UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

An information theoretic approach for the evaluation of surrogates of protection: A simulation study

Emmanuel Bache Bache

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ariel ALONSO ABAD

2019
2020

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

An information theoretic approach for the evaluation of surrogates of protection: A simulation study

Emmanuel Bache Bache

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ariel ALONSO ABAD

Contents

List of Figures

List of Tables

Acknowledgement

I wish thank the Lord Almighty for the gift of good health and the strength throughout my studies in Hasselt.

Special thanks to my supervisor Professor dr. Ariel Alonso Abad of the University of Hasselt and University of Leuven for the guidance and his patience with me throughout this thesis process. His detail review and precise input helped me to understand and conceptualize this project. I also want to thank the University of Hasselt, Department of Statistics for the substantial knowledge bank they have imparted on me for the past two years.

Furthermore, I am indebted to my partner Sophie, my daughter Ella and our youngest addition to the family Benjamin for the psychological and moral support and relentless motivation to go through my studies in Hasselt.

Abstract

The Individual Causal Association (ICA) metric of surrogacy based on the information theoretic approach (ITA), has been established as an efficient and reliable metric for the evaluation of a binary outcome as a putative surrogate marker for a true binary endpoint for chronic conditions in literature. In an attempt to apply this methodology in the context of evaluating surrogates of protection, this present simulation based study seeks to explore the behaviour of the ICA metric in evaluating surrogacy in a single-trial setting (STS) under different settings for continuous surrogate and true endpoint and dichotomization. The methodology is illustrated using the fine-tuned R Surrogate package, to generate datasets, analyse and extract results. The ICA was affected by the choice of surrogate, choice of endpoint dichotomization, underlying assumptions and measurement error. Since dichotomizing continuous endpoint and/or surrogate is common place in the context of evaluating surrogates of protection, when applying the ICA on real world data, the choice of cut-point, a sensitivity based analysis, field knowledge based assumptions and the presentation of results with more than one cut-point may be considered.

Keywords: Surrogate of Protection, Individual Causal Association, Single-Trial Setting, Information Theoretic Approach

1 Introduction

Vaccines are administered primarily to protect against disease, though sometimes protection against infection and even infectiousness is of interest (WHO, 2013). So far, vaccines remain the most cost effective method in preventing infectious diseases, amid a lengthy and costly clinical development process prior to regulatory approval. Traditionally, the licensure of vaccines often require proof of protective efficacy against clinical endpoint (protection of disease) in large phase III clinical trials, but recently with the Food and Drug Administration (FDA) "Accelerated Approval" (Katz, 2004) and the European Medicines Agency's (EMA) "PRIME scheme" for priority medicines (Detela and Lodge, 2019), permits the accelerated approval of certain treatments under specific conditions based on un-validated surrogates if "reasonably likely" to predict the clinical endpoint (Ensor et al., 2016). Establishing efficacy base on vaccine induced immune responses (immunogenicity) without measuring the clinical endpoint has significant advantages for most vaccines against pathogens causing outbreaks like *Ebolavirus* (Ebola virus Disease-EVD) Medaglini et al. (2018) , bacterium *Vibrio cholerae* (cholera) and human *Influenza viruses* (seasonal flu) (Auckland et al., 2006) or chronic infections with long incubation periods like *Human Immunodefieciency Virus* (HIV/AIDS) and *Mycobacterium tuberculosis* (TB) (Ensor et al., 2016). Generally, in outbreak situations, the goal is to curb the spread of the disease, the conduct of proper vaccine trials and the evaluations of the clinical endpoint of disease in such settings is very challenging. The development of immunological or biological endpoints that are surrogates replacing the true endpoint will accelerate vaccine evaluation and quickly brings effective vaccines for use in target populations (Buyse et al., 2016, WHO, 2013).

1.1 Endpoints in Vaccine Trials

The choice of the endpoints is one of the most critical aspects in clinical studies. Measuring clinically significant disease endpoints is paramount for vaccine research. Endpoints in vaccine clinical trials have been extensively described in (Hudgens et al., 2004), these endpoints may be continuous, binary, ordinal, or time-to-event measurements and do vary per phase of the study, vaccine type, and pathogen studied. During the preclinical stages of vaccine development, investigations are performed in animals (*invivo*) and in laboratories (*invitro*)

to assess safety and immunogenicity with focus on qualitative than quantitative outcomes. At this early stage of vaccine development, the concept of surrogate endpoints may come into play when ranking animals models corresponding to humans, or when assessing quantitative outcomes for lot-to-lot variability or stability and validation of immunogencity assays (Ensor et al., 2016). In phase I trials, evaluating the safety of first in humans use of a vaccine, immunogenicity outcomes may also be evaluated which is critical in underpinning the mechanism of potential correlates or surrogates of protection. Knowledge gathered from phase I trials do inform the design of Phase II trials. Phase II trials will be designed to further characterise the safety, immunogenicity and sometimes powered with the overall objective of finding the most promising dose or optimal dosing schedule to proceed with in phase III studies. Estimating vaccine effect is done by comparing the proportions of vaccinees with a predefined threshold antibody responses to the target pathogen believed to correlate with protection (dichotomization) or comparison of geometric mean titres of these antibodies per group. Recently, with the introduction of Controlled Human Infection Model(CHIM) challenge studies, the efficacy of a vaccine can be evaluated in more controlled settings with endpoints for infection or disease as in testing malaria (McCall et al., 2018), cholera (Cohen et al., 2002), hookworm (Diemert et al., 2018), influenza (Sherman et al., 2019) and typhoid fever (Jin et al., 2017) vaccines. In addition, where antibody responses have been established to correlate with protection and are valid surrogates for infection or disease, phase II trials serve as a useful portal for the introduction of licensed vaccines in new populations, and/or for the evaluation of vaccine combinations. Findings from such studies will orientate phase III trial designs with main objective to estimate the efficacy of a vaccine in preventing disease or infection in a target population. The endpoint of choice depends primarily on the vaccine and disease. For chronic diseases with long incubation periods, the endpoint of clinical disease may not be feasible thus infection endpoints may be considered. Furthermore, for some conditions rather than evaluating the overall effect of the vaccine in preventing disease, the vaccine effect on the clinical course of the infection may be evaluated. This may imply dichotomizing the infection endpoint into severity thresholds (mild, moderate, or severe) or into other informative classifications base on the clinical pathology of the disease under investigation (Hudgens et al., 2004).

1.2 Rationale for Surrogates of Protection

Nowadays, with the advancements in vaccinology, the use of endpoints base on immunogenicity permitted the fast track licensure of influenza(Van Els et al., 2014) and meningococcal vaccines (Auckland et al., 2006), without the classical demonstration of efficacy against the clinical endpoint. This is relevant in situations where phase III clinical trials may seem unrealistic or when new combinations of existing vaccines are being tested. Establishing a surrogate of protection (SoP) is important in the evaluating vaccines to be used as combinations, changes in immunization schedules of licensed vaccines and in alterations in manufacturing (Goldenthal et al., 2001). Regulatory agencies will fast tract the licensing of vaccines or approve use of combinations of vaccines when proven correlates or valid surrogates markers of protection exist. However, the use of these immune markers depends entirely on the relationship between the immunological markers and protective immunity (Plotkin, 2010, Thakur et al., 2012). A surrogate can be used to predict protection or clinical benefit when an absolute immune correlate is unknown. The protective threshold of vaccine induced antibodies against infections have been well established for several vaccines using sophisticated laboratory assays (Thakur et al., 2012). Generally, both surrogates and correlates are statistically associated with the true endpoint, but a surrogate also has causal link (McCall et al., 2018, WHO, 2013). Thus , the use of surrogate endpoints that predicts the effect of a vaccine on an unobserved true outcome may have enormous economic and ethical advantages in terms of reducing the duration and size of vaccine clinical trials (Ensor et al., 2016). The availability of validated and reliable surrogate endpoints which are comparatively easy, cheap and quick to measure, will not only offer logistical advantages, reduce design complexity, sample size, trial duration and potential cost in conducting the vaccine studies, it will definitely push more effective vaccines faster into the market (Ensor et al., 2016, McCall et al., 2018)

1.3 Validations of Surrogate Endpoints

Several statistical approaches have been proposed for validating surrogate endpoints with varying degree of merits (Ensor et al., 2016). Notably, besides the association between a potential surrogate endpoint and corresponding true clinical endpoint, for a surrogate endpoint to replace a true clinical endpoint, the effect of the vaccine on the surrogate should reliably predict the effect on the clinical endpoint (Buyse et al., 2016, WHO, 2013). Correlation

between the surrogate and the true endpoint without considering the vaccine effect is not a reliable indicator of the validity of the surrogate to replace the true endpoint. When the surrogate doesn't fully lie in the causal pathway and picks up only a proportion of the vaccine effect on the true endpoint, the relationship between the surrogate and the true endpoint may not exist, after the vaccine is administered. To enable the approval of vaccines with reliable surrogates, sound methodology to establish surrogacy is required.

Landmark formative proposals for assessing surrogacy in clinical trials has been extensively studied (Ensor et al., 2016). Statistical evaluation of surrogacy in the single trial setting was first recognized by (Prentice, 1989). Later (Freedman et al., 1992) proposed the proportion of treatment explained (PTE) to quantify the level of surrogacy. (Buyse and Molenberghs, 1998) proposed the relative effect (RE) as a measure of surrogacy derived from the Prentice criteria in the single trial setting . Subsequently, the meta-analytical approach (Buyse et al., 2000), and various model fitting extensions of the meta-analytical approach accommodating time-to-event (Burzykowski et al., 2001), binary and ordinal Molenberghs et al. (2001), and repeatedly measured (Alonso et al., 2016a) outcomes were proposed for the evaluation of surrogacy in multi-trial settings. Further extensions of the meta-analytical approach let to the development of surrogate threshold effect (STE) (Burzykowski and Buyse, 2006), likelihood reduction factor (LRF) (Alonso et al., 2004), and information theoretic approach (ITA) (Alonso and Molenberghs, 2007). Adaptations of the ITA were developed to incorporate time-to-event (Alonso and Molenberghs, 2008), binary (Pryseley et al., 2007) and repeated measures endpoints (Alonso et al., 2006). Alternatively, causal evaluation approaches such as principal stratification (Frangakis and Rubin, 2002) and the measurement of the direct and indirect effect of treatment on surrogate and true endpoint have also been widely used (Ensor et al., 2016).

The focus of this project is to use the ITA which primarily considers the quantification of the level of uncertainty in a random variable to measure surrogate association. The ITA provides a unified framework in validating surrogacy with computational benefits (no joint models or latent variables) , intuitive interpretational capabilities across different types of outcomes and provides narrower confidence intervals (Alonso and Molenberghs, 2007).

Throughout this paper, we will refer to the following notations: T and S as random variables that denote the true and surrogate endpoints, respectively and Z an indicator for the treatment (vaccine) variable.

1.4 Objectives

Explore statistical frameworks for the validation of surrogates in vaccine trials.

1. Generate a continuous outcome and a continuous surrogate in a single trial setting
2. Dichotomize surrogate and true endpoint by given specific cut-off
3. Estimate the Pearson (ρ_Δ) correlation measure of association for case of continuous-continuous surrogate and true endpoint
4. Estimate and describe the behavior of the individual causal association (ICA, R^2_H), which captures the association between the individual causal effects of the treatment on surrogate and true endpoint using the information-theoretic approach
5. Compare the information loss in estimating Pearson (ρ_Δ)correlation measure of association and the individual causal association (ICA, R^2_H) in the case of continuous -continuous and dichotomization respectively
6. Investigate the behavior Pearson (ρ_Δ)correlation measure of association and the individual causal association (ICA, R^2_H) in the presence of measurement error

In this paper, we discuss the statistical validation for continuous surrogate and a continuous endpoint using the individual causal association approach. Also we examine the behavior and information loss under various data driven dichotomization schemes, in an attempt to translate observations when analysing real world data on validating surrogate of protection.

2 Methodology

2.1 The Information Theoretic Approach

2.1.1 Information Theoretic Measure of Association

The concepts underpinning the information theory has been discussed at length in (Alonso and Molenberghs, 2007, Alonso et al., 2016b). The concept of entropy is focal to the information theory and quantifies aptly the randomness or uncertainty associated with a random variable. Base on the entropy power concept, for continuous and normally distributed variables, the information theoretic measure of association is given by equation 1 (Alonso and Molenberghs, 2007, Alonso et al., 2017, Cover et al., 1991).

$$R_h^2 = \frac{EP(T) - EP(T|S)}{EP(T)} \tag{1}$$

Where EP(T) is the entropy power of T and EP(T|S) is the entropy power of T given S , R_h^2 satisfies: (i) $0 \leq R_h^2 \leq 1$, (ii) $R_h^2 = 0$ if T and S are independent, (iii) R_h^2 is symmetric in (S,T), (iv) R_h^2 is invariant by bijective transformations of S and T and (v) R_h^2 approaches 1 for continuous models (Alonso and Molenberghs, 2007, Ensor et al., 2016).

2.2 Causal Inference Framework

To describe the causal inference framework, we will assume the following: the treatment indicator Z has values (Z=0/1) indicating the control and active vaccine, the standard stable unit treatment assumption (SUTVA)(Rubin, 1980). Following the Rubin's framework for causal inference it is assumed that each subject has four potential outcomes Y= $(T_0, T_1, S_0, S_1)'$. T_0 and T_1 represent potential outcomes of the true endpoint (T) while S_0 and S_1 represent potential outcomes of surrogate (S) endpoint for subjects who received the control or active treatment respectively. Usually, only one T_0 or T_1 is observed in practice. Under the SUTVA, if we denote T_j as the observed outcome of subject j then $T_j = Z_j T_{1j} + (1 - Z_j)T_{0j}$. Thus, based on the vector of potential outcomes, $(T_{0j}T_{1j})$, the individual causal effect of the treatment on the subject is estimated as $\Delta T = T_1 - T_0$, and the expected causal effect on Z in the target population is $\beta = E(T_{1j} - T_{0j})$. For clarity sub-index j will be omitted in this expression such that ΔT_j is denoted only as ΔT . Also,the individual causal effect of the treatment on the surrogate for a subject is estimated as $\Delta S = S_1 - S_0$ and the expected causal effect on Z in the target population is $\alpha = E(S_{1j} - S_{0j})$

The distribution of the vector of potential outcomes play an important role in surrogacy evaluation (Alonso et al., 2015, Alonso Abad et al., 2016).

2.3 The Individual Casual Association (ICA)

The Individual Causal Association (ICA), defined as the association between the individual causal treatment effects ΔT and ΔS has been extensively used as a measure of surrogacy (Alonso et al., 2017). For the case of a continuous, normally distributed surrogate and true endpoint, the ICA is depicted by $\rho_{\Delta} = corr(\Delta T, \Delta S)$ (Alonso et al., 2017, 2015). For a binary surrogate and true endpoint, within the ITA, the measure of surrogacy is given by:

$$R^2_H(\Delta T, \Delta S) = \frac{I(\Delta T, \Delta S)}{\min[H(\Delta T), H(\Delta S)]} \tag{2}$$

Where the numerator is termed the mutual information is connoted as:

$$I(\Delta T, \Delta S) = \sum_{i,j=-1}^{1} \pi^{\Delta}_{i,j} \left(\frac{\pi^{\Delta}_{i,j}}{\pi^{\Delta T}_{i} \pi^{\Delta S}_{j}} \right)$$

The denominator in equation 2 represents the minimum entropies of the individual casual treatment effects, which are:

$$H(\Delta T) = \sum_{i=-1}^{1} \pi^{\Delta T}_{i} log(\pi^{\Delta T}_{i}),$$

$$H(\Delta S) = \sum_{j=-1}^{1} \pi^{\Delta S}_{j} log(\pi^{\Delta S}_{j})$$

Where $R^2_H(\Delta T, \Delta S)$ is invariant under one-to-one transformations and lies in the unit interval, $R^2_H(\Delta T, \Delta S) = 0$ when ΔT and ΔS are independent and one when there is a nontrivial transformation ψ so that the P[ΔT=$\psi\Delta S$]=1.

3 Simulation Studies and Analysis

3.1 Individual Causal Association (ICA) in the Single Trial Setting

3.1.1 Continuous Surrogate and True Endpoint

Given the vector of potential outcomes $Y_j = (T_{0j}, T_{1j}, S_{0j}, S_{1j})^{'}$ has a distribution $N(\mu, \Sigma)$, where $\mu = (\mu_{T_0}, \mu_{T_1}, \mu_{S_0}, \mu_{S_1})^{'}$ with

$$\Sigma = \left(\begin{array}{cc|cc} \sigma_{T_0T_0} & \sigma_{T_0T1} & \sigma_{T_0S_0} & \sigma_{T_0S_1} \\ \sigma_{T_0T_1} & \sigma_{T_1T1} & \sigma_{T_1S_0} & \sigma_{T_1S_1} \\ \hline \sigma_{T_0S_0} & \sigma_{T_1S_0} & \sigma_{S_0S_0} & \sigma_{S_0S_1} \\ \sigma_{T_0S_1} & \sigma_{T_1S1} & \sigma_{S_0S_1} & \sigma_{S_0S_1} \end{array}\right)$$

This implies

$$\Delta = AY = \left(\begin{array}{c} \Delta T_1 - \Delta T_0 \\ \Delta S_1 - \Delta S_0 \end{array}\right) \sim N(\mu, \Sigma), \tag{3}$$

Where $\Sigma_\Delta = A\Sigma A^{'}, \mu_\Delta = (\beta, \alpha)^{'}$ and A is the contrast matrix Van der Elst et al. (2016). The measure of surrogacy in the STS using the ICA $\rho_\Delta = corr(\Delta_T, \Delta_S)$ is given by

$$\rho_\Delta = \frac{\sqrt{\sigma_{T_0T_0}}\sigma_{S_0S_0}\rho_{T_0S_0} + \sqrt{\sigma_{T_1T_1}}\sigma_{S_1S_1}\rho_{T_1S_1} - \sqrt{\sigma_{T_1T_1}}\sigma_{S_0S_0}\rho_{T_1S_0} - \sqrt{\sigma_{T_0T_0}}\sigma_{S_1S_1}\rho_{T_0S_1}}{\sqrt{(\sigma_{T_0T_0} + \sigma_{T_1T1} - 2\sqrt{\sigma_{T_0T_0}\sigma_{T_1T1}\rho_{T_0T1}})(\sigma_{S_0S_0} + \sigma_{S_1S1} - 2\sqrt{\sigma_{S_0S_0}\sigma_{S_1S1}\rho_{S_0S1}})}} \tag{4}$$

Assuming homoscedascity (the variability in the true and surrogate endpoint is constant, $\sigma_{T_0T_0} = \sigma_{T_1T1} = \sigma_T$ and $\sigma_{S_0S_0} = \sigma_{S_1S1} = \sigma_S$, the expression in 4 becomes:

$$\rho_\Delta = \frac{\rho_{S_0T_0} + \rho_{S_1T_1} - \rho_{S_0T_1} - \rho_{S_1T_0}}{2\sqrt{(1 - \rho_{T_0T1})(1 - \rho_{S_0S1})}} \tag{5}$$

The implementation of ICA in practice is challenging because $\rho_{S_0T_1}, \rho_{S_1T_0}, \rho_{T_0T1}$ and ρ_{S_0S1} in equation 5 is not estimable making ρ_Δ not identifiable. In this paper, we will use the simulation-based sensitivity analysis by which ρ_Δ is estimated over a set of plausible values of the unidentified correlations mentioned above. With this approach, we begin by setting a grid G=$(g_1, g_2, ..., g_k)$ with grid values (-1, -0.90,...1) for sensitivity-based identification of the unidentified correlations between the potential outcomes. Then, generate several Σ matrices which fixes the identifiable correlations $\rho_{S_0T_0}, \rho_{S_1T_1}$ at their estimated values and use combinations obtained from the specified grid for the unidentified correlations $\rho_{S_0T_1}, \rho_{S_1T_0}, \rho_{T_0T1}$

and ρ_{S_0S1}. Among the several Σ matrices, only valid correlations (positive definites) are used to obtain a vector of ρ_Δ(Alonso et al., 2015, Van der Elst et al., 2016). We used the Single.Trial.RE.AA() to obtain unidentifiable correlations and the ICA.ContCont() to estimate ρ_Δ in a single trial setting.

3.1.2 Generating data for Single-trial Setting

We simulated theoretical datasets for a continuous surrogate and true endpoint for a single trial setting by keeping Y_j with $\mu = (0, 0, 0, 0)'$, and varying the adjusted association(γ) = $\{0.0, 0.4, 0.8\}$ and for trial sizes N=(50, 200, 2000). Varying γ yielded different values of the identifiable correlations $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$ shown in Table 2. The γ=corr$(T, S|Z)$ can be defined as a metric of surrogacy in a single trial setting that quantifies the accuracy by which T can be predicted based on S taking into account treatment. This was done using the Sim.Data.STS () function in the Surrogate package in R (Van der Elst et al., 2020).

3.1.3 Dichotomizing the Surrogate and True endpoint

From datasets generated for the case of continuous surrogate and true endpoints, we will dichotomize the surrogate and true endpoint using various cut-off and estimate the R^2_H measure of surrogacy. Using the surrogate package, we applied the MarginalProbs() function to obtain descriptive statistics and the identifiable marginal probabilities. The ICA measure of surrogacy R^2_H is obtained using the function ICA.BinBin(). In this framework, the ICA.BinBin() functions estimates R^2_H as a function of parameters π characterizing the distribution of potential outcomes Y, implementing a two step Monte-Carlo algorithm to sample π vectors from a parametric space compatible with the data (Γ_D)(Alonso et al., 2017). We performed the analysis considering a G=(0.0, 0.01,...0.99), M=10,000 for the number of runs on the specified G and accounting for all possible monotonicty scenarios. Under the monotonicity assumption $P(T_0 \leq T_1) = P(S_0 \leq S_1) = 1$ and $\pi^T_{10} = \pi^S_{10} = 0$ (Alonso et al., 2016b)
The following data driven cut-off were used to dichotomize the continuous true and surrogate endpoint:

- Mean split
- Median split: dividing the overall trial population into two halves

- 75th percentile split

These data driven cut-points were chosen due to the high reflex of usage in clinical literature and the theoretical basis of this simulated exercise.

4 Results

We generated 9 theoretical datasets with common means (0,0,0,0), varying sample size N (50,200, 2000) and adjusted association (γ= 0.0, 0.4, 0.8). This yielded nine datasets with continuous surrogate and true endpoints with descriptive statistics represented in Table 1. To investigate effect of random contamination, we introduced random measurement error on the surrogate endpoint by inflating a batch of entries for datasets generated γ= 0.8 and sample sizes N(50, 200, 2000).

Table 1: Summary characteristics of generated datasets

Dataset	Endpoint	Sample Size (N)	Adjusted Association (γ)	Mean(SD)	Median	Percentiles [25%,75%]
1A	Surrogate	2000	0.8	0.004(0.99)	-0.006	[-0.65,0.66]
	True	2000	0.8	0.004(1.00)	-0.013	[-0.68,0.65]
1B	Surrogate	2000	0	-0.025(1.01)	-0.023	[-0.68,0.63]
	True	2000	0	-0.004(0.98)	0.016	[-0.66,0.66]
1C	Surrogate	2000	0.4	-0.009(1.01)	-0.002	[-0.69,0.69]
	True	2000	0.4	-0.002(0.99)	0.002	[-0.69,0.71]
2A	Surrogate	50	0.8	-0.15(1.1)	-0.13	[-0.55,0.41]
	True	50	0.8	-0.11(0.93)	-0.13	[-0.81,0.59]
2B	Surrogate	50	0	-0.3(0.83)	-0.29	[-0.84,0.15]
	True	50	0	0.27(0.93)	0.45	[-0.27,0.91]
2C	Surrogate	50	0.4	0.13(1.16)	0.2	[-0.51,0.77]
	True	50	0.4	0.07(1.09)	-0.09	[-0.65,0.70]
3A	Surrogate	200	0.8	-0.038(1.02)	-0.015	[-0.78,0.69]
	True	200	0.8	-0.015(1.00)	-0.02	[-0.73,0.73]
3B	Surrogate	200	0	-0.015(1.00)	0.002	[-0.63,0.61]
	True	200	0	-0.015(1.00)	0.085	[-0.64,0.61]
3C	Surrogate	200	0.4	0.05(1.14)	0.15	[-0.73,0.80]
	True	200	0.4	0.07(0.93)	0.11	[-0.54,0.62]

Adjusted association (γ)

N: Sample size, SD: Standard deviation

Table 2: Data derived identifiable correlations for $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$

Scenario	N	γ	$\hat{\rho}_{S_0T_0}$(95% CI:LL,UL)	$\hat{\rho}_{S_1T_1}$(95%CI:LL,UL)	**$\hat{\gamma}$(95%CI:LL,UL)
1A	2000	0.8	0.78(0.76,0.80)	0.80(0.78,0.81)	0.80(0.77,0.81)
1B	2000	0	0.02(0.0,0.06)	-0.04(-0.002,0)	-0.01(-0.06,-0.03)
1C	2000	0.4	0.42(0.38,0.45)	0.44(0.41,0.48)	0.43(0.40,0.47)
2A	50	0.8	0.65(0.45,0.79)	0.80(0.68,0.88)	0.74(0.59,0.85)
2B	50	0	-0.09(0.0,0.20)	0.27(0.0,0.50)	0.07(-0.21,0.31)
2C	50	0.4	0.38(0.11,0.60)	0.51(0.27,0.69)	0.43(0.15,0.69)
3A	200	0.8	0.76(0.76,0.81)	0.77(0.70,0.82)	0.76(0.70,0.82)
3B	200	0	0.03(0.0, 0.17)	-0.02(0,0.11)	0.0(-0.13,0.15)
3C	200	0.4	0.40(0.27,0.51)	0.57(0.47,0.66)	0.48(0.38,0.58)
4A	2000	0.8	0.65(0.62,0.67)	0.63(0.61,0.66)	0.64(0.62,0.66)
4B	200	0.8	0.63(0.54,0.70)	0.64(0.55,0.71)	0.63(0.57,0.70)
4C	50	0.8	0.54(0.30,0.71)	0.64(0.45,0.78)	0.60(0.47,0.77)

N: Sample size, γ: User defined adjusted association

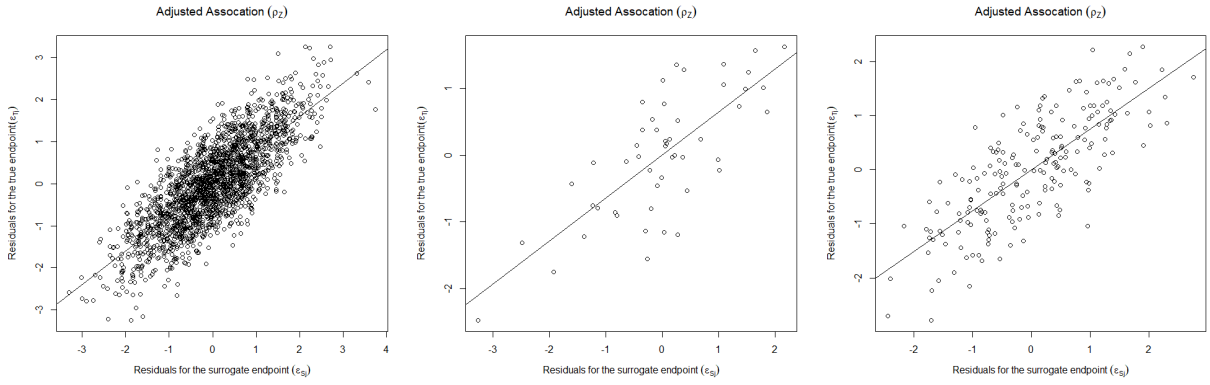
(95%CI:LL,UL): 95% confidence interval lower limit (LL) and upper limit (UL)

$\hat{\rho}_{S_1T_1}$:Identifiable correlation between S and T in experimental group

$\hat{\rho}_{S_0T_0}$:Identifiable correlation between S and T in control group

$\hat{\gamma}$:Pooled adjusted association from both groups

**Bootstrap-based confidence interval



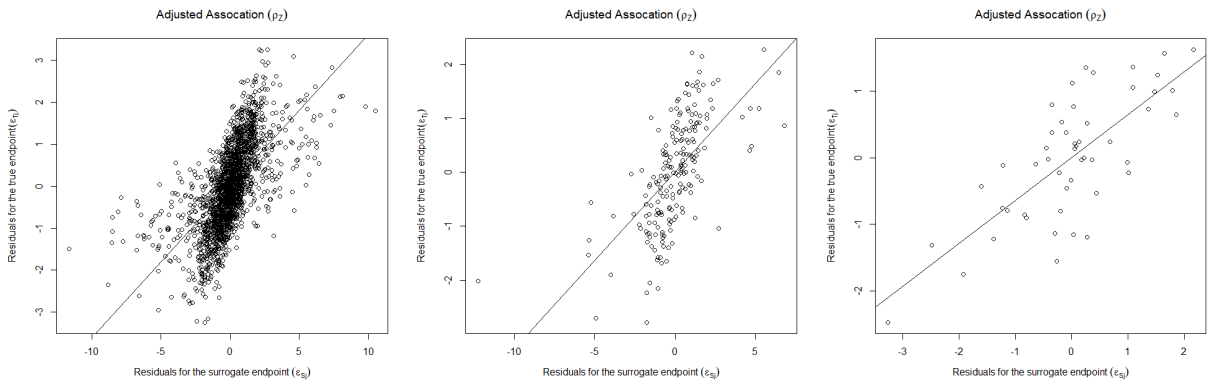
(a) Plot of γ scenario 1A

(b) Plot of γ scenario 2A

(c) Plot of γ scenario 3A

Figure 1: Adjusted association for scenario 1A,2A & 3A adjusted for Z

As indicated in Table 3 to 5, the choice of cut-point yielded different values for odds ratio-$\theta_{T_1S_1}$ and $\theta_{T_0S_0}$ in the experimental and control arms. In some situations, strong associations were observed between S and T for the experimental than the control group and vice versa. For example in Table 3, the association between S and T was stronger in the control group $\theta_{T_0S_0}$ =16.55(95%CI:11.69-23.43) than in the experimental group, we observed



(a) Plot of γ scenario 4A

(b) Plot of γ scenario 4B

(c) Plot of γ scenario 4C

Figure 2: Adjusted association for scenario 4A,4B & 4C with measurement error adjusted for Z

$\theta_{T_1S_1}$=13.84(95%CI:9.82-19.51) for percentile split with N=2000, and γ=0.4. Given sample size N=50 and γ=0.8 with median split, we observed a stronger association between S and T in the experimental group $\theta_{T_1S_1}$=7.7(95%CI:1.16-51.17) than in the control group $\theta_{T_0S_0}$ =3.33(95%CI:0.60-18.54). As depicted in Table 2, Figure 1 shows the graphical representation of the data derived adjusted association after adjusting for the treatment effect for scenario 1A, 2A, and 3A without measurement error. We observed a slight reduction in the adjusted association after adjusting for treatment effect(Z) for scenarios 4A, 4B, and 4C with measurement error highlighted in Figure 2.

Table 3: Data driven dichotomization for T and S endpoints with sample size N:2000

Parameter	Cut-off		OR Experiment (95%CI:)	OR Control (95%CI:)
	True Endpoint	Surrogate Endpoint	$\theta_{T_1S_1}$	$\theta_{T_0S_0}$
N=2000, γ=0.8 , dataset 1A				
Mean	0.004	0.004	12.26(9.10-16.52)	12.71(9.40-17.16)
Median	-0.013	-0.025	12.28(7.50-13.4)	12.51 (11.46-25.25)
75th percentile	0.65	0.66	13.84(9.82-19.51)	16.55(11.69-23.43)
N=2000, γ=0.0 , dataset 1B				
Mean	-0.004	-0.025	1.04(0.81-1.33)	1.04(0.81-1.33)
Median	0.016	-0.023	1.1(0.84-1.38)	1.0(0.78-1.28)
75th percentile	0.66	0.63	0.71(0.50-1.01)	1.06(0.76-1.47)
N=2000, γ= 0.4 , dataset 1C				
Mean	-0.002	-0.009	3.17(2.45-4.11)	3.20(2.47-4.14)
Median	0.002	-0.002	3.32(2.56-4.31)	3.03(2.34-3.91)
75th percentile	0.71	0.69	3.39(2.48-4.63)	3.60(2.64-4.90)

γ: Adjusted association, N:Sample size, CI: Wald Confidence Intervals, OR: Odds Ratio

$\theta_{T_1S_1}$: Odds ratio-clinically meaningful change of T on S in experimental group

$\theta_{T_0S_0}$: Odds ratio -clinically meaningful change of T on S in control group

Table 4: Data driven dichotomization for T and S endpoints with sample size N:50

Parameter	Cut-off		OR Experiment (95%CI:)	OR Control (95%CI:)
	True Endpoint	Surrogate Endpoint	$\theta_{T_1S_1}$	$\theta_{T_0S_0}$
N=50, γ=0.8 , dataset 2A				
Mean	-0.11	-0.15	4.4(0.77-25.15)	3.02(2.34-3.91)
Median	-0.013	-0.013	7.7(1.16-51.17)	3.33 (0.60-18.54)
75th percentile	0.59	0.41	5.50(0.84-36.2)	3.33(0.60-18.54)
N=50, γ=0.0 , dataset 2B				
Mean	0.27	-0.3	0.63(0.13-3.10)	1.2(0.24-5.84)
Median	0.45	-0.29	0.63(0.13-3.10)	0.88(0.18-4.34)
75th percentile	0.91	0.15	0.63(0.09-4.17)	–
N=50, γ= 0.4 , dataset 2C				
Mean	0.07	0.13	9.78(1.55-61.65)	4.67(0.83-26.24)
Median	-0.09	0.2	16.00(2.16-118.3)	3.00(0.5-15.77)
75th percentile	0.70	0.77	18.75(1.68-209.5)	3.2(0.35-28.94)

γ: Adjusted association, N:Sample size, CI: Wald Confidence Intervals, OR: Odds Ratio

$\theta_{T_1S_1}$: Odds ratio-clinically meaningful change of T on S in experimental group

$\theta_{T_0S_0}$: Odds ratio -clinically meaningful change of T on S in control group

Table 5: Data driven dichotomization for T and S endpoints with sample size N:200

Parameter	Cut-off		OR Experiment (95%CI:)	OR Control (95%CI:)
	True Endpoint	Surrogate Endpoint	$\theta_{T_1S_1}$	$\theta_{T_0S_0}$
N=200, γ=0.8 , dataset 3A				
Mean	-0.015	-0.038	15.75(5.58-44.49)	14.4(5.29-39.2)
Median	-0.02	-0.015	18.94(6.61-54.26)	14.4(5.29-39.2)
75th percentile	0.73	0.69	29.75(8.64-102.5)	12.36(4.3-35.55)
N=200, γ=0.0 , dataset 3B				
Mean	-0.015	-0.015	0.65(0.29-1.43)	0.84(0.38-1.86)
Median	0.085	0.002	0.53(0.23-1.17)	1.01(0.46-2.27)
75th percentile	0.61	0.61	0.69(0.23-2.08)	1.78(0.67-4.70)
N=200, γ= 0.4, dataset 3C				
Mean	0.07	0.05	3.42(1.48-7.91)	6.80(2.79- 16.56)
Median	0.11	0.15	3.58(2.16-118.3)	5.21(2.20-13.35)
75th percentile	0.62	0.80	3.16(1.17-8.58)	3.37(1.30-8.78)

γ: Adjusted association, N:Sample size, CI: Wald Confidence Intervals, OR: Odds Ratio

$\theta_{T_1S_1}$: Odds ratio-clinically meaningful change of T on S in experimental group

$\theta_{T_0S_0}$: Odds ratio -clinically meaningful change of T on S in control group

We introduced random measurement error for about 23% of the surrogate endpoint in datasets 1A, 2A and 3A, and used the same cut-off as the original dataset to mimic clinical practice (where a fix clinical data driven reference cut-off point is adopted for a given condition) to perform surrogacy evaluation. We observed a stronger association of S and T in the experimental group than in the control group for all split types given sample size 50 and 200. The association was slightly similar given sample size N=2000 as indicated in Table 6.

Table 6: Data driven dichotomization with measurement error

Parameter	#Cut-off		OR Experiment (95%CI:)	OR Control (95%CI:)
	True Endpoint	Surrogate Endpoint	$\theta_{T_1S_1}$	$\theta_{T_0S_0}$
Scenario 4A: N=2000, γ=0.8, dataset 1A				
Mean	0.004	0.004	12.26(9.1-16.52)	12.71(9.42-17.16)
Median	-0.013	-0.006	12.56(9.31-16.94)	12.56(9.31-16.95)
75th percentile	0.65	0.66	13.68(9.75-19.21)	14.74(10.49-20.72)
Scenario 4B: N=200,γ = 0.8, dataset 3A				
Mean	-0.015	-0.054	15.75(5.57-44.49)	14.40(5.29-39.2)
Median	-0.02	-0.015	18.94(6.61-54.26)	14.40(2.20-13.35)
75th percentile	0.73	0.69	17.00(5.32-54.3)	8.75(3.27-23.41)
Scenario 4C: N=50, γ=0.8 , dataset 2A				
Mean	-0.11	-0.15	7.2(0.60-18.54)	3.33(1.10-48.64)
Median	-0.13	-0.13	5.5(0.83-36.2)	3.33(0.60-18.54)
75th percentile	0.61	0.61	8.5(0.97-74.42)	30(2.58-348.8)

γ: Adjusted association, N:Sample size, CI: Wald Confidence Intervals, OR: Odds Ratio

$\theta_{T_1S_1}$: Odds ratio-clinically meaningful change of T on S in experimental group

$\theta_{T_0S_0}$: Odds ratio -clinically meaningful change of T on S in control group

#:Measurement error using same cut-off as original data

In scenario 1A highlighted in Table 7, ρ_Δ (mean=0.76, Sd=0.19, range=[-0.81,1.00] was very wide indicating that values assume for the unidentified correlations through the sensitivity analysis have high impact on the estimate, this indicates that validating the surrogate on this basis is highly sensitive to assumptions(Van der Elst et al., 2016). Also, larger estimates were obtained for ρ^2_Δ with mean=0.62, median=64, and range [0.00-1.00] after considering 9711 valid results compatible with the data. In this scenario, the largest estimates upon dichotomization were observed with mean split,R^2_H1 mean= 0.33, median=32 and range [0.03-0.76] under the assumption of no monotonicity reflected by the data. The range for median and percentile splits R^2_H2 range [0.03-0.64] and R^2_H3 range [0.00-0.65] respectively were shorter compared to mean dichotomization under the same assumption. After considering 32593 valid results for mean split, 21653 for median split and 169920 for percentile split compatible with the data in scenario 1A, the highest R^2_H observed was 0.76 as oppose to values close to 1 for ρ^2_Δ in 9711 valid results.

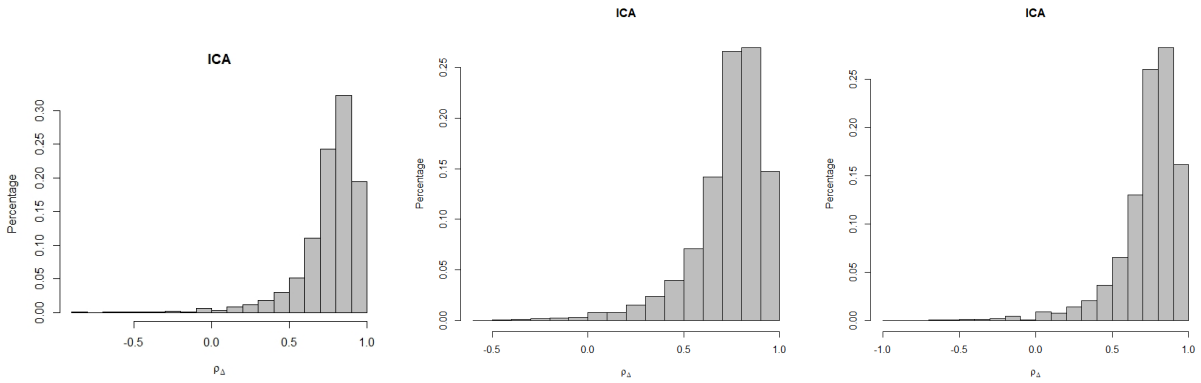
Table 7: Scenario 1A-Descriptive statistics of ρ_Δ and R^2_H

Parameter	*No. Positive Definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 1A: N=2000 and adjusted association (γ)=0.80							
ρ_Δ	*9711	NA	0.76	0.8	0.83	0.19	-0.81-1.00
ρ^2_Δ	*9711	NA	0.62	0.64	0.67	0.22	0.00-1.00
R^2_H1: No monotonicity	32593	Mean	0.33	0.32	0.32	0.06	0.03-0.76
R^2_H2: No monotonicity	21653	Median	0.32	0.32	0.30	0.06	0.03-0.64
R^2_H3: No monotonicity	169920	75th percentile	0.27	0.26	0.26	0.05	0.00-0.65
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							

In scenario 1B highlighted in Table 8, it was observed that R^2_H for mean and median dichotomization was mean=0.21 and 0.19 respectively under the assumption that monotonicity holds for true endpoint, were higher than mean ρ^2_Δ= 0.15 in the continuous outcome case. The range of R^2_H mean split [0.00-0.89] and median split [0.00-0.81] remained narrower than that of ρ^2_Δ [0.00-0.99]. This indicates that when all 32162 and 51283 valid results compatible with the data were considered,maximum values for R^2_H mean and median split were 0.89 and 0.81 respectively. For percentile split the maximum value was 0.57 after considering 147,932 valid results compatible with the data. This implies that depending on the dataset, surrogate and choice of dichotomization, monotonicity assumption may have an impact on ICA metric.

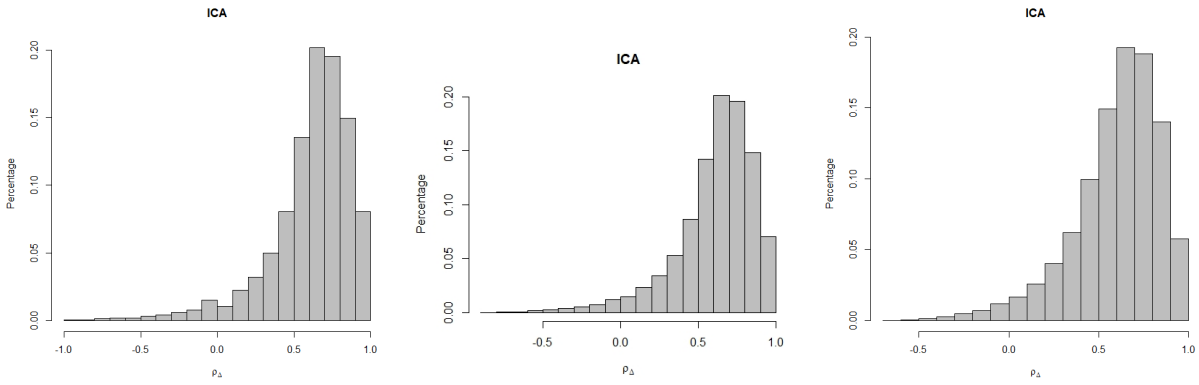
Table 8: Scenario 1B-Descriptive statistics of ρ_Δ and R^2_H

Parameter	*No. Positive definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 1B: N=2000 and adjusted association (γ)=0.00							
ρ_Δ	*65069	NA	-0.011	-0.01	-0.03	0.38	-1.0-1.0
ρ^2_Δ	*65069	NA	0.15	0.06	0.00	0.00	0.00-0.99
R^2_H1:Monotonicity S & T	200	Mean	0.10	0.08	0.03	0.09	0.00-0.35
R^2_H1:Monotonicity T	**17591**	**Mean**	**0.21**	**0.18**	**0.10**	**0.14**	**0.00-0.89**
R^2_H1:Monotonicity S	17	Mean	0.06	0.03	0.02	0.06	0.00-0.23
R^2_H1:No monotonicity	14354	Mean	0.10	0.09	0.07	0.06	0.00-0.53
Total	32162	-	-	-	-	-	-
R^2_H2:Monotonicity S & T	214	Median	0.08	0.05	0.02	0.09	0.00-0.30
R^2_H2:Monotonicity T	**32502**	**Median**	**0.19**	**0.16**	**0.02**	**0.14**	**0.00- 0.81**
R^2_H2:Monotonicity S	27	Median	0.06	0.04	0.02	0.06	0.00-0.29
R^2_H2:No monotonicity	18540	Median	0.10	0.09	0.07	0.06	0.00-0.53
Total	51283	-	-	-	-	-	-
R^2_H3:Monotonicity S & T	789	75th percentile	0.12	0.06	0.18	0.14	0.00-0.57
R^2_H3:Monotonicity T	468	75th percentile	0.07	0.04	0.07	0.07	0.00- 0.43
R^2_H3:Monotonicity S	982	75th percentile	0.05	0.04	0.02	0.05	0.00-0.40
R^2_H3:No monotonicity	145693	75th percentile	0.12	0.13	0.16	0.05	0.00-0.45
Total	147932	-	-	-	-	-	-
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							



(a) ρ_Δ scenario 1A (b) ρ_Δ scenario 2A (c) ρ_Δ scenario 3A

Figure 3: Distribution of ICA,ρ_Δ for scenario 1A,2A & 3A



(a) ρ_Δ scenario 4A (b) ρ_Δ scenario 4B (c) ρ_Δ scenario 4C

Figure 4: Distribution of ICA,ρ_Δ for scenario 4A,4B & 4C

Table 9: Scenario 1C-Descriptive statistics of ρ_Δ and R^2_H

Parameter	*No. Positive definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 1C: N=2000 and adjusted association (γ)=0.4							
ρ_Δ	*43171	NA	0.41	0.45	0.46	0.33	-1.00-1.00
ρ^2_Δ	*43171	NA	0.28	0.22	0.02	0.23	0.00-1.00
R^2_H1:Monotonicity S & T	1192	Mean	0.11	0.07	0.02	0.11	0.00-0.42
R^2_H1:Monotonicity T	297	Mean	0.05	0.04	0.03	0.05	0.00-0.23
R^2_H1:Monotonicity S	70	Mean	0.06	0.04	0.03	0.05	0.00-0.20
R^2_H1:No monotonicity	32125	Mean	0.15	0.14	0.13	0.07	0.00-0.61
Total	33684	-	-	-	-	-	-
R^2_H2:Monotonicity S & T	1193	Median	0.13	0.09	0.02	0.14	0.00-0.56
R^2_H2:Monotonicity T	232	Median	0.05	0.05	0.02	0.04	0.00- 0.20
R^2_H2:Monotonicity S	46	Median	0.05	0.03	0.02	0.04	0.00-0.16
R^2_H2:No monotonicity	32393	Median	0.14	0.14	0.12	0.07	0.00-0.62
Total	33864	-	-	-	-	-	-
R^2_H3:Monotonicity T	459	75th percentile	0.04	0.03	0.02	0.04	0.00- 0.29
R^2_H3:No monotonicity	231414	75th percentile	0.09	0.09	0.08	0.04	0.00-0.57
Total	231873	-	-	-	-	-	-
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							

Considering scenarios 1A, 1B and 1C presented in Tables 7,8 and 9 respectively, ρ_Δ was affected by the choice of γ. The ICA measure ρ_Δ (mean=0.76, SD=0.19) for scenario 1A was highest than corresponding values for ρ_Δ (mean=-0.11, SD=0.38) and ρ_Δ (mean=0.41,

SD=0.33) for scenarios 1B and 1C respectively though the minimum and maximum bounds for the ICA measure were broad. Similar patterns were observed for ρ_Δ in scenarios 2A,2B and 2C with N=50 represented in Tables 10-12 and for scenarios 3A-C shown in Tables 13-15 for N=200. Overall, the minimum and maximum bounds for this surrogacy metric was broad indicating that the values assumed for the unidentifiable correlations had a huge impact on the results. From Figure 3, the distribution of ρ_Δ for scenario 1A,2A and 3A indicates higher values occurring between 0.5-1.0.

Table 10: Scenario 2A-Descriptive statistics of ρ_Δ and R^2_H

Parameter	*No. Positive definites/ Valid probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 2A: N=50, and adjusted association (γ)=0.8							
ρ_Δ	*12367	NA	0.73	0.77	0.79	0.19	-0.58-1.00
ρ^2_Δ	*12367	NA	0.57	0.59	0.60	0.22	0.00-1.00
R^2_H1:No monotonicity	22281	Mean	0.17	0.16	0.14	0.07	0.00-0.52
R^2_H2:No monotonicity	9856	Median	0.18	0.17	0.17	0.07	0.00-0.57
R^2_H3:No Monotonicity	182297	75th percentile	0.22	0.22	0.22	0.05	0.00-0.62
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							

Table 11: Scenario 2B-Descriptive statistics of ρ_Δ and R^2_H

Parameter	*No. Positive definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 2B: N=50, and adjusted association (γ)=0.0							
ρ_Δ	*63429	NA	0.05	0.36	0.05	0.38	-0.99-1.0
ρ^2_Δ	*63429	NA	0.14	0.06	0.00	0.19	0.00-1.0
R^2_H1:Monotonicity S	16604	Mean	0.09	0.07	0.03	0.08	0.00-0.61
R^2_H1:No monotonicity	9200	Mean	0.10	0.09	0.06	0.06	0.01-0.49
Total	16604	-	-	-	-	-	-
R^2_H2:Monotonicity S	30389	Median	0.10	0.08	0.02	0.08	0.00-0.67
R^2_H2:No monotonicity	8314	Median	0.10	0.08	0.06	0.06	0.00-0.41
Total	38703	-	-	-	-	-	-
R^2_H3:Monotonicity S & T	3993	75th percentile	0.05	0.02	0.03	0.07	0.00-0.31
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							

Table 12: Scenario 2C-Descriptive statistics of ρ_Δ and R^2_H

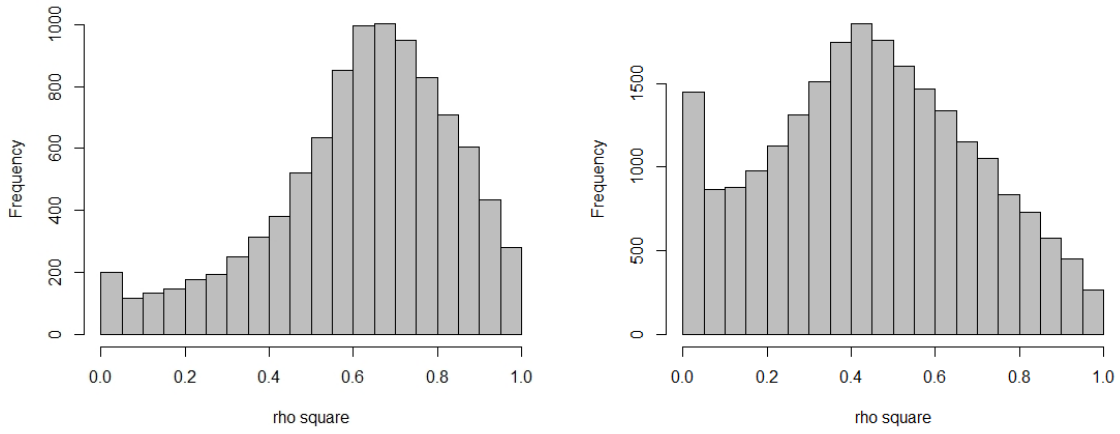
Parameter [5%,95%]	*No. Positive definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 2C: N=50, and adjusted association (γ)=0.4							
ρ_Δ	*41143	NA	0.41	0.46	0.33	0.51	-0.96-1.0
ρ^2_Δ	*41143	NA	0.28	0.23	0.02	0.23	0.00-1.0
R^2_H1: No monotonicity	20514	Mean	0.23	0.22	0.22	0.06	0.00-0.54
R^2_H2:No monotonicity	17690	Median	0.21	0.20	0.19	0.07	0.00-0.61
R^2_H3:Monotonicity T	152510	75th percentile	0.18	0.18	0.21	0.10	0.00-0.80
R^2_H3:No monotonicity	122807	75th percentile	0.22	0.22	0.23	0.04	0.00-0.57
Total	275317	-	-	-	-	-	-
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							

Table 13: Scenario 3A-Descriptive statistics of ρ_Δ and R^2_H

Parameter	*No. Positive definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 3A: N=200, and adjusted association (γ)=0.8							
ρ_Δ	*11659	NA	0.74	0.78	0.80	0.21	-0.94-1.0
ρ^2_Δ	*11659	NA	0.58	0.61	0.62	0.23	0.00-1.0
R^2_H1:Monotonicity S	21	Mean	0.08	0.07	0.03	0.08	0.00-0.32
R^2_H1:No monotonicity	8601	Mean	0.33	0.33	0.33	0.06	0.00-0.59
Total	8622	-	-	-	-	-	-
R^2_H2:No monotonicity	31428	Median	0.31	0.31	0.30	0.06	0.00-0.59
R^2_H3:No monotonicity	137313	75th percentile	0.30	0.29	0.29	0.04	0.01-0.62
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							

From Tables 6, 9 and 11 , R^2_H was higher for scenario 1A, 2A and 3A with γ=0.8 compared to scenarios with γ= 0.0 and 0.4 under same assumption of no monotonicity, although the mean and range of R^2_H remained smaller and narrower compared to the corresponding ρ^2_Δ in these scenarios irrespective of choice of dichotomization. This may be related to the fact that the relationship between ICA and γ may be hugely determined by the correlation between potential outcomes for surrogate and true endpoint(Alonso et al., 2017). Generally, in scenarios 1- 3, R^2_H mean values and range were smaller than ρ^2_Δ values for the corresponding datasets with the exception of scenario 1B for mean and median split when the assumption

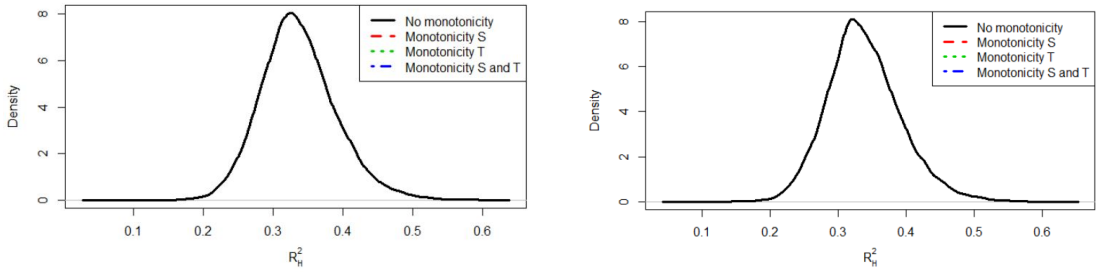
of monotonicity holds for True endpoint.



(a) ρ^2_Δ scenario 1A

(b) ρ^2_Δ scenario 4A with measurement error

Figure 5: Distribution of ICA,ρ^2_Δ for scenario 1A & 4A



(a) R^2_H median split scenario 1A

(b) R^2_H median split scenario 4A

Figure 6: Distribution of ICA,R^2_H for scenario 1A& 4A median split

Table 14: Scenario 3B-Descriptive statistics of ρ_Δ and R^2_H

Parameter	*No. Positive definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 3B: N=200, and adjusted association (γ)=0.0							
ρ_Δ	* 64989	NA	0.003	0.003	0.003	0.38	-0.99-1.0
ρ^2_Δ	*64989	NA	0.15	0.06	0.01	0.19	0.00-0.99
R^2_H1: Monotonicity T	5194	Mean	0.11	0.08	0.04	0.09	0.00-0.51
R^2_H1:No monotonicity	4674	Mean	0.10	0.09	0.07	0.06	0.00-0.45
Total	9868	-	-	-	-	-	-
R^2_H2:Monotonicity T	1067	Median	0.12	0.10	0.05	0.10	0.00-0.49
R^2_H2:No monotonicity	16027	Median	0.10	0.08	0.06	0.06	0.00-0.43
Total	17094	-	-	-	-	-	-
R^2_H3:No monotonicity	157916	75th percentile	0.12	0.12	0.12	0.05	0.00-0.44
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							

Table 15: Scenario 3C-Descriptive statistics of ρ_Δ and R^2_H

Parameter	*No. Positive definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 3C: N=200, and adjusted association (γ)=0.4							
ρ_Δ	*36983	NA	-0.05	-0.04	-0.02	0.34	-0.99-0.96
ρ^2_Δ	*36983	NA	0.12	0.05	0.00	0.15	0.00-0.98
R^2_H1:Monotonicity S	147	Mean	0.07	0.04	0.03	0.07	0.00-0.45
R^2_H1:No monotonicity	7212	Mean	0.17	0.17	0.07	0.06	0.00-0.49
Total	7359	-	-	-	-	-	-
R^2_H2:Monotonicity S	251	Median	0.07	0.05	0.03	0.06	0.00-0.35
R^2_H2:No monotonicity	24743	Median	0.17	0.16	0.15	0.07	0.00-0.54
Total	24994	-	-	-	-	-	-
R^2_H3:No monotonicity	233295	75th percentile	0.09	0.08	0.08	0.04	0.00-0.49
Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)							
Simulation-based sensitivity analysis for R^2_H with grid values G=(0.0, 0.01,...0.99)							

The impact of randomly introducing 23% of measurement error on the surrogate, was directly reflected on the data derived correlation estimates of $\hat{\rho}_{S_0T_0}$ and $\hat{\rho}_{S_1T_1}$. As depicted in Table 2 scenario 1A, values for $\hat{\rho}_{S_0T_0}$=0.78(95%CI:0.76,0.80) and $\hat{\rho}_{S_1T_1}$= 0.80(95%CI:0.78,0.81) were larger compared to 0.65(95%:0.62,0.67) and 0.63(95%0:0.61,0.66) respectively in scenario 4A with measurement error. Also, ρ_Δ (mean=0.61, SD=0.26) for scenario 4A with measure-

ment error than ρ_Δ (mean=0.76, SD=0.38). Similar findings observed for ρ_Δ scenarios 2A versus 4C and scenarios 3A versus 4B. The effect of measurement in 4A-C scenarios did not drastically affect R_H^2, as indicated with distribution in Figure 6 (median split under the assumption of no monotonicity). All R_H^2 values regardless of monotonicity assumption had smaller values and narrower range compared to ρ_Δ^2. As indicated in Figure 4, the distribution of ρ_Δ was similar to that in Figure 3 but for the fact that slightly more values were out of 0.5-1 range, this is also supported by Figure 5 of ρ_Δ^2 for scenario 1A versus 4A. Dichotomizing an endpoint could be preferred when errors are large, since contamination errors often have larger influence on the estimators based continuous responses than that of dichotomized responses (Shentu and Xie, 2010).

Table 16: Scenario 4A-C-Descriptive statistics of ρ_Δ and R_H^2

Parameter	*No. Positive definites/ Valid Probabilities	Cut-off Type	Mean	Median	Mode	SD	Range Min-Max
Scenario 4A: Measurement error, N=2000 and γ=0.8							
ρ_Δ	*22947	NA	0.61	0.66	0.68	0.26	-0.98-1.0
ρ_Δ^2	*22947	NA	0.45	0.44	0.42	0.24	0.00-1.0
$R_H^2$1:No monotonicity	32888	Mean	0.33	0.32	0.32	0.06	0.03-0.69
$R_H^2$2:No monotonicity	26479	Median	0.33	0.32	0.32	0.06	0.06-0.68
$R_H^2$3: No monotonicity	176562	75th percentile	0.26	0.25	0.25	0.05	0.00-0.65
Scenario 4B: Measurement error, N=200 and γ=0.8							
ρ_Δ	*23243	NA	0.61	0.66	0.68	0.26	-0.88-1.0
ρ_Δ^2	*23243	NA	0.43	0.44	0.42	0.24	0.00-0.99
$R_H^2$1: Monotonicity S	21	Mean	0.08	0.07	0.03	0.08	0.00-0.32
$R_H^2$1: No monotonicity	8601	Mean	0.33	0.33	0.33	0.05	0.06-0.59
Total	8622	-	-	-	-	-	-
$R_H^2$2: Monotonicity S	6	Median	0.11	0.05	0.03	0.13	0.00-0.34
$R_H^2$2: No monotonicity	9657	Median	0.35	0.35	0.33	0.05	0.08-0.62
Total	9663	-	-	-	-	-	-
$R_H^2$3:No monotonicity	153148	75th percentile	0.24	0.24	0.25	0.04	0.00-0.59
Scenario 4C: Measurement error, N=50, and γ=0.8							
ρ_Δ	* 24673	NA	0.60	0.64	0.69	0.24	-0.63-1.0
ρ_Δ^2	*24673	NA	0.42	0.41	0.40	0.24	0.00-1.0
$R_H^2$1:No Monotonicity	19045	Mean	0.19	0.18	0.16	0.07	0.00-0.63
$R_H^2$2: No Monotonicity T	11320	Median	0.18	0.17	0.16	0.07	0.02- 0.59
$R_H^2$3: No monotonicity	182318	75th percentile	0.24	0.23	0.24	0.05	0.01-0.63

Simulation-based sensitivity analysis for ρ_Δ, with grid values G=(-1, -0.90,...1)

Simulation-based sensitivity analysis for R_H^2 with grid values G=(0.0, 0.01,...0.99)

23% measurement error on the surrogate endpoint

5 Discussion

The focus of this project is to apply the information theoretic approach (ITA) considered as the unifying framework in validating surrogacy with computational benefits (no joint models or latent variables) , intuitive deciphering capabilities across different types of outcomes and yielding narrower confidence intervals (Alonso and Molenberghs, 2007) to the field of evaluating surrogates of protection. In vaccine development, the evaluation of surrogate endpoints may start at the pre-clinical (*invivo* and *invitro*) phases of the study, generating data with endpoints on cell activity, growth and cellular or humoral mechanism of actions. The discovery and statistical validation of reliable surrogates of vaccine protection may lead to accelerated approval of vaccines by regulatory agencies. Through a simulation study approach, we evaluated the behavior of the individual causal association (ICA) metric ρ_Δ and R^2_H of surrogacy validation in the context of continuous surrogate and true endpoint and binary dichotomization.

Generally, if S is a good surrogate for T, ΔS should convey a substantial amount of information on ΔT. Thus, in practice, for a good surrogate, an increase/decrease of the individual causal effect on the surrogate should be indicative of an increase/decrease of the individual causal effect on the true endpoint, implying ICA should be positive (Alonso et al., 2015). The range of ρ_Δ estimated in this exercise were wide (spanning from negative to positive intervals) suggestive of potential impact of unidentifiable correlations, but (Van der Elst et al., 2016) highlighted high values of $\rho_\Delta = corr(\Delta T, \Delta S) \approx 1$ with narrow range for continuous-continuous surrogate and true endpoint suggestive of a good surrogate. Similarly, for binary-binary surrogate and true endpoint, $R^2_H \approx 1$ is indicative of a good surrogate (Alonso et al., 2016b). It is hypothesized that ρ_Δ or corresponding ρ^2_Δ values will be higher and have narrower range whereas R^2_H values might be lower with broader range indicating loss of information (accuracy) upon dichotomizing continuous-continuous endpoints. Given our theoretical scenarios we highlighted lower values of R^2_H compared to their corresponding ρ^2_Δ estimates.

Dichotomizing continuous endpoints in clinical research is very common and may be done for clinical or statistical reasons, though not advisable from a statistical view point due to the

potential loss of information (Alonso and Molenberghs, 2007), sometimes it aids interpretation. For example it is easy to classify an individual as protected against a disease or not after attaining a specified threshold of post vaccination antibody titres, cellular or humoral activity akin to most vaccines used in the expanded program immunization (EPI)(Plotkin, 2010, WHO, 2013). The challenge of converting continuous endpoint to a dichotomous form is thus weighted on the amount of information loss and interpretability of the endpoint. Usually the process of dichotomizing a continuous endpoint requires the setting of clinical meaningful splits of the endpoint which is not easy to determine. We have highlighted some loss of information in estimating the individual causal association (ICA) metrics of surrogacy when dichotomizing a continuous true endpoint and continuous surrogate using data driven cut-points. Sometimes clinical meaningful threshold (Alonso and Molenberghs, 2007) are required to define protection, statistical optimal and/or unifying clinical cut-points should be evaluated when translating this approach into real world data. Also, using these theoretical scenarios it was observed that data dependent methods of dichotmomizing do impact the behavior of R^2_H individual causal association matric of surrogacy in the binary-binary setting under different assumptions of monotonicity, this is consistent with the work of (Alonso et al., 2017). The type of monotonicity assumption adopted have a big impact on R^2_H in some scenarios, which may be reflected in real world vaccine studies where new vaccines or a new combination of vaccines are evaluated against an active control or other active combinations. Although, we have loss of information dichotomizing continuous outcomes, the choice of dichotomizing endpoints may proof useful in data with contamination or measurement error (Shentu and Xie, 2010). The implementation of a sensitivity-based analysis in surrogacy validation under varying scenarios (sample size, adjusted association and inducing measurement error in surrogate measurement) was useful in highlighting the behavior of ICA given theoretical endpoints. In addition , simulation studies may be bias and may not directly translate to real world situation, the merits of this exercise is that a sensitivity based approach, making certain assumptions about the endpoints, the choice of surrogate, and the level of unidentifiable correlation of the potential outcomes should be considered in future evaluation. Future work on real world data on this topic should evaluate the effect of optimal statistical cut-points based on tailored models and reputed clinical thresholds on the behavior of ρ_Δ and R^2_H individual causal association metric of surrogacy.

6 Conclusion

The effect of cut-off points in surrogacy evaluation when dichotomizing a continuous endpoint remains an area of interest. The individual causal association (ICA) metric of surrogacy based on the information theoretic approach has been proven as a reliable metric for the evaluation surrogacy for binary-binary endpoints. Where there is a need to dichotomize a continuous surrogate and true endpoints; the choice of cut-points, underlining assumptions, a sensitivity-based analysis should be implemented and the authors should consider presenting results of more than one cut-off point.

References

Alonso, A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N. J., Shkedy, Z., and Van der Elst, W. (2016a). *Applied surrogate endpoint evaluation methods with SAS and R*. CRC Press.

Alonso, A. and Molenberghs, G. (2007). Surrogate marker evaluation from an information theory perspective. *Biometrics*, 63(1):180–186.

Alonso, A. and Molenberghs, G. (2008). Evaluating time to cancer recurrence as a surrogate marker for survival from an information theory perspective. *Statistical Methods in Medical Research*, 17(5):497–504.

Alonso, A., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Abrahantes, J. C., and Buyse, M. (2004). Prentice's approach and the meta-analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. *Biometrics*, 60(3):724–728.

Alonso, A., Molenberghs, G., Geys, H., Buyse, M., and Vangeneugden, T. (2006). A unifying approach for surrogate marker validation based on prentice's criteria. *Statistics in medicine*, 25(2):205–221.

Alonso, A., Van der Elst, W., and Meyvisch, P. (2017). Assessing a surrogate predictive value: a causal inference approach. *Statistics in medicine*, 36(7):1083–1098.

Alonso, A., Van der Elst, W., Molenberghs, G., Buyse, M., and Burzykowski, T. (2015). On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics*, 71(1):15–24.

Alonso, A., Van der Elst, W., Molenberghs, G., Buyse, M., and Burzykowski, T. (2016b). An information-theoretic approach for the evaluation of surrogate endpoints based on causal inference. *Biometrics*, 72(3):669–677.

Alonso Abad, A. et al. (2016). A causal-inference approach for the validation of surrogate endpoints based on information theory and sensitivity analysis. *Biometrics*, 10.

Auckland, C., Gray, S., Borrow, R., Andrews, N., Goldblatt, D., Ramsay, M., Miller, and Elizabeth (2006). Clinical and immunologic risk factors for meningococcal c conjugate

vaccine failure in the united kingdom. *The Journal of infectious diseases*, 194(12):1745–1752.

Burzykowski, T. and Buyse, M. (2006). Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 5(3):173–186.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001). Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(4):405–422.

Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, pages 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1):49–67.

Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., Alonso, A., Van der Elst, W., and Burzykowski, T. (2016). Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal*, 58(1):104–132.

Cohen, M. B., Giannella, R. A., Bean, J., Taylor, D. N., Parker, S., Hoeper, A., Wowk, S., Hawkins, J., Kochi, S. K., Schiff, G., et al. (2002). Randomized, controlled human challenge study of the safety, immunogenicity, and protective efficacy of a single dose of peru-15, a live attenuated oral cholera vaccine. *Infection and immunity*, 70(4):1965–1970.

Cover, T., Thomas, J., and Wiley, J. (1991). W. interscience. In *Elements of Information Theory*. Wiley New York.

Detela, G. and Lodge, A. (2019). Eu regulatory pathways for atmps: standard, accelerated and adaptive pathways to marketing authorisation. *Molecular therapy. Methods & clinical development*, 13:205.

Diemert, D., Campbell, D., Brelsford, J., Leasure, C., Li, G., Peng, J., Zumer, M., Younes, N., Bottazzi, M. E., Mejia, R., et al. (2018). Controlled human hookworm infection: accelerating human hookworm vaccine development. In *Open Forum Infectious Diseases*, volume 5, page ofy083. Oxford University Press US.

Ensor, H., Lee, R. J., Sudlow, C., and Weir, C. J. (2016). Statistical approaches for evaluating surrogate outcomes in clinical trials: a systematic review. *Journal of biopharmaceutical statistics*, 26(5):859–879.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Freedman, L. S., Graubard, B. I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine*, 11(2):167–178.

Goldenthal, K. L., Falk, L. A., Ball, L., and Geber, A. (2001). Prelicensure evaluation of combination vaccines. *Clinical Infectious Diseases*, 33(Supplement_4):S267–S273.

Hudgens, M. G., Gilbert, P. B., and Self, S. G. (2004). Endpoints in vaccine trials. *Statistical methods in medical research*, 13(2):89–114.

Jin, C., Gibani, M. M., Moore, M., Juel, H. B., Jones, E., Meiring, J., Harris, V., Gardner, J., Nebykova, A., Kerridge, S. A., et al. (2017). Efficacy and immunogenicity of a vi-tetanus toxoid conjugate vaccine in the prevention of typhoid fever using a controlled human infection model of salmonella typhi: a randomised controlled, phase 2b trial. *The Lancet*, 390(10111):2472–2480.

Katz, R. (2004). Biomarkers and surrogate markers: an fda perspective. *NeuroRx*, 1(2):189–195.

McCall, M. B., Kremsner, P. G., and Mordmüller, B. (2018). Correlating efficacy and immunogenicity in malaria vaccine trials. In *Seminars in immunology*, volume 39, pages 52–64. Elsevier.

Medaglini, D., Santoro, F., and Siegrist, C.-A. (2018). Correlates of vaccine-induced protective immunity against ebola virus disease. In *Seminars in immunology*, volume 39, pages 65–72. Elsevier.

Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in medicine*, 20(20):3023–3038.

Plotkin, S. A. (2010). Correlates of protection induced by vaccination. *Clin. Vaccine Immunol.*, 17(7):1055–1065.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440.

Pryseley, A., Tilahun, A., Alonso, A., and Molenberghs, G. (2007). Information-theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate endpoints. *Clinical Trials*, 4(6):587–597.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593.

Shentu, Y. and Xie, M. (2010). A note on dichotomization of continuous response variable in the presence of contamination and model misspecification. *Statistics in medicine*, 29(21):2200–2214.

Sherman, A. C., Mehta, A., Dickert, N. W., Anderson, E. J., and Rouphael, N. (2019). The future of flu: a review of the human challenge model and systems biology for advancement of influenza vaccinology. *Frontiers in cellular and infection microbiology*, 9.

Thakur, A., Pedersen, L. E., and Jungersen, G. (2012). Immune markers and correlates of protection for vaccine induced immune responses. *Vaccine*, 30(33):4907–4920.

Van der Elst, W., Meyvisch, P., Poveda, A. F., Alonso, A., and Van der Elst, M. W. (2020). Package 'surrogate'.

Van der Elst, W., Molenberghs, G., and Alonso, A. (2016). Exploring the relationship between the causal-inference and meta-analytic paradigms for the evaluation of surrogate endpoints. *Statistics in medicine*, 35(8):1281–1298.

Van Els, C., Mjaaland, S., Næss, L., Sarkadi, J., Gonczol, E., Smith Korsholm, K., Hansen, J., De Jonge, J., Kersten, G., Warner, J., et al. (2014). Fast vaccine design and development based on correlates of protection (cops): Influenza as a trendsetter. *Human vaccines & immunotherapeutics*, 10(7):1935–1948.

WHO (2013). Correlates of vaccine-induced protection: methods and implications. Technical report, World Health Organization.

7 Appendix

R Code

```
setwd("C:/Users/User/Documents/TheThesis")
library(Surrogate)# load the Surrogate library
library(ggplot2)
library(OptimalCutpoints)
library(glm)
library(lawstat)
library(epitools) # OR calculation

## SIM STS cont-cont###
# Simulate a dataset:
Sim.Data.STS(N.Total=2000, R.Indiv.Target=0.8, Means=c(0, 0, 0, 0), Seed=12)
View(Data.Observed.STS)
str(Data.Observed.STS)
simsts1<-Data.Observed.STS
names(simsts1)

#write.table(simsts1,file = "contcont2000.csv")# saved dataset for contcont2000#
as.factor(simsts1$Treat)
summary(simsts1)
#mean(simsts1$Surr)
#var(simsts1$Surr)
names(simsts1)
head(simsts1)
# Overall descriptive stats
(mean_sim<-apply(simsts1[,1:2], MARGIN = 2, FUN = mean))

(min_sim<-apply(simsts1[,1:2], MARGIN = 2, FUN = min))
(max_sim<-apply(simsts1[,1:2], MARGIN = 2, FUN = max))
(var_sim<-apply(simsts1[,1:2], MARGIN = 2, FUN = var))
(var_sim<-apply(simsts1[,1:2], MARGIN = 2, FUN = sd))
(sd_sim<-apply(simsts1[,1:2], MARGIN = 2, FUN = median))
(quantiles_sim<-apply(simsts1[,1:2], MARGIN = 2, FUN = quantile))
# Overall descriptive stats per treatment group
(mean_simcont<-apply(simsts1[simsts1$Treat==-1,1:2], MARGIN = 2, FUN = mean))
(mean_simtreat<-apply(simsts1[simsts1$Treat==1,1:2], MARGIN = 2, FUN = mean))
(min_simcont<-apply(simsts1[simsts1$Treat==-1,1:2], MARGIN = 2, FUN = min))
(min_simtreat<-apply(simsts1[simsts1$Treat==1,1:2], MARGIN = 2, FUN = min))
(max_simcont<-apply(simsts1[simsts1$Treat==-1,1:2], MARGIN = 2, FUN = max))
(max_simtreat<-apply(simsts1[simsts1$Treat==1,1:2], MARGIN = 2, FUN = max))
(var_simcont<-apply(simsts1[simsts1$Treat==-1,1:2], MARGIN = 2, FUN = var))
(var_simtreat<-apply(simsts1[simsts1$Treat==1,1:2], MARGIN = 2, FUN = var))
(var_simcont<-apply(simsts1[simsts1$Treat==-1,1:2], MARGIN = 2, FUN = sd))
(var_simtreat<-apply(simsts1[simsts1$Treat==1,1:2], MARGIN = 2, FUN = sd))
(median_simcont<-apply(simsts1[simsts1$Treat==-1,1:2], MARGIN = 2, FUN = median))
(median_simtreat<-apply(simsts1[simsts1$Treat==1,1:2], MARGIN = 2, FUN = median))
```

```
(quantile_simcont<-apply(simsts1[simsts1$Treat==-1,1:2], MARGIN = 2, FUN = quantile))
(quantile_simtreat<-apply(simsts1[simsts1$Treat==1,1:2], MARGIN = 2, FUN = quantile))

# Seed is used for reproducibility##
# Obtain identifiable correlation#
sts_surr_true2000 <- Single.Trial.RE.AA(Dataset = simsts1, Surr = Surr, True = True,
         Treat = Treat, Pat.ID = Pat.ID, Seed = 112)
summary(sts_surr_true2000)
# Plot Adjusted association for 1A
plot(sts_surr_true2000, Indiv.Level = TRUE, Trial.Level = FALSE)
### Scenario 1A: Obtaining ICA rho for cont-cont##
sts_ica_surr_true2000 <- ICA.ContCont(T0S0 = 0.78, T1S1 = 0.80, S0S0 = 0.945,
        S1S1 = 1.016, T0T0 = 1.002, T1T1 = 1.05, T0T1 = seq(-1, 1, by = 0.1),
        T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(sts_ica_surr_true2000) # obtain ICA cont-cont scenario 1A
## Explore graphical plots for ICA cont-cont scenario 1A##
plot(sts_ica_surr_true2000)
plot(sts_ica_surr_true2000, Labels = TRUE)
plot(sts_ica_surr_true2000, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true2000, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true2000, Min = -1, Max = 0.9)
###### Obtain rho square distribution for ICA cont-cont scenario 1A##
rho1<-sts_ica_surr_true2000$ICA
View(rho1)
summary(rho1)
sqrho1<-(rho1)^2
summary(sqrho1)
sd(sqrho1)
hist(sqrho1)
Modes(sqrho1)
str(sqrho1)# number of positive definites 9711
quantile(sqrho1,probs = seq(0, 1, 0.05) )

############################################################
### Dichotomization of T and S Binary-Binary endpoint###
# Scenario 1A Make T binary-mean cut-off#
simsts1_Bin <- simsts1
simsts1_Bin$True_Bin <- simsts1_Bin$True
simsts1_Bin$True_Bin[simsts1_Bin$True>0.004] <- 1
simsts1_Bin$True_Bin[simsts1_Bin$True<=0.004] <- 0
# Making S binary
simsts1_Bin$Surr_Bin <- simsts1_Bin$Surr
simsts1_Bin$Surr_Bin[simsts1_Bin$Surr>0.004] <- 1
simsts1_Bin$Surr_Bin[simsts1_Bin$Surr<=0.004] <- 0

names(simsts1_Bin)
head(simsts1_Bin)
#write.table(simsts1_Bin,file = "Binbin2000.csv")
```

```
# saved dataset for Scenario 1A BinBin2000 with individualassociation 0.8#
####RH1 Obtaining the Identifiable counterfactual probabilities####
MarginalProbs(Dataset=simsts1_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
#### Calculating OR theta_T0S0 and theta_T1S1   ####
OR1atable_con<-matrix(c(383,109,110,398), nrow = 2, ncol = 2)
OR1atable_con
oddsratio.wald(OR1atable_con) # control group theta_T0S0
OR1atable_treat<-matrix(c(408,113,109,370), nrow = 2, ncol = 2)
OR1atable_treat
oddsratio.wald(OR1atable_treat) # treatment group theta_T1S1
# Scenario 1A-RH1 Obtaining ICA using  ICA-BIN-BIN function plugin marginal probs #
start.time <- Sys.time()
simsts_Bin_ICA <- ICA.BinBin(pi1_1_=0.398, pi0_1_=0.109, pi1_0_=0.111,
  pi_1_1=0.373, pi_1_0=0.109, pi_0_1=0.113, Seed=112, Monotonicity=c("General"),
  Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA)


summary(simsts_Bin_ICA1)
#### Plots and Obtaining the distribution of RH1)####
plot(simsts_Bin_ICA1a)
View(simsts_Bin_ICA1a$R2_H)
df_mean_082000<-simsts_Bin_ICA1a$R2_H
head(df_mean_082000)
quantile(df_mean_082000,probs = seq(0, 1, 0.05) )
write.table(df_mean_082000,file = "df_mean_082000.csv")
str(df_mean_082000)
#####################################################################
### 2-Binary-Binary endpoint-median cut-off###
# Make T binary
simsts2_Bin <- simsts1
simsts2_Bin$True_Bin <- simsts2_Bin$True
simsts2_Bin$True_Bin[simsts2_Bin$True>-0.013] <- 1
simsts2_Bin$True_Bin[simsts2_Bin$True<=-0.013] <- 0
# Making S binary
simsts2_Bin$Surr_Bin <- simsts2_Bin$Surr
simsts2_Bin$Surr_Bin[simsts2_Bin$Surr>-0.006] <- 1
simsts2_Bin$Surr_Bin[simsts2_Bin$Surr<=-0.006] <- 0
head(simsts2_Bin)
View(simsts2_Bin)
####RH2 Obtaining the Identifiable counterfactual probabilities####
MarginalProbs(Dataset=simsts2_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
#### Calculating OR theta_T0S0 and theta_T1S1   ####
start.time <- Sys.time()
# Scenario 1A-RH12 Obtaining ICA using  ICA-BIN-BIN function plugin marginal probs #
start.time <- Sys.time()
```

```
simsts_Bin_ICA2a <- ICA.BinBin(pi1_1_=0.401, pi0_1_=0.108, pi1_0_=0.112,
             pi_1_1=0.378, pi_1_0=0.109, pi_0_1=0.113, Seed=112, Monotonicity=c("General"),
            Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime
summary(simsts_Bin_ICA2a) # Obtaining RH2 for median split
df_med_082000<-simsts_Bin_ICA2a$R2_H
head(df_med_082000)
quantile(df_med_082000,probs = seq(0, 1, 0.05) )
write.table(df_med_082000,file = "df_median_082000.csv")
str(df_med_082000)

### RH3-Binary-Binary endpoint-75% percentile cut-off###
# Make T binary#
simsts3_Bin <- simsts1
simsts3_Bin$True_Bin <- simsts3_Bin$True
simsts3_Bin$True_Bin[simsts3_Bin$True>0.65] <- 1
simsts3_Bin$True_Bin[simsts3_Bin$True<=0.65] <- 0
# Making S binary
simsts3_Bin$Surr_Bin <- simsts3_Bin$Surr
simsts3_Bin$Surr_Bin[simsts3_Bin$Surr>0.66] <- 1
simsts3_Bin$Surr_Bin[simsts3_Bin$Surr<=0.66] <- 0
head(simsts3_Bin)
View(simsts3_Bin)
####RH3 Obtaining the Identifiable counterfactual probabilities####
MarginalProbs(Dataset=simsts3_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)

# Scenario 1A-RH3 Obtaining ICA using  ICA-BIN-BIN function plugin marginal probs #
start.time <- Sys.time()
simsts_Bin_ICA3a <- ICA.BinBin(pi1_1_= 0.174, pi0_1_= 0.089, pi1_0_=0.08,
         pi_1_1=0.157, pi_1_0=0.091, pi_0_1=0.081, Seed=112, Monotonicity=c("General"),
        Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime
summary(simsts_Bin_ICA3a)
plot(simsts_Bin_ICA3a)
df_per_082000<-simsts_Bin_ICA3a$R2_H
head(df_per_082000)
quantile(df_per_082000,probs = seq(0, 1, 0.05) )
write.table(df_per_082000,file = "df_per_082000.csv")

#################################################################
#################################################################
# Scenario 1B simulation with adjusted association of 0.00)
Sim.Data.STS(N.Total=2000, R.Indiv.Target=0.0, Means=c(0, 0, 0, 0), Seed=121)
head(Data.Observed.STS)
simsts2<-Data.Observed.STS
# saved dataset for contcont2000 adjusted association 0.0 #
#write.table(simsts2,file = "contcont02000.csv")
# summary statistics for overall dataset#
```

```
(mean_sim<-apply(simsts2[,1:2], MARGIN = 2, FUN = mean))
(min_sim<-apply(simsts2[,1:2], MARGIN = 2, FUN = min))
(max_sim<-apply(simsts2[,1:2], MARGIN = 2, FUN = max))
(var_sim<-apply(simsts2[,1:2], MARGIN = 2, FUN = var))
(sd_sim<-apply(simsts2[,1:2], MARGIN = 2, FUN = sd))
(median_sim<-apply(simsts2[,1:2], MARGIN = 2, FUN = median))
(var_sim<-apply(simsts2[,1:2], MARGIN = 2, FUN = quantile))
# # rho scenario 1B Obtain identifiable correlations#
sts_surr_true2000b <- Single.Trial.RE.AA(Dataset = simsts2, Surr = Surr, True = True,
     Treat = Treat, Pat.ID = Pat.ID, Seed = 212)
summary(sts_surr_true2000b)
# plot adjusted association scenario 1B
plot(sts_surr_true2000b, Indiv.Level = TRUE, Trial.Level = FALSE)
### Scenario 1A: Obtaining ICA rho for cont-cont##
sts_ica_surr_true2000b <- ICA.ContCont(T0S0 = -0.035, T1S1 = 0.011, S0S0 = 1.044,
            S1S1 = 0.998, T0T0 = 0.959, T1T1 = 0.963, T0T1 = seq(-1, 1, by = 0.1),
        T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(sts_ica_surr_true2000b) #ICA 1B
# rho scenario 1B plots and diagrams##
plot(sts_ica_surr_true2000b)
plot(sts_ica_surr_true2000b, Labels = TRUE)
plot(sts_ica_surr_true2000b, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true2000b, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true2000b, Min = -1, Max = 0.9)
# Obtaining rho_square distribution
rho2<-sts_ica_surr_true2000b$ICA
sqrho2<-(rho2)^2
summary(sqrho2)
Modes(sqrho2)
str(rho2)
quantile(sqrho2,probs = seq(0, 1, 0.05) )

#### Binary dichotomization for adjusted correlation of 0.00##
### RH1 Scenario 1B-Binary-Binary endpoint-Mean dichotomization###
# Make T binary-mean
simsts2b1_Bin <- simsts2
simsts2b1_Bin$True_Bin <- simsts2b1_Bin$True
simsts2b1_Bin$True_Bin[simsts2b1_Bin$True>-0.004] <- 1
simsts2b1_Bin$True_Bin[simsts2b1_Bin$True<=-0.004] <- 0
# Making S binary
simsts2b1_Bin$Surr_Bin <- simsts2b1_Bin$Surr
simsts2b1_Bin$Surr_Bin[simsts2b1_Bin$Surr>-0.025] <- 1
simsts2b1_Bin$Surr_Bin[simsts2b1_Bin$Surr<=-0.025] <- 0

head(simsts2b1_Bin)
# saved dataset for scenario 1B BinBin2000 with AA=0.0#
#write.table(simsts2b1_Bin,file = "Binbin002000b1.csv")
```

```
# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=simsts2b1_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
##Calculating OR theta_T0S0 and theta_T1S1  ###
OR1btable_con<-matrix(c(241,242,253,264), nrow = 2, ncol = 2)
OR1btable_con
oddsratio.wald(OR1btable_con)
OR1btable_treat<-matrix(c(256,246,249,249), nrow = 2, ncol = 2)
OR1btable_treat
oddsratio.wald(OR1btable_treat)
# Scenario 1B RH1 Obtaining the ICA-BIN-BIN#
#start.time <- Sys.time()
start.time <- Sys.time()
simsts_Bin_ICA2b <- ICA.BinBin(pi1_1_=0.241, pi0_1_=0.254, pi1_0_=0.257,
 pi_1_1=0.272, pi_1_0=0.257, pi_0_1=0.234, Seed=112, Monotonicity=c("General"),
     Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime
summary(simsts_Bin_ICA2b)
# plotting RH2-for mean split AA=0.00#
plot(simsts_Bin_ICA2b)
df_mean_002000<-simsts_Bin_ICA2b$R2_H
head(df_per_002000)
View(df_per_002000)
quantile(df_per_002000,probs = seq(0, 1, 0.05) )
write.table(df_per_002000,file = "df_mean_002000.csv")
### Scenario 1b RH2-Median-Binary-Binary endpoint###
# Make T binary-median
simsts2b2_Bin <- simsts2
simsts2b2_Bin$True_Bin <- simsts2b2_Bin$True
simsts2b2_Bin$True_Bin[simsts2b2_Bin$True>0.016] <- 1
simsts2b2_Bin$True_Bin[simsts2b2_Bin$True<=0.016] <- 0
# Making S binary
simsts2b2_Bin$Surr_Bin <- simsts2b2_Bin$Surr
simsts2b2_Bin$Surr_Bin[simsts2b2_Bin$Surr>-0.023] <- 1
simsts2b2_Bin$Surr_Bin[simsts2b2_Bin$Surr<=-0.023] <- 0
head(simsts2b2_Bin)
View(simsts2b2_Bin)
## Obtaining conterfactual probabilities##
MarginalProbs(Dataset=simsts2b2_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
#Calculating OR median split AA=0.00 N:2000##
OR2btable_con<-matrix(c(264,259,241,236), nrow = 2, ncol = 2)
OR2btable_con
oddsratio.wald(OR2btable_con)
OR2btable_treat<-matrix(c(240,237,254,269), nrow = 2, ncol = 2)
OR2btable_treat
oddsratio.wald(OR2btable_treat)
# Scenario 1B median split RH2-2 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA1b2 <- ICA.BinBin(pi1_1_=0.236, pi0_1_=0.259, pi1_0_=0.241,
```

```
        pi_1_1=0.269, pi_1_0=0.254, pi_0_1=0.237, Seed=112, Monotonicity=c("General"),
        Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time

summary(simsts_Bin_ICA1b2)
# Obtaining the distribution of RH2#
df_med_002000<-simsts_Bin_IC1b2A$R2_H
quantile(df_med_002000,probs = seq(0, 1, 0.05) )
write.table(df_med_002000,file = "df_medi_002000.csv")

### Scenario 1B Percentile split-Binary-Binary endpoint###
# Make T binary-75% percentile split
simsts3b3_Bin <- simsts2
simsts3b3_Bin$True_Bin <- simsts3b3_Bin$True
simsts3b3_Bin$True_Bin[simsts3b3_Bin$True>0.66] <- 1
simsts3b3_Bin$True_Bin[simsts3b3_Bin$True<=0.66] <- 0
# Making S binary
simsts3b3_Bin$Surr_Bin <- simsts3b3_Bin$Surr
simsts3b3_Bin$Surr_Bin[simsts3b3_Bin$Surr>0.63] <- 1
simsts3b3_Bin$Surr_Bin[simsts3b3_Bin$Surr<=0.63] <- 0
head(simsts3b3_Bin)
View(simsts3b3_Bin)
## Obtain counterfactual probabilities##
MarginalProbs(Dataset=simsts3b3_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
# Calculate OR_T0S0 and OR_T1S1
OR3btable_con<-matrix(c(558,191,184,67), nrow = 2, ncol = 2)
OR3btable_con
oddsratio.wald(OR3btable_con)
OR3btable_treat<-matrix(c(558,194,199,49), nrow = 2, ncol = 2)
OR3btable_treat
oddsratio.wald(OR3btable_treat)
## RH2-3 ICA for percentile split##
start.time <- Sys.time()
simsts_Bin_ICA4b <- ICA.BinBin(pi1_1_= 0.05, pi0_1_= 0.196, pi1_0_=0.179,
 pi_1_1=0.066, pi_1_0=0.204, pi_0_1=0.196, Seed=112, Monotonicity=c("General"),
 Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime
summary(simsts_Bin_ICA4b)
# plot RH2-3 for percentile split##
plot(simsts_Bin_ICA4b)
df_per_002000<-simsts_Bin_ICA4b$R2_H
head(df_per_002000)
View(df_per_002000)
quantile(df_per_002000,probs = seq(0, 1, 0.05) )
write.table(df_per_002000,file = "df_per_002000.csv")
### 4b4-Binary-Binary endpoint###
# Make T binary
```

```
####### ##############################################################################
##############################################################################################
# Scenario 1C Generating data for simulation with adjusted association(AA)= 0.4)
Sim.Data.STS(N.Total=2000, R.Indiv.Target=0.4, Means=c(0, 0, 0, 0), Seed=122)
head(Data.Observed.STS)
simsts3<-Data.Observed.STS
#write.table(simsts3,file = "contcont042000.csv")
# saved dataset 1C for contcont2000 adjusted association 0.4#
head(simsts3)
## Obtaining summary statistics##
(mean_sim3<-apply(simsts3[,1:2], MARGIN = 2, FUN = mean))
(min_sim3<-apply(simsts3[,1:2], MARGIN = 2, FUN = min))
(max_sim3<-apply(simsts3[,1:2], MARGIN = 2, FUN = max))
(var_sim3<-apply(simsts3[,1:2], MARGIN = 2, FUN = var))
(sd_sim3<-apply(simsts3[,1:2], MARGIN = 2, FUN = sd))
(median_sim3<-apply(simsts3[,1:2], MARGIN = 2, FUN = median))
(quantiles_sim3<-apply(simsts3[,1:2], MARGIN = 2, FUN = quantile))

# Obtain unidentified correlation#
sts_surr_true2000c <- Single.Trial.RE.AA(Dataset = simsts3, Surr = Surr, True = True,
     Treat = Treat, Pat.ID = Pat.ID, Seed = 312)
summary(sts_surr_true2000c)
### Plot Adjusted association for N:200 AA=0.4
plot(sts_surr_true2000c, Indiv.Level = TRUE, Trial.Level = FALSE)
# Scenario 1C: Obtain ICA-rho Adjusted association for N:200 AA=0.4
sts_ica_surr_true2000c <- ICA.ContCont(T0S0 = 0.45, T1S1 = 0.41, S0S0 = 1.04,
 S1S1 = 1.04, T0T0 = 1.04, T1T1 = 0.94, T0T1 = seq(-1, 1, by = 0.1),
T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(sts_ica_surr_true2000c) # rhop scenario 1C
Plots for for ICA 1C###
plot(sts_ica_surr_true2000c)
plot(sts_ica_surr_true2000c, Labels = TRUE)
plot(sts_ica_surr_true2000c, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true2000c, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true2000c, Min = -1, Max = 0.9)
## Obtaining rho square distribution##
rho3<-sts_ica_surr_true2000c$ICA
str(rho3)
sqrho3<-(rho3)^2
summary(sqrho3)
Modes(sqrho3)
sd(sqrho3)
hist(sqrho3)
quantile(sqrho3,probs = seq(0, 1, 0.05) )
# Scenario 1C Binary dichotomization for adjusted correlation of 0.4##
### 1c-Binary-Binary endpoint###
# Make T binary-mean split
simsts3c1_Bin <- simsts3
```

```
simsts3c1_Bin$True_Bin <- simsts3c1_Bin$True
simsts3c1_Bin$True_Bin[simsts3c1_Bin$True>-0.002] <- 1
simsts3c1_Bin$True_Bin[simsts3c1_Bin$True<=-0.002] <- 0
# Making S binary
simsts3c1_Bin$Surr_Bin <- simsts3c1_Bin$Surr
simsts3c1_Bin$Surr_Bin[simsts3c1_Bin$Surr>-0.009] <- 1
simsts3c1_Bin$Surr_Bin[simsts3c1_Bin$Surr<=-0.009] <- 0
head(simsts3c1_Bin)
#write.table(simsts3c1_Bin,file = "Binbin042000c1.csv")
# saved dataset for mean BinBin2000 with AA=0.4#
# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=simsts3c1_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
# calculating OR_T0S0 and OR_T1S1
OR1ctable_con<-matrix(c(316,188,171,325), nrow = 2, ncol = 2)
OR1ctable_con
oddsratio.wald(OR1ctable_con)
OR1ctable_treat<-matrix(c(317,169,191,323), nrow = 2, ncol = 2)
OR1ctable_treat
oddsratio.wald(OR1ctable_treat)
# Scenario 1C RH2-1 Obtaininmg the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA2c <- ICA.BinBin(pi1_1_=0.306, pi0_1_=0.183, pi1_0_=0.184,
     pi_1_1=0.342, pi_1_0=0.178, pi_0_1=0.174, Seed=112, Monotonicity=c("General"),
    Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA2c)
plot(simsts_Bin_ICA2c)
df_mean_042000<-simsts_Bin_ICA3c$R2_H
head(df_mean_042000)
quantile(df_mean_042000,probs = seq(0, 1, 0.05) )
write.table(df_mean_042000,file = "df_mean_042000.csv") # save distribution of RH2-1

### Scenario 1C-Median split-Binary-Binary endpoint###
# Make T binary-median split
simsts3c2_Bin <- simsts3
simsts3c2_Bin$True_Bin <- simsts3c2_Bin$True
simsts3c2_Bin$True_Bin[simsts3c2_Bin$True>0.002] <- 1
simsts3c2_Bin$True_Bin[simsts3c2_Bin$True<=0.002] <- 0
# Making S binary
simsts3c2_Bin$Surr_Bin <- simsts3c2_Bin$Surr
simsts3c2_Bin$Surr_Bin[simsts3c2_Bin$Surr>-0.002] <- 1
simsts3c2_Bin$Surr_Bin[simsts3c2_Bin$Surr<=-0.002] <- 0
head(simsts3c2_Bin)
View(simsts3c2_Bin)
###Obtaining counterfactual probabilities
MarginalProbs(Dataset=simsts3c2_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## Calculating OR_T0S0 and OR_T1S1
```

```
OR1ctable_con<-matrix(c(316,188,171,325), nrow = 2, ncol = 2)
OR1ctable_con
oddsratio.wald(OR1ctable_con)
OR1ctable_treat<-matrix(c(317,169,191,323), nrow = 2, ncol = 2)
OR1ctable_treat
oddsratio.wald(OR1ctable_treat)
# Obtain RH2-2-ICA from median split#
start.time <- Sys.time()
simsts_Bin_ICA3c <- ICA.BinBin(pi1_1_=0.306, pi0_1_=0.182, pi1_0_=0.183,
             pi_1_1=0.338, pi_1_0=0.18, pi_0_1=0.174, Seed=312, Monotonicity=c("General"),
            Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime
summary(simsts_Bin_ICA3c)
df_med_042000<-simsts_Bin_ICA3c$R2_H
head(df_med_042000)
View(df_med_042000)
quantile(df_med_042000,probs = seq(0, 1, 0.05) )
write.table(df_med_042000,file = "df_med_002000.csv")
### RH2-3 75% percentile split-Binary-Binary endpoint###
# Make T binary-75% split
simsts3c3_Bin <- simsts3
simsts3c3_Bin$True_Bin <- simsts3c3_Bin$True
simsts3c3_Bin$True_Bin[simsts3c3_Bin$True>0.71] <- 1
simsts3c3_Bin$True_Bin[simsts3c3_Bin$True<=0.71] <- 0
# Making S binary
simsts3c3_Bin$Surr_Bin <- simsts3c3_Bin$Surr
simsts3c3_Bin$Surr_Bin[simsts3c3_Bin$Surr>0.69] <- 1
simsts3c3_Bin$Surr_Bin[simsts3c3_Bin$Surr<=0.69] <- 0
head(simsts3c3_Bin)
View(simsts3c3_Bin)
### Obtaining counterfactual probabilities
MarginalProbs(Dataset=simsts3c3_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## Calculating OR_T0S0 and OR_T1S1
OR3ctable_con<-matrix(c(608,146,132,114), nrow = 2, ncol = 2)
OR3ctable_con
oddsratio.wald(OR3ctable_con)
OR3ctable_treat<-matrix(c(614,132,147,107), nrow = 2, ncol = 2)
OR3ctable_treat
oddsratio.wald(OR3ctable_treat)
###RH2-3 Obbtaining ICA
start.time <- Sys.time()
simsts_Bin_ICA4c <- ICA.BinBin(pi1_1_= 0.112, pi0_1_= 0.141, pi1_0_=0.132,
        pi_1_1=0.109, pi_1_0=0.147, pi_0_1=0.137, Seed=412, Monotonicity=c("General"),
         Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime
summary(simsts_Bin_ICA4c)
Plot(simsts_Bin_ICA4c)
df_per_042000<-simsts_Bin_ICA4c$R2_H
```

```
head(df_per_042000)
quantile(df_per_042000,probs = seq(0, 1, 0.05) )
write.table(df_per_042000,file = "df_median_042000.csv")

##############################################################################
##############################################################################
# Sample size small N=50#
Sim.Data.STS(N.Total=50, R.Indiv.Target=0.8, Means=c(0, 0, 0, 0), Seed=512)
head(Data.Observed.STS)
simsts501<-Data.Observed.STS
## Summary statistics
# saved dataset for contcont50 AA=0.8#
#write.table(simsts501,file = "contcont0850.csv")
(mean_sim2a<-apply(simsts501[,1:2], MARGIN = 2, FUN = mean))
(min_sim2a<-apply(simsts501[,1:2], MARGIN = 2, FUN = min))
(max_sim2a<-apply(simsts501[,1:2], MARGIN = 2, FUN = max))
(var_sim2a<-apply(simsts501[,1:2], MARGIN = 2, FUN = var))
(var_sim2a<-apply(simsts501[,1:2], MARGIN = 2, FUN = sd))
(median_sim2a<-apply(simsts501[,1:2], MARGIN = 2, FUN = median))
(quantiles_sim2a<-apply(simsts501[,1:2], MARGIN = 2, FUN = quantile))

## Obtaining identifiable correlations
sts_surr_true50a <- Single.Trial.RE.AA(Dataset = simsts501, Surr = Surr, True = True,
            Treat = Treat, Pat.ID = Pat.ID, Seed = 511)
summary(sts_surr_true50a)
plot(sts_surr_true50a, Indiv.Level = TRUE, Trial.Level = FALSE)
####Scenario 2A obtaining ICA rho
sts_ica_surr_true50a <- ICA.ContCont(T0S0 = 0.65, T1S1 = 0.80, S0S0 = 1.54,
    S1S1 = 0.88, T0T0 =  1.05, T1T1 = 0.73, T0T1 = seq(-1, 1, by = 0.1),
    T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(sts_ica_surr_true50a)
plot(sts_ica_surr_true50a)
plot(sts_ica_surr_true50a, Labels = TRUE)
plot(sts_ica_surr_true50a, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true50a, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true50a, Min = -1, Max = 0.9)
#Obtaining the distribution of rho-sqaure
rho4<-sts_ica_surr_true50a$ICA
str(rho4)
sqrho4<-(rho4)^2
sd(sqrho4)
summary(sqrho4)
Modes(sqrho4)
hist(sqrho4)
quantile(sqrho4,probs = seq(0, 1, 0.05) )

# Scenario 2A RH2-1 Mean dichotomization for N=50 and adjusted correlation of 0.8##
### 501a-Binary-Binary endpoint###
```

```
# Make T and S binary mean dichotomisation
simsts501a_Bin <- simsts501
simsts501a_Bin$True_Bin <- simsts501a_Bin$True
simsts501a_Bin$True_Bin[simsts501a_Bin$True>-0.11] <- 1
simsts501a_Bin$True_Bin[simsts501a_Bin$True<=-0.11] <- 0
simsts501a_Bin$Surr_Bin <- simsts501a_Bin$Surr
simsts501a_Bin$Surr_Bin[simsts501a_Bin$Surr>-0.15] <- 1
simsts501a_Bin$Surr_Bin[simsts501a_Bin$Surr<=-0.15] <- 0
head(simsts501a_Bin)
#write.table(simsts501a_Bin,file = "Binbin08501a.csv")
# saved dataset for BinBin501a with AA= 0.8#
# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=simsts501a_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_con2a<-matrix(c(6,6,3,10), nrow = 2, ncol = 2)
OR_table_con2a
oddsratio.wald(OR_table_con2a)
OR_table_treat2a<-matrix(c(11,3,5,6), nrow = 2, ncol = 2)
OR_table_treat2a
oddsratio.wald(OR_table_treat2a)
# Scenario 2A RH2-1 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA501a <- ICA.BinBin(pi1_1_=0.4, pi0_1_=0.24, pi1_0_=0.12,
 pi_1_1=0.24, pi_1_0=0.2, pi_0_1=0.12, Seed=512, Monotonicity=c("General"),
 Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA501a)
df_mean_0850<-simsts_Bin_ICA501a$R2_H
head(df_mean_0850)
quantile(df_mean_0850,probs = seq(0, 1, 0.05) )
write.table(df_mean_0850,file = "df_mean_0850.csv")

### Scenario 2a Median dichotomization with adjusted association 0.8
# Make Tand S binary median dichotomisation
simsts501b_Bin <- simsts501
simsts501b_Bin$True_Bin <- simsts501b_Bin$True
simsts501b_Bin$True_Bin[simsts501b_Bin$True>-0.13] <- 1
simsts501b_Bin$True_Bin[simsts501b_Bin$True<=-0.13] <-0
simsts501b_Bin$Surr_Bin <- simsts501b_Bin$Surr
simsts501b_Bin$Surr_Bin[simsts501b_Bin$Surr>-0.13] <-1
simsts501b_Bin$Surr_Bin[simsts501b_Bin$Surr<=-0.13] <- 0
head(simsts501b_Bin)

# obtaining counterfactual probabilities
MarginalProbs(Dataset=simsts501b_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_con2b<-matrix(c(6,6,3,10), nrow = 2, ncol = 2)
```

```
OR2ctable_con2b
oddsratio.wald(OR_table_con2b)
OR_table_treat2b<-matrix(c(11,2,5,7), nrow = 2, ncol = 2)
OR_table_treat2b
oddsratio.wald(OR_table_treat2b)
# Scenario 2A RH2-2 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA502 <- ICA.BinBin(pi1_1_=0.4, pi0_1_=0.24, pi1_0_=0.12,
         pi_1_1=0.28, pi_1_0=0.2, pi_0_1=0.08, Seed=513, Monotonicity=c("General"),
         Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

#summary(simsts_Bin_ICA502)
plot(simsts_Bin_ICA502)
df_median_0850<-simsts_Bin_ICA$R2_H
head(df_median_0850)
quantile(df_mean_0850,probs = seq(0, 1, 0.05) )
write.table(df_median_0850,file = "df_median_0850.csv")


###Scenario 2A :  501c-Binary-Binary endpoint###
# Make Tand S binary 75% dichotomisation
simsts501c_Bin <- simsts501
simsts501c_Bin$True_Bin <- simsts501c_Bin$True
simsts501c_Bin$True_Bin[simsts501c_Bin$True>0.59] <- 1
simsts501c_Bin$True_Bin[simsts501c_Bin$True<=0.59] <- 0
simsts501c_Bin$Surr_Bin <- simsts501c_Bin$Surr
simsts501c_Bin$Surr_Bin[simsts501c_Bin$Surr>0.41] <- 1
simsts501c_Bin$Surr_Bin[simsts501c_Bin$Surr<=0.41] <- 0
head(simsts501c_Bin)
## Obtaining counterfactual probabilities#
MarginalProbs(Dataset=simsts501c_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_con2c<-matrix(c(6,6,3,10), nrow = 2, ncol = 2)
OR_table_con2c
oddsratio.wald(OR_table_con2c)
OR_table_treat2c<-matrix(c(11,2,6,6), nrow = 2, ncol = 2)
OR_table_treat2c
oddsratio.wald(OR_table_treat2c)
### RH2-3 ICABinBin percentile dichtomization
start.time <- Sys.time()
simsts_Bin_ICA503 <- ICA.BinBin(pi1_1_=0.2, pi0_1_=0.12, pi1_0_=0.08,
 pi_1_1=0.12, pi_1_0=0.12, pi_0_1=0.08, Seed=514, Monotonicity=c("General"),
  Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

#summary(simsts_Bin_ICA503)
plot(simsts_Bin_ICA503)
```

```
df_per_0850<-simsts_Bin_ICA$R2_H
head(df_per_0850)
quantile(df_per_0850,probs = seq(0, 1, 0.05) )
write.table(df_per_0850,file = "df_per_0850.csv")
start.time <- Sys.time()
modsts50c_Bin_ICA <- ICA.BinBin.Grid.Sample(pi1_1_=0.24, pi1_0_=0.04,
 pi_1_1=0.12, pi_1_0=0.12, pi0_1_=0.12, pi_0_1=0.08,
 Monotonicity=c("General"), M=100000, Seed=331)

Sys.time() - start.time
summary(modsts50c_Bin_ICA)
############################################################
# saved dataset for contcont50 N=50 and adjusted association 0.0#
Sim.Data.STS(N.Total=50, R.Indiv.Target=0.0, Means=c(0, 0, 0, 0), Seed=522)
head(Data.Observed.STS)
simsts502<-Data.Observed.STS
#write.table(simsts502,file = "contcont0050.csv")
# saved dataset for contcont50 adjusted association 0.0#
# summmary statisitics
(mean_sim2b<-apply(simsts502[,1:2], MARGIN = 2, FUN = mean))
(min_sim2b<-apply(simsts502[,1:2], MARGIN = 2, FUN = min))
(max_sim2b<-apply(simsts502[,1:2], MARGIN = 2, FUN = max))
(var_sim2b<-apply(simsts501[,1:2], MARGIN = 2, FUN = var))
(sd_sim2b<-apply(simsts502[,1:2], MARGIN = 2, FUN = sd))
(median_sim2b<-apply(simsts502[,1:2], MARGIN = 2, FUN = median))
(quantiles_sim2b<-apply(simsts502[,1:2], MARGIN = 2, FUN = quantile))

# obtain Identifiable correlations
sts_surr_true502a <- Single.Trial.RE.AA(Dataset = simsts502, Surr = Surr, True = True,
               Treat = Treat, Pat.ID = Pat.ID, Seed = 511)
summary(sts_surr_true502a)
plot(sts_surr_true502a, Indiv.Level = TRUE, Trial.Level = FALSE)
# Scenario 2B rho for ICA.contcont
sts_ica_surr_true502a <- ICA.ContCont(T0S0 = 0.14, T1S1 = -0.05, S0S0 = 0.69,
    S1S1 = 0.68, T0T0 =  0.98, T1T1 = 0.80, T0T1 = seq(-1, 1, by = 0.1),
    T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(sts_ica_surr_true502a)
plot(sts_ica_surr_true502a)
plot(sts_ica_surr_true502a, Labels = TRUE)
plot(sts_ica_surr_true502a, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true502a, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true502a, Min = -1, Max = 0.9)
# Obtaining rho square distribution
rho5<-sts_ica_surr_true502a$ICA
str(rho5)
sqrho5<-(rho5)^2
sd(sqrho5)
summary(sqrho5)
```

```
Modes(sqrho5)
quantile(sqrho5,probs = seq(0, 1, 0.05) )

# Binary dichotomization for N=50 and adjusted correlation of 0.0##
### Scenario 2B RH2-1 mean split Binary-Binary endpoint###
# Make T and S binary mean dichotomisation
simsts502a_Bin <- simsts502
simsts502a_Bin$True_Bin <- simsts502a_Bin$True
simsts502a_Bin$True_Bin[simsts502a_Bin$True>0.27] <- 1
simsts502a_Bin$True_Bin[simsts502a_Bin$True<=0.27] <- 0
simsts502a_Bin$Surr_Bin <- simsts502a_Bin$Surr
simsts502a_Bin$Surr_Bin[simsts502a_Bin$Surr>-0.3] <- 1
simsts502a_Bin$Surr_Bin[simsts502a_Bin$Surr<=-0.3] <- 0
head(simsts502a_Bin)
#write.table(simsts501a_Bin,file = "Binbin08501a.csv")
# saved dataset for BinBin501a with individualassociation 0.0#
# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=simsts502a_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# scenario 2B RH2-1 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA502a <- ICA.BinBin(pi1_1_=0.28, pi0_1_=0.2, pi1_0_=0.28,
    pi_1_1=0.24, pi_1_0=0.2, pi_0_1=0.32, Seed=513, Monotonicity=c("General"),
    Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA502a)

### scenario 2B RH2-2 501b-Binary-Binary endpoint Median split with AA= 0.0##
# Make Tand S binary median dichotomization
simsts502b_Bin <- simsts502
simsts502b_Bin$True_Bin <- simsts502b_Bin$True
simsts502b_Bin$True_Bin[simsts502b_Bin$True>0.45] <- 1
simsts502b_Bin$True_Bin[simsts502b_Bin$True<=0.45] <-0
simsts502b_Bin$Surr_Bin <- simsts502b_Bin$Surr
simsts502b_Bin$Surr_Bin[simsts502b_Bin$Surr>-0.29] <-1
simsts502b_Bin$Surr_Bin[simsts502b_Bin$Surr<=-0.29] <- 0
head(simsts502b_Bin)

#3 Obtain counterfactual probabilities
MarginalProbs(Dataset=simsts502b_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# RH2-2 Obtaininmg the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA502b <- ICA.BinBin(pi1_1_=0.2, pi0_1_=0.2, pi1_0_=0.32,
         pi_1_1=0.24, pi_1_0=0.24, pi_0_1=0.32, Seed=513, Monotonicity=c("General"),
        Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time #
```

```
summary(simsts_Bin_ICA502b)

### Scenario 2B RH2-3 percentile split-Binary-Binary endpoint###
# Make Tand S binary 75% dichotomisation
simsts502c_Bin <- simsts502
simsts502c_Bin$True_Bin <- simsts502c_Bin$True
simsts502c_Bin$True_Bin[simsts502c_Bin$True>0.91] <- 1
simsts502c_Bin$True_Bin[simsts502c_Bin$True<=0.91] <- 0
simsts502c_Bin$Surr_Bin <- simsts502c_Bin$Surr
simsts502c_Bin$Surr_Bin[simsts502c_Bin$Surr>0.15] <- 1
simsts502c_Bin$Surr_Bin[simsts502c_Bin$Surr<=0.15] <- 0
head(simsts502c_Bin)
# Obtain counterfacual probabilities
MarginalProbs(Dataset=simsts502c_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# scenario 2B RH2-3 ICA.BinBin
start.time <- Sys.time()
simsts_Bin_ICA503c <- ICA.BinBin(pi1_1_=0.0, pi0_1_=0.16, pi1_0_=0.24,
 pi_1_1=0.08, pi_1_0=0.2, pi_0_1=0.28, Seed=514, Monotonicity=c("General"),
 Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA503c)
#############################################################################
# saved dataset for contcont50 N=50 and adjusted association 0.4#
Sim.Data.STS(N.Total=50, R.Indiv.Target=0.4, Means=c(0, 0, 0, 0), Seed=532)
head(Data.Observed.STS)
simsts503<-Data.Observed.STS
# saved dataset for contcont50 adjusted association 0.4#
#write.table(simsts503,file = "contcont0450.csv")
## summary statisitics
(mean_sim2c<-apply(simsts503[,1:2], MARGIN = 2, FUN = mean))
(min_sim2c<-apply(simsts503[,1:2], MARGIN = 2, FUN = min))
(max_sim2c<-apply(simsts503[,1:2], MARGIN = 2, FUN = max))
(var_sim2c<-apply(simsts503[,1:2], MARGIN = 2, FUN = var))
(sd_sim2c<-apply(simsts503[,1:2], MARGIN = 2, FUN = sd))
(median_sim2c<-apply(simsts503[,1:2], MARGIN = 2, FUN = median))
(quantiles_sim2c<-apply(simsts503[,1:2], MARGIN = 2, FUN = quantile))
## Obtain indentiable correlations
sts_surr_true50_04 <- Single.Trial.RE.AA(Dataset = simsts503, Surr = Surr, True = True,
        Treat = Treat, Pat.ID = Pat.ID, Seed = 511)
summary(sts_surr_true50_04)
plot(sts_surr_true50_04, Indiv.Level = TRUE, Trial.Level = FALSE)
##### Scenario 2c Rho ICA cont-cont###
sts_ica_surr_true50_04 <- ICA.ContCont(T0S0 = 0.38, T1S1 = 0.51, S0S0 = 0.91,
   S1S1 = 1.68, T0T0 =  1.37, T1T1 = 1.01, T0T1 = seq(-1, 1, by = 0.1),
    T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
```

```
summary(sts_ica_surr_true50_04)
plot(sts_ica_surr_true50_04)
plot(sts_ica_surr_true50_04, Labels = TRUE)
plot(sts_ica_surr_true50_04, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true50_04, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true50_04, Min = -1, Max = 0.9)
## obtain rho-square distribution
rho6<-sts_ica_surr_true50_04$ICA
sqrho6<-(rho6)^2
str(rho6)
sd(sqrho6)
Modes(sqrho6)
summary(sqrho6)
hist( sqrho6)
quantile(sqrho6,probs = seq(0, 1, 0.05) )

# Scenario 2C Binary dichotomization for N=50 and adjusted correlation of 0.4##
### 503a-Binary-Binary endpoint###
# Make T and S binary mean dichotomisation
simsts503a_Bin <- simsts503
simsts503a_Bin$True_Bin <- simsts503a_Bin$True
simsts503a_Bin$True_Bin[simsts503a_Bin$True>0.07] <- 1
simsts503a_Bin$True_Bin[simsts503a_Bin$True<=0.07] <- 0
simsts503a_Bin$Surr_Bin <- simsts503a_Bin$Surr
simsts503a_Bin$Surr_Bin[simsts503a_Bin$Surr>0.13] <- 1
simsts503a_Bin$Surr_Bin[simsts503a_Bin$Surr<=0.13] <- 0
head(simsts503a_Bin)

# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=simsts503a_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# RH2-1 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA3 <- ICA.BinBin(pi1_1_=0.4, pi0_1_=0.2, pi1_0_=0.12,
         pi_1_1=0.32, pi_1_0=0.12, pi_0_1=0.12, Seed=513, Monotonicity=c("General"),
        Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA3)
### Scenario 2C RH2-2 Median split with adjusted association 0.4
# Make Tand S binary median dichotomisation
simsts503b_Bin <- simsts503
simsts503b_Bin$True_Bin <- simsts503b_Bin$True
simsts503b_Bin$True_Bin[simsts503b_Bin$True>-0.09] <- 1
simsts503b_Bin$True_Bin[simsts503b_Bin$True<=-0.09] <-0
simsts503b_Bin$Surr_Bin <- simsts503b_Bin$Surr
simsts503b_Bin$Surr_Bin[simsts503b_Bin$Surr>0.2] <-1
simsts503b_Bin$Surr_Bin[simsts503b_Bin$Surr<=0.2] <- 0
```

```
head(simsts503b_Bin)

#Obtain counterfactual probabilities
MarginalProbs(Dataset=simsts503b_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# RH2-2 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA503b <- ICA.BinBin(pi1_1_=0.4, pi0_1_=0.2, pi1_0_=0.12,
    pi_1_1=0.28, pi_1_0=0.2, pi_0_1=0.08, Seed=513, Monotonicity=c("General"),
    Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA503b)
####Scenario 2C 75 percentile split 503c-Binary-Binary endpoint###
# Make Tand S binary 75% dichotomization
simsts503c_Bin <- simsts503
simsts503c_Bin$True_Bin <- simsts503c_Bin$True
simsts503c_Bin$True_Bin[simsts503c_Bin$True>0.70] <- 1
simsts503c_Bin$True_Bin[simsts503c_Bin$True<=0.70] <- 0
simsts503c_Bin$Surr_Bin <- simsts503c_Bin$Surr
simsts503c_Bin$Surr_Bin[simsts503c_Bin$Surr>0.77] <- 1
simsts503c_Bin$Surr_Bin[simsts503c_Bin$Surr<=0.77] <- 0
head(simsts503c_Bin)
# Obtain counterfactual probabilities
MarginalProbs(Dataset=simsts503c_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
## RH2-3 75% percentile split
start.time <- Sys.time()
simsts_Bin_ICA503c <- ICA.BinBin(pi1_1_=0.08, pi0_1_=0.2, pi1_0_=0.08,
    pi_1_1=0.2, pi_1_0=0.16, pi_0_1=0.04, Seed=514, Monotonicity=c("General"),
    Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time

summary(simsts_Bin_ICA503c)

#########################################################################

########## Sample size small N=200####################################
Sim.Data.STS(N.Total=200, R.Indiv.Target=0.8, Means=c(0, 0, 0, 0), Seed=212)
head(Data.Observed.STS)
simsts2001<-Data.Observed.STS
# saved dataset for contcont200 adjusted association 0.8#
#write.table(simsts2001,file = "contcont08200.csv")
(mean_sim3a<-apply(simsts2001[,1:2], MARGIN = 2, FUN = mean))
(min_sim3a<-apply(simsts2001[,1:2], MARGIN = 2, FUN = min))
(max_sim3a<-apply(simsts2001[,1:2], MARGIN = 2, FUN = max))
(var_sim3a<-apply(simsts2001[,1:2], MARGIN = 2, FUN = var))
(sd_sim3a<-apply(simsts2001[,1:2], MARGIN = 2, FUN = sd))
```

```
(median_sim3a<-apply(simsts2001[,1:2], MARGIN = 2, FUN = median))
(quantiles_sim3a<-apply(simsts2001[,1:2], MARGIN = 2, FUN = quantile))

# Scenario 3A rho- Pearson correlation
## Obtain indentifiable correlations
sts_surr_true200a <- Single.Trial.RE.AA(Dataset = simsts2001, Surr = Surr, True = True,
                                        Treat = Treat, Pat.ID = Pat.ID, Seed = 511)
summary(sts_surr_true200a)
plot(sts_surr_true200a, Indiv.Level = TRUE, Trial.Level = FALSE)
### Rho scenario 3A
sts_ica_surr_true200a <- ICA.ContCont(T0S0 = 0.76, T1S1 = 0.77, S0S0 = 1.00,
    S1S1 =1.07, T0T0 =  0.87, T1T1 = 1.14, T0T1 = seq(-1, 1, by = 0.1),
T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(sts_ica_surr_true200a)
plot(sts_ica_surr_true200a)
plot(sts_ica_surr_true200a, Labels = TRUE)
plot(sts_ica_surr_true200a, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true200a, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true200a, Min = -1, Max = 0.9)
# Obtaining rho square distribution
rho7<-sts_ica_surr_true200a$ICA
str(rho7)
sqrho7<-(rho7)^2
sd(sqrho7)
Modes(sqrho7)
summary(sqrho7)
quantile(sqrho7,probs = seq(0, 1, 0.05) )

# RH2-1 Mean split Binary dichotomization for N=200 and adjusted correlation of 0.8##
### 2001a-Binary-Binary endpoint###
# Make T and S binary mean dichotomization
simsts201a_Bin <- simsts2001
simsts201a_Bin$True_Bin <- simsts201a_Bin$True
simsts201a_Bin$True_Bin[simsts201a_Bin$True>-0.015] <- 1
simsts201a_Bin$True_Bin[simsts201a_Bin$True<=-0.015] <- 0
simsts201a_Bin$Surr_Bin <- simsts201a_Bin$Surr
simsts201a_Bin$Surr_Bin[simsts201a_Bin$Surr>-0.038] <- 1
simsts201a_Bin$Surr_Bin[simsts201a_Bin$Surr<=-0.038] <- 0
head(simsts201a_Bin)

# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=simsts201a_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_con31c<-matrix(c(36,7,15,42), nrow = 2, ncol = 2)
OR_table_con31c
oddsratio.wald(OR_table_con31c)
OR_table_treat31c<-matrix(c(42,16,6,36), nrow = 2, ncol = 2)
OR_table_treat31c
```

```
oddsratio.wald(OR_table_treat31c)
# RH2-1 Obtaininmg the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA201a <- ICA.BinBin(pi1_1_=0.42, pi0_1_=0.07, pi1_0_=0.15,
    pi_1_1=0.36, pi_1_0=0.06, pi_0_1=0.16, Seed=512, Monotonicity=c("General"),
    Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA201a)

### Scenario 3A-RH2-2 median split  201b-Binary-Binary endpoint with adjusted association 0.8
## ## Calculating OR_TOS0 and OR_T1S1
# Make Tand S binary median dichotomisation
simsts201b_Bin <- simsts2001
simsts201b_Bin$True_Bin <- simsts201b_Bin$True
simsts201b_Bin$True_Bin[simsts201b_Bin$True>-0.02] <- 1
simsts201b_Bin$True_Bin[simsts201b_Bin$True<=-0.02] <-0
simsts201b_Bin$Surr_Bin <- simsts201b_Bin$Surr
simsts201b_Bin$Surr_Bin[simsts201b_Bin$Surr>-0.015] <-1
simsts201b_Bin$Surr_Bin[simsts201b_Bin$Surr<=-0.015] <- 0
head(simsts201b_Bin)

# Obtaining counterfactual probabilities#
MarginalProbs(Dataset=simsts201b_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_TOS0 and OR_T1S1
OR_table_con32b<-matrix(c(36,7,15,42), nrow = 2, ncol = 2)
OR_table_con32b
oddsratio.wald(OR_table_con32b)
OR_table_treat32b<-matrix(c(43,14,6,37), nrow = 2, ncol = 2)
OR_table_treat32b
oddsratio.wald(OR_table_treat32b)
# Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA202b <- ICA.BinBin(pi1_1_=0.42, pi0_1_=0.07, pi1_0_=0.12,
     pi_1_1=0.28, pi_1_0=0.2, pi_0_1=0.08, Seed=513, Monotonicity=c("General"),
     Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA202b)

# Scenario 3A-RH2-2  75% dichotomisation adjusted association 0.8
simsts203c_Bin <- simsts2001
simsts203c_Bin$True_Bin <- simsts203c_Bin$True
simsts203c_Bin$True_Bin[simsts203c_Bin$True>0.73] <- 1
simsts203c_Bin$True_Bin[simsts203c_Bin$True<=0.73] <- 0
simsts203c_Bin$Surr_Bin <- simsts203c_Bin$Surr
simsts203c_Bin$Surr_Bin[simsts203c_Bin$Surr>0.69] <- 1
simsts203c_Bin$Surr_Bin[simsts203c_Bin$Surr<=0.69] <- 0
```

```
head(simsts203c_Bin)
# Obtain counterfactual probabilities
MarginalProbs(Dataset=simsts203c_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_con33c<-matrix(c(64, 8,11,17), nrow = 2, ncol = 2)
OR_table_con33c
oddsratio.wald(OR_table_con33c)
OR_table_treat33c<-matrix(c( 70,8,5,17), nrow = 2, ncol = 2)
OR_table_treat33c
oddsratio.wald(OR_table_treat33c)
# Obtain RH2-3 ICA.BinBin
start.time <- Sys.time()
#simsts_Bin_ICA203 <- ICA.BinBin(pi1_1_=0.17, pi0_1_=0.08, pi1_0_=0.11,
   pi_1_1=0.17, pi_1_0=0.05, pi_0_1=0.08, Seed=524, Monotonicity=c("General"),
   Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

#summary(simsts_Bin_ICA203)
##############################################################################
# Scenario 3B-saved dataset for contcont N=200 and adjusted association 0.0#
Sim.Data.STS(N.Total=200, R.Indiv.Target=0.0, Means=c(0, 0, 0, 0), Seed=222)
head(Data.Observed.STS)
simsts2002<-Data.Observed.STS
# Obtaining summary statisitics
(mean_sim3b<-apply(simsts2002[,1:2], MARGIN = 2, FUN = mean))
(min_sim3b<-apply(simsts2002[,1:2], MARGIN = 2, FUN = min))
(max_sim3b<-apply(simsts2002[,1:2], MARGIN = 2, FUN = max))
(var_sim3b<-apply(simsts2002[,1:2], MARGIN = 2, FUN = var))
(sd_sim3b<-apply(simsts2002[,1:2], MARGIN = 2, FUN = sd))
(median_sim3b<-apply(simsts2002[,1:2], MARGIN = 2, FUN = median))
(quantiles_sim3b<-apply(simsts2002[,1:2], MARGIN = 2, FUN = quantile))

# Obtaining identifiable correlations
sts_surr_true2002 <- Single.Trial.RE.AA(Dataset = simsts2002, Surr = Surr, True = True,
         Treat = Treat, Pat.ID = Pat.ID, Seed = 511)
summary(sts_surr_true2002)
plot(sts_surr_true2002, Indiv.Level = TRUE, Trial.Level = FALSE)
## Rho
sts_ica_surr_true2002 <- ICA.ContCont(T0S0 = 0.03, T1S1 =- 0.02, S0S0 = 1.04,
S1S1 =0.97, T0T0 =  0.78, T1T1 = 1.18, T0T1 = seq(-1, 1, by = 0.1),
T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(sts_ica_surr_true2002)
plot(sts_ica_surr_true2002)
plot(sts_ica_surr_true2002, Labels = TRUE)
plot(sts_ica_surr_true2002, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true2002, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true2002, Min = -1, Max = 0.9)
# Obtaining rho square distibution
```

```
rho8<-sts_ica_surr_true2002$ICA
str(rho8)
sqrho8<-(rho8)^2
sd(sqrho8)
Modes(sqrho8)
summary(sqrho8)
quantile(sqrho8,probs = seq(0, 1, 0.05) )


### Scenario 3B Mean split for N=200 and adjusted correlation of 0.0##
### 2002a-Binary-Binary endpoint###
# Make T and S binary mean dichotomisation
simsts202a_Bin <- simsts2002
simsts202a_Bin$True_Bin <- simsts202a_Bin$True
simsts202a_Bin$True_Bin[simsts202a_Bin$True>-0.015] <- 1
simsts202a_Bin$True_Bin[simsts202a_Bin$True<=-0.015] <- 0
simsts202a_Bin$Surr_Bin <- simsts202a_Bin$Surr
simsts202a_Bin$Surr_Bin[simsts202a_Bin$Surr>-0.015] <- 1
simsts202a_Bin$Surr_Bin[simsts202a_Bin$Surr<=-0.015] <- 0
head(simsts202a_Bin)

# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=simsts202a_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# RH2-2 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA202a <- ICA.BinBin(pi1_1_=0.27, pi0_1_=0.28, pi1_0_=0.24,
    pi_1_1=0.23, pi_1_0= 0.34, pi_0_1=0.22, Seed=212, Monotonicity=c("General"),
    Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

#summary(simsts_Bin_ICA202a)

### RH2-2 median -Binary-Binary endpoint with adjusted association 0.0
# Make T and S binary median dichotomization
simsts202b_Bin <- simsts2002
simsts202b_Bin$True_Bin <- simsts202b_Bin$True
simsts202b_Bin$True_Bin[simsts202b_Bin$True>0.085] <- 1
simsts202b_Bin$True_Bin[simsts202b_Bin$True<=0.085] <-0
simsts202b_Bin$Surr_Bin <- simsts202b_Bin$Surr
simsts202b_Bin$Surr_Bin[simsts202b_Bin$Surr>0.002] <-1
simsts202b_Bin$Surr_Bin[simsts202b_Bin$Surr<=0.002] <- 0
head(simsts202b_Bin)

# Obtaining counterfactual probabilities#

MarginalProbs(Dataset=simsts202b_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
```

```
# Obtaininmg the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA202b <- ICA.BinBin(pi1_1_=0.27, pi0_1_=0.22, pi1_0_=0.22,
 pi_1_1=0.19, pi_1_0=0.32, pi_0_1=0.26, Seed=523, Monotonicity=c("General"),
 Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA202b)

### 202c-Binary-Binary endpoint###
# Make Tand S binary 75% dichotomisation adjusted association 0.0
simsts202c_Bin <- simsts2002
simsts202c_Bin$True_Bin <- simsts202c_Bin$True
simsts202c_Bin$True_Bin[simsts202c_Bin$True>0.61] <- 1
simsts202c_Bin$True_Bin[simsts202c_Bin$True<=0.61] <- 0
simsts202c_Bin$Surr_Bin <- simsts202c_Bin$Surr
simsts202c_Bin$Surr_Bin[simsts202c_Bin$Surr>0.61] <- 1
simsts202c_Bin$Surr_Bin[simsts202c_Bin$Surr<=0.61] <- 0
head(simsts202c_Bin)
# Obtain counterfactual probabilities
MarginalProbs(Dataset=simsts202c_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
start.time <- Sys.time()
simsts_Bin_ICA202c <- ICA.BinBin(pi1_1_=0.09, pi0_1_=0.17, pi1_0_=0.17,
 pi_1_1=0.05, pi_1_0=0.2, pi_0_1=0.2, Seed=524, Monotonicity=c("General"),
Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA202c)
#####################################################################
# saved dataset for contcont5 N=200 and adjusted association 0.4#
Sim.Data.STS(N.Total=200, R.Indiv.Target=0.4, Means=c(0, 0, 0, 0), Seed=232)
head(Data.Observed.STS)
simsts2003<-Data.Observed.STS
# saved dataset for contcont200 adjusted association 0.4#
#write.table(simsts2003,file = "contcont04200.csv")
# Obtaining summary statistics
(mean_sim3c<-apply(simsts2003[,1:2], MARGIN = 2, FUN = mean))
(min_sim3c<-apply(simsts2003[,1:2], MARGIN = 2, FUN = min))
(max_sim3c<-apply(simsts2003[,1:2], MARGIN = 2, FUN = max))
(var_sim3c<-apply(simsts2003[,1:2], MARGIN = 2, FUN = var))
(sd_sim3c<-apply(simsts2003[,1:2], MARGIN = 2, FUN = sd))
(median_sim3c<-apply(simsts2003[,1:2], MARGIN = 2, FUN = median))
(quantiles_sim3c<-apply(simsts2003[,1:2], MARGIN = 2, FUN = quantile))
# Pearson correlation, adjusted correlation 0.4
sts_surr_true2003 <- Single.Trial.RE.AA(Dataset = simsts2003, Surr = Surr, True = True,
            Treat = Treat, Pat.ID = Pat.ID, Seed = 511)
summary(sts_surr_true2003)
```

```
plot(sts_surr_true2002, Indiv.Level = TRUE, Trial.Level = FALSE)
###  rho scenario 3C ICA.BinBin
sts_ica_surr_true2003 <- ICA.ContCont(T0S0 = 0.40, T1S1 =0.57, S0S0 = 1.47,
        S1S1 =1.11, T0T0 =  0.83, T1T1 = 1.91, T0T1 = seq(-1, 1, by = 0.1),
    T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(sts_ica_surr_true2003)
plot(sts_ica_surr_true2003)
plot(sts_ica_surr_true2003, Labels = TRUE)
plot(sts_ica_surr_true2003, Type = "CumPerc")
CausalDiagramContCont(sts_ica_surr_true2003, Min = 0.9, Max = 1)
CausalDiagramContCont(sts_ica_surr_true2003, Min = -1, Max = 0.9)
###Obtain rho square
rho9<-sts_ica_surr_true2003$ICA
str(rho9)
sqrho9<-(rho9)^2
sd(sqrho9)
Modes(sqrho9)
summary(sqrho9)
quantile(sqrho9,probs = seq(0, 1, 0.05) )

# Scenario 3C Mean split N=200 and adjusted correlation of 0.4##
### 203a-Binary-Binary endpoint###
# Make T and S binary mean dichotomization
simsts203a_Bin <- simsts2003
simsts203a_Bin$True_Bin <- simsts203a_Bin$True
simsts203a_Bin$True_Bin[simsts203a_Bin$True>0.07] <- 1
simsts203a_Bin$True_Bin[simsts203a_Bin$True<=0.07] <- 0
simsts203a_Bin$Surr_Bin <- simsts203a_Bin$Surr
simsts203a_Bin$Surr_Bin[simsts203a_Bin$Surr>0.05] <- 1
simsts203a_Bin$Surr_Bin[simsts203a_Bin$Surr<=0.05] <- 0
head(simsts203a_Bin)

# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=simsts203a_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# RH2-1 Obtaininmg the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA203a <- ICA.BinBin(pi1_1_=0.41, pi0_1_=0.11, pi1_0_=0.17,
 pi_1_1=0.33, pi_1_0= 0.13, pi_0_1=0.23, Seed=312, Monotonicity=c("General"),
    Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA203a)

###RH2-2 Median split with adjusted association 0.4
# Make T and S binary median dichotomisation
simsts203b_Bin <- simsts2003
simsts203b_Bin$True_Bin <- simsts203b_Bin$True
```

```
simsts203b_Bin$True_Bin[simsts203b_Bin$True>0.11] <- 1
simsts203b_Bin$True_Bin[simsts203b_Bin$True<=0.11] <-0
simsts203b_Bin$Surr_Bin <- simsts203b_Bin$Surr
simsts203b_Bin$Surr_Bin[simsts203b_Bin$Surr>0.15] <-1
simsts203b_Bin$Surr_Bin[simsts203b_Bin$Surr<=0.15] <- 0
head(simsts203b_Bin)

# Obtaining counterfactual probabilities#
MarginalProbs(Dataset=simsts203b_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA203b <- ICA.BinBin(pi1_1_=0.36, pi0_1_=0.12, pi1_0_= 0.19,
           pi_1_1=0.31, pi_1_0=0.14, pi_0_1=0.21, Seed=523, Monotonicity=c("General"),
         Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

#summary(simsts_Bin_ICA203b)


### RH2-3 75% dichotomisation -Binary-Binary endpoint###
# Make T and S binary 75% dichotomization adjusted association 0.4
simsts203c_Bin <- simsts2003
simsts203c_Bin$True_Bin <- simsts203c_Bin$True
simsts203c_Bin$True_Bin[simsts203c_Bin$True>0.62] <- 1
simsts203c_Bin$True_Bin[simsts203c_Bin$True<=0.62] <- 0
simsts203c_Bin$Surr_Bin <- simsts203c_Bin$Surr
simsts203c_Bin$Surr_Bin[simsts203c_Bin$Surr>0.80] <- 1
simsts203c_Bin$Surr_Bin[simsts203c_Bin$Surr<=0.80] <- 0
head(simsts203c_Bin)
##Obtaining marginal probabilities
MarginalProbs(Dataset=simsts203c_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA203c <- ICA.BinBin(pi1_1_=0.12, pi0_1_=0.15, pi1_0_=0.14,
        pi_1_1=0.1, pi_1_0=0.14, pi_0_1= 0.15, Seed=324, Monotonicity=c("General"),
         Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time

summary(simsts_Bin_ICA203c)
#####################################################################
####### ICA evaluation with Measurement Error###############
# Scenario 4A N:2000 and AA=0.8
modsts1<-simsts1
modsts1$Surr[1:100]<-modsts1$Surr[1:100]*5
modsts1$Surr[514:614]<-modsts1$Surr[514:614]*2
modsts1$Surr[1000:1160]<-modsts1$Surr[1000:1160]*3
```

```
modsts1$Surr[1901:2000]<-modsts1$Surr[1901:2000]*3
summary(modsts1$Surr)
summary(modsts1$True)
# Descriptive statisitics #
# summary with 23% measuremenrt error
(mean_sim<-apply(modsts1[,1:2], MARGIN = 2, FUN = mean))
(min_sim<-apply(modsts1[,1:2], MARGIN = 2, FUN = min))
(max_sim<-apply(modsts1[,1:2], MARGIN = 2, FUN = max))
(var_sim<-apply(modsts1[,1:2], MARGIN = 2, FUN = var))
(sd_sim<-apply(modsts1[,1:2], MARGIN = 2, FUN = sd))
(median_sim<-apply(modsts1[,1:2], MARGIN = 2, FUN = median))
(quantile_sim<-apply(modsts1[,1:2], MARGIN = 2, FUN = quantile))
# Measurement error descriptive stats per treatment group
(mean_simcont2<-apply(modsts1[modsts1$Treat==-1,1:2], MARGIN = 2, FUN = mean))
(mean_simtreat2<-apply(modsts1[modsts1$Treat==1,1:2], MARGIN = 2, FUN = mean))
(min_simcont2<-apply(modsts1[modsts1$Treat==-1,1:2], MARGIN = 2, FUN = min))
(min_simtreat2<-apply(modsts1[modsts1$Treat==1,1:2], MARGIN = 2, FUN = min))
(max_simcont2<-apply(modsts1[modsts1$Treat==-1,1:2], MARGIN = 2, FUN = max))
(max_simtreat2<-apply(modsts1[modsts1$Treat==1,1:2], MARGIN = 2, FUN = max))
(var_simcont2<-apply(modsts1[modsts1$Treat==-1,1:2], MARGIN = 2, FUN = var))
(var_simtreat2<-apply(modsts1[modsts1$Treat==1,1:2], MARGIN = 2, FUN = var))
(var_simcont2<-apply(modsts1[modsts1$Treat==-1,1:2], MARGIN = 2, FUN = sd))
(var_simtreat2<-apply(modsts1[modsts1$Treat==1,1:2], MARGIN = 2, FUN = sd))
(median_simcont2<-apply(modsts1[modsts1$Treat==-1,1:2], MARGIN = 2, FUN = median))
(median_simtreat2<-apply(modsts1[modsts1$Treat==1,1:2], MARGIN = 2, FUN = median))
(quantile_simcont2<-apply(modsts1[modsts1$Treat==-1,1:2], MARGIN = 2, FUN = quantile))
(quantile_simtreat2<-apply(modsts1[modsts1$Treat==1,1:2], MARGIN = 2, FUN = quantile))

# Scenario 4A rho with measurment error
# Obtain identifiable  correlations#
mea_sts2000 <- Single.Trial.RE.AA(Dataset = modsts1, Surr = Surr, True = True,
                                      Treat = Treat, Pat.ID = Pat.ID, Seed = 111)
summary(mea_sts2000)
plot(mea_sts2000, Indiv.Level = TRUE, Trial.Level = FALSE)
# scenario 4A rho N:2000 and AA=0.8
mea_sts2000_ica <- ICA.ContCont(T0S0 =0.65, T1S1 = 0.63, S0S0 = 3.30,
     S1S1 =  3.00, T0T0 = 1.00, T1T1 = 1.01, T0T1 = seq(-1, 1, by = 0.1),
     T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(mea_sts2000_ica)
plot(mea_sts2000_ica)
plot(mea_sts2000_ica, Labels = TRUE)
plot(mea_sts2000_ica, Type = "CumPerc")
CausalDiagramContCont(mea_sts2000_ica, Min = 0.9, Max = 1)
CausalDiagramContCont(mea_sts2000_ica, Min = -1, Max = 0.9)
###obtain rho square
mearho1<-mea_sts2000_ica$ICA
summary(mearho1)
sqmearho1<-(mearho1)^2
```

```
summary(sqmearho1)
Modes(mearho1)
Modes(sqmearho1)
sd(sqmearho1)
str(sqmearho1)
quantile(sqrho1,probs = seq(0, 1, 0.05) )

## Dichotomization with measurement error
#scenario 4A mean  T and S binary-mean cut-off#
modsts1_Bin <- modsts1
modsts1_Bin$True_Bin <- modsts1_Bin$True
modsts1_Bin$True_Bin[modsts1_Bin$True>0.004] <- 1
modsts1_Bin$True_Bin[modsts1_Bin$True<=0.004] <- 0
modsts1_Bin$Surr_Bin <- modsts1_Bin$Surr
modsts1_Bin$Surr_Bin[modsts1_Bin$Surr>0.004] <- 1
modsts1_Bin$Surr_Bin[modsts1_Bin$Surr<=0.004] <- 0
head(modsts1_Bin)
# Obtaining counterfactual probabilities
MarginalProbs(Dataset=modsts1_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_conma1<-matrix(c(383,109,110,398), nrow = 2, ncol = 2)
OR_table_conma1
oddsratio.wald(OR_table_conma1)
OR_table_treatma1<-matrix(c(408,113,109,370), nrow = 2, ncol = 2)
OR_table_treatma1
oddsratio.wald(OR_table_treatma1)
#RH2-1 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
modsts_Bin_ICA21 <- ICA.BinBin(pi1_1_=0.398, pi0_1_=0.109, pi1_0_=0.11,
     pi_1_1=0.37, pi_1_0=0.109, pi_0_1=0.113, Seed=212, Monotonicity=c("General"),
        Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time

summary(modsts_Bin_ICA21)
plot(modsts_Bin_ICA21)
df_meanmeas_082000<-modsts_Bin_ICA21$R2_H
head(df_meanmeas_082000)
quantile(df_meanmeas_082000,probs = seq(0, 1, 0.05) )
write.table(df_meanmeas_082000,file = "df_meamea_082000.csv")
###RH2-2 Measurement error 2-Binary-Binary endpoint-median cut-off###
# Make T and Sbinary
modsts2_Bin <- modsts1
modsts2_Bin$True_Bin <- modsts2_Bin$True
modsts2_Bin$True_Bin[modsts2_Bin$True>-0.013] <- 1
modsts2_Bin$True_Bin[modsts2_Bin$True<=-0.013] <- 0
modsts2_Bin$Surr_Bin <- modsts2_Bin$Surr
modsts2_Bin$Surr_Bin[modsts2_Bin$Surr>-0.006] <- 1
modsts2_Bin$Surr_Bin[modsts2_Bin$Surr<=-0.006] <- 0
```

```
head(modsts2_Bin)
View(modsts2_Bin)
MarginalProbs(Dataset=modsts2_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_TOS0 and OR_T1S1
OR_table_conma2<-matrix(c(379,108,112,401), nrow = 2, ncol = 2)
OR_table_conma2
oddsratio.wald(OR_table_conma2)
OR_table_treatma2<-matrix(c(402,111,109,378), nrow = 2, ncol = 2)
OR_table_treatma2
oddsratio.wald(OR_table_treatma2)
#RH2-2 Obtaining the ICA-BIN-BIN#
start.time <- Sys.time()
modsts_Bin_ICA2 <- ICA.BinBin(pi1_1_=0.401, pi0_1_=0.108, pi1_0_=0.112,
     pi_1_1=0.378, pi_1_0=0.109, pi_0_1=0.111, Seed=222, Monotonicity=c("General"),
            Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time

summary(modsts_Bin_ICA2)
plot(modsts_Bin_ICA2)
df_medmeas_082000<-modsts_Bin_ICA2$R2_H
head(df_medmeas_082000)
quantile(df_medmeas_082000,probs = seq(0, 1, 0.05) )
write.table(df_medmeas_082000,file = "df_medmea_082000.csv")

###Scenario 4A RH2-3 Measurement Error Make T and S binary-75% percentile cut-off#

modsts3_Bin <- modsts1
modsts3_Bin$True_Bin <- modsts3_Bin$True
modsts3_Bin$True_Bin[modsts3_Bin$True>0.65] <- 1
modsts3_Bin$True_Bin[modsts3_Bin$True<=0.65] <- 0
modsts3_Bin$Surr_Bin <- modsts3_Bin$Surr
modsts3_Bin$Surr_Bin[modsts3_Bin$Surr>0.66] <- 1
modsts3_Bin$Surr_Bin[modsts3_Bin$Surr<=0.66] <- 0
head(modsts3_Bin)
# Obtaining counterfactual probabilities
MarginalProbs(Dataset=modsts3_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_TOS0 and OR_T1S1
# RH2-3 Obtaining the ICA-BIN-BIN#
#start.time <- Sys.time()
modsts_Bin_ICA3 <- ICA.BinBin(pi1_1_=0.183, pi0_1_=0.111, pi1_0_= 0.077,
     pi_1_1=0.171, pi_1_0=0.077, pi_0_1=0.105, Seed=322, Monotonicity=c("General"),
         Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time

summary(modsts_Bin_ICA3)
plot(modsts_Bin_ICA3)
df_permeas_082000<-modsts_Bin_ICA2$R2_H
head(df_permeas_082000)
```

```
quantile(df_permeas_082000,probs = seq(0, 1, 0.05) )
write.table(df_permeas_082000,file = "df_permea_082000.csv")


#####################################################################
#####Scenario 4C Introducing  Measurement error with AA=0.8 N=50###############
modsts50<-simsts501
modsts50$Surr[1:4]<-modsts50$Surr[1:4]*5
modsts50$Surr[14:19]<-modsts50$Surr[14:19]*2
modsts50$Surr[47:50]<-modsts50$Surr[47:50]*3
summary(modsts50$Surr)
head(modsts50)
# ICA concont for sample size 50
# Obtaining identifiable correlations
modsts_surr_true50 <- Single.Trial.RE.AA(Dataset = modsts50, Surr = Surr, True = True,
                                      Treat = Treat, Pat.ID = Pat.ID, Seed = 555)
summary(modsts_surr_true50)
plot(sts_surr_true50a, Indiv.Level = TRUE, Trial.Level = FALSE)
#Scenario 4C ICA concont for sample size 50
modsts50_ica <- ICA.ContCont(T0S0 = 0.54, T1S1 = 0.64, S0S0 = 6.98,
         S1S1 = 1.84, T0T0 =  1.05, T1T1 = 0.73, T0T1 = seq(-1, 1, by = 0.1),
            T0S1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), S0S1 = seq(-1,1, by = 0.1))
summary(modsts50_ica)
# Obtaining rho-square
mearho2<-modsts50_ica$ICA
plot(modsts50_ica)
sqmearho2<-(mearho2)^2
sd(sqmearho2)
summary(sqmearho2)
Modes(sqmearho2)
quantile(sqmearho2,probs = seq(0, 1, 0.05) )

# Measurement Binary dichotomization for N=50 and adjusted correlation of 0.8##
### 50a-Binary-Binary endpoint###
# Make T and S binary mean dichotomisation
modsts50a_Bin <- modsts50
modsts50a_Bin$True_Bin <- modsts50a_Bin$True
modsts50a_Bin$True_Bin[modsts50a_Bin$True>-0.11] <- 1
modsts50a_Bin$True_Bin[modsts50a_Bin$True<=-0.11] <- 0
modsts50a_Bin$Surr_Bin <- modsts50a_Bin$Surr
modsts50a_Bin$Surr_Bin[modsts50a_Bin$Surr>-0.15] <- 1
modsts50a_Bin$Surr_Bin[modsts50a_Bin$Surr<=-0.15] <- 0

head(modsts50a_Bin)

# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=modsts50a_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
```

```
OR_table_conmb1<-matrix(c(6,6,3,10), nrow = 2, ncol = 2)
OR_table_conmb1
oddsratio.wald(OR_table_conmb1)
OR_table_treatmb1<-matrix(c(12,2,5,6), nrow = 2, ncol = 2)
OR_table_treatmb1
oddsratio.wald(OR_table_treatmb1)
# RH2-1 scenario 4C Obtaining the ICA-BIN-BIN#
#start.time <- Sys.time()
#modsts_Bin_ICA1 <- ICA.BinBin(pi1_1_=0.4, pi0_1_=0.24, pi1_0_=0.12,
     pi_1_1=0.24, pi_1_0=0.2, pi_0_1=0.08, Seed=513, Monotonicity=c("General"),
        Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time

summary(modsts_Bin_ICA1)

### Measurement error 50b-Binary-Binary endpoint with adjusted association 0.8
# Make Tand S binary median dichotomisation
modsts50b_Bin  <- modsts50
modsts50b_Bin$True_Bin <- modsts50b_Bin$True
modsts50b_Bin$True_Bin[modsts50b_Bin$True>-0.13] <- 1
modsts50b_Bin$True_Bin[modsts50b_Bin$True<=-0.13] <-0
modsts50b_Bin$Surr_Bin <- modsts50b_Bin$Surr
modsts50b_Bin$Surr_Bin[modsts50b_Bin$Surr>-0.13] <-1
modsts50b_Bin$Surr_Bin[modsts50b_Bin$Surr<=-0.13] <- 0
head(modsts50b_Bin)


MarginalProbs(Dataset=modsts50b_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
# Obtaininmg the ICA-BIN-BIN#
OR_table_conmb2<-matrix(c(6,6,3,10), nrow = 2, ncol = 2)
OR_table_conmb2
oddsratio.wald(OR_table_conmb2)
OR_table_treatmb2<-matrix(c(11,2,6,6), nrow = 2, ncol = 2)
OR_table_treatmb2
oddsratio.wald(OR_table_treatma2)
## RH2-2 scenario 4c median split
start.time <- Sys.time()
modst50bs_Bin_ICA2c <- ICA.BinBin(pi1_1_=0.4, pi0_1_=0.24, pi1_0_=0.12,
        pi_1_1=0.24, pi_1_0=0.24, pi_0_1=0.08, Seed=522, Monotonicity=c("General"),
            Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(modst50bs_Bin_ICA2c)

### Measurement error 50c-Binary-Binary endpoint###
# Make Tand S binary 75% dichotomisation
modsts50c_Bin  <- modsts50
```

```
modsts50c_Bin$True_Bin <- modsts50c_Bin$True
modsts50c_Bin$True_Bin[modsts50c_Bin$True>0.59] <- 1
modsts50c_Bin$True_Bin[modsts50c_Bin$True<=0.59] <- 0
modsts50c_Bin$Surr_Bin <- modsts50c_Bin$Surr
modsts50c_Bin$Surr_Bin[modsts50c_Bin$Surr>0.41] <- 1
modsts50c_Bin$Surr_Bin[modsts50c_Bin$Surr<=0.41] <- 0
head(modsts50c_Bin)
# Obtaining identifiable correlations
MarginalProbs(Dataset=modsts50c_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_conmc3<-matrix(c(15, 3,1,6), nrow = 2, ncol = 2)
OR_table_conmc3
oddsratio.wald(OR_table_conmc3)
OR_table_treatmc3<-matrix(c( 17,2,3,3), nrow = 2, ncol = 2)
OR_table_treatmc3
oddsratio.wald(OR_table_treatmc3)
start.time <- Sys.time()
modsts50c_Bin_ICA <- ICA.BinBin(pi1_1_=0.24, pi0_1_=0.12, pi1_0_=0.08,
      pi_1_1=0.12, pi_1_0=0.12, pi_0_1=0.08, Seed=514, Monotonicity=c("General"),
        Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(modsts50c_Bin_ICA)
############################################################
#Scenario 4B Measurement error N=200
modsts200<-simsts2001
modsts200$Surr[1:10]<-modsts200$Surr[1:10]*5
modsts200$Surr[54:64]<-modsts200$Surr[54:64]*2
modsts200$Surr[100:116]<-modsts200$Surr[100:116]*3
modsts200$Surr[191:200]<-modsts200$Surr[191:200]*3

summary(modsts200$True)
head(modsts200)
#summary messurment error
(mean_mod3a<-apply(modsts200[,1:2], MARGIN = 2, FUN = mean))
(min_mod3a<-apply(modsts200[,1:2], MARGIN = 2, FUN = min))
(max_mod3a<-apply(modsts200[,1:2], MARGIN = 2, FUN = max))
(var_mod3a<-apply(modsts200[,1:2], MARGIN = 2, FUN = var))
(sd_mod3a<-apply(modsts200[,1:2], MARGIN = 2, FUN = sd))
(median_mod3a<-apply(modsts200[,1:2], MARGIN = 2, FUN = median))
(quantiles_mod3a<-apply(modsts200[,1:2], MARGIN = 2, FUN = quantile))
# Measurement error Pearson correlation
modsts200_surr_true <- Single.Trial.RE.AA(Dataset =modsts200, Surr = Surr, True = True,
            Treat = Treat, Pat.ID = Pat.ID, Seed = 211)
summary(modsts200_surr_true)
plot(modsts200_surr_true, Indiv.Level = TRUE, Trial.Level = FALSE)
modsts200_ica <- ICA.ContCont(T0S0 =  0.62, T1S1 = 0.65, S0S0 = 2.56,
            S1S1 =5.06, T0T0 =  0.87, T1T1 = 1.14, T0T1 = seq(-1, 1, by = 0.1),
```

```
           TOS1 = seq(-1, 1, by = 0.1), T1S0 = seq(-1, 1, by = 0.1), SOS1 = seq(-1,1, by = 0.1))
summary(modsts200_ica)
mearho3<-modsts200_ica$ICA
sqmearho3<-(mearho3)^2
str(mearho3)
sd(sqmearho3)
Modes(sqmearho3)
summary(sqmearho3)
quantile(sqmearho3,probs = seq(0, 1, 0.05) )
plot(modsts200_ica)
plot(modsts200_ica, Labels = TRUE)
plot(modsts200_ica, Type = "CumPerc")
CausalDiagramContCont(modsts200_ica, Min = 0.9, Max = 1)
CausalDiagramContCont(modsts200_ica, Min = -1, Max = 0.9)
# Measurement error Binary dichotomization for N=200 and adjusted correlation of 0.8##
### 200a-Binary-Binary endpoint###
# Make T and S binary mean dichotomisation
modsts2001_Bin <- modsts200
modsts2001_Bin$True_Bin <- modsts2001_Bin$True
modsts2001_Bin$True_Bin[modsts2001_Bin$True>-0.015] <- 1
modsts2001_Bin$True_Bin[modsts2001_Bin$True<=-0.015] <- 0
modsts2001_Bin$Surr_Bin <- modsts2001_Bin$Surr
modsts2001_Bin$Surr_Bin[modsts2001_Bin$Surr>-0.04] <- 1
modsts2001_Bin$Surr_Bin[modsts2001_Bin$Surr<=-0.04] <- 0
head(modsts2001_Bin)
View(modsts2001_Bin)

# Obtaining the counterfactual probabilities#
MarginalProbs(Dataset=modsts2001_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_TOS0 and OR_T1S1
OR_table_conme1<-matrix(c(36,7,15,42), nrow = 2, ncol = 2)
OR_table_conme1
oddsratio.wald(OR_table_conme1)
OR_table_treatme1<-matrix(c(42,16,6,36), nrow = 2, ncol = 2)
OR_table_treatme1
oddsratio.wald(OR_table_treatme1)
# Obtaininmg the ICA-BIN-BIN#
start.time <- Sys.time()
simsts_Bin_ICA201a <- ICA.BinBin(pi1_1_=0.42, pi0_1_=0.07, pi1_0_=0.15,
         pi_1_1=0.36, pi_1_0=0.06, pi_0_1=0.16, Seed=512, Monotonicity=c("General"),
            Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(simsts_Bin_ICA201a)


summary(modsts_Bin_ICA201a)
#Measurement error  Make T and S binary median dichotomization
```

```
modsts2001b_Bin <- modsts200
modsts2001b_Bin$True_Bin <- modsts2001b_Bin$True
modsts2001b_Bin$True_Bin[modsts2001b_Bin$True>-0.02] <- 1
modsts2001b_Bin$True_Bin[modsts2001b_Bin$True<=-0.02] <-0
modsts2001b_Bin$Surr_Bin <- modsts2001b_Bin$Surr
modsts2001b_Bin$Surr_Bin[modsts2001b_Bin$Surr>-0.015] <-1
modsts2001b_Bin$Surr_Bin[modsts2001b_Bin$Surr<=-0.015] <- 0
head(modsts2001b_Bin)
# Obtaining counterfactual probabilities#

MarginalProbs(Dataset=modsts2001b_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_conme2<-matrix(c(36,7,15,42), nrow = 2, ncol = 2)
OR_table_conme2
oddsratio.wald(OR_table_conme2)
OR_table_treatme2<-matrix(c(43,14,6,37), nrow = 2, ncol = 2)
OR_table_treatme2
oddsratio.wald(OR_table_treatme2)
# Obtaininmg the ICA-BIN-BIN#
start.time <- Sys.time()
modsts200b_Bin_ICA <- ICA.BinBin(pi1_1_=0.42, pi0_1_=0.07, pi1_0_=0.15,
      pi_1_1=0.37, pi_1_0=0.06, pi_0_1=0.14, Seed=333, Monotonicity=c("General"),
                           Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time
#summary(modsts200b_Bin_ICA)

### Measurement error 202c-Binary-Binary endpoint###
# Make Tand S binary 75% dichotomisation adjusted association 0.8
modsts203c_Bin<- modsts200
modsts203c_Bin$True_Bin <- modsts203c_Bin$True
modsts203c_Bin$True_Bin[modsts203c_Bin$True>0.73] <- 1
modsts203c_Bin$True_Bin[modsts203c_Bin$True<=0.73] <- 0
modsts203c_Bin$Surr_Bin <- modsts203c_Bin$Surr
modsts203c_Bin$Surr_Bin[modsts203c_Bin$Surr>0.69] <- 1
modsts203c_Bin $Surr_Bin[modsts203c_Bin$Surr<=0.69] <- 0
head(modsts203c_Bin)

MarginalProbs(Dataset=modsts203c_Bin, Surr=Surr_Bin, True=True_Bin, Treat=Treat)
## ## Calculating OR_T0S0 and OR_T1S1
OR_table_conme3<-matrix(c(58, 14,9,19), nrow = 2, ncol = 2)
OR_table_conme3
oddsratio.wald(OR_table_conme3)
OR_table_treatme3<-matrix(c( 65,13,5,17), nrow = 2, ncol = 2)
OR_table_treatme3
oddsratio.wald(OR_table_treatme3)
#start.time <- Sys.time()
modsts203_Bin_ICA <- ICA.BinBin(pi1_1_=0.19, pi0_1_=0.14, pi1_0_=0.09,
   pi_1_1=0.17, pi_1_0=0.05, pi_0_1=0.13, Seed=334, Monotonicity=c("General"),
```

```
                              Sum_Pi_f = seq(from=0.01, to=.99, by=.01), M=10000)
Sys.time() - start.time # runtime

summary(modsts203_Bin_ICA)
```