Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Testing hypotheses under covariate-adaptive randomisation and additive models

Ting Ye

To cite this article: Ting Ye (2018) Testing hypotheses under covariate-adaptive randomisation and additive models, Statistical Theory and Related Fields, 2:1, 96-101, DOI: 10.1080/24754269.2018.1477005

To link to this article: https://doi.org/10.1080/24754269.2018.1477005

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 1, 96–101
https://doi.org/10.1080/24754269.2018.1477005

Taylor & Francis
Taylor & Francis Group

Testing hypotheses under covariate-adaptive randomisation and additive models

Ting Ye

Department of Statistics, University of Wisconsin Madison, WI, USA

ABSTRACT

Covariate-adaptive randomisation has a long history of applications in clinical trials. Shao, Yu, and Zhong [(2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, *97*, 347–360] and Shao and Yu [(2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics*, *69*, 960–969] showed that the simple *t*-test is conservative under covariate-adaptive biased coin (CABC) randomisation in terms of type I error, and proposed a valid test using the bootstrap. Under a general additive model with CABC randomisation, we construct a calibrated *t*-test that shares the same property as the bootstrap method in Shao et al. (2010), but do not need large computation required by the bootstrap method. Some simulation results are presented to show the finite sample performance of the calibrated *t*-test.

ARTICLE HISTORY
Received 5 April 2018
Accepted 13 May 2018

KEYWORDS
Biased coin; clinical trials; robust test; *t*-test; type I error; variance estimator

1. Introduction

In clinical trials and medical studies, patients arrive sequentially and must be treated immediately. When two treatments are compared under simple randomisation (SR), patients are allocated randomly into two treatment groups. The statistical inference may suffer from the disadvantage of not balancing patients' prognostic factors such as the age category, gender, disease stage, prior chemotherapy and geographical region that may influence the outcomes, although simple randomisation still produces valid statistical tests. Various randomisation methods have been proposed in the literature and they have advantages such as minimising imbalance between treatment groups, reducing selection bias, minimising accidental bias and improving efficiency in inference; see, for example, Efron (1971), Taves (1974), Pocock Simon (1975), Kalish Begg (1985), Aickin (2001), Weir Lees (2003), Shao, Yu, Zhong (2010), Shao Yu (2013) and Ma, Hu, Zhang (2015). A common characteristic of these methods is the use of a randomised treatment allocation that depends on covariates or prognostic factors but is conditionally independent of the outcomes given the covariates used in randomisation. Thus, they are called covariate-adaptive randomisation methods. The current paper focuses on one such method that applies the biased coin method (Efron, 1971) to patients grouped by prognostic factors, which is referred to as the covariate-adaptive biased coin (CABC) method by Shao et al. (2010). Similar results can be obtained for the minimisation procedure (Pocock & Simon, 1975; Taves, 1974) and the stratified block randomisation (Kalish & Begg, 1985), which together with the CABC are the most popular covariate-adaptive randomisation methods in clinical trials.

For any given randomisation method, statistical tests valid under the particular randomisation scheme should be used for testing the possible treatment effect. A statistical test is said to be valid if the type I error rate of the test is at most α, a given significance level, at least in the limiting case when the total sample size increases to infinity. The validity of various statistical tests under SR has been extensively studied in the statistical literature. For covariate-adaptive randomisation, however, there only exist a few theoretical results about the validity of statistical tests (e.g. Ma et al., 2015; Shao et al., 2010 and Shao & Yu, 2013), although covariate-adaptive randomisation has been used in clinical trials for a long time and there are many empirical results regarding properties of tests under covariate-adaptive randomisation (e.g. Aickin, 2002; Brikett, 1985; Forsythe, 1987; Hagino et al., 2004; Weir & Lees, 2003). As Rosenberger and Sverdlov (2008, Section 4) pointed out in their review, 'Very little theoretical work has been done in this area, despite the proliferation of papers. The original source papers are fairly uninformative about theoretical properties of the procedures'.

Under linear and generalised linear models, Shao et al. (2010) and Shao and Yu (2013), respectively, derived valid tests for comparing two treatments under CABC. Their tests are based on a modification of the

CONTACT Ting Ye Tye27@wisc.Edu Department of Statistics, University of Wisconsin 1300 University Ave., Madison, WI 53706, USA

tests developed under SR, where the modification is to apply a bootstrap variance estimation method that has a CABC component to address the variation in CABC randomisation. This bootstrap test was shown to be valid asymptotically and robust against misspecification of model and link function.

The purpose of this paper is to show that we can construct an asymptotically valid test under CABC without using the bootstrap by directly providing a consistent variance estimator in a general additive model that includes both linear and generalised linear models as special cases. The new test shares the robustness property with the bootstrap, but does not need the large computation required by the bootstrap. The same idea can be applied to the other two popular covariate-adaptive randomisation methods in clinical trials, the minimisation and stratified block randomisation.

2. Notation and preliminaries

Let N be the number of patients under two treatments, I_i be the treatment indicator that equals j if patient i is assigned to treatment j, $j=0,1$ and Y_{ij} be the outcome of patient i under treatment j. For patient i, $Y_i = I_i Y_{i1} + (1 - I_i) Y_{i0}$ is observed. Associated with patient i, let $\boldsymbol{X}_i$ be a vector of covariates and prognostic factors and Z_i be a function of $\boldsymbol{X}_i$ used in CABC, where Z_i is discrete with values z_k, $k = 1, \ldots, K$, and K is a fixed integer ≥ 2. We assume that $(Y_{i0}, Y_{i1}, \boldsymbol{X}_i)$, $i = 1, \ldots, N$, are independent and identically distributed random vectors from some distribution.

Under SR, I_i's are independent with $P(I_i = 1) = 1/2$ for all i and are independent of $(Y_i, \boldsymbol{X}_i)$. With a fixed constant $p > 1/2$, the biased coin method in Efron (1971) assigns the ith patient according to

$$P(I_i = 1) = \begin{cases} p, & D_{i-1} < 0, \\ 1/2, & D_{i-1} = 0, \\ 1 - p, & D_{i-1} > 0, \end{cases}$$

$i = 1, \ldots, N$, where $D_0 = 0$ and D_{i-1} is the difference between the number of patients in treatment 1 and the number of patients in treatment 0 after $i-1$ assignments have been made. This assignment rule tends to achieve balance between the numbers of patients in two treatment groups, since $p > 1/2$ and D_{i-1} is an imbalance metric. The CABC method applies the biased coin within each category of patients with $Z_i = z_k$, $k = 1, \ldots, K$. The motivation is to achieve balance between treatment groups for each prognostic factor. A characteristic of CABC, which is common for all covariate-adaptive randomisation methods, is that I_i's and (Y_{i1}, Y_{i0})'s are conditionally independent given Z_i's, although unconditionally I_i's and (Y_{i1}, Y_{i0})'s are dependent.

A statistical test T is a function of observed $(I_i, Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$, constructed such that we reject a given null hypothesis H_0 if and only if $T > c_\alpha$, where α is a given significance level and c_α is a quantile of the standard normal distribution or a t-distribution. T is said to be (asymptotically) valid if, when H_0 holds,

$$\lim_{N \to \infty} P(T > c_\alpha) \leq \alpha \tag{1}$$

with equality holds for at least some cases.

One of the main results in Shao et al. (2010), followed by Shao and Yu (2013), is that if a test T is constructed using covariates $\boldsymbol{X}_i$'s under a correctly specified model between Y_i and $\boldsymbol{X}_i$, and T is valid according to Equation (1) under SR, then T is still valid under CABC.

However, there are practical considerations under which some covariates are not included in the construction of the test T. For example, including all covariates may lead to changing a simple test procedure to a complicated one, such as from one-way analysis of variance to two-way analysis of variance; data in some discrete covariate categories may be sparse so that including these covariates may result in some bad behaviour of the test. When Z_i is not included in the construction of T and CABC is used, the result in Shao et al. (2010) indicates that the test is conservative in the sense that $\lim_{N \to \infty} P(T > c_\alpha) \leq \alpha_0 < \alpha$ with a fixed α_0. The reason for this is that typically T is a ratio of an estimated effect $\hat{\theta}$ under SR divided by the standard error of $\hat{\theta}$; although $\hat{\theta}$ is still asymptotically valid under CABC, the standard error of $\hat{\theta}$ valid under SR overestimates that under CABC.

To obtain a valid test under CABC, it suffices to derive a standard error of $\hat{\theta}$ that is asymptotically consistent, or equivalently a consistent variance estimator of $\hat{\theta}$. Shao et al. (2010) proposed a bootstrap variance estimator with a re-assigning treatment indicators in bootstrapping. This bootstrap method, however, requires a large amount of computation.

3. The main result

We consider the following general additive model:

$$E(Y_{ij} | \boldsymbol{X}_i) = \mu_j + \psi(\boldsymbol{X}_i), \tag{2}$$

where $\psi(\cdot)$ is an unknown function satisfying $E\{\psi(\boldsymbol{X}_i)\} = 0$ and $E\{\psi(\boldsymbol{X}_i)^2\} < \infty$, and μ_j is the response mean under treatment $j=0,1$. We consider either the two-sided hypotheses $H_0 : \mu_1 = \mu_0$ versus $H_1 : \mu_1 \neq \mu_0$, or the one-sided hypotheses $H_0 : \mu_1 \leq \mu_0$ versus $H_1 : \mu_1 > \mu_0$.

The two sample t-test is

$$T_S = \frac{\bar{Y}_1 - \bar{Y}_0}{(S_1^2/n_1 + S_0^2/n_0)^{1/2}} \tag{3}$$

or the absolute value of T_S, where $n_1 = \sum_{i=1}^N I_i$ and $n_0 = \sum_{i=1}^N (1 - I_i)$ are, respectively, the numbers of

patients in treatment groups 1 and 0, and $\bar{Y}_j$ and S_j^2 are, respectively, the sample mean and sample variance within treatment j.

Suppose that CABC is applied within each group formed by Z_i, which is a discrete function of $\boldsymbol{X}_i$ taking values $z_1, \ldots, z_K$ with a fixed $K \geq 2$. As proved in Shao et al. (2010), T_S is conservative under CABC because CABC does not introduce any bias and the variance estimator $S_1^2/n_1 + S_0^2/n_0$ in Equation (3) does not account for the correlation between $\bar{Y}_1$ and $\bar{Y}_0$. They then suggested applying a particular bootstrap method to construct a consistent variance estimator of $\text{var}(\bar{Y}_1 - \bar{Y}_0)$ under CABC, which leads to a valid bootstrap t-test, denoted as T_B.

Explicitly, as shown in the appendix, under model (2) and CABC,

$$\frac{(\bar{Y}_1 - \bar{Y}_0) - (\mu_1 - \mu_0)}{2\tau_{\psi|\mathcal{Z}}/N^{1/2}} \to_D N(0,1), \tag{4}$$

where $\to_D$ is convergence in distribution,

$$\tau_{\psi|\mathcal{Z}}^2 = E[\text{var}\{\psi(\boldsymbol{X}_i)|Z_i\}] + \sigma_\varepsilon^2, \tag{5}$$

and $\sigma_\varepsilon^2 = \text{var}(Y_{ij} - E(Y_{ij}|\boldsymbol{X}_i))$. An interesting observation is that, under model (2) and the null hypothesis,

$$E[\text{var}(Y_{ij}|Z_i)] = \tau_{\psi|\mathcal{Z}}^2, \tag{6}$$

which can be consistently estimated by

$$\hat{\tau}_{\psi|\mathcal{Z}}^2 = \frac{1}{N}\sum_{k=1}^{K} m_k S_k^2, \tag{7}$$

where S_k^2 is the sample variance of Y_i within $Z = z_k$ and m_k is the number of subjects in the data set with $Z = z_k$, $k = 1, \ldots, K$. The proof is given in the appendix. This alternative way of obtaining a consistent variance estimator is not only computationally easy but also robust against any model misspecification. The two sample t-test with variance estimated by (7) is

$$T_{SC} = \frac{(\bar{Y}_1 - \bar{Y}_0)}{2\hat{\tau}_{\psi|\mathcal{Z}}/N^{1/2}}, \tag{8}$$

which is named as a calibrated t-test.

Consider the following working model,

$$E(Y_{ij}|Z_i) = \mu_j + \beta Z_i. \tag{9}$$

This model is a special case of model (2) but it is not necessary correct. Wald's test statistic under SR is

$$T_W = \frac{\hat{Y}_1 - \hat{Y}_0}{(\hat{S}_1^2/n_1 + \hat{S}_0^2/n_0)^{1/2}}, \tag{10}$$

where $\hat{Y}_j$ and $\hat{S}_j^2$ are, respectively, the sample mean and sample variance based on $(Y_i - \hat{\beta}Z_i)$'s under treatment j, and $\hat{\beta}$ is the least square estimator of β assuming model (9). As shown in Shao Yu (2013), under CABC and model (2),

$$\hat{\beta} = \beta_0 + o_p(1), \quad \beta_0 = \frac{\text{cov}\{Z_i, \psi(\boldsymbol{X}_i)\}}{\text{var}(Z_i)}, \tag{11}$$

and

$$\frac{(\hat{Y}_1 - \hat{Y}_0) - (\mu_1 - \mu_0)}{2\tau_{\psi|\mathcal{Z}}/N^{1/2}} \to_D N(0,1). \tag{12}$$

Under model (2) and CABC,

$$\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_0^2}{n_0} = \frac{4\tau_\psi^2}{N} + o_p\left(\frac{1}{N}\right), \tag{13}$$

where

$$\tau_\psi^2 = \text{var}\{\psi(\boldsymbol{X}_i) - \beta_0 Z_i\} + \sigma_\varepsilon^2. \tag{14}$$

Since τ_ψ^2 in Equation (14) and $\tau_{\psi|\mathcal{Z}}^2$ in Equation (5) are related by

$$\tau_\psi^2 = \text{var}\{E[\psi(\boldsymbol{X}_i) - \beta_0 Z_i|Z_i]\} + \tau_{\psi|\mathcal{Z}}^2, \tag{15}$$

results (12)–(15) show that Wald's test T_W is conservative under CABC unless $E\{\psi(\boldsymbol{X}_i) - \beta_0 Z_i|Z_i\}$ is a constant, i.e. $\psi(\boldsymbol{X}_i) - \beta_0 Z_i$ is independent of Z_i. Thus, Wald's test T_W is not valid in the sense of Equation (1), unless the working model (9) is a correct model.

If we borrow the idea of consistently estimating the variance of $\hat{Y}_1 - \hat{Y}_0$ under H_0, a calibrated Wald's test can also be constructed as

$$T_{WC} = \frac{(\hat{Y}_1 - \hat{Y}_0)}{2\hat{\tau}_{\psi|\mathcal{Z}}/N^{1/2}}, \tag{16}$$

which is valid and asymptotically equivalent to its counterpart T_{SC} in Equation (8).

This calibrated variance idea can also be extended to the case where working model (9) is replaced by a more complicated one.

4. Simulation results

4.1. Linear model

A simulation study was carried out to examine the type I error of the calibrated t-test T_{SC} and Wald's test T_{WC} under CABC along with five other tests: the two sample t-test T_S under SR, Wald's test T_W under SR, the two sample t-test under CABC, Wald's test under CABC and the bootstrap t-test T_B under CABC.

In the simulation study, the significance level is $\alpha = 5\%$; ε_{ij} is $N(0,1)$; the probability p in CABC is 2/3; the sample size N is 200; the bootstrap variance estimator V_B is approximated by Monte Carlo with $B = 200$; and the simulated type I error and power are based on 10,000 runs and 2000 runs, respectively. The simulation setting is $Y_{ij} = (\mu_1 - \mu_0)I_i + Z_{i1} + 2Z_{i2} - 2Z_{i1} * Z_{i2} + \varepsilon_{ij}$, where Z_{i1} and Z_{i2} are both binary with

Table 1. Simulation power in % under linear model ($\alpha = 5\%$, $N = 200$, 10,000 simulation runs when $\mu_1 - \mu_0 = 0$, 2000 simulation runs when $\mu_1 - \mu_0 \neq 0$).

	SR		CABC				
$\mu_1 - \mu_0$	T_S	T_W	T_S	T_W	T_B	T_{SC}	T_{WC}
0	4.97	4.96	1.91	3.06	5.37	5.49	5.35
0.1	8.40	10.15	5.00	7.40	11.50	11.40	11.35
0.2	19.82	24.62	16.82	22.12	27.83	28.38	28.88
0.3	40.90	48.00	37.65	46.55	53.85	54.70	54.75
0.4	62.85	70.70	64.00	72.40	78.05	78.55	78.40
0.5	81.30	88.35	85.10	90.50	93.20	93.25	93.55
0.6	92.95	97.10	96.50	97.85	98.65	98.80	98.90
0.7	98.05	99.25	99.30	99.75	99.80	99.85	99.75
0.8	99.55	99.90	99.85	99.85	99.90	99.90	99.95
0.9	99.90	100.00	100.00	100.00	100.00	100.00	100.00

Notes: T_S: two sample t-test; T_B: bootstrap t-test; T_W: Wald's test, in this case, one-way analysis of covariance test; T_{SC}: calibrated t-test, with variance estimated by $\hat{\tau}^2_{\psi|\mathcal{Z}}$; T_{WC}: calibrated Wald's test, with variance estimated by $\hat{\tau}^2_{\psi|\mathcal{Z}}$; SR: simple randomisation; CABC: covariate-adaptive biased coin.

$P(Z_{i1} = 1) = P(Z_{i2} = 1) = 1/2$. Both Z_{i1} and Z_{i2} are used in the CABC and in the construction of Wald's test, but the interaction term is ignored in the construction of Wald's test.

The simulation results and values of $\mu_1 - \mu_0$ are shown in Table 1. A few conclusions from Table 1 are:

1. The two sample t-test T_S and Wald's test T_W derived under the simplified working model are conservative under CABC.
2. The type I errors of the bootstrap t-test T_B, calibrated t-test T_{SC} and calibrated Wald's test T_{WC} under CABC are reasonably close to the nominal level 5%, depicting the validity of all three tests, and the consistency of $\hat{\tau}^2_{\psi|\mathcal{Z}}$.
3. T_B, T_{SC} and T_{WC} have almost the same empirical power, which agrees with the asymptotic equivalence of T_B, T_{SC} and T_{WC} under CABC.

The advantage of the proposed bootstrap t-test is that it directly estimates the variance of $\bar{Y}_1 - \bar{Y}_0$ by Monte Carlo sampling, which performs well under small sample size and is robust against any model misspecification. The one-way analysis of covariance test is invalid under CABC if model is misspecified. But the calibrated one-way analysis of covariance test is robust against model misspecification, computationally easy and performs well with regard to both type I error and power. The calibrated t-test is computationally easy, but has certain requirement on sample size for the gap between variance estimator and $\text{var}(\bar{Y}_1 - \bar{Y}_2)$ to be ignorable.

4.2. Logistic model

The second simulation setting is $\text{logit}(p_{ij}) = -1.5 + (\mu_1 - \mu_0)I_i + Z_{i1} + 3Z_{i2} + 2Z_{i1}Z_{i2}$, where Z_{i1} and Z_{i2} are both binary with $P(Z_{i1} = 1) = P(Z_{i2} = 1) = 1/2$. Both Z_{i1} and Z_{i2} are used in the CABC and in the construction of Wald's test, but the interaction term is ignored in the analysis. The rest of the parameters are the same as in Table 1.

Table 2. Simulation power in % under logistic model ($\alpha = 5\%$, $N = 200$, 10,000 simulation runs when $\mu_1 - \mu_0 = 0$, 2000 simulation runs when $\mu_1 - \mu_0 \neq 0$).

	SR		CABC			
$\mu_1 - \mu_0$	T_S	T_W	T_S	T_W	T_B	T_{SC}
0.0	5.25	5.01	1.13	5.03	5.24	5.75
0.1	5.57	6.00	1.44	5.82	5.79	6.41
0.2	6.84	8.44	2.27	8.05	8.42	8.98
0.3	9.13	12.34	4.21	12.08	12.15	12.90
0.4	12.72	17.53	6.93	17.91	17.84	18.67
0.5	17.82	25.35	10.39	24.90	23.54	25.05
0.6	20.80	33.30	18.05	35.15	33.25	34.35
0.7	27.51	43.22	25.10	44.68	42.17	43.32
0.8	35.34	53.16	33.33	55.27	52.76	54.27
0.9	43.30	63.20	44.35	66.30	64.25	65.80
1.0	52.05	73.95	55.50	75.40	73.70	74.60
1.1	60.59	81.88	66.01	82.43	80.87	81.93
1.2	69.80	87.75	74.90	88.30	86.90	87.50
1.3	77.36	92.87	82.83	92.72	91.37	91.92
1.4	83.05	95.55	88.10	95.40	94.35	94.80
1.5	88.55	97.29	91.67	97.34	96.59	96.84

Notes: T_S: two sample t-test; T_B: bootstrap t-test; T_W: Wald's test; T_{SC}: calibrated t-test, with variance estimated by $\hat{\tau}^2_{\psi|\mathcal{Z}}$; SR: simple randomisation; CABC: covariate-adaptive biased coin.

The simulation results and values of $\mu_1 - \mu_0$ are shown in Table 2. A few conclusions from Table 2 are:

1. The two sample t-test is conservative under CABC, while Wald's test is valid though derived under the simplified working model.
2. T_{SC} is valid under CABC, indicating that under the generalised linear model, the new variance estimator $\hat{\tau}^2_{\psi|\mathcal{Z}}$ is still valid.
3. T_B and T_{SC} have almost the same power as Wald's test T_W.

Acknowledgements

The author would like to thank two referees for their helpful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Ting Ye is a Ph.D. student in Department of Statistics in University of Wisconsin-Madison. Her research interests focus on clinical trial design, survival analysis and missing data.

ORCID

Ting Ye http://orcid.org/0000-0001-6009-641X

References

Aickin M. (2001). Randomization, balance, and the validity and efficiency of design-adaptive allocation methods. *Journal of Statistical Planning and Inference*, *94*, 97–119.

Aickin M. (2002). Beyond randomization. *The Journal of Alternative Medicine*, *8*, 765–772.
Brikett N. J. (1985). Adaptive allocation in randomized controlled trials. *Controlled Clinical Trials*, *6*, 146–155.
Efron B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, *58*, 403–417.
Forsythe A. B. (1987). Validity and power of tests when groups have been balanced for prognostic factors. *Computational Statistics and Data Analysis*, *5*, 193–200.
Hagino A., Hamada C., Yoshimura I., Ohashi Y., Sakamoto J., & Nakazato H. (2004). Statistical comparison of random allocation methods in cancer clinical trials. *Controlled Clinical Trials*, *25*, 572–584.
Kalish L. A., & Begg C. B. (1985). Treatment allocation methods in clinical trials: A review. *Statistics in Medicine*, *4*, 129–144.
Ma W., Hu F., & Zhang L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association*, *110*(510), 669–680.
Pocock S. J., & Simon R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, *31*, 103–115.
Shao J., & Yu X. (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics*, *69*, 960–969.
Shao J., Yu X., & Zhong B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, *97*, 347–360.
Taves D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics*, *15*, 443–453.
Weir C. J., & Lees K. R. (2003). Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Statistics in Medicine*, *22*, 705–726.

Appendix. Proofs of (4)–(7)

Proof of (4)**:** Applying result (7.9) in Efron (1971) to each category defined by Z_i and using the fact that $E(I_i|\mathcal{Z}) = 1/2$ and $E(n_j|\mathcal{Z}) = N/2$, where $\mathcal{Z} = \{Z_1, \ldots, Z_N\}$, we obtain that

$$\frac{n_j}{N} - \frac{1}{2} = o_p(N^{-1/2}) \quad \text{conditionally on } \mathcal{Z}, \ j = 0, 1. \quad \text{(A1)}$$

Applying (A1), we obtain

$$\bar{Y}_1 - \bar{Y}_0 = \mu_1 - \mu_0 + \frac{2}{N}\sum_{i=1}^{N}\{I_i - (1 - I_i)\}\psi(\boldsymbol{X}_i) + \frac{2}{N}\sum_{i=1}^{N}\{I_i\varepsilon_{i1} - (1 - I_i)\varepsilon_{i0}\} + o_p(N^{-1/2}).$$

Letting $\Delta_i = \psi(\boldsymbol{X}_i) - E\{\psi(\boldsymbol{X}_i)|Z_i\}$, we obtain that

$$\frac{1}{N}\sum_{i=1}^{N}\{I_i - (1 - I_i)\}\psi(\boldsymbol{X}_i) = \frac{1}{N}\sum_{i=1}^{N}\{I_i - (1 - I_i)\}\Delta_i + \frac{1}{N}\sum_{i=1}^{N}\{I_i - (1 - I_i)\}E\{\psi(\boldsymbol{X}_i)|Z_i\}.$$

Applying result (7.9) in Efron (1971) to each category defined by Z_i and using the fact that $E\{\psi(\boldsymbol{X}_i)|Z_i\}$ is discrete, we conclude that the last term in the previous expression is $o_p(N^{-1/2})$ conditionally on $\mathcal{Z}$. Thus,

$$\bar{Y}_1 - \bar{Y}_0 = \mu_1 - \mu_0 + \frac{2}{N}\sum_{i=1}^{N}\{I_i - (1 - I_i)\}\Delta_i + \frac{2}{N}\sum_{i=1}^{N}\{I_i\varepsilon_{i1} - (1 - I_i)\varepsilon_{i0}\} + o_p(N^{-1/2}).$$

The asymptotic mean of $\bar{Y}_1 - \bar{Y}_0$ is $\mu_1 - \mu_0$, which follows from the fact that $(\Delta_i, \varepsilon_{i1}, \varepsilon_{i0})$'s are conditionally independent of $\mathcal{I} = \{I_1, \ldots, I_N\}$ given $\mathcal{Z}$, $E(\varepsilon_{ij}|\mathcal{Z}) = E(\varepsilon_{ij}) = 0$, and $E(\Delta_i|\mathcal{Z}) = E\{\Delta_i|Z_i\} = 0$ by the definition of Δ_i.

Since ε_{ij}'s are of mean 0 and independent of $(\mathcal{Z}, \mathcal{I})$,

$$\operatorname{cov}\left(\frac{2}{N}\sum_{i=1}^{N}\{I_i - (1 - I_i)\}\Delta_i, \times\frac{2}{N}\sum_{i=1}^{N}\{I_i\varepsilon_{i1} - (1 - I_i)\varepsilon_{i0}\}\middle|\mathcal{Z}\right) = 0$$

and

$$\begin{aligned}
&\operatorname{var}\left(\frac{2}{N}\sum_{i=1}^{N}\{I_i\varepsilon_{i1} - (1 - I_i)\varepsilon_{i0}\}\middle|\mathcal{Z}\right)\\
&= \left(\frac{4}{N^2}\sum_{i=1}^{N}\{I_i^2\operatorname{var}(\varepsilon_{i1}) + (1 - I_i)^2\operatorname{var}(\varepsilon_{i0})\}\middle|\mathcal{Z}\right)\\
&= \sigma_\varepsilon^2 E\left(\frac{4}{N^2}\sum_{i=1}^{N}I_i + \frac{4}{N^2}\sum_{i=1}^{N}(1 - I_i)\middle|\mathcal{Z}\right)\\
&= \frac{4\sigma_\varepsilon^2}{N}.
\end{aligned}$$

Since Δ_i's and $\mathcal{I}$ are conditionally independent given $\mathcal{Z}$ and $E\{\Delta_i|Z_i\} = 0$, we obtain that

$$\begin{aligned}
&\operatorname{var}\left(\frac{2}{N}\sum_{i=1}^{N}\{I_i - (1 - I_i)\}\Delta_i\middle|\mathcal{Z}\right)\\
&= \frac{4}{N^2}E\left\{\operatorname{var}\left(\sum_{i=1}^{N}\{I_i - (1 - I_i)\}\Delta_i\,|\mathcal{Z}, \mathcal{I}\right)\middle|\mathcal{Z}\right\}\\
&= \frac{4}{N^2}E\left\{\sum_{i=1}^{N}\{I_i - (1 - I_i)\}^2\operatorname{var}\{\Delta_i|Z_i\}\middle|\mathcal{Z}\right\}\\
&= \frac{4}{N^2}\sum_{i=1}^{N}\operatorname{var}\{\Delta_i|Z_i\}\\
&= \frac{4V_\Delta}{N},
\end{aligned}$$

where $V_\Delta = N^{-1}\sum_{i=1}^{N}\operatorname{var}\{\Delta_i|Z_i\}$. Therefore,

$$\operatorname{var}(\bar{Y}_1 - \bar{Y}_0|\mathcal{Z}) = \frac{4\left(V_\Delta + \sigma_\varepsilon^2\right)}{N} + o_p(N^{-1})$$

and

$$\operatorname{var}(\bar{Y}_1 - \bar{Y}_0) = \frac{4\tau_{\psi|\mathcal{Z}}^2}{N} + o(N^{-1})$$

Given $\mathcal{Z}$, I_i's and $(\Delta_i, \varepsilon_{i1}, \varepsilon_{i0})$'s are conditionally independent. Hence, by the central limit theorem and the above results, the conditional distribution of

$$\frac{2}{N^{1/2}}\sum_{i=1}^{N}[b\{I_i - (1 - I_i)\}\Delta_i + I_i\varepsilon_{i1} - (1 - I_i)\varepsilon_{i0}]$$

given $(\mathcal{Z}, \mathcal{I})$, is asymptotically normal with mean 0 and variance $4(V_\Delta + \sigma_\varepsilon^2)$, which converges to $4\tau^2_{\psi|\mathcal{Z}}$ by the law of large number. Thus, conditionally on $\mathcal{Z}$ or unconditionally, the quantity in (??) is asymptotically normal with mean 0 and variance $4\tau^2_{\psi|\mathcal{Z}}$. ■

Proof of (7)**:** Without loss of generality, we assume that under H_0, $\mu_1 = \mu_0 = 0$ in the proof. From the fact that $\bar{Y}_j = \mu_j + o_p(1)$,

$$S_k^2 = \frac{1}{m_k} \sum_{i=1}^{N} Y_{ij}^2 I(Z_i = z_k) + o_p(1),$$

where m_k is the number of subjects satisfying $Z = z_k$. Recall that Y_{ij}'s and Z_i's are independent and identically distributed. By the law of large numbers,

$$S_k^2 = \frac{N}{m_k} E[Y_{ij}^2 I(Z_i = z_k)] + o_p(1) = E[Y_{ij}^2 | Z_i = z_k] + o_p(1).$$

Now that $\hat{\tau}^2_{\psi|\mathcal{Z}}$ can be expressed as $N^{-1} \sum_{i=1}^{N} E[Y_{ij}^2|Z_i] + o_p(1)$, which together with the dominated convergence theorem and the fact that $N^{-1} \sum_{i=1}^{N} E\{Y_{ij}^2|Z_i\} = \tau^2_{\psi|\mathcal{Z}} + o_p(1)$ imply that $\hat{\tau}^2_{\psi|\mathcal{Z}} = \tau^2_{\psi|\mathcal{Z}} + o_p(1)$. ■