



Published in final edited form as:

Stat Interface. 2010 April 1; 3(2): 169–176.

Within-Cluster Resampling for Analysis of Family Data: Ready for Prime-Time?

Hemant K. Tiwari^{1,*}, Amit Patki¹, and David B. Allison^{1,2,3}

¹Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294

²Department of Nutrition Sciences, University of Alabama at Birmingham, Birmingham, Alabama 35294

³Clinical Nutrition Research Center, University of Alabama at Birmingham, Birmingham, Alabama 35294

Abstract

Hoffman et al. [1] proposed an elegant resampling method for analyzing clustered binary data. The focus of their paper was to perform association tests on clustered binary data using within-cluster-resampling (WCR) method. Follmann et al. [2] extended Hoffman et al.'s procedure more generally with applicability to angular data, combining of p-values, testing of vectors of parameters, and Bayesian inference. Follmann et al. [2] termed their procedure *multiple outputation* because all “*excess*” data within each cluster is thrown out multiple times. Herein, we refer to this procedure as WCR-MO. For any statistical test to be useful for a particular design, it must be robust, have adequate power, and be easy to implement and flexible. WCR-MO can be easily extended to continuous data and is a computationally intensive but simple and highly flexible method. Considering family as a cluster, one can apply WCR to familial data in genetic studies. Using simulations, we evaluated WCR-MO's robustness for analysis of a continuous trait in terms of type I error rates in genetic research. WCR-MO performed well at the 5% α -level. However, it provided inflated type I error rates for α -levels less than 5% implying the procedure is liberal and may not be ready for application to genetic studies where α levels used are typically much less than 0.05.

Keywords

Correlated residuals; WCR; Multiple Outputation; familial data; genetic research; Type I error

1. INTRODUCTION

Data sets in which the residuals from a fitted model cannot be expected or assumed to be independent across all cases analyzed because cases are grouped into clusters are common. This is especially so in genetic research where data from related individuals are included. In such cases, families constitute clusters. Naively analyzing such data with methods that assume the residuals are independent can lead to inefficient estimation and low power when the alternative hypothesis is true and type I error rate inflation when the null hypothesis is true.

*Correspondence to: Hemant K. Tiwari, Department of Biostatistics, Ryals Public Health Building, 420D, University of Alabama at Birmingham, 1665 University Boulevard, Birmingham, Alabama 35294. Phone: 205-934-4907, Fax: 205-975-2541.

Many methods have been proposed to accommodate such data (e.g., [1-24]) and each has advantages and disadvantages. WCR-MO has been mentioned as a technique of interest for the problem of correlated residuals in at least 6 other methodologic papers [8,20-24].

A detailed description of WCR-MO appears in section 2 below. In brief, WCR-MO entails identifying clusters of observations -- in this context the families would constitute the clusters -- and then creating a new pseudo-dataset by randomly selecting one observation from within each cluster. The new dataset is analyzed with a data analyst's preferred 'standard' procedure that is valid for data with independent residuals and the results recorded. Then, one repeats the process multiple times and finally compiles the results using formulae that account for the dependency among the multiple pseudo-datasets created. This technique potentially takes any test that is legitimate if the observations are all independent and converts it to a test that is legitimate even if the observations are not independent. Conveniently, it does so with no adjustment to the essential modeling procedure but only by 'wrapping' that procedure in a re-sampling based method for quantifying uncertainty.

In principle, WCR-MO has several very highly desirable advantages. As Follmann et al. ([2]; p. 421) wrote "Multiple outputation is very simple and only requires an appropriate statistical procedure for independent data." Therefore, one might expect that it could be adapted to virtually any testing situation as Wu and Huang [8] conjecture. This is perhaps its greatest potential advantage. Second, programming WCR-MO to couple it with existing statistical packages is quite easy. Third, WCR-MO does not require any knowledge or specification of the covariance structure among residuals nor does it require that there is a constant covariance among all pairs of observations within clusters. Finally, WCR is both easy to understand and explain to colleagues.

A number of investigators have recognized the utility of WCR-MO and utilized it in genetic and other analyses (e.g., [20-24]). From these applied papers, several things are notable. First, none of the authors mention and perhaps may not be aware that the results supporting the use of WCR-MO are asymptotic results. The extent to which WCR-MO is valid (in terms of maintaining the type I error rate to the nominal α level) across a variety of situations and, importantly, with the small α levels often used in genetic studies due to frequent massive multiple testing is unknown. This is important because, as Mehta et al. [25] note, procedures that work well at one α level, may not work well at others, particularly at more stringent levels. Second, there was no discussion of how challenges that may be encountered, such as negative estimates of variance were dealt with. Third, re-sample sizes in the applied studies ranged from 1,000 to 10,000 with no justification for the choice of resample number suggesting a need for guidance on this point. Finally, the analyses conducted are important ones including some that may immediately affect patient care and the lives of many individuals (e.g., [20]). Therefore, taking the time to evaluate the performance of the method in finite sample sizes, with realistic situations, and at small α levels seems vitally important. The purpose of this paper is to conduct such an evaluation.

The remainder of this paper is divided into four sections. In Section 2, we provide an exposition of WCR-MO and its connection to certain GEE methods. Section 3 explains our simulation methodology; Section 4 offers results, and Section 5 general discussion and conclusions.

2. THE WCR-MO METHOD

The steps are described below for WCR-MO procedure.

Step 1: From each of C clusters, select one individual with replacement to create a new dataset that has exactly C independent observations. Repeat this m times to create m such data sets.

Step 2: Analyze each of the m data sets using standard complete-data methods such as SAS or SPSS.

Step 3: The last step is to integrate or combine m analyses to get a single estimate of parameters and their variances. This involves averaging the values of the parameter estimates across the m samples to produce a single point estimate and variance. Formally, we can describe it as follows:

Let m = the number of data sets analyzed,

\hat{Q}_i = Estimate of the parameter of interest from the i^{th} set,

\hat{T}_i = Variance Estimate of the \hat{Q}_i from the i^{th} set.

The point estimate from the WCR method is the average of the estimates from m analyses and is given by

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i \tag{1}$$

The total variance estimate of the point estimate is the difference of the average within-

replicate variance $\left(U = \frac{1}{m} \sum_{i=1}^m \hat{T}_i \right)$ and the among-replicate variance $\left(B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2 \right)$ and is given by

$$T = U - (1+m^{-1})B = \frac{1}{m} \sum_{i=1}^m \hat{T}_i - (1+m^{-1}) \left(\frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2 \right). \tag{2}$$

The statistic:

$$t^* = \tilde{t} \equiv \frac{(Q - \bar{Q})}{\sqrt{T}} \tag{3}$$

is approximately distributed as t with v_m degrees of freedom, where

$$v_m = (m-1) \left(1 + \frac{U}{(1+m^{-1})B} \right)^2.$$

When testing for effects of genetic loci in complex traits, "...the locus-specific effects on complex and quantitative traits cannot *a priori* be assumed to be additive and can even be over-dominant ... For this reason, many investigators wisely choose to test for genotypic effects in two degrees of freedom models ...rather than restricting themselves to allelic (additive) effects" [26]. This necessitates that, in the case of a di-allelic locus, we conduct 2 degree of freedom (*df*) tests. The original formulation of WCR was not set up to do tests with more than 1 *df*. Follmann et al. [2] offer a straightforward method for adapting WCR-MO from simple 1 *df* tests to any testing situation in which one can obtain a legitimate p-

values under the assumption of independence of observations for each replicate and then combine the p-values as described by Follmann et al. [2].

WCR-MO is simple to implement in any standard statistical packages, but the moment-based variance given in equation (2) can be negative as we will show below (see Table 1). To resolve the issue of negative variance, Follmann et al. [2] proposed using an ML approach, where the estimate of the variance is replaced by the average of the efficient score divided by the Fisher information for data points of the within cluster. However, deriving the likelihood of the data may be challenging. Moreover, if one could derive the likelihood of the data and be confident that the data were sampled from the distribution specified, then one usually could instead rely on existing ML-based procedures, obviating the need for WCR-MO.

Connection to GEE methods

Williamson et al. [6] and Benhin et al. [9] have shown that, asymptotically in the number of resamples, WCR with a dichotomous outcome and a logistic regression framework is identical to a particular form of GEE when cluster size is informative. Also, Follmann et al. [2] showed that GEE is similar to the WCR-MO using simulation corresponding to rectangular, Triangular, and L-shaped data structures. Furthermore, Datta and Satten [10] had extended GEE to handle non-normal data via a modified rank test. The beauty of WCR-MO is that (in principle) it can be used in virtually any situation in which one has a legitimate method for deriving tests if the observations were all independent [2].

3. SIMULATION METHODOLOGY AND ANALYSIS MODELS

To examine performance under the null hypothesis for WCR-MO, we first simulated correlated clusters consisting of 50 nuclear families with both parents and two offspring for a total of 400 individuals in each simulated dataset (see Figure 1) and subsequently, datasets composed of unequal correlated clusters of varying family size also with a total of 400 individuals per dataset (see figure 2).

Equal Cluster Size

Data were simulated for 50 and 100 nuclear families with both parents and 2 offspring (see Figure 1). A phenotype for each individual was sampled from a continuous distribution generated using a linear model consisting of SNP effect, polygenic effect, and random non-shared residual effect. The SNP markers were simulated for all founders randomly and then the markers for offspring were simulated using Mendel's law of segregation assuming population minor allele frequency of 0.2. Polygenic data for founders were simulated from a standard normal distribution. Polygenic values for offspring were simulated from a normal distribution with mean equal to average of parents' polygenic values and variance equal to 0.5 (Elston et al., 1992). To evaluate type I error rates we set SNP specific heritabilities to

zero. $phenotype = \sqrt{h_{polygenic}^2} \times polygenic + \sqrt{1 - h_{polygenic}^2} \times \varepsilon$, where ε is a random stochastic error and $\varepsilon \sim N(0, 1)$

We assumed polygenic heritability $h_{polygenic}^2 = 0.3$ and allele frequency = 0.2.

Unequal Cluster Size

We also simulated unequal family data for 50 and 100 families all parameters the same as described above (see Figure 2).

We simulated 100,000 replicates (datasets) of both equal and unequal families. We used a total of 10,000 bootstrap samples for each replicate for WCR-MO testing of the association between SNP and phenotype. We tested at nominal α -levels of 0.05, 0.01, 0.001, and 0.0001.

Analysis Models

We can analyze the data either testing only additive effect of the SNP (i.e. one degree of freedom test) or jointly testing additive and dominance effect of the SNP (i.e. two degrees of freedom test). The following analysis models were used to determine the behavior of null distribution.

Analysis Model 1:

$$Y = \alpha + \beta_1 A + \varepsilon, \tag{1}$$

where A is additive effect of the SNP and is given by

$$A = \begin{cases} 1 & \text{if individual has AA genotype} \\ 0 & \text{if individual has Aa genotype} \\ -1 & \text{if individual has aa genotype} \end{cases} .$$

Analysis Model 2:

$$Y = \alpha + \beta_1 A + \beta_2 D + \varepsilon \tag{2}$$

where A is additive effect of the SNP given as above and dominance effect D is given by

$$D = \begin{cases} 0 & \text{if individual has AA genotype} \\ 1 & \text{if individual has Aa genotype} \\ 0 & \text{if individual has aa genotype} \end{cases} .$$

The p-values corresponding to association tests were calculated using three ways as follows.

1. **Testing β_1 from Model 1 with 1 WCR-MO \tilde{t} -test:** Using a \tilde{t} -test in analysis model 1 implemented via equation 3.;
2. **Testing β_1 from Model 1 *df* Using Z-scores:** Testing β_1 by combining p-values in analysis model 1 as described in Follmann et al. [2]. In short, we describe here method of combining p-values in WCR-MO procedure [2].

Except for non-parametric procedures in finite sample sizes, a valid statistical procedure produces p-values that follow a Uniform(0,1) distribution under the null hypothesis. Hence, if valid, WCR-MO produces the exchangeable p-values p_1, p_2, \dots, p_M , where M is the number of outputation re-samples, and each is marginally Uniform (0, 1) under the null hypothesis. Applying the transformation $Z_i = \Phi^{-1}(p_i)$ produces Z_1, Z_2, \dots, Z_M of exchangeable random variables that are marginally standard normal under the null

hypothesis, assuming p-values are independent [27]. Let $\bar{Z} = \sum_{i=1}^M Z_i / M$. The variance of Z can be obtained by variance decomposition formula. If M is large enough, then $Z \approx E(Z_i | X)$ and $\text{var}(Z) \approx \text{var}\{E(Z_i | X)\}$, where X is the raw data from that p-values were derived. Since $\text{var}(Z_i) = 1$ under the null hypothesis, we can approximate $\text{var}(\bar{Z}) \approx 1 - S_z^2$ where S_z^2 is the

sample variance of $M Z_i$'s. We use the following statistic to get the final p-value for each replicate: $\bar{Z} / \sqrt{1 - S_z^2}$, that is distributed as standard normal approximately.

- 3) **2 df test using Z-scores:** In this method we jointly tested β_1 and β_2 and used Follmann et al.'s procedure as described above to combine p-values in analysis model 2 [2].

4. RESULTS

The simulated data were analyzed using the methods described above and the p-values were calculated for each replicate, by testing β_1 with 1 *df* test, 1 *df* test using Z-scores, and 2 *df* test using Z-scores when testing additive and dominance effects together in the model.

Tables 1-4 show the empirical Type 1 error rates for the WCR-MO procedure, allowing only testing for the additive effect of the SNP (i.e. testing β_1 with one *df* test), the additive effect of the SNP by converting p-values to Z-scores, and by allowing tests for both additive and dominance effects of the SNP with 2 *df* test using Z-scores. These values serve as an evaluation of the conformity of the WCR-MO procedure to its asymptotic for the cluster sizes of 50 and 100 with equal and unequal cluster size. As can be seen, the empirical Type I error rates are nearly correct at the 5% α -level, but inflated at levels less than or equal to 1% for 50 families of equal size (Tables 1). For 100 unequally sized families, the test procedures becomes conservative at 5% α -levels, specifically corresponding to 1 *df* test using z-scores and 2 *df* test using z-scores (Table 2). Similar results follow as in equal size families at α -levels of greater than or equal to 0.01. Tables 3 and 4 show similar pattern of type I error rates distribution that the test procedures are conservative at the 5% α -level and liberal levels less than or equal to 1% (Tables 3-4). Also, we observed that approximately 1/3rd of the time; variance was estimated to be negative and could not be included in the association test leading to missing p-values. Exact numbers are given in column 4 of Tables 1-4.

5. DISCUSSION & CONCLUSIONS

The WCR-MO is an attractive simple to use procedure for analyzing clustered data such as pedigree data in genomic studies. The advantages of WCR-MO include simplicity to program; ability to use standard software; and use of statistical procedure for independent data sets. We investigated the utility of the WCR-MO procedure for association studies in pedigree data. The pedigrees can be considered as clusters and usually there is no between clusters correlations, but within cluster correlations exist due to relationships among individuals within the pedigree. We examined the null distribution of the test-statistic for the WCR-MO procedure using simulations with 100,000 replicates for 50 and 100 pedigrees of equal and varying size. We observed that Type I error rate was close to the nominal level at 95% confidence for most of the situations, but was inflated for confidence levels above 99%. Genetic studies, including genome-wide association studies (GWAS), typically require testing a large number of markers. The α -level to declare significance in GWAS is usually less than or equal to 10^{-7} , necessitating study of the behavior of test statistics at very small α -levels. In spite of simplicity of WCR-MO procedure, we have shown that WCR-MO is not yet ready to be used to analyze GWAS with pedigree data since it is very liberal for small α -levels. Further research is needed to modify the WCR-MO to offer valid tests, especially at very small α -levels. Note that, we have only tested WCR-MO for a specific type of genetic study. We further advise other researchers to perform simulation study before analyzing the real data to test the validity of WCR-MO for their specific application. To facilitate the programming of WCR-MO procedure for other applications, we have provided two programs (1) to simulate pedigree data and (2) analyze it with WCR-MO procedure as an

example. Both modules of the program were written in Java and can be accessed at <http://www.soph.uab.edu/ssg/software/wcr>.

Acknowledgments

This study is supported in part by R21LM008791, R01DK052431, P30DK056336, R01DK074842, R01GM077490, U54CA100949, and R01ES09912. The opinions expressed herein are those of the authors and not necessarily those of the NIH or any other organization with which the authors are affiliated. We thank Dr. Follman for valuable discussions on WCR-MO procedure.

References

- Hoffman EB, Sen PK, Weinberg C. Within-cluster resampling. *Biometrika* 2001;88:1121–1134.
- Follmann D, Proschan M, Leifer E. Multiple outputation: Inference for complex clustered data by averaging analyses from independent data. *Biometrics* 2003;59(2):420–9. [PubMed: 12926727]
- Meester SG, MacKay J. A parametric model for cluster correlated categorical data. *Biometrics* 1994;50(4):954–63. [PubMed: 7787008]
- Rieger RH, Weinberg CR. Analysis of clustered binary outcomes using within-cluster paired resampling. *Biometrics* 2002;58(2):332–41. [PubMed: 12071406]
- Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* 2003;157(4):364–75. [PubMed: 12578807]
- Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics* 2003;59(1):36–42. [PubMed: 12762439]
- Wu CO, Zheng G, Leifer E, Follmann D, Lin JP. A statistical method for adjusting covariates in linkage analysis with sib pairs. *BMC Genet* 2003;4(Suppl 1):S51. [PubMed: 14975119]
- Wu CO, Huang JHZ. Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis – Comment. *Journal of American Statistical Association* 2003;98(463):588–591.
- Benhin E, Rao JNK, Scott AJ. Mean estimating equation approach to analyzing cluster-correlated data with nonignorable cluster sizes. *Biometrika* 2005;92(2):435–450.
- Datta S, Satten GA. Rank-sum tests for clustered data. *Journal of Statistical Association* 2005;100(471):908–915.
- Kistner EO, Weinberg CR. A method for identifying genes related to a quantitative trait, incorporating multiple siblings and missing parents. *Genet Epidemiol* 2005 Sep;29(2):155–65. [PubMed: 16025442]
- Lu W. Marginal regression of multivariate event times based on linear transformation models. *Lifetime Data Analysis* 2005;11:389–404. [PubMed: 16133886]
- Matthews AG, Finkelstein DM, Betensky RA. Multivariate logistic regression for familial aggregation in age at disease onset. *Lifetime Data Anal* 2007;13(2):191–209. [PubMed: 17410428]
- Gilbert PB, Rossini AJ, Shankarappa R. Two-sample tests for comparing intra-individual genetic sequence diversity between populations. *Biometrics* 2005;61:106–117. [PubMed: 15737083]
- Faes C, Hens N, Aerts M, Shkedy Z. Estimating herd-specific force of infection by using random-effects models for clustered binary data and monotone fractional polynomials. *Appl Statist* 2006;55:595–613.
- Panageas KS, Schrag D, Localio AR, Venkatraman ES, Begg CB. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statist Med* 2007;26:2017–2035.
- Larocque D, Nevalainen J, Oja H. A weight multivariate sign test for cluster-correlated data. *Biometrika* 2007;94:267–283.
- Williamson JM, Kim HY, Manatunga A, Addiss DG. Modeling survival data with informative cluster size. *Stat Med* 2008;27(4):543–55. [PubMed: 17640035]

19. Shin J, Darlington GA, Cotton C, Corey M, Bull SB. Confidence intervals for candidate gene effects and environmental factors in population-based association studies of families. *Ann Hum Genet* 2007;71:421–32. [PubMed: 17346258]
20. Ballard RA, Truog WE, Cnaan A, Martin RJ, Ballard PL, Merrill JD, Walsh MC, Durand DJ, Mayock DE, Eichenwald EC, Null DR, Hudak ML, Puri AR, Golombek SG, Courtney SE, Stewart DL, Welty SE, Phibbs RH, Hibbs AM, Luan X, Wadlinger SR, Asselin JM, Coburn CE, Coburn CE. NO CLD Study Group. Inhaled nitric oxide in preterm infants undergoing mechanical ventilation. *N Engl J Med* 2006;355:343–53. [PubMed: 16870913]
21. Monti P, Ciribilli Y, Jordan J, Menichini P, Umbach DM, Resnick MA, Luzzatto L, Inga A, Fronza G. Transcriptional functionality of germ line p53 mutants influences cancer phenotype. *Clin Cancer Res* 2007;13:3789–95. [PubMed: 17606709]
22. Di Mascio M, Markowitz M, Louie M, Hurley A, Hogan C, Simon V, Follmann D, Ho DD, Perelson AS. Dynamics of intermittent viremia during highly active antiretroviral therapy in patients who initiate therapy during chronic versus acute and early human immunodeficiency virus type 1 infection. *J Virol* 2004;78(19):10566–73. [PubMed: 15367623]
23. Highbarger HC, Hu Z, Kottlil S, Metcalf JA, Polis MA, Vasudevachari MB, Lane HC, Dewar RL. Comparison of the Abbott 7000 and the Bayer 340 Systems for the Measurement of HCV Viral Load. *J Clin Microbiol* 2007;45:2808–2812. [PubMed: 17596352]
24. Wong WT, Agrón E, Coleman HR, Reed GF, Csaky K, Peterson J, Glenn G, Linehan WM, Albert P, Chew EY. Genotype-phenotype correlation in von Hippel-Lindau disease with retinal angiomas. *Arch Ophthalmol* 2007;125:239–45. [PubMed: 17296901]
25. Mehta T, Tanik M, Allison DB. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat Genet* 2004 Sep;36(9):943–7. [PubMed: 15340433]
26. Redden DT, Allison DB. The effect of assortative mating upon genetic association studies: spurious associations and population substructure in the absence of admixture. *Behav Genet* 2006;36(5):678–86. [PubMed: 16514473]
27. Hedges, LV.; Olkin, I. *Statistical methods for meta-analysis*. Orlando: Academic Press; 1985.

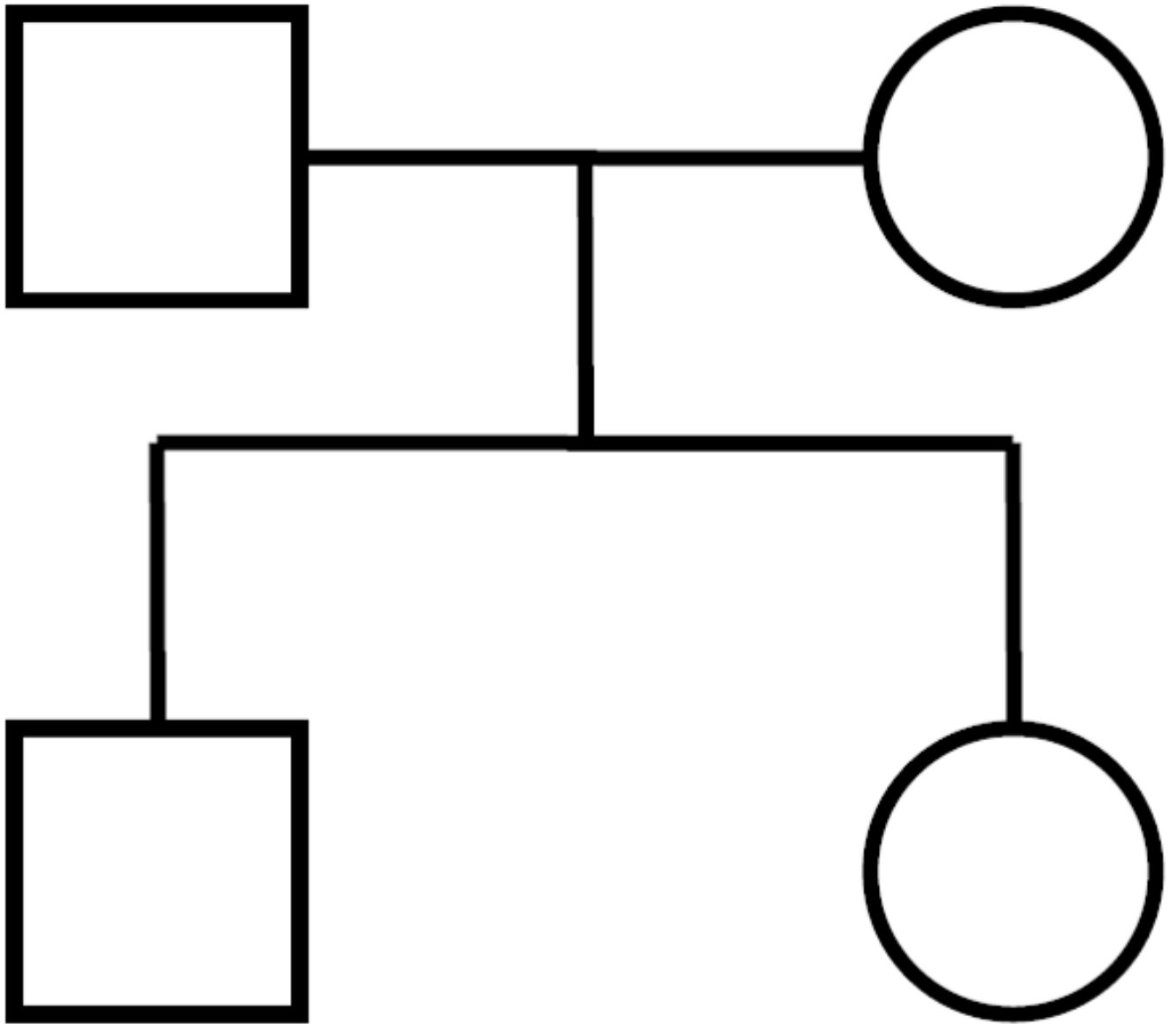


Figure 1.
Equal size pedigrees used for simulation.

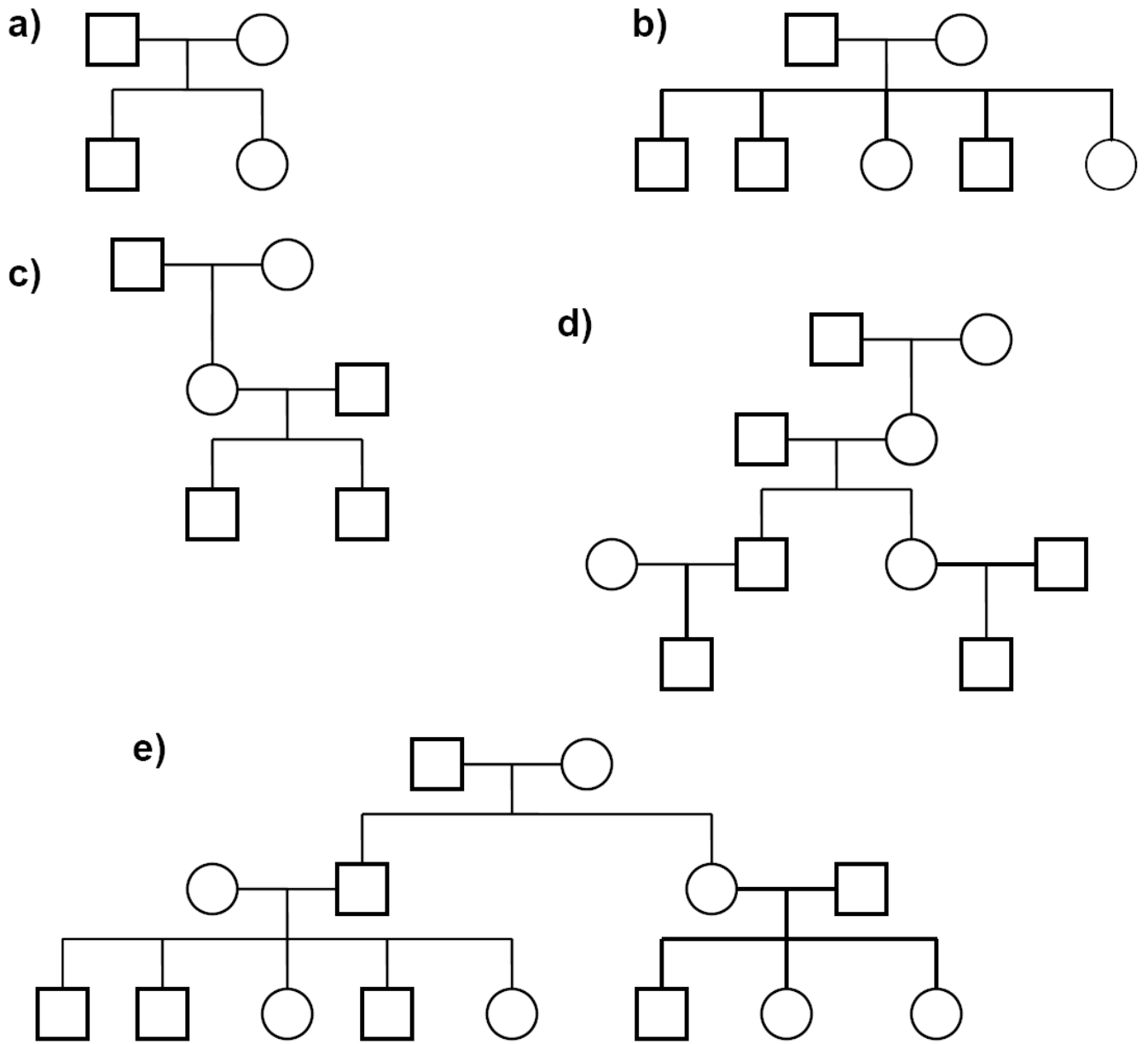


Figure 2.
Unequal size pedigrees used for simulation.

Table 1

The null distribution of the test statistic for WCR-MO procedure using 100,000 replicates of 50 equal size families (i.e. two siblings and both parents in the family). For each replicate we used 10,000 bootstrap samples.

Test Procedure	Nominal α -level	Total number of times the total variance estimate was positive	Number of times total variance was negative resulting in missing p-values	Number of times p-values less than given α -level	Type I error rate using only non-missing p-values
Testing β_1 with 1 df test	0.05	95,906	4,094	6,571	0.0685
	0.01	95,906	4,094	2,674	0.0279
	0.001	95,906	4,094	1,183	0.0123
	0.0001	95,906	4,094	713	0.0074
1 df test using Z-scores	0.05	92,129	7,871	4,601	0.0499
	0.01	92,129	7,871	1,896	0.0206
	0.001	92,129	7,871	901	0.0098
	0.0001	92,129	7,871	583	0.0063
2 df test using Z-scores	0.05	97,431	2,569	4,904	0.0503
	0.01	97,431	2,569	1,491	0.0153
	0.001	97,431	2,569	536	0.0055
	0.0001	97,431	2,569	305	0.0031

Table 2

The null distribution of the test statistic for WCR-MO procedure using 100,000 replicates of 50 unequal size families given in Figure 2. For each replicate we used 10,000 bootstrap samples.

Test Procedure	Nominal α -level	Total number of times the total variance estimate was positive	Number of times negative resulting in missing p-values	Number of times total variance was less than given α -level	Type I error rate using only non-missing p-values
Testing β_1 with 1 df test	0.05	67,318	32,682	3,023	0.0449
	0.01	67,318	32,682	1,730	0.0257
	0.001	67,318	32,682	1,092	0.0162
1 df test using Z-scores	0.0001	67,318	32,682	788	0.0117
	0.05	68,220	31,780	2,199	0.0292
	0.01	68,220	31,780	1,303	0.0175
2 df test using Z-scores	0.001	68,220	31,780	740	0.0109
	0.0001	68,220	31,780	541	0.0077
	0.05	64,747	35,253	2,161	0.0334
2 df test using Z-scores	0.01	64,747	35,253	1,280	0.0198
	0.001	64,747	35,253	817	0.0126
	0.0001	64,747	35,253	581	0.0090

Table 3

The null distribution of the test statistic for WCR-MO procedure using 100,000 replicates of 100 equal size families (i.e. two siblings and both parents in the family). For each replicate we used 10,000 bootstrap samples.

Test Procedure	Nominal α -level	Total number of times the total variance estimate was positive	Number of times total variance was negative resulting in missing p-values	Number of times p-values less than given α -level	Type I error rate using only non-missing p-values
Testing β_1 with 1 df test	0.05	71,223	28,777	4,576	0.06425
	0.01	71,223	28,777	2,552	0.03583
	0.001	71,223	28,777	1,529	0.02147
1 df test using Z-scores	0.0001	71,223	28,777	1,082	0.01519
	0.05	68,220	31,780	1,989	0.02916
	0.01	68,220	31,780	1,193	0.01749
2 df test using Z-scores	0.001	68,220	31,780	743	0.01089
	0.0001	68,220	31,780	528	0.00774
	0.05	72,578	27,422	3,064	0.04222
2 df test using Z-scores	0.01	72,578	27,422	1,904	0.02623
	0.001	72,578	27,422	1,239	0.01707
	0.0001	72,578	27,422	888	0.01223

Table 4

The null distribution of the test statistic for WCR-MO procedure using 100,000 replicates of 100 unequal size families given in Figure 2. For each replicate we used 10,000 bootstrap samples.

Test Procedure	Nominal α -level	Total number of times the total variance estimate was positive	Number of times total variance was negative resulting in missing p-values	Number of times p-values less than given α -level	Type I error rate using only non-missing p-values
Testing β_1 with 1 df test	0.05	64,555	35,445	3,304	0.0512
	0.01	64,555	35,445	1,903	0.0295
	0.001	64,555	35,445	1,172	0.0185
	0.0001	64,555	35,445	834	0.0129
1 df test using Z-scores	0.05	62,695	37,305	1,723	0.0275
	0.01	62,695	37,305	1,005	0.016
	0.001	62,695	37,305	584	0.0093
	0.0001	62,695	37,305	409	0.0065
2 df test using Z-scores	0.05	64,191	35,809	1,903	0.0296
	0.01	64,191	35,809	1,120	0.0174
	0.001	64,191	35,809	676	0.0105
	0.0001	64,191	35,809	502	0.0078