in the high dose group. Although the result was non-significant, it was more reliable because it was based on similar profiles, inadequate sample size could be an issue. The re-sampling technique was utilized to generate 80 more samples from the original 23 subject data from low dose. The dose effect was again estimated and found to be significant at Week 16 and Week 24.

Conclusion: The PS method was able to find subjects with similar profiles across the dose groups and provide reliable results regarding dose differences. After re-sampling and making covariate adjustments, the dose difference was found significant.

KEY WORDS: PS, PSM, IPW, stratum, non-randomized, RWE, confounding, covariate balance

1. Introduction

The objective of this paper is to outline in detail the challenges and potential solutions for analyze the studies without the ability to randomize subjects into treatment groups. The randomization technique suggested by Sir Ronald Aylmer Fisher (1890-1962) has been accepted as the gold standard method whenever need to assess the treatment effect (mathematical comparison) of two or more treatments to obtain an unbiased estimate of results, independent of confounding variables. Randomization is a process of assigning treatments to subjects, assuming that each subject has an equal chance of participating in any of the treatments. It is assumed that in a randomized experiment, the random allocation of subjects to treatment arms approximately balances subject characteristics, removes selection bias, and promotes comparability of treatments. However, in the below situation, need to consider alternate methods of randomization.

1. Randomization through true experimental designs is not always possible, practical, logistical, even desirable in every situation.
2. Even in a randomized study, there are several factors that could affect the validity of results:
    a. The randomization process could be broken or compromised in some way, or
    b. Randomization may fail to recruit a similar profile of subjects across different treatments, or
    c. There may be a new group difference that was not accounted for in the randomization process or
3. Estimating treatment effects in Real-World Evidence (RWE) observational studies can be challenging due to the lack of randomization and other sources of bias.

4. There are various other scenarios in which one may encounter randomization. Some of them are listed below.
    - Ethical considerations: If the therapeutic benefit is already known, randomization may not be useful.
    - Lack of representativeness of the target population of interest in the selected sample, or:
        - Inadequate sample size (i.e, rare disease, small population), or
        - The disease is not under control, as in the case of pandemic (e.g., public health emergencies, war).
    - Randomization may not be useful in complex interventions and implementation research.
    - Randomized trials are useful only when studying narrowly defined groups of subjects under artificial laboratory conditions, and external validity is questionable.
    - Randomization may not be useful in "health policy research" (community trials). Resource allocation decisions, policy changes, or public health campaigns may be based on population needs, political considerations, or community input, rather than random assignment(Brownson, Chriqui et al. 2009).
    - Natural experiments: The introduction of a new policy or the implementation of a program in different locations can be used to study the causal effects without randomization.
5. There are some challenges with randomized studies as following.
    - Randomized trials are time consuming, and cost intensive, and require careful planning, recruitment, and allocation of subjects.
    - Randomization is not based on disease or subject characteristics but rather on the rule of probability and may fail to measure unmeasured confounding.
    - Randomized trial may fail to follow compliance due to subjects' refusal /drop out for any reason, leading to bias and potential confounding that undermines the effectiveness of randomization.

In 1998, the World Health Organization (Health Promotion Evaluation) made a recommendation to the policymakers that "The use of randomized controlled trials to evaluate health promotion is, in most cases, inappropriate, misleading, and unnecessarily expensive."((WHO) 1998) The International Union for Health Promotion went even further in 1999 and send a message to

researchers (challenging current evaluation approaches) stating that "randomized controlled trials or corresponding experimental designs should not be used to measure the effectiveness of health promotion interventions."

Drawing conclusions in non – randomized situations was challenging. In such cases, researcher found an alternative methods that yield results similar to those obtained from randomized trials. One such method is Propensity Score Matching (PSM). PSM is a statistical technique that can be used to identify pairs of subjects across treatments (not necessarily identical, but as similar as possible) based on subject characteristics.

In the case of two comparison groups, the Propensity Score (PS) for subject i (=1,2,-,-,n , is conditional probability of that subject being in one of the two treatments (Z_i=1 or Z_i=0), given a vector of observed covariates such as demographics, baseline and clinical characteristics, socioeconomic covariates or any other variable

$$exp(Xi) \; = \; Pr(Zi = 1|X = xi)$$

where it is assumed that, given value the variable X_i's the Z_i's are independent.
If two subjects have the same value of PS, it is assumed that they have the same distribution of covariates.

The PS represents a vector of covariates for each subject in a study, which balances the covariates across groups. When a pair of subjects from the treated and control group have a similar PS, they are considered comparable in terms of treatment, even if they may differ in the value of a specific characteristic. If the characteristics at baseline are as similar as possible in the comparable treatment, then the data was considered comparable. Any estimates obtained from this data was attributed to the treatment, rather than to any know/unknown confounding factors. The PS is a useful tool for controlling confounding variables in observational studies.

Paul R Rosenbaum and Donald B. Rubin proposed the central role of PS in observational studies to determine causal effects (Rosenbaum and Rubin 1983, Rubin 1997). PS are typically unknown in non-randomized studies, it was estimated with logistic regression model with potential known

covariates for each subject. Scalar PS is sufficient to remove bias caused by all observed covariates (demographics, Baseline characteristics, clinical characteristics, socio-economic status and other variables). Logistic regression employed to generate PS incorporating the selected covariates. Logistic regression is a data analysis technique used to establish a relationship between dependent variables (binary or multinomial) and independent variables.

The primary question of interest was how to choose pairs? How to define pairs are "similar"? If there are many such pairs then the comparison between two treatment groups is likely to be fair, approached that of a randomized trial. The matched groups of subjects can be utilized to compare the two treatments groups.

Matching can be performed with or without stratification. The central idea of matching is to match each treated subject to each control subject based on observed subject characteristics (covariates). Subject's characteristics is defined as vector of variables that have notable impact on treatment effect.

The key feature of the PSM method is that it compares each treated ($x_i/z_i=1$) subject to all control ($x_i/z_i=0$) subjects based on distance between PS, where x is vector of matching variable and z is received treatment(i=1) or received control (i=0).

To select comparable subjects, an appropriate PS matching method such as optimal matching, exact matching, nearest neighbor matching or other matching techniques can be selected. Rosenbaum (2002b) recommended the use of optimal matching to address the matching problem(Guo 2014). Optimal matching is defined as the smallest difference between the PS of treated subjects and comparator subjects. If subjects from the two groups have the same/approximately equal PS, it indicates that they have similar properties/profile. After generating PS, create a dataset with the Cartesian product of both treatment groups and select the paired subjects with the smallest difference in PS/logit score.

$$\min \; ||E((Xi/Ti = 1)) \; - \; (Xi/Ti = 0))||$$

There was a randomized study conducted to evaluate the effectiveness of Treatment (A) and Control (B) for a disease. The primary objective of the study was to determine that Treatment is not inferior to Control in terms of the change from Baseline to Week 24 of the specified parameter ADAS-Cog.

ADAS-Cog is a test that measures cognitive and behavioural impairments in patients with AD (Rosen, Mohs et al. 1984). The cognitive subscale of ADAS (ADAS-Cog) includes 11 items that are summed up to create a score ranging from 0 to 70, with lower scores indicating less severe impairment.

To test the hypothesis, a total of 501 subjects were recruited and randomly assigned to receive either Treatment (N=248) or Control (N=253). Out of these subjects, 380 individuals were included in the per-protocol set (A=192 and B=188). The per-protocol set was used to assess the primary endpoint (Table 1).

Table 1: Number (%) of patients in analysis population by treatment group (Randomized Set)

Population	**Treatment**	**Control 253**	**Total 501**
Randomized	248 (100.0)	253 (100.0)	501 (100.0)
Safety (SAF)	247 (99.6)	251 (99.2)	498 (99.4)
Intent to treat (ITT)	236 (95.2)	238 (94.1)	474 (94.6)
Per protocol (PP)	192 (77.4)	188 (74.3)	380 (75.8)
SAF: patients who received at least one dose of study drug, and had at least one post baseline safety assessment. ITT: patients who received at least one dose of study drug, and had a baseline assessment and at least one post baseline assessment of the primary efficacy variable ADAS-Cog. PP: patients who received at least one dose of study drug, had a baseline assessment and at least one post-baseline assessment on treatment (after Day 140 and not more than 2 Day after last known date of study drug) of primary efficacy variable and have no major protocol deviation.			

The demographics and Baseline characteristics indicate that the study successfully recruited patients with similar profiles (in both treatment groups (p-value ≥0.05) on randomized set (Table 2), with exception of the Baseline total MMSE score and the time since the first symptoms noticed (AD).

Table 2: Patients demographic and baseline characteristics by treatment group (Randomized Set)

Characteristic Statistics	**Treatment N=248**	**Control N=253**	**Overall N=501**	**P-value**
Sex -n(%)				
Male	108 (43.5)	114 (45.1)	222 (44.3)	0.7336
Female	140 (56.5)	139 (54.9)	279 (55.7)	

Characteristic Statistics	Treatment N=248	Control N=253	Overall N=501	P-value
Race -n(%)				
Asian	248 (100.0)	253 (100.0)	501 (100.0)	NE
Age (yrs (1))				
n	248	253	501	0.3946
Mean	70.3	69.7	70.0	
SD	8.12	8.15	8.13	
Median	72.0	72.0	72.0	
Range	(52, 87)	(52, 82)	(52, 87)	
Age group (yrs)(1) -n(%)				
< 65	60 (24.2)	75 (29.6)	135 (26.9)	0.2766
65 - < 75	92 (37.1)	97 (38.3)	189 (37.7)	
75 - < 85	95 (38.3)	81 (32.0)	176 (35.1)	
>= 85	1 (0.4)		1 (0.2)	
Family history of AD -n(%)				
No	212 (85.5)	223 (88.1)	435 (86.8)	0.3790
Yes	36 (14.5)	30 (11.9)	66 (13.2)	
Time since 1st symptoms noticed (AD) (yrs)				
n	248	253	501	0.0321
Mean	3.3	2.9	3.1	
SD	2.75	2.03	2.42	
Median	3.0	2.0	3.0	
Range	(0, 23)	(0, 10)	(0, 23)	
Time since 1st symptoms diagnosed (yrs)				
n	248	253	501	0.1636
Mean	1.0	0.8	0.9	
SD	2.22	1.52	1.90	
Median	0.0	0.0	0.0	

Characteristic Statistics	Treatment N=248	Control N=253	Overall N=501	P-value
Range	(0, 22)	(0, 9)	(0, 22)	
Number of years of formal education (yrs)				
n	248	253	501	0.5985
Mean	10.9	10.7	10.8	
SD	4.16	4.12	4.14	
Median	12.0	12.0	12.0	
Range	(2, 22)	(2, 22)	(2, 22)	
Baseline total MMSE score				
n	248	253	501	0.0392
Mean	16.0	16.6	16.3	
SD	3.46	3.08	3.28	
Median	16.0	17.0	17.0	
Range	(10, 24)	(10, 21)	(10, 24)	
Baseline NPI for Items 1-10				
n	247	252	499	0.4980
Mean	9.6	8.8	9.2	
SD	13.19	12.56	12.87	
Median	4.0	4.0	4.0	
Range	(0, 67)	(0, 77)	(0, 77)	
Baseline NPI distress Items 1-12				
n	247	252	499	0.7363
Mean	5.1	4.9	5.0	
SD	6.91	7.20	7.05	
Median	3.0	2.0	3.0	
Range	(0, 37)	(0, 45)	(0, 45)	
Baseline total Basic ADL				
n	248	252	500	0.6128

Characteristic Statistics	Treatment N=248	Control N=253	Overall N=501	P-value
Mean	20.2	20.3	20.2	
SD	3.05	2.59	2.82	
Median	22.0	21.0	22.0	
Range	(5, 22)	(4, 22)	(4, 22)	
Baseline total Instrumental ADL				
n	248	252	500	0.1053
Mean	30.9	32.6	31.8	
SD	11.90	10.96	11.46	
Median	31.0	34.0	32.5	
Range	(4, 54)	(5, 53)	(4, 54)	
Baseline for total ADAS-Cog score				
n	240	247	487	0.2192
Mean	29.1	28.0	28.6	
SD	9.53	9.17	9.36	
Median	28.0	27.0	28.0	
Range	(9, 53)	(8, 54)	(8, 54)	

Demographic characteristics are collected at Screening visit. (1) Age and weight reported at Baseline vi sit.
N: Number of patients in the Randomized population.
n: Number of patients meeting the criterion (categorical) and number of patients with a non-missing assessment (continuous).
P-valuie derived from t-test for continuous variable and chi-Square for categorial variable.
SD: Standard deviation, NE: Not applicable.

The results of the primary endpoint, specifically the change from Baseline in ADAS-Cog by treatment group, fail to reject the null hypothesis (p-value ≥0.05) at all evaluated time points. This indicates that Treatment has a similar effect as Control. However, it's noted that Control exhibited a slightly greater effect compared to Treatment (Table 3).

Table 3: Change from baseline in ADAS-Cog by Treatment (Per protocol population)

		Treatment N=192			**Control N=188**			**TRT A Vs TRT B**		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
Week 8	Baseline	184	28.9	9.15	183	28.1	9.29			
	Post baseline	184	28.2	9.56	183	27.1	9.32			
	Change	184	-0.7	5.58	183	-1.0	4.83			
	LS mean for change	184	-0.6	0.37*	183	-1.0	0.38*	0.4	(-0.6, 1.5)	0.438
Week 16	Baseline	179	28.9	9.22	181	28.2	9.39			
	Post baseline	179	27.5	9.58	181	26.7	9.72			
	Change	179	-1.4	5.99	181	-1.5	5.31			
	LS mean for change	179	-1.3	0.41*	181	-1.6	0.41*	0.2	(-0.9, 1.4)	0.696
Week 24	Baseline	182	29.0	9.19	185	28.2	9.33			
	Post baseline	182	28.4	10.42	185	27.4	10.50			
	Change	182	-0.5	6.70	185	-0.7	6.74			
	LS mean for change	182	-0.5	0.49*	185	-0.8	0.49*	0.3	(-1.1, 1.7)	0.653
*=Standard error, DLSM=Difference in least square mean, CI=Confidence interval.										

After above result (Table 3) there was an ad-hoc question on the performance of low dose (N=23) and high dose (N=186) within Treatment. The unequal distribution of subject between dose group had raised following question.

1. Is data in the dose groups low and high have balanced and comparable, considering that the data was not randomized for dose groups? Is there a concern regarding the balancing of subjects' attributes in the dose group?
2. Is the comparison of the dose groups valid without adjusting for covariates?
3. How can it be ensured that the results are unbiased?
4. Is the estimated difference between the dose groups solely due to the dose group factor, or is it influenced by other covariates?

These questions highlight the need to address potential biases and confounding factors when comparing the low and high dose groups. It is important to take into account relevant covariates and employ appropriate statistical techniques to ensure the validity and accuracy of the results.

Based on the above discussion, it can be concluded that comparison a low dose with N=23 vs. a high dose with N=186 cannot be justified since:

- Insufficient sample size.
- There is no clarity on balancing of data.
- Un equal proportion, it’s typically less power.
    - Power =.44, with effect size 2 and pooled SD=5, A=23, B=186 (Table S 4).

Table 4: Two-Sample T-Test Power Analysis

Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1≠Mean2
The standard deviations were assumed to be known and equal.

Power	N1	N2	Allocation Ratio	Alpha	Beta	Mean1	Mean2	S1	S2
0.43993	23	184	8.000	0.05000	0.56007	30.0	28.0	5.0	5.0

The estimated difference in ADAS-Cog scores for the dose groups was found to be insignificant at all assessed time points (Table 5). A higher value of ADAS-Cog score indicates cognitive impairment. The results highlight need for a better approach to select comparable subjects, considering the reasons mentioned above.

Table 5: Change from baseline in ADAS-Cog by dose group for treatment before matching (Safety population)

		Low Dose N=23			High Dose N=186			Low Dose Vs High Dose		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
Week 8	Baseline	22	26.7	7.94	177	29.3	9.40			
	Post baseline	22	27.5	9.27	177	28.6	9.87			
	Change	22	0.8	6.18	177	-0.7	5.64			
	LS mean for change	22	0.5	1.20*	177	-0.7	0.42*	1.2	(-1.3, 3.7)	0.342
Week 16	Baseline	23	27.4	8.55	171	29.3	9.49			
	Post baseline	23	28.3	8.85	171	27.7	9.96			
	Change	23	0.8	4.20	171	-1.6	6.20			
	LS mean for change	23	0.6	1.22*	171	-1.6	0.45*	2.2	(-0.4, 4.7)	0.095
Week 24	Baseline	20	27.3	8.11	169	29.2	9.42			
	Post baseline	20	28.6	10.02	169	28.5	10.62			
	Change	20	1.4	5.84	169	-0.7	6.78			
	LS mean for change	20	1.2	1.48*	169	-0.7	0.51*	1.8	(-1.3, 4.9)	0.244

In order to address the effectiveness of dose group the new null hypothesis (H_0): both dose group have equal effect in ADAS-Cog score vs alternative hypothesis(H_1): There is difference between dose group in ADAS-Cog score. To assess the new hypothesis, a minimum of 200 subjects will be required (100 subjects in each dose group). This calculation is based on a baseline mean of ADAS-Cog of 30 at Baseline and a post-baseline mean of 28 at Week 24 (Cummings, Froelich et al. 2012), assuming that both dose groups have equal standard deviations of 5, 80% power, and an alpha level of 5% (level of significance). The sample size requirement should be determined accordingly.

$$n=n*W/(\Delta/\sigma)^2$$

W= Depend on power (1-β) and level of significance.

Δ=Effect size. Depend on difference of null (from reference or source data) and alternative hypothesis (expected value in current value).

σ= Population variability from source data.

$$2*(Z_{1-\alpha/2} + Z_{1-\beta})^2/ (\Delta/\sigma)^2$$

$$= \quad 2*(.84+1.96)^2/ (30-28/5)^2$$

$$2*7.84/ (2/5)^2$$

$$= 2*7.84/ (4/25)$$

$$= 15.68/.16$$

= 98~100 subjects per dose groups

In non-randomized studies, the ignorability of treatment allocation assumption holds, because the creation of treatment comparison groups follows a natural process that confounds treatment assignment with outcomes. With strongly ignorable treatment allocation, pair matching on PS can provide an unbiased estimate of the treatment effect. Therefore, in any evaluation, it is important to assess the tenability/plausibility of independence between the treatment assignment and outcome under different conditions. For this purpose, bivariate analysis can be conducted using the treatment groups obtained from the PSM method. A Chi square test can be applied to assess the association between the treatment group for categorical variables, testing the null hypothesis H_0: there is no association between the treatment group against the alternative hypothesis H_1: there is an association between treatment group. A t-test can be used to assess equality of two treatment group for continuous variable with null hypothesis H_0: treatment group have equal profile against H_1: both treatments have different profile. The PS can be utilized to assess overall similarity between treatment. It's better to performed covariate balance check before and after matching. This can be estimated using below formula:

$$dx = |Mxt - Mxc|/ Sx$$

where *Mxt*= mean of covariate in treated group

Mxc= mean of covariate in comparator group

Sx= standard deviation of mean difference (known) or

pooled standard deviation of treated and comparator group (unknown).

The standardized difference in case of dichotomous response variable.

$$dx = |\hat{P}xt - \hat{P}xc| / sqrt((\hat{P}xt\ (1 - \hat{P}xt) + \hat{P}xc\ (1 - \hat{P}xc))/2$$

Pxt= Proportion of covariate in treated group

Pxc= Proportion of covariate in comparator group

The balance check can be performed in following way.

1. If dx after matching is lower than before matching, then it can be concluded that sample balance improved after matching.
2. The strongest of dx (covariate balance) can be categorized in fallowing way:

Range for balance score (dx)

Range	**Group**
0 < |dx| < 0.1	Balance
0.1 > |dx| < 0.2	Moderate Balance
|dx|=> 0.2	Unbalanced

In case of unbalanced situation, the specific variable advised to drop from the model or re define the model.

2. Methodology

To assess the effectiveness of treatments in a study, that is not randomized or where randomization is not ethical, a series of analysis needs to be performed through matching of subject profiles based on PS, PS can be defined as subject participant's chance (based on their characteristics) of receiving one of two treatments/groups. Comparable subjects can be identified using the following steps:

1. Select potential variables, from demographics, baseline and clinical characteristics, socioeconomic factors and other relevant variables. In case of unknown independent variables, an appropriate selection method (such as Backward, Forward or Stepwise) can be used with logistic regression.

2. Carry out PS/logit score matching
    a. Generate PS from logistic regression using the variables selected in S-1.
    b. Generate the Cartesian product of PS from both treatment groups and calculate the absolute difference in PS across the treated and comparison subjects.
3. Select an appropriate matching method such as optimum matching or nearest neighbour matching.
4. Select subjects with similar profiles from treated group to match with subjects from control group. The smallest difference in PS between two subjects across the treatments can be considered as a similar profile of subjects.
5. Repeat the matching process until all subjects in the treated group are matched to subjects in the control group.
6. Perform a covariate balance test on the variables for the selected subjects from PSM between treatments. The balance check should be performed before and after matching. If the difference in covariate balance after matching is smaller than before matching, then PSM has successfully reduced bias and confounding.
7. Merge PS with the selected similar profile subjects at the subject level.
8. Perform the expected analysis based on the selected subjects from PSM as usual.
9. Compare the results with and without applying PS as covariate. If the results obtained from PSM method are less biased than the results obtained from a simple comparison of two treatments, then the method can be considered successful.

2.1 Matching method/matrices

There are many matching methods available, and researchers may select a matching method according to their needs. The following are popular matching methods:

2.1.1 Matching method

1) Optimum Matching: This method matches units based on their covariates, specifically the PS. The aim is to minimize differences in PS between treated and control units. Optimum matching aims to create a complete matched pair sample, where every treatment unit is matched with at least one control unit. This method ensures that the matched pairs are as similar as possible in terms of their covariates.

2) Greedy Nearest Neighbor Matching: This method selects control units that have covariate values closest to those of the treated unit. It is called "greedy" because it chooses the nearest neighbor without considering the impact on future matches. The matching is done without replacement, meaning that a control unit can only be matched to one treated unit. Greedy nearest neighbor matching can result in both complete and incomplete matched-pair samples, depending on the availability of suitable control units.
3) Replacement Matching**:** In replacement matching, a control unit can be matched with multiple treated units. This method is useful when there are limited control units available relative to the number of treated units. Replacement matching allows for a larger sample size in the matched pairs, as the same control unit can be matched with different treated units. This method can increase the efficiency of the matching process but may also introduce more variation between the matched pairs.

These matching methods offer different approaches to creating matched pairs based on the similarity of covariates between treated and control units. The choice of method depends on the specific research question, available data, and the desired balance between precision and sample size.

2.1.2 Matching metrics

The following matching matrices can be utilized.

1. Smallest difference in logit of the propensity score: This method involves matching units based on the smallest difference in the logit of their PS. The PS is the estimated probability of receiving treatment and transforming it into a logit scale allows for a continuous scale. Matching based on the smallest difference in logit of PS ensures that units with similar probabilities of treatment are paired together.
2. Smallest difference in PS**:** This method matches units based on the smallest difference in their PS, without any transformations. The PS is estimated based on observed covariates, and matching is done to pair units with similar probabilities of receiving treatment. This method is straightforward and easy to implement.
3. Smallest Mahalanobis distance for continuous variables: This method matches units based on the smallest Mahalanobis distance, which takes into account the covariance structure of the continuous covariates. The Mahalanobis distance measures the distance between two

units in the multivariate space of the continuous variables. Matching based on the smallest Mahalanobis distance ensures that units with similar patterns of continuous covariates are paired together. This distance can only be applied in bivariate analysis. If more than two variables are involved, then this distance may not apply for a vector of variables.

Matching metrices can be specified as

Distance= PS/

LPS/

MAH (lps var= (var1, var2))

These matching methods aim to minimize differences between units in terms of their PS or continuous covariates. The selection of a specific method depends on the factors such as research question, the characteristics of the data, and the desired level of precision in matching.

To further enhance the matching process, these matching metrics can be paired with appropriate calipar. A calipar determines the maximum allowable difference in covariate values between matched units. For example, if a calipar value of 0.02 is specified, potential matches will only be considered if their covariate values are within this fixed distance of each other. However, if CALIPAR=. specified, the calipar requirement will be ignored. This calipar restriction can help ensure that the matched units are as similar as possible in terms of their covariates.

2.2 Inverse Probability of Weighting (IPW)

In addition to matching using PS, IPW method could also be used to analyse the data. IPW can be utilized to achieve a balanced distribution of covariates; these weights adjust the covariate values to create a pseudo-population in which the treated and control groups have a similar distribution of covariates. In our case, we have a smaller and larger group, it can use for each patient the weights as given below: -

IPW=1/PS, (smaller group)

or

1/ (1- PS), (larger group)

We will follow these steps to use IPW.

1. Specify the PS model.
2. Convert the PS to IPW

3. Convert each patients' covariates using the weights as given above. Check the weighted covariates for balance.
4. Estimate the treatment effect based on weighted covariates
5. Compare IPW method to PS method using the parameters given in section 6

3. Results

The objective of this paper was to discuss the challenges and solution in non-randomized situations. There was various scenario discussed where randomization was not supportive/ethical and top of that health agency also raised similar concerns. This situation had triggered to researcher to search for method which can give unbiased and unconfounded result in non-randomized situations.

3.1 PSM

As discussed, PS refers to the conditional probability of receiving treatment based on individual's background characteristics. PS score acts as a balancing score that represents a vector of covariates, such as demographics, Baseline characteristics, clinical characteristics, socio-economic variables, and others. PSM can be helpful when randomization is not feasible.

For each subject i (i = 1, …, N), the PS represents the conditional probability of being assigned to the high group ($Z_i = 1$) or the low group ($Z_i = 0$) given a vector x_i of observed covariates. These covariates include sex, age, family history of Alzheimer's disease (AD), time since the first symptoms were noticed (AD) (years), time since the first symptoms were diagnosed (years), number of years of formal education (years), Baseline total MMSE score, Baseline NPI for items 1-10, Baseline total basic ADL, Baseline total instrumental ADL, and Baseline total ADAS-Cog score.

$$exp(Xi) \; = \; Pr\,(Zi = 1|X = xi)$$

where it is assumed that, given the X_i's the Z_i's are independent.

With the same value of propensity score, it is assumed to have the same distribution of covariates.

Table 6: Estimation of variable and PS for a subject

Obs	Variable		Class Value	Estimate (β_i)	Aval (X_i)	Estimate* Aval (β_i* X_i)
1	Intercept			-2.77		-2.77
2	AGE_1N	Age		0.05	77	4.09
3	FAMHIS1C	Family history of AD	No	0.74		0.74
4	SEX1C	Gender	Female	0.02		
5	TIMSYM1N	Time since 1st symptoms noticed (AD) (yrs)		-0.18	2	-0.36
6	TIMSYM2N	Time since 1st symptoms diagnosed (yrs)		0.13	0	0
7	TOT_BSMS	Baseline total MMSE score		-0.05	18	-0.90
8	TOTBADLB	Baseline total Basic ADL		-0.04	22	-0.91
9	TOTIADLB	Baseline total Instrumental ADL		0.01	35	0.44
10	TOT_BSAD	Baseline for total ADAS-Cog score		-0.04	15	-0.57
11	WGT_1N	Weight (kg)		-0.003	69.4	-0.23
12	YRSEDU1N	Number of years of formal education (yrs)		-0.09	7	-0.59
13	NPIBS10	Baseline NPI for Items 1-10		-0.05	0	0
14	DISBS12	Baseline NPI distress Items 1-12		0.11	0	0
Total Sum						-1.39

$$PS = \frac{\exp(\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \cdots + \beta nXn)}{1+\exp(\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \cdots + \beta nXn)}$$

$$= \exp(-1.4)/1+\exp(-1.4)$$

$$= (2.81)^{-1.4}/(1+(2.81)^{-1.4}$$

$$= 0.24 /(1+0.24)$$

$$= 0.\ 24 /1.24$$

$$= 0.19$$

Similar way PS for all subjects can be calculated.

3.2 Selection of Comparable subjects

In order to get comparable subjects, the following steps are expected to be followed:

From Table 5, it is observed that the low dose group consists of 23 subjects, while the high dose group consists of 186 subjects. Therefore, the goal is to select 23 subjects from the high dose group who possess a similar profile (near PS) to that of the low dose group. Various matching methods are available, and researcher should choose a method based on their specific requirement. The following are popular matching methods with appropriate matric:

1. Optimum
2. Exact
3. Nearest available

In this manuscript, the optimal matching method was employed to identify subjects with similar profiles from the high dose group. The rationale behind selecting of the optimal matching method is to ensure the inclusion of all subjects from the low dose group in the analysis. Before proceeding with the matching process, it is a necessary to create a Cartesian product of PS, combining all subjects from both the low and high dose groups. This allows for the calculation of the difference in PS among the merged subjects. After the merged dataset is created, it is important to sort the data by the subject identifier from the low dose group, the absolute difference of stratum (if applicable) and absolute difference in PS. Then, select the first record for each subject from low dose group. In case any subjects were selected multiple times from the high dose group, then select the subject with a smaller difference in PS.

Table 7: Selection of comparable subjects based nearest PS

Obs	Subject ID Low dose	Subject ID High dose	Propensity score for Low dose	Propensity score for High dose	Absolute difference in propensity score
1	**0028**	**0025**	0.12275	0.12296	0.00022
2	0028	0012	0.12275	0.12316	0.00041
3	0028	0003	0.12275	0.12321	0.00046
4	0028	0007	0.12275	0.12733	0.00459
5	0028	0012	0.12275	0.12756	0.00481
6	0028	0024	0.12275	0.12809	0.00534
7	0028	0013	0.12275	0.11549	0.00726
8	0028	0032	0.12275	0.11535	0.00740
173	**0007**	**0005**	0.36668	0.36963	0.00294
174	0007	0003	0.36668	0.35794	0.00875
175	0007	0020	0.36668	0.35338	0.01330
176	0007	0016	0.36668	0.40318	0.03650
177	0007	0006	0.36668	0.31902	0.04766
178	0007	0042	0.36668	0.30055	0.06613
179	0007	0004	0.36668	0.43724	0.07055
180	0007	0006	0.36668	0.28187	0.08481
181	0007	0006	0.36668	0.28155	0.08513

The above table are based on author assumption.

3.3 Before and after PSM

Balancing score

After PSM, it was observed that the balancing score improved for variables Age, number of years of formal education (yrs), Baseline total MMSE score, Baseline total instrumental ADL and Baseline for total ADAS-Cog score (Table 8).

This improvement in balancing scores indicates that after matching, the distribution of these variables became more similar between the dose groups. This suggests that the PSM method successfully reduced the imbalance in these variables, which is important for ensuring comparability between the groups and obtaining reliable treatment effect estimates.

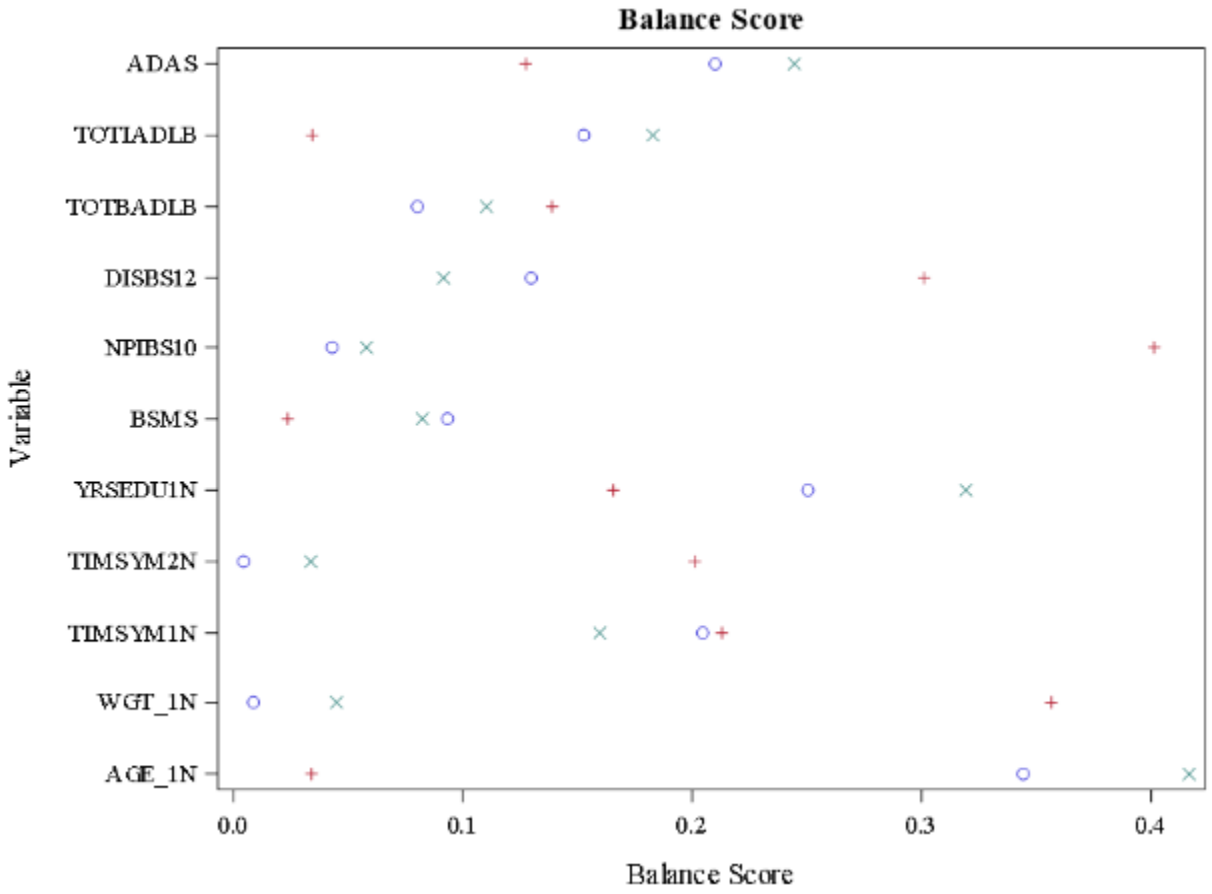
Table 8: Balancing score of dose group for treatment before, after matching and re-sampling

	Pre-matching			Post matching				Post Sampling			
Variable	**Mean**	**SD**	**Score**	**Mean**	**SD**	**Score**	**Balance status**	**Mean**	**SD**	**Score**	**Balance status**
Age (yrs)	2.8	7.99	0.344	0.2	6.39	0.034	IMP	3.0	7.31	0.417	Non-IMP
Weight (kg)	0.1	10.91	0.009	3.9	10.95	0.356	Non-IMP	0.5	11.10	0.045	Non-IMP
Time since 1st symptoms noticed (AD) (yrs)	0.6	2.85	0.205	0.4	2.04	0.213	Non-IMP	0.4	2.52	0.160	IMP
Time since 1st symptoms diagnosed (yrs)	0.0	2.35	0.004	0.4	1.95	0.201	Non-IMP	0.1	2.15	0.034	Non-IMP
Number of years of formal education (yrs)	1.0	4.17	0.250	0.7	3.94	0.166	IMP	1.3	4.22	0.319	Non-IMP
Baseline total MMSE score	0.3	3.37	0.093	0.1	3.68	0.024	IMP	0.3	3.42	0.082	IMP
Baseline NPI for Items 1-10	0.6	13.41	0.043	6.1	15.17	0.401	Non-IMP	0.7	12.77	0.058	Non-IMP
Baseline NPI distress Items 1-12	0.9	7.05	0.130	2.6	8.52	0.301	Non-IMP	0.6	6.82	0.092	IMP
Baseline total Basic ADL	0.3	3.20	0.080	0.5	3.44	0.139	Non-IMP	0.3	2.87	0.110	Non-IMP
Baseline total Instrumental ADL	1.8	11.67	0.153	0.4	11.41	0.034	IMP	2.1	11.40	0.183	Non-IMP
Baseline for total ADAS-Cog score	2.0	9.30	0.210	1.1	8.36	0.127	IMP	2.2	9.17	0.245	Non-IMP

Means are absolute mean difference between low and high dose group.
Score is balancing score derived on absolute mean difference between low and high dose group, divided by polled SD
If balance score decrease after matching then it's consider balancing between dose group had improved.
Balance: IMP- Improvement, Non-IMP- Not Improvement, SD- Stable Disease

After propensity score matching, it was observed that the balancing score improved for variables Age, number of years of formal education (yrs), Baseline total MMSE score, Baseline total instrumental ADL Baseline for total ADAS-Cog score (Table 8).

This improvement in balancing scores indicates that after matching, the distribution of these variables became more similar between the dose groups. This suggests that the PSM method successfully reduced the imbalance in these variables, which is important for ensuring comparability between the groups and obtaining reliable treatment effect estimates.



Source: Table S8

Analysis of ADAS-Cog score after matching

After conducting the analysis, it was observed that the difference in least square mean for ADAS-Cog (Table 9) was similar to before matching but more trustable because it based on similar profile of subjects. However, it should be noted that the results were not statistically significant, suggesting that the observed differences may be due to chance and additionally, the sample size might be a problem. In low dose group, only 23 subjects were available, which falls short of the minimum desired sample size of 100 subjects for each dose group. To address this limitation, a re-sampling technique could be employed to increase the sample size.

Table 9: Change from baseline in ADAS-Cog by dose group for treatment after matching (Safety population)

		Low Dose N=23			High Dose N=186			Low Dose Vs High Dose		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
Week 8	Baseline	22	26.7	7.94	22	28.5	8.15			
	Post baseline	22	27.5	9.27	22	27.9	10.12			
	Change	22	0.8	6.18	22	-0.6	5.21			
	LS mean for change	22	0.8	1.24*	22	-0.6	1.24*	1.4	(-2.1, 5.0)	0.427
Week 16	Baseline	23	27.4	8.55	20	28.2	8.32			
	Post baseline	23	28.3	8.85	20	27.9	9.50			
	Change	23	0.8	4.20	20	-0.4	4.40			
	LS mean for change	23	0.8	0.90*	20	-0.3	0.97*	1.1	(-1.5, 3.8)	0.393
Week 24	Baseline	20	27.3	8.11	19	27.9	8.47			
	Post baseline	20	28.6	10.02	19	27.6	9.90			
	Change	20	1.4	5.84	19	-0.4	5.54			
	LS mean for change	20	1.3	1.29*	19	-0.4	1.33*	1.7	(-2.0, 5.5)	0.362

3.4 Re-sampling

Given the limitation of having only 23 subjects in the low dose group and considering the minimum sample size requirement of 100 each dose group, it becomes necessary to obtain additional 80 records for ADAS-Cog, based 23 subjects' records.

The analysis revealed that the differences in least square means of change from Baseline for ADAS-Cog were to be statistically significant at Week 16 and Week 24 (Table 10). However, this difference was insignificant at Week 8. It was observed that the high dose group exhibited better performance than the low dose group at both Week 16 and Week 24.

Table 10: Change from baseline in ADAS-Cog by dose group for treatment after re-sampling (Safety population)

		Low Dose N=23			High Dose N=186			Low Dose Vs High Dose		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
Week 8	Baseline	102	26.4	7.86	177	29.3	9.40			
	Post baseline	102	27.5	9.37	177	28.6	9.87			
	Change	102	1.1	6.83	177	-0.7	5.64			
	LS mean for change	102	0.9	0.60*	177	-0.6	0.45*	1.4	(-0.0, 2.9)	0.056
Week 16	Baseline	103	27.1	8.77	171	29.3	9.49			
	Post baseline	103	28.3	8.83	171	27.7	9.96			
	Change	103	1.1	4.47	171	-1.6	6.20			
	LS mean for change	103	0.9	0.54*	171	-1.5	0.42*	2.4	(1.1, 3.8)	<.001
Week 24	Baseline	100	27.0	7.76	169	29.2	9.42			
	Post baseline	100	28.3	9.71	169	28.5	10.62			
	Change	100	1.4	5.67	169	-0.7	6.78			
	LS mean for change	100	1.3	0.64*	169	-0.6	0.49*	1.9	(0.3, 3.5)	0.020

3.5 Covariate adjustments

The change from Baseline for ADAS-Cog were to be statistically significant at all assessed time point with PS as covariate (Table 11) and only PS as covariate (Table 12).

These findings suggest that the high dose had a more positive impact on improvement of ADAS-Cog scores compared to the low dose group.

Table 11: Change from baseline in ADAS-Cog with as a covariate by dose for treatment after re-sampling (Safety population)

		Low Dose N=23			High Dose N=186			Low Dose Vs High Dose		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
Week 8	Baseline	102	26.4	7.86	177	29.3	9.40			
	Post baseline	102	27.5	9.37	177	28.6	9.87			
	Change	102	1.1	6.83	177	-0.7	5.64			
	LS mean for change	102	1.0	0.63*	177	-0.7	0.47*	1.7	(0.1,3.3)	0.042
Week 16	Baseline	103	27.1	8.77	171	29.3	9.49			
	Post baseline	103	28.3	8.83	171	27.7	9.96			
	Change	103	1.1	4.47	171	-1.6	6.20			
	LS mean for change	103	0.7	0.56*	171	-1.4	0.43*	2.0	(0.6,3.5)	0.006
Week 24	Baseline	100	27.0	7.76	169	29.2	9.42			
	Post baseline	100	28.3	9.71	169	28.5	10.62			
	Change	100	1.4	5.67	169	-0.7	6.78			
	LS mean for change	100	2.0	0.64*	169	-1.1	0.48*	3.1	(1.5,4.8)	<.001

Table 12: Change from baseline in ADAS-Cog with only PS as covariate by dose group for Treatment after re-sampling on PS as covariate (Safety population)

		Low Dose N=23			High Dose N=186			Low Dose Vs High Dose		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
Week 8	Baseline	102	26.4	7.86	177	29.3	9.40			
	Post baseline	102	27.5	9.37	177	28.6	9.87			
	Change	102	1.1	6.83	177	-0.7	5.64			
	LS mean for change	102	1.1	0.64*	177	-0.7	0.47*	1.8	(0.2,3.5)	0.028
Week 16	Baseline	103	27.1	8.77	171	29.3	9.49			

		Low Dose N=23			High Dose N=186			Low Dose Vs High Dose		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
	Post baseline	103	28.3	8.83	171	27.7	9.96			
	Change	103	1.1	4.47	171	-1.6	6.20			
	LS mean for change	103	0.7	0.58*	171	-1.4	0.44*	2.1	(0.6,3.6)	0.006
Week 24	Baseline	100	27.0	7.76	169	29.2	9.42			
	Post baseline	100	28.3	9.71	169	28.5	10.62			
	Change	100	1.4	5.67	169	-0.7	6.78			
	LS mean for change	100	2.1	0.65*	169	-1.1	0.49*	3.2	(1.6, 4.9)	<.001

3.6 IPW

Apart from PSM, IPW was used to estimate difference of the dose group for change from Baseline of ADAS-Cog score, but similar non-significant results were observed (Table 13).

Table 13: Change from baseline in ADAS-Cog with IPW of covariate by dose group for Treatment (Safety population)

		Low Dose N=23			High Dose N=186			Low Dose Vs High Dose		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
Week 8	Baseline	22	26.7	7.94	22	28.5	8.15			
	Post baseline	22	27.5	9.27	22	27.9	10.12			
	Change	22	0.8	6.18	22	-0.6	5.21			
	LS mean for change	22	0.8	1.26*	22	-0.6	1.26*	1.4	(-2.3, 5.1)	0.463
Week 16	Baseline	23	27.4	8.55	20	28.2	8.32			
	Post baseline	23	28.3	8.85	20	27.9	9.50			
	Change	23	0.8	4.20	20	-0.4	4.40			
	LS mean for change	23	0.9	0.92*	20	-0.4	1.00*	1.3	(-1.5, 4.1)	0.350
Week 24	Baseline	20	27.3	8.11	19	27.9	8.47			
	Post baseline	20	28.6	10.02	19	27.6	9.90			
	Change	20	1.4	5.84	19	-0.4	5.54			
	LS mean for change	20	1.1	1.31*	19	-0.1	1.34*	1.2	(-2.7, 5.0)	0.552

		Low Dose N=23			High Dose N=186			Low Dose Vs High Dose		
Visit		**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**DLSM**	**95% CI**	**P-value**
IPW: Inverse probability weighting										

3.7 Stratum of PS

The stratification of the PS is an effective technique for reducing bias and confounding in observational studies, ultimately improving the reliability of treatment evaluations. This methodology facilitates the balancing of covariate distribution across treatment and control groups within each stratum, thereby mitigating bias. By grouping individuals with similar propensities together, stratification ensures that comparisons are made between subjects who possess similar characteristics. As a result, the estimation of treatment effects becomes more accurate, facilitating a more precise evaluation of the impact of the treatment. In the pilot study stratification was not an issue.

Table 14 Selection of subject based on nearest stratum and PS

Subject ID from Low dose	Subject ID from Low dose	PS from low Dose	PS High Dose	Stratum low dose	Stratum High dose	Absolute stratum difference	Absolute difference in PS
21	30	0.48	0.37	10	8	2	**0.1**
21	45	0.48	0.34	10	7	3	0.13
21	51	0.48	0.32	10	7	3	0.15

Table 15: Selection of comparable subjects within stratum

Subject ID from Low dose	Subject ID from Low dose	PS from low Dose	PS High Dose	Stratum low dose	Stratum High dose	Absolute stratum difference	Absolute difference in PS

31	40	0.178	0.1743	4	4	0	**0.003**
31	38	0.178	0.1723	4	4	0	0.005
31	55	0.178	0.1846	4	4	0	0.006

The above table based on the author assumption.

Stratification was not the issue for exercised study, similar pairs of subjects were observed after stratification, so similar results should appear after applying the stratification.

3.8 Model Variance

The analysis revealed that the model variances were found to be significant at all assessment time points, indicating unequal variances. After matching lower variance were observed, indicates that the estimated parameters are more precise and stables.

Table 16 Analysis of covariance for ADAS-Cog for Treatment

Visit	Stage	Variance	Standard error	95% CI	P-value
Week 8	Pre-matching	31.2	3.151	(25.8 - 38.43)	<.0001
	Post-matching	33.4	7.385	(22.6 - 54.37)	<.0001
	Sampling	35.8	3.045	(30.5 - 42.58)	<.0001
	PS as a covariate	35.8	3.056	(30.5 - 42.67)	<.0001
	Only PS as a covariate	33.4	7.387	(22.6 - 54.39)	<.0001
	IPW of a covariate	37.3	3.176	(31.8 - 44.41)	<.0001
	IPW and PS as covariate	34.2	7.652	(23.1 - 56.02)	<.0001
Week 16	Pre-matching	34.1	3.490	(28.2 - 42.13)	<.0001
	Post-matching	18.1	4.038	(12.2 - 29.57)	<.0001
	Sampling	29.7	2.553	(25.3 - 35.44)	<.0001
	PS as a covariate	29.6	2.546	(25.2 - 35.28)	<.0001
	Only PS as a covariate	18.8	4.213	(12.7 - 30.85)	<.0001
	IPW of a covariate	31.5	2.707	(26.8 - 37.57)	<.0001
	IPW and PS as covariate	18.5	4.194	(12.4 - 30.53)	<.0001
Week 24	Pre-matching	43.9	4.548	(36.1 - 54.33)	<.0001
	Post-matching	29.2	6.886	(19.3 - 49.30)	<.0001
	Sampling	40.4	3.507	(34.4 - 48.30)	<.0001
	PS as a covariate	37.4	3.250	(31.8 - 44.70)	<.0001
	Only PS as a covariate	32.5	7.666	(21.5 - 54.88)	<.0001
	IPW of a covariate	40.5	3.512	(34.4 - 48.37)	<.0001

Visit	Stage	Variance	Standard error	95% CI	P-value
	IPW and PS as covariate	30.0	7.162	(19.7 - 50.98)	<.0001

3.9 Information Criteria

Fit statistics are metrics used to evaluate how well a statistical model fits the observed data. Lower fit statistics were observed after matching and/or where only PS was used as a covariate. This indicates an improvement in the overall fit of the model after matching, as lower value of model fit statistics generally indicates a better fit between the observed data and fitted model.

Table 17: Analysis of information criteria (fit statistics) for ADAS-Cog for Treatment

Visit	Stage	Neg2LogLike	AIC	AICC	CAIC	BIC	HQIC
Week 8	Pre-matching	1248.5	1250.5	1250.6	1254.8	1253.8	1251.9
	Post-matching	266.8	268.8	268.9	271.6	270.6	269.5
	Re- sampling without PS as covariate	1790.3	1792.3	1792.4	1797.0	1796.0	1793.8
	Re- sampling PS as Covariate	1786.2	1788.2	1788.3	1792.9	1791.9	1789.7
	Re - sampling only PS as covariate	282.4	284.4	284.5	287.1	286.1	285.0
	IPW of a covariate	1809.1	1811.1	1811.1	1815.7	1814.7	1812.6
	IPW and PS as covariate	277.3	279.3	279.4	282.0	281.0	279.9
Week 16	Pre-matching	1234.2	1236.2	1236.2	1240.4	1239.4	1237.5
	Post-matching	235.8	237.8	237.9	240.4	239.4	238.4
	Re-sampling	1708.1	1710.1	1710.1	1714.7	1713.7	1711.5

Visit	Stage	Neg2LogLike	AIC	AICC	CAIC	BIC	HQIC
	Re- sampling without PS as covariate	1702.4	1704.4	1704.4	1709.0	1708.0	1705.9
	Re-sampling only PS as covariate	253.1	255.1	255.2	257.8	256.8	255.7
	IPW of a covariate	1731.6	1733.6	1733.6	1738.2	1737.2	1735.1
	IPW and PS as covariate	246.9	248.9	249.0	251.6	250.6	249.5
Week 24	Pre-matching	1248.9	1250.9	1250.9	1255.1	1254.1	1252.2
	Post-matching	229.9	231.9	232.0	234.5	233.5	232.4
	Re- sampling without PS as covariate	1758.7	1760.7	1760.7	1765.3	1764.3	1762.1
	Re-sampling PS as Covariate	1733.2	1735.2	1735.2	1739.8	1738.8	1736.6
	Re - sampling only PS as covariate	249.4	251.4	251.5	254.0	253.0	252.0
	IPW of a covariate	1766.3	1768.3	1768.3	1772.9	1771.9	1769.7
	IPW and PS as covariate	240.4	242.4	242.5	245.0	244.0	243.0

AIC: Akaike Information Criterion, AICC: Akaike Information Criterion Corrected, CAIC: Consistent Akaike Information Criterion, BIC: Bayesian Information Criterion, HQIC: Hannan-Quinn Information Criterion.

4. Discussion

The randomization technique has been accepted as Gold Standard for estimate the treatment difference between two or more treatments, However, randomization is not feasible in every situation, and even in randomized studies, it can sometimes be considered unethical. In these cases, it cannot be guaranteed that the estimate of the treatment difference is unbiased and unaffected by confounded variables. This situation has motivated statisticians to search for a method that can provide identical results as to a randomized trial without conducting a new trial. There have been

many discussions on this topic, but statisticians had not reached a consensus yet. Statistician Rosenbaum & Rubin and econometrician Heckman have made substantial contributions by developing and refining new approaches for estimating treatment effects from observational study (Heckman 1978, Heckman 1979, Rosenbaum and Rubin 1983, Rubin 1997). These approaches collectively are known as PSM. PSM helps to minimize the differences between treatments in regard to known factors or specific threats to experimental validity.

In a randomized trial, it is generally assumed that random allocation of treatment balances the attribute of the subjects and allows for causal inference, as it helps control for nuisance factors at bay. However, this assumption is not always true, as randomization can sometimes lead to imbalance. Imbalances in observed variables can indicate imbalances in unobserved variables. This situation has once again motivated statisticians to search for a method that can be used when randomization is not feasible. PSM help to address this issue by identifying pair of subjects that have similar characteristics. This method can, be particularly useful when randomization fails or is impossible (Barth, Greeson et al. 2007). PSM aims to reduce the bias caused by confounding variables between subjects who received the treatment and those who did not.

Sample size can pose a challenge in PSM method particularly when there is significant discrepancy in the number of subjects between treatment groups. In the example of an observational study, researchers have two dose groups: Group A (N=23) and Group B (N=186). Due to the unequal proportion of subjects between the treatment groups, the many questions arise.

This limitation in sample size can also impact the statistical power of the analysis. PSM tends to work better with large sample sizes, as it allows for a more robust matching process. Therefore, sample size estimation in PSM can be approached in a similar way to traditional practices for estimating sample sizes.

In this research, patients randomized to the treated group will be compared with patients from the control group selected using the PSM method. The success of the method will be determined by comparing the result obtained from the PSM method with the result obtained from a simple comparison of the two treatments without PSM. The following parameter will be used to assess the success of the method.

a) Root Mean Square Error (RMSE) - Model variance: The RMSE, which reflects the variance of the model, will be compared before and after matching. The success of the method will be indicated by a lower RMSE after matching.

Other parameters as given below will also be compared.

a) Tenability of matching: The feasibility and adequacy of the matching process will be assessed to ensure that suitable matches are achieved.
b) Reduction of covariate imbalance: The imbalance of covariates between the treatment and comparator groups will be evaluated before and after matching. The success of the method will be determined by a reduction in the imbalance of covariates after matching.
c) Estimation of treatment effect: The estimation of treatment effect will be compared between the matched groups before and after matching. The success of the method will be determined by similar or improved estimation of treatment effect after matching.
d) Model Fit statistics: Model Fit statistics such as Neg2LogLike, AIC, AICC, CAIC, BIC, and HQIC will be evaluated before and after matching. The success of the method will be determined by lower values of these statistics after matching.

By comparing the analysis results using the listed parameters, PSM method, IPW method can be assessed with respect to their ability in comparing groups which are unmatched with respect to covariates.

If the results of using PSM method are positive, it can provide the following benefits:

- It will allow for answering to research objectives in non-randomized situations, thus saving both time and costs savings.
- Comparative studies can be concluded using RWE such as observational studies without the need for a randomized trial. This approach can lead to significant time and costs savings.
- Even if randomization is compromised, the PSM method allows for the selection of comparable subjects. This prevents the entire study from becoming infeasible and contributes to cost savings.
- In situations where there is a need to estimate treatment differences for non-randomized comparisons. The PSM method can be used to achieve cost saving. By utilizing existing

data and matching comparable subjects, researchers can obtain valuable insights without the need for additional costly interventions or experiments.

- Overall, the benefits of utilizing the PSM method include cost and time savings, the ability to conduct comparative studies using real-world evidence, and obtaining treatment difference estimates in non-randomized scenarios.

The risks and benefits of PSM are as follows.

- The advantage of PSM is that it allows for all subjects from a treatment (where number of subjects are smaller) to be utilized in the analysis.
- The disadvantage is that if a subject does not have a matching data, their data cannot be used for the analysis.
- Limitations of PSM
    - It cannot control for unobserved selection bias, meaning it cannot adjust for hidden biases in subject selection within the original study (Rubin 1997).
    - PSM cannot handle covariates that are related to treatment and outcomes. It is based on subject's characteristics (Demographics, Baseline & Clinical characteristics, Socioeconomic and other variables) collected before treatment allocation and unrelated with treatment and outcome.
    - If there is missing data for any covariate, the PS cannot be computed for that specific subject/unit.
    - PSM works better with large samples, and insufficient sample size may be problematic.
    - This method is not useful if there is substantial variability in PS.

5. Conclusion

The objective of the proposed research was to find a method that can yield better results when randomization is not feasible. The results indicate that the PSM method is a supportive approach for achieving the following:

a) Finding similar subjects from different groups.
b) Reducing the imbalance in some of variables.

c) Lowering the DSLM after matching, this results more reliable, as they are based on subjects with similar profiles.

d) P-values at all assessed timepoints were ≥ 0.05 before and after matching, suggesting equality of null hypothesis failed to reject and concluded that both dose group have equal effect. However, it may be due to smaller sample size.

e) By utilizing re-sampling technique, the significant differences observed in ADAS-Cog at Week 16 and Week 24 leading to the conclusion that the high dose performed better.

f) Reduction in model variance.

g) Decrease in fit statistic values after matching.

These findings support the use of PSM as a reliable method for obtaining more accurate results in situations when randomization is not feasible.

The researcher in this paper discusses problems from a clinical trial study. However, this approach can also be applied in other domains such as finance, education, social problems, and other areas where the requirement is to find similar units across a group. Currently, there is a lot of observational data being collected due to the development of IT infrastructure. In the future, there will be an increased demand for this type of approach to select comparable units and obtain unbiased and unconfounded results across the group.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors have no competing interests with any organization or entity with the subject matter discussed in the manuscript. This includes but not limited to consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

The authors declare no competing interests. The views and opinions expressed in this article are purely those of the authors and do not necessarily reflect the official policy of or position of author's employer (Novartis) or any Novartis officers.

Acknowledgements

Authors sincerely thank Manipal Academy of Higher Education, MAHE, Manipal, Karnataka, India and Novartis Healthcare Private Limited, Hyderabad, India for providing research opportunities. The authors would like to acknowledge the support from Prasanna School of Public Health and Centre for Doctoral Studies, MAHE, Manipal, India.

Reference

(WHO), W. H. O. (1998). Health Promotion Evaluation: Recommendations to Policymakers. Report of the WHO European Working Group on Health Promotion Evaluation.

Barth, R. P., J. K. Greeson, S. Guo, R. L. Green, S. Hurley and J. Sisson (2007). "Outcomes for youth receiving intensive in-home therapy or residential care: a comparison using propensity scores." Am J Orthopsychiatry **77**(4): 497-505.

Brownson, R. C., J. F. Chriqui and K. A. Stamatakis (2009). "Understanding evidence-based public health policy." Am J Public Health **99**(9): 1576-1583.

Cummings, J., L. Froelich, S. E. Black, S. Bakchine, G. Bellelli, J. L. Molinuevo, R. W. Kressig, P. Downs, A. Caputo and C. Strohmaier (2012). "Randomized, double-blind, parallel-group, 48-week study for efficacy and safety of a higher-dose rivastigmine patch (15 vs. 10 cm²) in Alzheimer's disease." Dement Geriatr Cogn Disord **33**(5): 341-353.

Guo, S., and Mark W. Fraser (2014). Propensity score analysis: Statistical methods and applications, SAGE publications.

Heckman, J. J. (1978). "Dummy Endogenous Variables in a Simultaneous Equation System." Econometrica, Econometric Society **46(4)**: 931-959.

Heckman, J. J. (1979). "Sample Selection Bias as a Specification Error." Econometrica, Econometric Society **47(1)**: 153-161.

Rosen, W. G., R. C. Mohs and K. L. Davis (1984). "A new rating scale for Alzheimer's disease." Am J Psychiatry **141**(11): 1356-1364.

Rosenbaum, P. R. and D. B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects." Biometrika **70**(1): 41-55.

Rubin, D. B. (1997). "Estimating causal effects from large data sets using propensity scores." Ann Intern Med **127**(8 Pt 2): 757-763.

Abbreviations

AD	Alzheimer's Disease
ADAS-Cog	Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog)
CI	Confidence Interval
IPW	Inverse Probability Weighting

LPS	Logit of Propensity Score
PS	Propensity Score
PSM	Propensity Score Matching
RWE	Read-World Evidence
SD	Standard Deviation
TRT	Treatment
P-value	Probability value