Empirical estimation of permutation-based metabolome-wide significance thresholds

Alina Peluso [*,‡], Timothy M. D. Ebbels [*,‡], and Robert Glen [*,‡]

November 24, 2018

Abstract

Motivation: A key issue in the omics literature is the search of statistically significant relationships between molecular markers and phenotype. The aim is to detect disease-related discriminatory features while controlling for false positive associations at adequate power. Metabolome-wide association studies have revealed significant relationships of metabolic phenotypes with disease risk by analysing hundreds to tens of thousands of molecular variables leading to multivariate data which are highly noisy and collinear. In this context, Bonferroni or Sidak correction are rather useful as these are valid for independent tests, while permutation procedures allow for the estimation of p-values from the null distribution without assuming independence among features. Nevertheless, under the permutation approach the distribution of p-values may presents systematic deviations from the theoretical null distribution which leads to biased adjusted threshold estimate, e.g. smaller than a Bonferroni or Sidak correction.

Methods: We make use of parametric approximation methods based on a multivariate Normal distribution to derive stable estimates of the metabolome-wide significance level within a univariate approach based on a permutation procedure which effectively controls the maximum overall type I error rate at the α level.

Results: We illustrate the results for different model parametrizations and distributional features of the outcome measure, as well as for diverse correlation levels within the features and between the features and the phenotype in real data and simulated studies.

Availability: `MWSL` is the open-source R software package for the empirical estimation of the metabolomic-wide significance level available at https://github.com/AlinaPeluso/MWSL. This include the original metabolomic dataset employed in this study and the main function for the MWSL estimation. A user guide tutorial is provided to detail the procedure.

[*]Imperial College London, Faculty of Medicine, Department of Surgery & Cancer, South Kensington Campus, London SW7 2AZ, UK. [‡]To whom correspondence should be addressed: {apeluso, t.ebbels, r.glen}@imperial.ac.uk.

List of Figures

1 Distributions of the clinical outcome measures. 11
2 BINNED data: ENT across clinical outcome measures and for different transformations of the features: identity, multivariate Normal, multivariate log-Normal. 12
3 CPMG data: ENT across clinical outcome measures and for different transformations of the features: identity, multivariate Normal, multivariate log-Normal. 13
4 NOESY data: ENT across clinical outcome measures and for different transformations of the features: identity, multivariate Normal, multivariate log-Normal. 14
5 ENTs from the estimation of uncorrelated (dx) and correlated (sx) outcome across correlated features. 15
6 R ratio (%) from the estimation of uncorrelated (dx) and correlated (sx) outcome across correlated features. 16
7 BINNED data: ENT, type I and type II error estimation from the permutation procedure for different type of simulated outcome measure: continuous, discrete-binary, discrete-count, survival. 17
8 PCA simulated features (M=655, n_t=1,500, PCs=350): ENT variation, type I and type II error estimation from the permutation procedure for different type of simulated outcome measure: continuous, discrete-binary, discrete-count, survival. 18

List of Tables

1 Descriptive statistics for the clinical outcome measures. 10
2 Descriptive statistics for the fixed effects confounders. 10

1 Introduction

One important goal in literature is the detection of statistically significant relationships between discriminatory features and disease phenotypes. The aim is to gauge each variable's strength of association with disease while minimising the risk of false positive associations at adequate power. In such studies many hundreds to tens of thousands of molecular variables are collected for each individual, leading to high-dimensional multivariate data which are highly collinear. In metabolomics, effective methods to adjust for multiple testing are a central topic in the context of Metabolome-Wide Association Studies (MWAS). The p-value from the statistical test of each discriminatory features generated via a univariate approach is employed to detect the association between metabolic variables and disease status, and the threshold required to control the family-wise error rate (FWER) is estimated through a permutation approach as previously described in [4], [2], and [1]. Although permutation procedures can yield the correct FWER even when the tests are dependent, it may happen that a permutation-based threshold could result in an estimate lower than a thresholds adjusted with more traditional method such as the Bonferroni or Sidak adjustment, which are known to be valid in the case of independent tests but they tend to be overly conservative in association studies in which the tests are correlated. A valid adjustment for multiple testing must account for the dependence between variables, and therefore guarantee that metabolomic-wide significance level (MWSL) estimate would be larger than that estimated from a Bonferroni or Sidak correction.

To address this, we make use of approximation methods within a permutation procedure to derive stable estimate of the MWSL while retaining the data structure up to the 2nd order. This guarantee that the complex correlation structure existing in the metabolic profiling is accounted for. We illustrate the results for different model parametrizations and diverse distributional properties of the outcome, as well as by investigating different correlation levels of the features and between the features and the phenotype in real data and simulated studies. Moreover, we show that the method effectively controls the maximum overall type I error rate at the α level.

2 Methods

2.1 Permutation-based MWSL estimation

Suppose the data consists of n observations, and let Y be the response variable, $X = (X_1, \ldots, X_M)^T$ the vector of M predictors or features, and $Z = (Z_1, \ldots, Z_P)^T$ the vector of P fixed effects confounders. The permutation procedure to estimate the MWSL can be described as follows.

- Step (1): Shuffle i.e. re-sample without replacement, the outcome variable Y together with the set of fixed effects confounders Z if any. In this way, the n subjects are resampled under the null hypothesis of no association.

- Step (2): Fit a regression model to the data in a univariate approach i.e. by including one feature

at times. From each model store the p-value associated to the feature of interest.

- Step (3): Extract the minimum of the M p-values.
- Step (4): Repeat Step (1)-(3) at least $K = n/2$ times [5]. K minimum p-values are the elements of the new vector *q*.
- Step (5): Sort the elements of *q*, and take the (αK)-value of this vector. This value is the MWSL estimate. An approximate confidence interval can be obtained by treating the true position of the MWSL estimate as a Binomial r.v. with parameters K and α. Then, using the Normal approximation to the Binomial, we obtain the (1-α)% confidence limits by extracting the elements of *q* in positions $(\alpha K) \pm \{(1-\alpha)\sqrt{\alpha K(1-\alpha)}\}$.
- Step (6): Compute the effective number of tests (ENT) defined as the number of independent tests that would be required to obtain the same significance level using the Bonferroni ENT$_B$=$\frac{\alpha}{\mathrm{MWSL}}$, or the Sidak ENT$_S$=$\frac{\log(1-\alpha)}{\log(1-\mathrm{MWSL})}$ correction. The ENT estimate measures the extent that the M markers are non-redundant. Therefore, the ratio of the effective and the actual number of tests as R=$\frac{\mathrm{ENT}}{M}\%$ is a measure of the dependence among features, which it is expected to be closer to 0% when highly correlated features are considered.

Similar versions of this procedure have been previously applied in different studies e.g. by [4] to approximate the genome-wide significance threshold for dense SNP and resequencing data, or by [2] to determine the MWSL based on the INTERMAP data set and provide a large-scale standardized set of urinary metabolic profiles which capture the variations between human populations in China, Japan, UK and USA. More recently in the context of NMR metabolic profiling studies, [1] employed the permutation algorithm to perform a series of MWAS for serum levels of glucose, and in this case the ENT estimate for raw glucose concentrations resulted beyond than the actual number of tests with an R ratio over 400%.

2.2 Parametric approximation methods

We investigate the properties of the permutation approach for significance level estimation, and we employ the multivariate Normal and log-Normal distributions to describe, at least approximately, the set of correlated features which clusters around their mean value to obtain stable estimates of the MWSL while effectively controlling the maximum overall type I error rate at the α level. Specifically, $X \sim \mathcal{N}_M(\mu, \Sigma^*)$ is the multivariate Normal distribution approximation to the set of features where $\mu = \mathrm{E}[X] = (\mathrm{E}[X_1], \ldots, \mathrm{E}[X_M])^T = \mathbf{0}$ is the M-dimensional mean vector of zero means, and Σ^* is the $(M \times M)$ shrinkage estimator of the covariance matrix as described by [7], which is always positive definite, well-conditioned, more efficient and therefore preferred to the unbiased estimator Σ, or to the related maximum likelihood estimator Σ_{ML}. When the simulation via the multivariate log-Normal distribution is considered the features are previously transformed i.e. the absolute value of their minimum plus one unit is added to their original value.

The algorithm is applied to real-data and simulated scenarios to illustrate the results for different model parametrizations and distributional features of the outcome, as well as to investigate different correlation levels across features and between the features and the phenotype.

3 Results

3.1 Real-data study

Univariate MWAS are employed to investigate the spectral features of human serum ^{1}H NMR spectroscopic profiles from participants of the Multi-Ethnic Study of Atherosclerosis (MESA). The data have been extensively described in [1].

Briefly, the cohort includes participants (51% females, 49% males), aged 44-84 years, (mean=63 years) from four different ethnic groups: Chinese-American, African-American, Hispanic, and Caucasian, all recruited between 2000-2002 at clinical centres in the United States and free of symptomatic cardiovascular disease at baseline. Demographic, medical history, anthropometric, lifestyle data, and serum samples were collected during the first examination, together with information on diabetes, and lipid and blood pressure treatment. The phenotypes of interest in this study are the glucose concentration levels and the body mass index (BMI) of the individuals in the study. Table 1 presents the descriptive statistics for the clinical outcome measures in Figure 1, while Table 2 reports the descriptive statistics for the fixed effects confounders used in the study.

Three set of features are considered: (1) a standard water suppressed one-dimensional spectrum termed **NOESY**, and (2) a Carr-Purcell-Meiboom-Gill spectrum as **CPMG**, and (3) a lower resolution of the CPMG data referred to as **BINNED** data. The BINNED version consists of M=655 features, while the NOESY and CPMG include M=30,590 metabolic profiles. By removing any outliers the BINNED data sample comprises of n=3,500 individuals, while dimension of the NOESY and CPMG data is n=3,867 participants.

From the analysis of the BINNED data as in Figure 2, when the real features are considered (sx) there is instability in the estimation of the ENT across the different outcomes, and in particular the ENT estimate of glucose is meaningless being beyond the ANT. When the data are approximated via a multivariate log-Normal (middle) and a multivariate Normal (dx) as described in Section 2.2 the ENT estimates are stable across the different outcomes and remain bounded below the total number of features with an average ENT of 355 and an R ratio around 55%. To assess the validity of this result in terms of redundancy of the set of features we consider a nonparametric method such as a principal component (PC) analysis (PCA) and we find that the cumulative proportion of variance explained by the first 350 PCs is around 99%. This means that around half of the features could be written as a linear transformation of the others, and this confirms our findings.

Figure 3 reports the ENT estimates for CPMG data across the clinical outcomes considered. Without any transformations applied to the set of features, there is very large variation across the ENT estimates for the different outcomes, and in particular very high and meaningless estimate for glucose levels which

goes beyond 400%. On the other hand, when the set of features is simulated from the multivariate (log-) Normal distribution the resulting ENT estimate bounded below the total number of features, and stable across different outcomes with an average ENT of around 16,000 features and an R ratio around 55%. A PCA analysis applied to the CPMG data would not be very informative as the ENTs estimates is greater than the total number of n PCs which account for the of 100% of the cumulative proportion of variance explained.

Figure 4 reports the ENT estimates for the NOESY data which are bounded below 100% but varies across outcomes when the original set of features is considered. When simulated features from the multivariate (log-) Normal distributions are considered we obtain lower ENT values than the ones from the CPMG data, with an average ENT of 2,700 features and an R ratio around 9%. This result was expected due to the reduced influence of broad signals which is linked to a reduction in the covariance structure of the data. By applying a PCA to the NOESY data the cumulative proportion of variance explained by the first 2,700 PCs is around 99%, and this is in line with our findings.

3.2 Parametric simulation study

We now investigate how the the correlation level among features and between outcome and features impact the ENTs estimation. To do so we simulate various sets of features with a defined correlation level as follows.

3.2.1 Feature simulation

- Step (1): Generate a $(M \times M)$ correlation matrix A assuming all variables have unit variance, i.e. the M elements on the diagonal are **1**s. The $[M(M-1)]/2$ elements of the upper triangular matrix are sampled from a Uniform distributions bounded to a certain interval, e.g. high correlation level within the interval [0.75,0.85], medium correlation in [0.45,0.55], or low correlations in [0.25,0.35].
- Step (2): As the eigenvalues of A are required to be greater than zero, compute S as the nearest positive definite matrix to the correlation matrix A as described by [3].
- Step (3): Derive the lower triangular matrix L via Cholesky decomposition of matrix S such that $S = LL'$.
- Step (4): M multivariate Normal features with zero means result from the product ZL between the $(n \times M)$ matrix Z of M random N(0,1) i.i.d. features, and the $(M \times M)$ lower triangular matrix L. The correlations of the simulated features are very close to these imputed in matrix A.

3.2.2 Outcome simulation

The set of features are analysed together with continuous uncorrelated and correlated outcome measures.

Uncorrelated outcomes of different shapes are easily simulated via parametric distributions such as the Normal distribution for a symmetric outcome, the Skew-Normal distribution for a left skewed

outcome, and a Weibull distribution for a right skewed outcome. Therefore we estimated the ENT for each simulated outcomes and all the set of simulated features as described in Section 3.2.1. As the results are similar for different outcome measure, in Figure 5 we present the averaged ENT estimates across the set of features simulated with diverse correlation levels.

Simulated correlated outcomes can be obtained as a linear combination of few features such as $\sum_{m=1}^{M_j} X_m + \epsilon$, with $\epsilon \sim U(0,1)$ and $M_j < M$, or via procedures based on Cholesky decomposition such as the method used to generate the simulated correlated features described in Section 3.2.1. Both methods lead to outcomes correlated with the features with the same correlation level among features. Both outcomes are tested and the average ENT estimate is showed in Figure 5. We observe that correlation to outcome makes no difference.

To describe the inverse relationship between the correlation levels of the features and the ENT estimates i.e. decreasing R(%) ratios for higher level of correlations within the features, we fit a nonlinear 3rd degree-polynomial curve to the ENT estimates as showed in Figure 6.

4 Validation of the approach

A type I error (false-positive) occurs when the true null hypothesis is being rejected, while a type II error (false-negative) consists in a failure to reject the null hypothesis which is actually false. The latter is directly linked to the statistical power of the procedure, and in particular for a type II error probability β, the corresponding statistical power is 1-β.

To check whether the procedure accounts for the family-wise error rate at a given α level we measured the the type I and type II error rate respectively as the occurrences of having a p-value for the considered feature less or equal (type I) or greater (type II) than the arbitrary threshold α. Moreover, we now extend the procedure to account for various types of outcomes. In particular, we simulate a total of four uncorrelated outcome measures i.e. a continuous outcome from a Normal distribution, a discrete-binary outcome from a Binomial distribution, a discrete-count outcome from a Poisson distribution, and a survival outcome as a time-to-event outcome. Figure 7 shows the ENT estimates for the BINNED set of features, across a range of simulated uncorrelated outcome measure: continuous, discrete-binary, discrete-count, survival. The procedure effectively controls the family-wise error rate at the (default) α-level of 5%.

4.1 Nonparametric simulation study

The procedure we described so far is based on a parametric multivariate simulation method and involve a re-sampling approach. To further validate the results we now employ a nonparametric method such as PCA to simulate the set of features and while accounting for the original structure of the data. The algorithm can be described as follow.

- Step (1): By randomly sampling n_t observations from the original data matrix of features X,

compute X_t the $(n_t \times M)$ *test* set of features, and $X_{\bar{t}}$ the $(n_{\bar{t}} \times M)$ *nontest* set, with $n_t < n$ and $n_{\bar{t}} = n - n_t$.

- Step (2): Standardise the *test* and the *nontest* set by subtracting its respective vector of column means i.e. μ_t and $\mu_{\bar{t}}$, and dividing by its standard deviations i.e. σ_t and $\sigma_{\bar{t}}$, to respectively obtain Z_t and $Z_{\bar{t}}$.

- Step (3): Compute PCA over the *nontest* set by applying singular value decomposition (SVD) methods such that $Z_{\bar{t}} = U_{\bar{t}}\Sigma_{\bar{t}}V_{\bar{t}}^T$, where $V_{\bar{t}}^T$ is the $(M \times M)$ matrix of loadings, while the PC scores are obtained as the product between the $(n_{\bar{t}} \times n_{\bar{t}})$ matrix $U_{\bar{t}}$ of eigenvectors of $Z_{\bar{t}}Z_{\bar{t}}^T$, and the $(n_{\bar{t}} \times M)$ diagonal matrix $\Sigma_{\bar{t}}$.

- Step (4): Use the *nontest* PC scores $U_{\bar{t}}\Sigma_{\bar{t}}$ from Step (3) to compute the $(n_t \times M)$ matrix $\hat{U}_t\hat{\Sigma}_t$ of PC predicted scores for the *test* set.

- Step (5): Build the $(n_t \times M)$ simulated *test* set of features $\hat{Z}_t$ as the product of the PC predicted scores from Step (4), and the matrix of loadings $V_{\bar{t}}^T$ from Step (3) such that $\hat{Z}_t = \hat{U}_t\hat{\Sigma}_t V_{\bar{t}}^T$. We note that S PCs, with $S \leq M$, can be selected to be used in the predictions, thus $\hat{Z}_t$ would result from the product of $(n_t \times S)$ matrix of PCs and the $(S \times M)$ matrix of loadings.

- Step (6): From the simulated *test* set of standardised features $\hat{Z}_t$ compute the $(M \times M)$ set of simulated features as $\hat{X}_t = \hat{Z}_t\sigma_t + \mu_t$.

To simulate features in this context the sample size of the data should be large enough, and no missing values are allowed.

We consider the BINNED data and define the *test* set by randomly sampling $n_t = 1,500$ individuals. From the PCA over the *nontest* set we select 350 PCs to be used to build the simulated *test* set of features $\hat{X}_t$ as these account for the 99% of variance explained and are in number equal to the average ENTs selected by the permutation procedure. The permutation procedure is therefore applied across diverse simulated uncorrelated outcome measure: continuous, discrete-binary, discrete-count, survival. As showed in Figure 8 the procedure effectively controls the family-wise error rate at the (default) α-level of 5%, and for the type II error rate at the (1-α)% level.

5 Conclusions

In this paper we consider the empirical estimation of a significance level threshold in the context of multiple testing adjustment when dealing with correlated data. The procedure is based on an iterative permutation approach via univariate regression models. Parametric approximation methods via the multivariate Normal and multivariate log-Normal distributions are employed beforehand to retain the dependence structure of the data. When permutation procedures are applied to the approximated data the MWSL is stable across outcome measures with diverse properties. Spectral variables from

metabolic profiles exhibit a high degree of collinearity, and this is supported by our finding that in all scenarios considered, when parametric methods are applied to approximate the structure of the data, the MWSL estimated through the permutation procedure is larger than the threshold obtained via a metabolome-wide Bonferroni or Sidak correction. Therefore, the corresponding ENTs is always less than the actual number of tests as it mainly depends on the extent of correlation within the data. The extent of collinearity is summarized by the ratio of effective to actual number of tests across the diverse spectral resolutions and, respectively, around 55% for the CPMG data (high-resolution and and BINNED version), and around 10% for the NOESY resolution suggesting that the latter shows lower correlations between spectral variables. The number of independent variables might give an indication of the number of independent metabolic processes exhibited by the system, since each independent process might be expected to manifest itself through multiple metabolic variables. If the data are interpreted this way, our analysis suggests that around 350 separate metabolic processes in the BINNED resolution, around 15,000 processes within the CPMG resolution and 3,000 processes for the NOESY resolution are being captured by humans NMR metabolic profiles of urine. The relationship between the ENT and the correlation level within the features is nonlinear. Via simulation and real data study it is showed how the permutation procedure satisfactorily performs in the task of discovering features providing reliable estimates in terms of false positive rates and statistical power.

Availability

FWERperm is the core function for the estimation of a permutation-based significance threshold for metabolomic data, and it is available within the R package under the name `MWSL`. Within this package it is also available the lower resolution CPMG data referred to in the text as BINNED data. Also, a user guide tutorial is provided to detail the procedure.

Funding

We thank Dr Marc Chadeau-Hyam for useful discussions and insights on this paper. This work has been supported by the PhenoMeNal project [grant agreement number 654241] [6].

References

[1] Raphaële Castagné, Claire Laurence Boulangé, Ibrahim Karaman, Gianluca Campanella, Diana L Santos Ferreira, Manuja R Kaluarachchi, Benjamin Lehne, Alireza Moayyeri, Matthew R Lewis, Konstantina Spagou, et al. Improving visualization and interpretation of metabolome-wide association studies: An application in a population-based cohort using untargeted 1H-NMR metabolic profiling. *Journal of proteome research*, 16(10):3623–3633, 2017.

[2] Marc Chadeau-Hyam, Timothy MD Ebbels, Ian J Brown, Queenie Chan, Jeremiah Stamler, Chiang Ching Huang, Martha L Daviglus, Hirotsugu Ueshima, Liancheng Zhao, Elaine Holmes, et al. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *Journal of proteome research*, 9(9):4620–4627, 2010.

[3] Nicholas J Higham. Computing the nearest correlation matrix - a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343, 2002.

[4] Clive J Hoggart, Taane G Clark, Maria De Iorio, John C Whittaker, and David J Balding. Genome-wide significance for dense snp and resequencing data. *Genetic epidemiology*, 32(2):179–185, 2008.

[5] Efstathios Paparoditis and Dimitris N Politis. The local bootstrap for periodogram statistics. *Journal of Time Series Analysis*, 20(2):193–222, 1999.

[6] PhenoMeNal. Phenomenal (phenome and metabolome analysis) - large scale computing for medical metabolomics.

[7] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

6 Appendix

Table 1: Descriptive statistics for the clinical outcome measures.

	mean	sd	median	min	max	skewness	kurtosis
Glucose (mg/dL)	98.28	31.10	90	38	507	4.17	28.89
Logarithm of Glucose	4.56	0.23	4.5	3.64	6.23	2.22	10.35
BMI (kg/m^2)	28.14	5.39	27.34	15.36	61.86	46.50	4.45
Logarithm of BMI	3.32	0.18	3.31	2.73	4.12	1.39	3.20

Table 2: Descriptive statistics for the fixed effects confounders.

	mean	sd
Age (years)	62.89	10.32
Gender	0.51	0.49
Height (cm)	166.43	10.23
Ethnicity: Caucasian	0.39	0.49
Ethnicity: Hispanic	0.23	0.42
Ethnicity: African-American	0.25	0.43
Ethnicity: Chinese-American	0.13	0.34
Smoking: Never	0.51	0.50
Smoking: Former	0.12	0.33
Smoking: Current	0.38	0.48
LDL cholesterol (mg/dL)	117.67	31.04
HDL cholesterol (mg/dL)	51.29	14.41
Systolic blood pressure (mmHg)	126.92	21.54
Blood pressure treatment	0.38	0.49
Diabetes	0.14	0.34
Lipids treatment	0.17	0.37

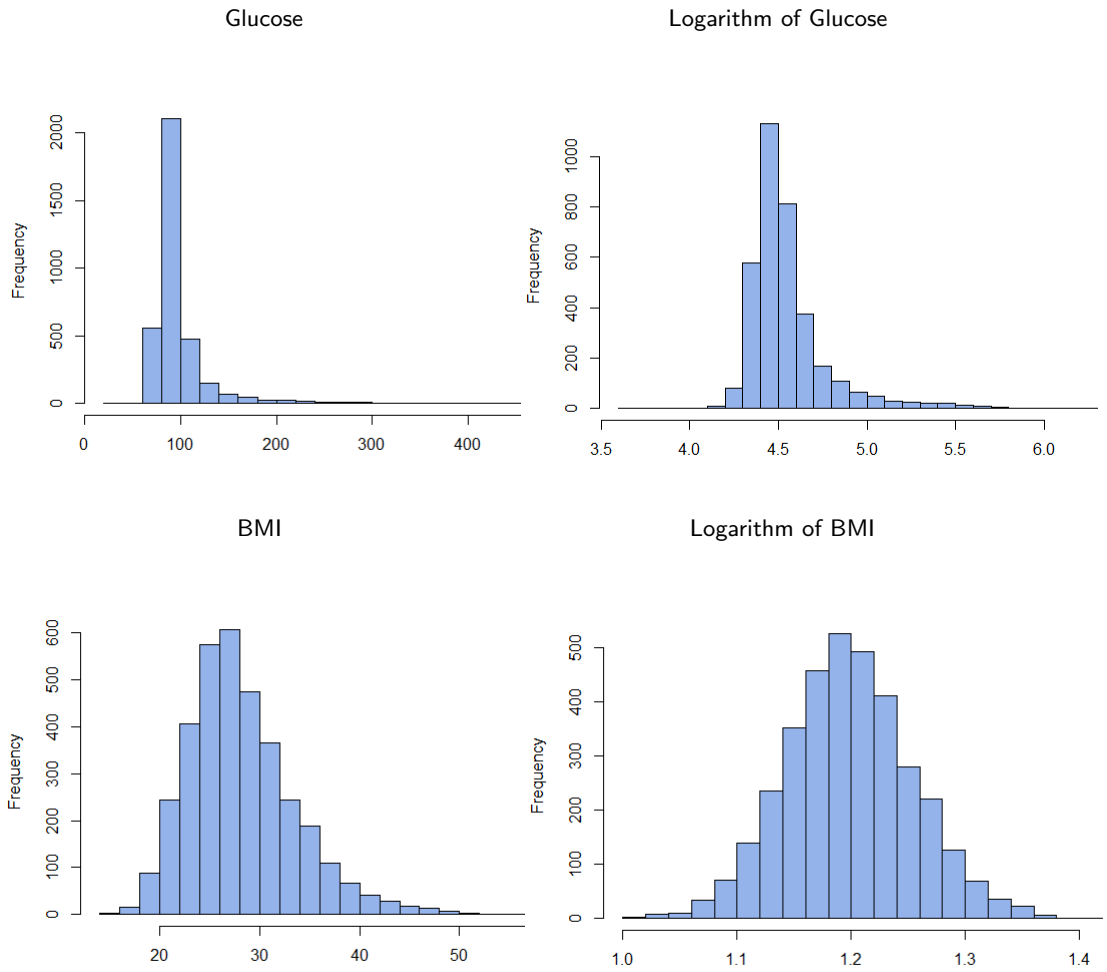


Figure 1: Distributions of the clinical outcome measures.

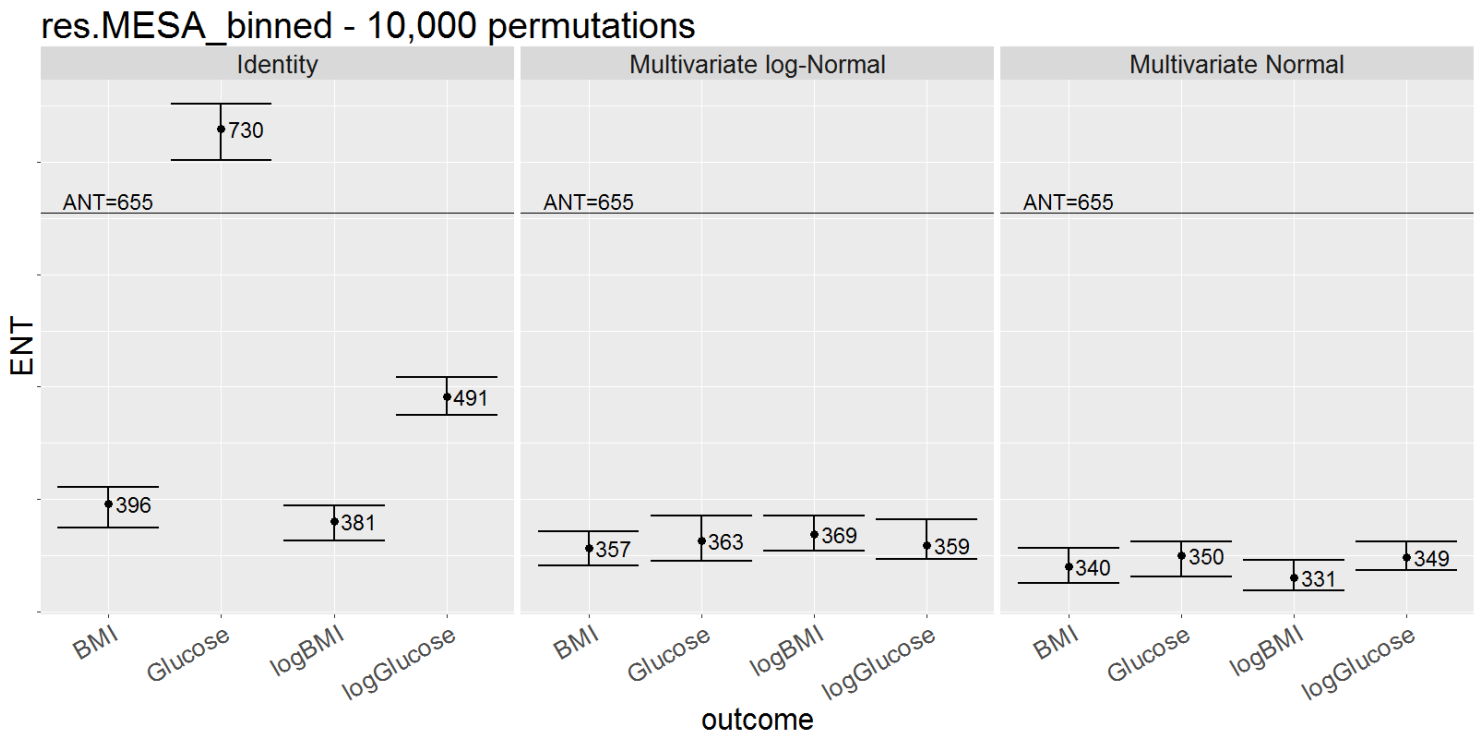


Figure 2: BINNED data: ENT across clinical outcome measures and for different transformations of the features: identity, multivariate Normal, multivariate log-Normal.

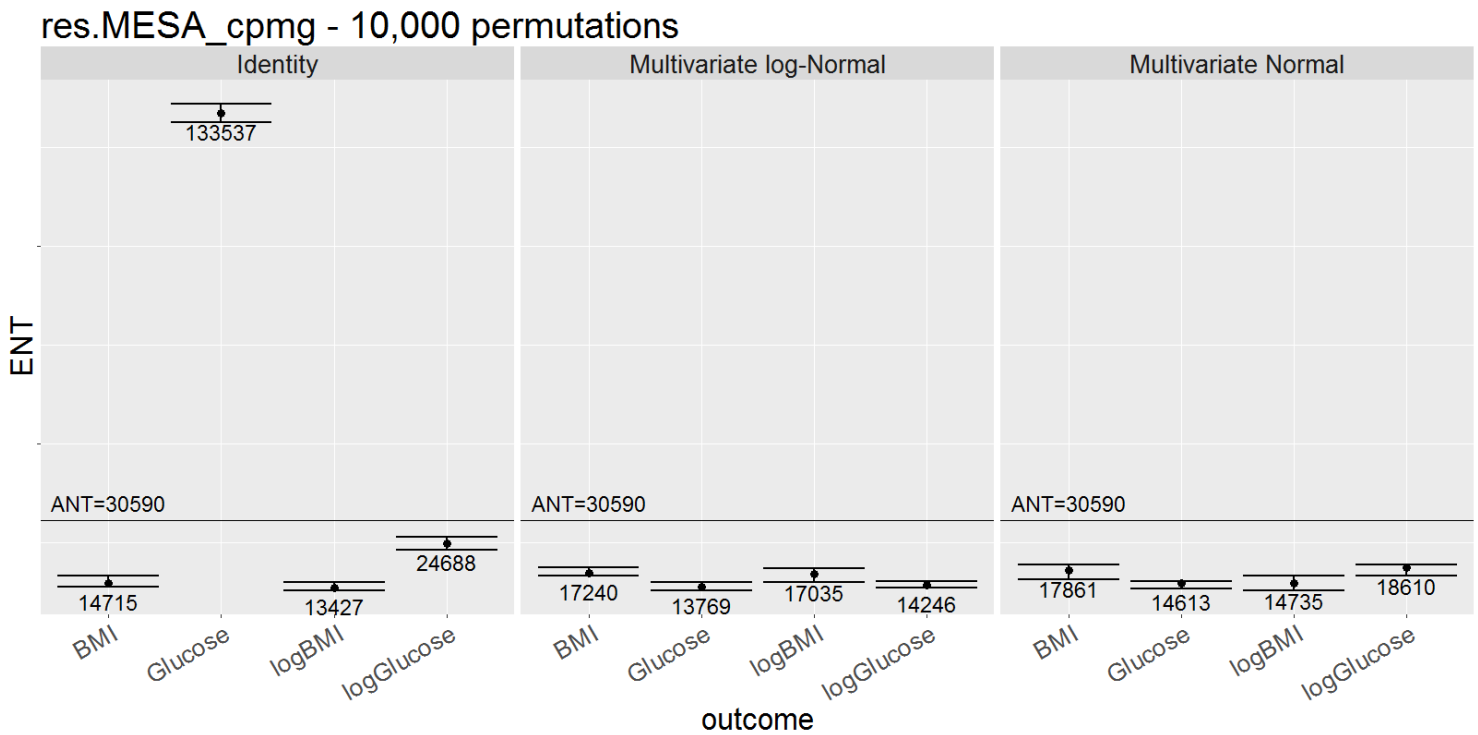


Figure 3: CPMG data: ENT across clinical outcome measures and for different transformations of the features: identity, multivariate Normal, multivariate log-Normal.

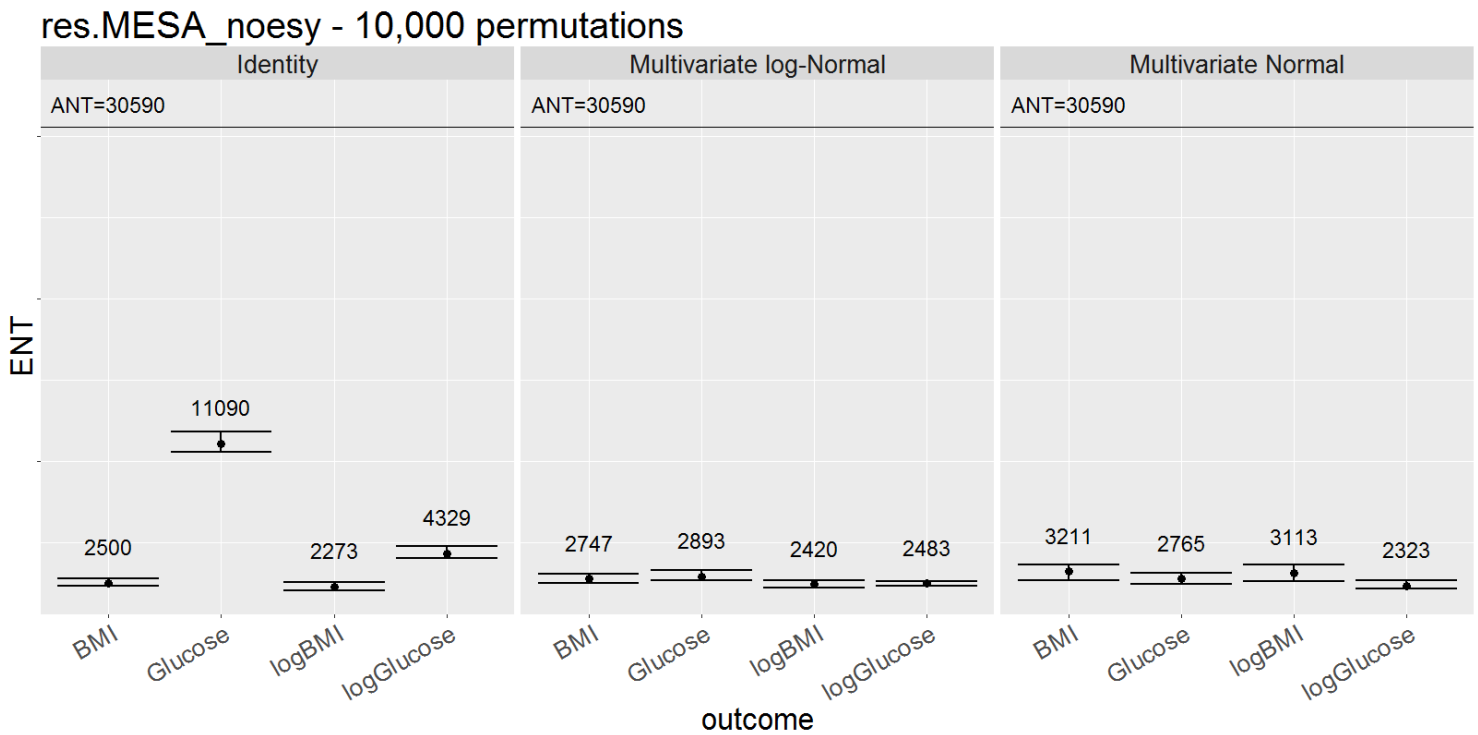


Figure 4: NOESY data: ENT across clinical outcome measures and for different transformations of the features: identity, multivariate Normal, multivariate log-Normal.

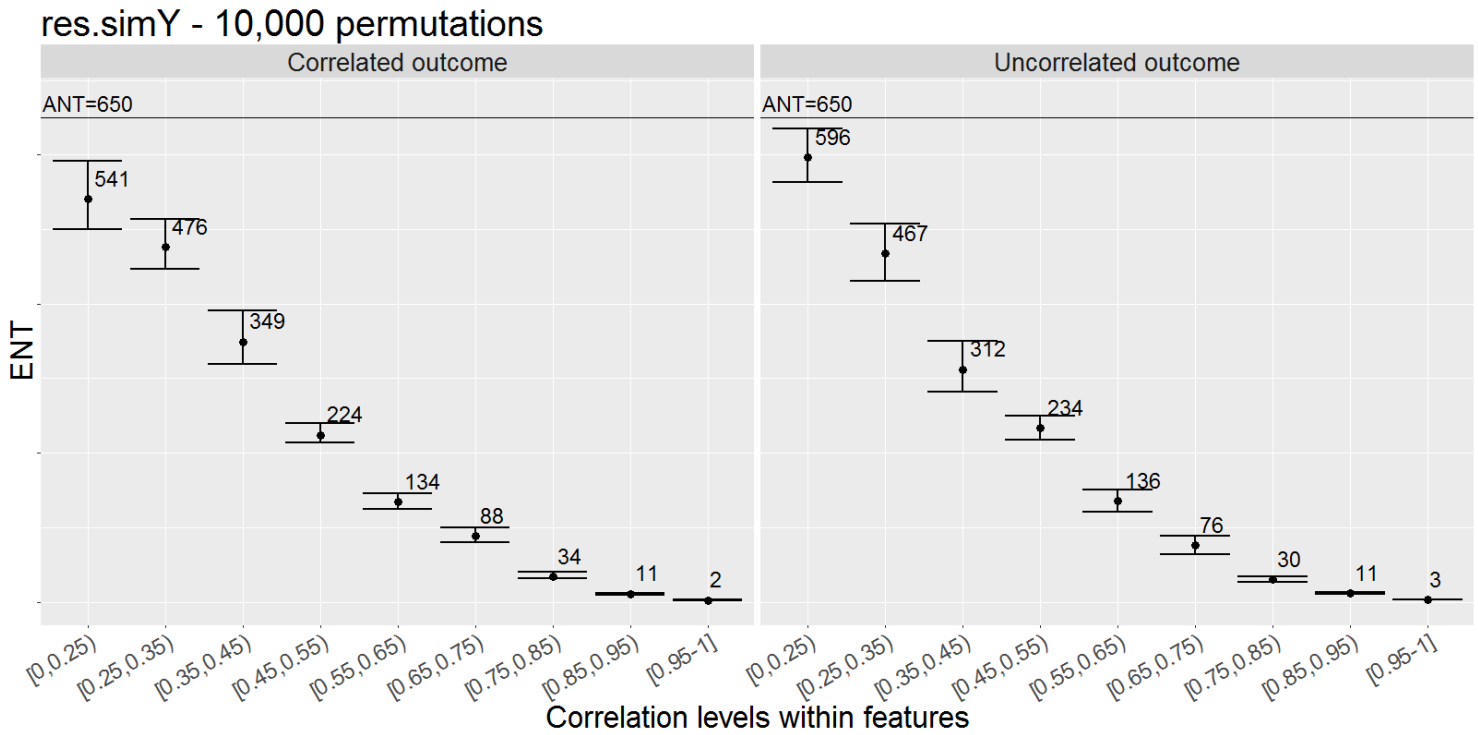


Figure 5: ENTs from the estimation of uncorrelated (dx) and correlated (sx) outcome across correlated features.

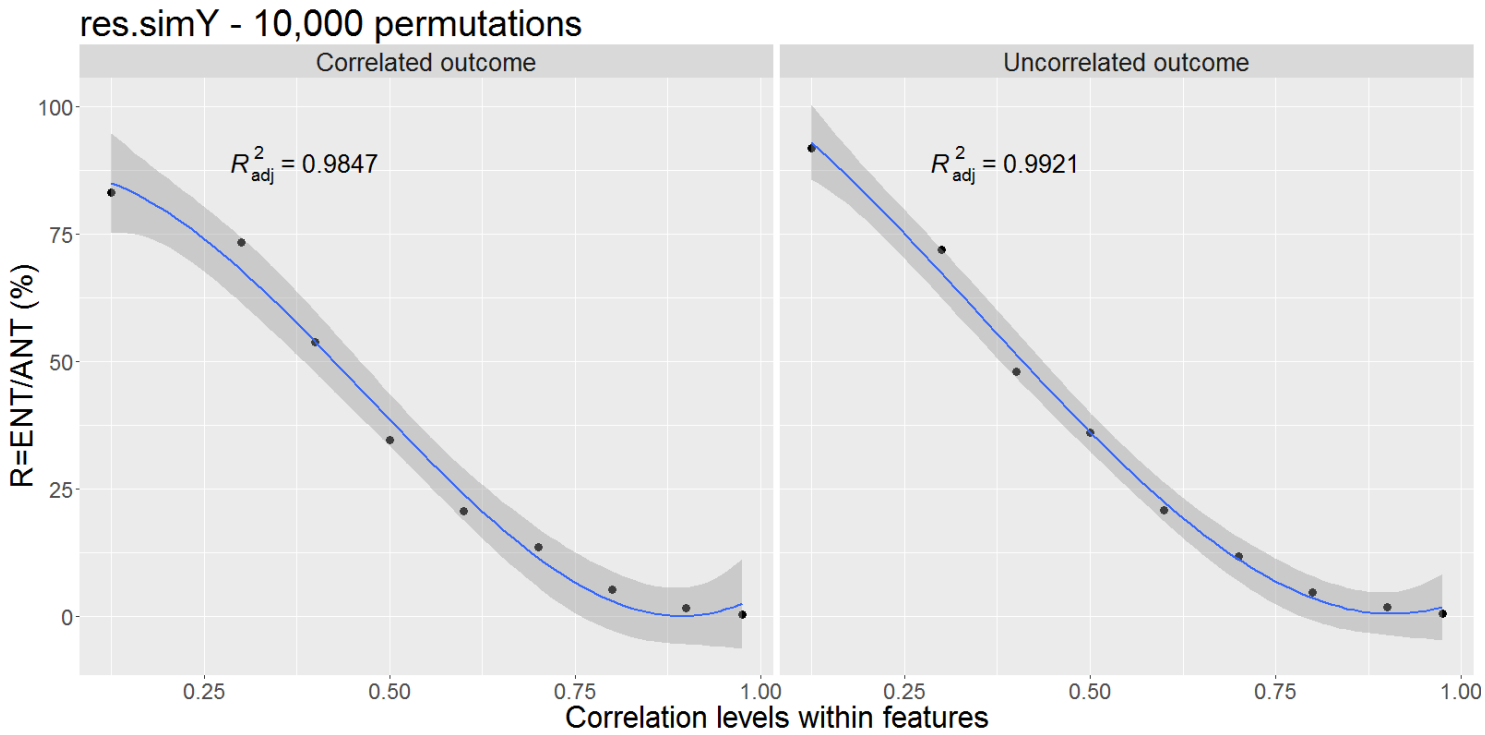
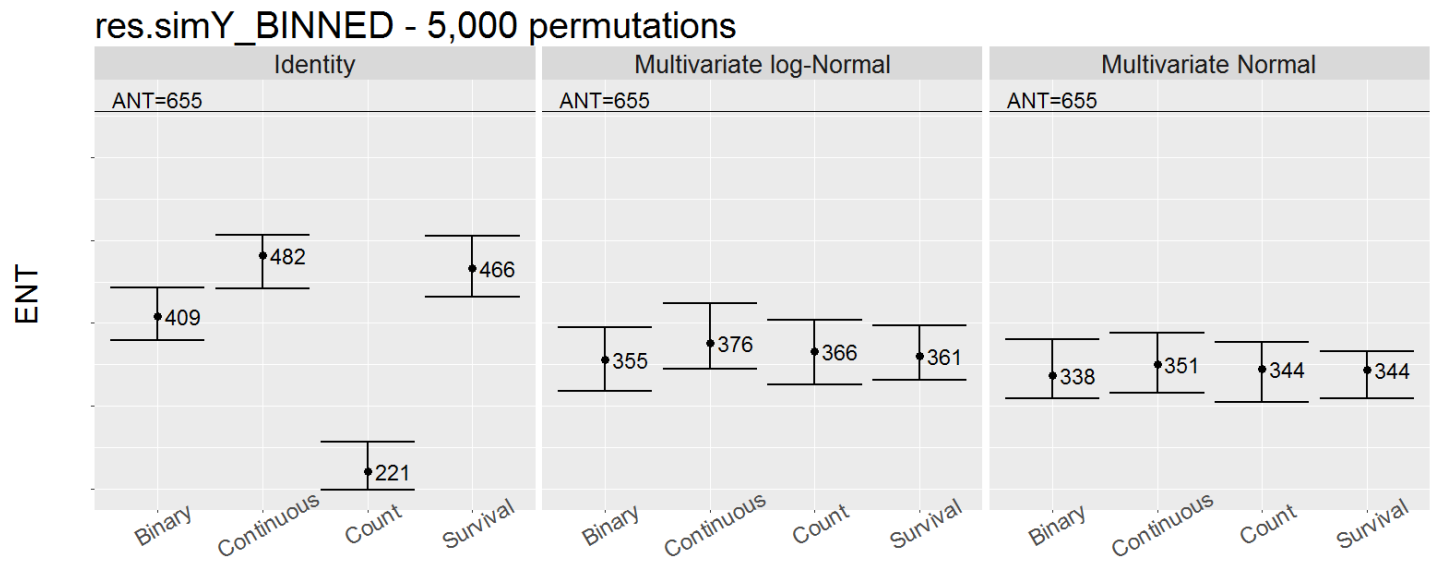
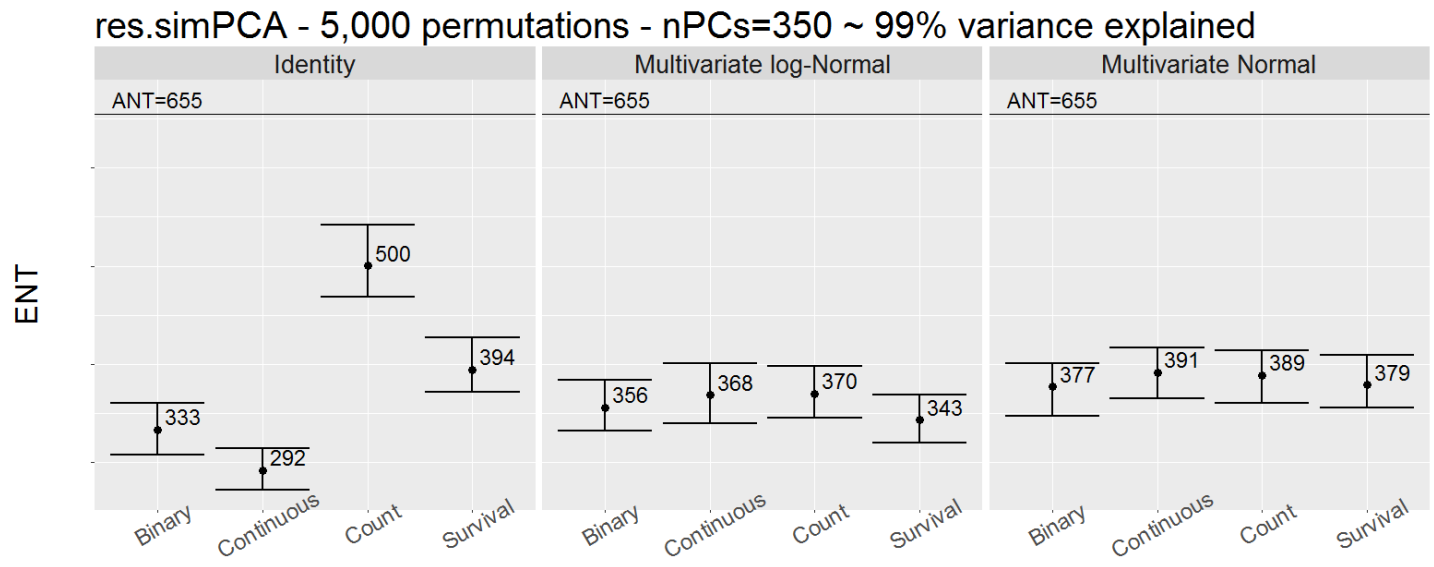


Figure 6: R ratio (%) from the estimation of uncorrelated (dx) and correlated (sx) outcome across correlated features.



	Identity				Multivariate log-Normal				Multivariate Normal			
Type I (%)	5.11	5.16	4.71	4.97	5.01	5.02	4.71	4.92	5.21	5.04	4.69	4.81
Type II (%)	94.89	94.84	95.29	95.03	94.99	94.98	95.29	95.08	94.79	94.96	95.31	95.19

Figure 7: BINNED data: ENT, type I and type II error estimation from the permutation procedure for different type of simulated outcome measure: continuous, discrete-binary, discrete-count, survival.



	Identity				Multivariate log-Normal				Multivariate Normal			
Type I (%)	4.93	5.11	5.05	4.95	5.09	4.99	4.98	4.93	5.01	4.98	5.11	5.19
Type II (%)	95.07	94.89	94.95	95.05	94.91	95.01	95.02	95.07	94.99	95.02	94.89	94.81

Figure 8: PCA simulated features (M=655, n_t=1,500, PCs=350): ENT variation, type I and type II error estimation from the permutation procedure for different type of simulated outcome measure: continuous, discrete-binary, discrete-count, survival.