

2016●2017

Faculty of Sciences

*Master of Statistics*

Master's Thesis

Evaluation of Surrogate Endpoints in Human Microbiome

Supervisor:

Prof. dr. Shekdy Ziv

Supervisor:

Dr. Van Der Elst, Wim

Olusoji Oluwafemi Daniel

*Thesis presented in fulfillment of the requirements for the degree of*

*Master of Statistics*

## Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Background &amp; Introduction</b>  | <b>3</b>  |
| 1.1      | The Human Microbiome . . . . .  | 3         |
| 1.2      | The Human Metabolism & Metabolic Disorder . . . . .                             | 4         |
| 1.3      | Surrogate Endpoints & its Evaluation . . . . .                                  | 4         |
| 1.4      | Aims & Objectives . . . . .   | 5         |
| 1.5      | Dataset Description, Source and Variables of Interest . . . . .                 | 5         |
| 1.6      | Notations . . . . .   | 5         |
| <b>2</b> | <b>Methods</b>  | <b>6</b>  |
| 2.1      | Earlier Methods for Surrogate Endpoint Evaluations and Their Pitfalls . . . . . | 6         |
| 2.2      | Joint Modelling Approach(JMA) . . . . .   | 8         |
| 2.3      | Causal Inference Approach (CIA) . . . . .                                       | 9         |
| 2.4      | Evaluating Use of Multiple Surrogates . . . . .                                 | 10        |
| 2.5      | Softwares and Tools . . . . .   | 10        |
| <b>3</b> | <b>Data Analysis</b>  | <b>11</b> |
| 3.1      | Exploratory Analysis . . . . .  | 11        |
| 3.2      | Estimation of Treatment Effects . . . . .                                       | 13        |
| 3.3      | Univariate Surrogate Evaluation . . . . .                                       | 15        |
| 3.3.1    | Joint Modelling Results . . . . .   | 15        |
| 3.3.2    | Causal Inference Results . . . . .  | 17        |
| 3.4      | Multiple Surrogates Evaluation . . . . .  | 19        |
| <b>4</b> | <b>Discussion &amp; Conclusion</b>  | <b>21</b> |

# 1 Background & Introduction

The sensitivity of some true endpoints (credible indicator of therapeutic response to an applied treatment<sup>9</sup>) as well as time taken for evaluation of treatment effect on these endpoints makes the search for surrogates (a biomarker intended to be used in replacement of true endpoint<sup>9</sup>) an important endeavour in modern medicine<sup>5</sup>. The use of surrogates actually isn't a wholly new idea<sup>14,19</sup>, however recent advances in genome sequencing technology which has enriched our understanding of human biology has led to the suggestion of several biomarkers (any objectively measured biological characteristic indicative of normal biological process<sup>9</sup>) that can be easily and cheaply measured as surrogates<sup>9</sup>. Notably, surrogates have led to early regulatory approval of some treatments<sup>6,7</sup>, but surrogates have failed on several other occasions<sup>9,14,15</sup>. One of such failure is the approval of zidovudine for HIV treatment. This treatment was approved by the regulatory board because of its effect on CD4 blood count which was used as a surrogate for time to clinical event and overall survival<sup>20</sup>. It was later observed that CD4 blood count was not a reliable predictor for both clinical endpoints of interest<sup>10</sup>. Another high profile incidence involving the use of surrogates, was the approval of three antiarrhythmic drugs (encainide, flecainide, and moricizine) by the U.S. Food and Drug Administration (FDA) based on their effects on arrhythmias suppression which was perceived as a surrogate for cardiac-related deaths<sup>9</sup>. This position was held because of the association between arrhythmias and cardiac complication related deaths. However, post marketing studies showed that these drugs rather increased the chances of cardiac related deaths compared to a placebo. In fact, death rate was almost four times higher in patients treated with anti-arrhythmias drug compared to patients treated with placebo<sup>11</sup>.

These high profile failures led to strong criticisms and huge skepticism, rightfully so, on the use of surrogate endpoints<sup>9</sup>. However, in the light of possible time shortening, cost reduction and early approval of needed treatments in clinical trials, makes the use of surrogates important<sup>5,9,35</sup>. This importance further highlights the need for critical validation of potential surrogates.

## 1.1 The Human Microbiome

The human microbiome is made up of trillions of bacteria cells in humans<sup>31</sup>. These cells range from symbiotic bacteria cells to parasitic ones and others that are neither of both<sup>31</sup>. Although their functions is not yet fully understood, they are associated with metabolism, immune function and human physiology<sup>4</sup>. The microbiome exist in supportive communities habiting from the human skin to the anus<sup>13</sup>. These communities are called microbiota while the cells and a collection of all their genes is referred to as the microbiome<sup>13,31</sup>. Although the Human Genome Project<sup>21</sup> really expanded our knowledge of the human microbiome, studies of the human microbiota dates back to 1680, a study where Antonie van Leewenhoek compared his oral and fecal micro biota<sup>31</sup>.

The human microbiota is majorly acquired through the environment and from parents, essentially the mother. The microbiota of babies given birth to vaginally, closely resembles the vagina mi-

crobiota, while the microbiota of those delivered through cesarean section closely resembles the skin microbiota<sup>31</sup>. Over the years, feeding habits, human interaction and contact with the environment, pets and other humans expand the microbiome as well as change its composition<sup>13,31</sup>.

As noted earlier, the full function of the microbiome is unknown, however scientists have been able to identify some essential functions of these bacteria colonies we carry about. They are actively involved in keep opportunist pathogens (who are members of the microbiome) at bay making them an active layer of defense in our immune system<sup>3,13</sup>. Furthermore, the gut microbiota or microflora has been heavily linked to almost all process in our body<sup>13</sup>, however more and more evidence points at they playing critical role in nutrition and digestion<sup>3,13</sup>. Changes in gut microbiota has been associated with obesity, autism, type 1 diabetes, celiac disease and atherosclerosis<sup>3</sup>. Specifically, it was observed that diet restriction in humans to either carbohydrate based or protein based diet leads to a change in the gut microbiota composition while the subjects lost weight<sup>31</sup>. This finding also made it clear that a healthy subject is not necessarily carrying a healthy microbiome, because the microbiome will continue to change its composition<sup>3</sup>. However, the ability of the microbiota community to maintain certain biological pathways and genes including the ability to remain stable in terms of biological function is essential to the health of the microbiome, since microbiota community essentially changes from one subject to the other but certain biological functions, genes and pathways are maintained<sup>13</sup>.

## 1.2 The Human Metabolism & Metabolic Disorder

### 1.3 Surrogate Endpoints & its Evaluation

A surrogate endpoint is an endpoint, usually a biomarker intended to replace the most clinically relevant primary endpoint<sup>1,5,9</sup>. Surrogate endpoints are used with the aim of measuring treatment effects earlier than it should have been, if the most clinically relevant primary endpoint was used<sup>9</sup>. Moreover, surrogates can be more sensitive to treatments compared to the clinically relevant primary endpoint<sup>1,5,9</sup>.

Essentially the the idea of surrogate endpoints is appealing but their evaluation is not a trivial task<sup>5</sup>. The non-triviality of the task of searching for and consequently validating a surrogate endpoint has not deferred research tailored at finding and developing methods for validating these endpoints<sup>1,5,35</sup>. The enormous advantage of shortening time required for a clinical trial, getting early approval for novel treatments, reducing the chances of having missing data, early safety and toxic signs amongst other benefits implies that surrogates will always be proposed, hence credible validation techniques are required<sup>9</sup>.

Early failures in studies involving use of surrogates was largely due to the fact that surrogacy was deduced from association between the proposed surrogate and the desired true endpoint<sup>9</sup>. However, this belief points in the direction of "correlation equals causation", which essentially

is not true, in the words of Fleming, "A Correlate does not a Surrogate Make"<sup>9,15,16</sup>. The idea of a correlate makes a surrogate is very natural, appealing and desirable for a true surrogate<sup>32</sup>, but more importantly the treatment effect on the desired true endpoint must be predicted as accurately as possible by the treatment effect on the surrogate<sup>5,9</sup>. Evaluating this concept was lacking in earlier use of surrogates and is largely responsible for the failures experienced, however, this gap in literature is being filled and still remains an active field of research over the past decade.

### 1.4 Aims & Objectives

Results from Vrieze A. et al, 2012<sup>37</sup> showed increase in insulin sensitivity as well as significant changes in elements of gut microbiota of patients suffering from metabolic disorder, that were treated with allogenic gut microbiota infusion. This support accumulating evidence (similar studies carried out using animal models<sup>12,22</sup>) that gut microbiota is associated with human metabolism. This association can be exploited for medical and clinical benefits if elements of the gut microbiota can be validated as surrogates for insulin sensitivity. Therefore, the chief aim of this thesis is to validate elements of the human microbiota, specifically the gut microbiota as potential surrogates for insulin sensitivity in patients with metabolic disorder. Individual elements of the gut microbiota (30 different bacteria cells) as well as combinations of these elements will be validated as surrogates. This study is essential since the full function of the human microbiome is unknown, hence being able to ascertain that some elements of the microbiome can serve as surrogates for insulin sensitivity will not only help in identifying one of the functions of the gut microbiome, it will help shed more light about the relationship between specific elements of the gut microbiome and the human host metabolism.

### 1.5 Dataset Description, Source and Variables of Interest

The dataset used for this research is simulated based on the trial designed and results obtained in Vrieze A. et al, 2012<sup>37</sup>. In other words, a trial involving 16 patients with metabolic disorder randomized between receiving allogenic and autologous gut microbiota infusion was simulated. The dataset contains measurements on insulin sensitivity at baseline and 6 months after treatment, difference in counts at baseline and 6 months after treatment, of 30 bacteria cells that are part of the gut microbiota.

The endpoint of interest (true endpoint) is the difference in insulin sensitivity at baseline and 6 months after treatment, while the difference in bacteria counts at baseline and 6 months will be used as surrogates.

### 1.6 Notations

The following notations will be used henceforth;

- $T$  = true endpoint, i.e. difference in insulin sensitivity

- $S_{ij}$  = surrogate endpoints, i.e. difference in bacteria counts,  $i = 1, \dots, 30$   $j = 1, 2$ ,
- $Z_j$  = Treatments, which are own feces or autologous gut microbiota infusion (control), donor feces or allogenic gut microbiota infusion (experimental),  $j = 1, 2$ .

Throughout the main body of this report, surrogates will be referred to as numbers, i.e.  $1, 2, \dots, 30$ , however the full names of each surrogate is presented in the appendix.

## 2 Methods

As identified in the background to this thesis, a gap in reasoning and methods for evaluating surrogate endpoints is responsible for earlier failures in use of surrogates. However, this gap in methodology is being filled and there are a plethora of methods to choose from presently, but as it is with most methods in statistics, each do have their pitfalls and criticisms.

Generally, surrogates evaluation methods can be broadly classified into two; Single Trial Approach (STA) and Multiple Trial Approach (MTA). Because the data for this research is simulated based on results from a single trial, only the STA methods will be considered, readers interested in MTA methods are therefore directed to Buurzykowski et al<sup>5</sup> and Alonso et al<sup>1</sup> for details on these methods.

### 2.1 Earlier Methods for Surrogate Endpoint Evaluations and Their Pitfalls

One of the earliest evaluation methods for surrogate endpoint dates back to Prentice, 1989<sup>1,5,28</sup>. The method involves a surrogate satisfying a set of conditions which are;

1. treatment must have a significant effect on surrogate, i.e.  $f(S|Z) \neq f(S)$ ,
2. treatment must have a significant effect on true endpoint, i.e.  $f(T|Z) \neq f(T)$ ,
3. surrogate must have a significant effect on true endpoint, i.e.  $f(T|S) \neq f(T)$ ,
4. effect of treatment on true endpoint should be captured by the surrogate, i.e.  $f(T|S, Z) = f(T|S)$

where  $f(\cdot)$  indicates the probability distribution of the variable in question. The criteria essentially boils down to requiring  $S$  to capture any existing relationship between the  $Z$  and  $T$ <sup>10</sup>. The first two criteria can be evaluated by fitting either two separate regression models or fitting the following bivariate linear regression model;

$$S_j = \mu_S + \alpha Z_j + \epsilon_{Sj} \quad (1)$$

$$T_j = \mu_T + \beta Z_j + \epsilon_{Tj} \quad (2)$$

where  $\alpha$  and  $\beta$  are treatment effect on S and T respectively,  $\begin{pmatrix} \epsilon_{Sj} \\ \epsilon_{Tj} \end{pmatrix} \sim N(0, \Sigma)$ ,  $0^T = (0, 0)$ ,  $\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}$  provided both endpoints are normally distributed<sup>1,9</sup>. The third and fourth criteria can be evaluated by fitting the following linear regression models;

$$T_j = \mu + \gamma S_j + \epsilon_j \quad (3)$$

$$T_j = \tilde{\mu}_T + \beta_S Z_j + \gamma_Z S_j + \tilde{\epsilon}_{Tj} \quad (4)$$

where the null hypothesis  $H_0 : \beta_S = 0$  is of interest in the last equation. The fourth criteria requires this hypothesis not to be rejected which is one of the major drawback as well as criticism for this approach to validating surrogates. The fourth criteria can never really be ascertained for several reasons amongst which are; poorly designed study or lack of power<sup>5,9</sup>, the other criticism of this approach arises from the fact that a statistical evaluation of the fourth criteria does not show that the effect of  $Z$  on  $T$  is fully captured by  $S$ , even if the study is adequately powered<sup>5,9</sup>.

In light of the drawbacks of the Prentice approach led to the development of the proportion explained approach<sup>17</sup>. It is an approach that measures surrogacy by comparing relatively, the effect of  $Z$  on  $T$  to the effect of  $Z$  on  $T$  adjusting for  $S$ <sup>5,9</sup>. In other words, proportion explained measures the proportion of the effect of  $Z$  on  $T$  that is captured by  $S$ <sup>9</sup>. The proportion explained is defined as;

$$PE(T, S, Z) = 1 - \frac{\beta_S}{\beta} \quad (5)$$

where  $\beta$  and  $\beta_S$  remain as defined previously. This setting implies that a perfect surrogate would have  $PE = 1$  and the farer PE is to 1, the worse the surrogate. In fact, if the effect of  $Z$  on  $T$  is entirely due to  $S$ , then  $\beta_S = 0$  and  $PE = 1$ , while if it is the other way round,  $PE = 0$  because  $\beta = \beta_S$ <sup>32</sup> which implies  $0 \leq PE \leq 1$ . This intuition is however a problem because it is entirely not true that  $\beta_S = 0$  when the effect of  $Z$  on  $T$  is fully captured by  $S$ , nor  $\beta = \beta_S$ <sup>32</sup> necessarily in the opposite as well. This imply that the assertion that PE is constrained between 0 and 1 is not necessarily true<sup>32,36</sup>. Furthermore, the uncertainty about PE tends to be high due to wide confidence intervals<sup>10</sup>, and PE is also shown not to be mathematically acceptable<sup>9,24</sup>.

Two other quantities were further proposed in response to the problems identified over the use of Freedman's proportion explained. They are the relative effect and adjusted association<sup>8</sup>. The relative effect compares relatively, the effect of  $Z$  on  $T$  to its effect on  $S$ . It is defines as;

$$RE = \frac{\beta}{\alpha} \quad (6)$$

where  $\beta$  and  $\alpha$  remains as before. This measure of surrogacy clearly allows for the prediction of the effect of  $Z$  on  $T$  from the effect of  $Z$  on  $S$ , but this requires the assumption that its value remains constant from trial to trial<sup>32</sup>. This assumption is not only strong, it is rather unverifiable if

the data available comes from a single clinical trial. The problem of multiple trial data needed for validation of RE becomes clearer, if (6) is rewritten in a linear regression form  $\beta = \Delta + RE\alpha + \epsilon$  with  $\epsilon \sim N(0, \sigma^2)$ , assuming both  $S$  and  $T$  are normally distributed endpoints. this formulation makes it clear that multiple  $\alpha$ 's and  $\beta$ 's is needed to establish the accuracy of RE<sup>23</sup>. Another advantage with the RE is the fact that it also has a direct causal interpretation, in fact  $\alpha$  and  $\beta$  are direct average individual causal effects of  $Z$  on  $S$  and  $T$ <sup>32</sup>. Any of the methods discussed above can easily be implemented per surrogate, i.e. surrogate by surrogate analysis but extending them to a combination of surrogates isn't direct and clear except in the case of the adjusted association<sup>33</sup>. The discussion about adjusted association will be delayed postponed to the joint modelling subsection.

While this section gave an overview of how surrogate endpoints evaluation has evolved over time, the remaining next two sections will be devoted to presenting the methods applied in this thesis.

## 2.2 Joint Modelling Approach(JMA)

The joint modelling approach(JMA) to surrogate endpoint evaluation is majorly based on the use of adjusted association and relative effect to evaluate surrogacy<sup>1,5</sup>. In a pictorial sense, the problem of surrogate endpoint evaluation from a joint modelling perspective could be viewed as;

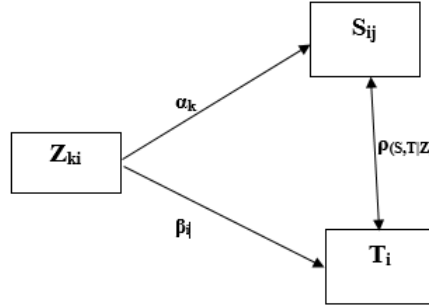


Figure 1: Model Framework

$\alpha$  and  $\beta$  remains as defined above but  $\rho_{S,T|Z}$  is the adjusted association. The model framework above was also used to depict the Quantitative Structure Transcription-Assay Relationship(QSTAR)<sup>1</sup>. Generally, the method involves fitting the bivariate linear regression model (1) and (2) which can as well be defined as;

$$\begin{pmatrix} S_{ij} \\ T_i \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu_S + \alpha_j X_i \\ \mu_T + \beta X_i \end{pmatrix}, \Sigma_j \right], \quad \Sigma_j = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix} \quad (7)$$

Equation 7 above is for a case of multiple surrogates. From the  $\Sigma$  matrix in (7), the adjusted association<sup>1,5</sup> can easily be computed as;



$$\rho_{S,T|Z} = \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}}} \quad (8)$$

where  $\sigma_{ST}$  is the covariance between the residuals  $\epsilon_{Sj}$  and  $\epsilon_{Tj}$  of (1) and (2),  $\sigma_{SS}$  is the variance of  $\epsilon_{Sj}$  and  $\sigma_{TT}$  is the variance of  $\epsilon_{Tj}$ . This association details the relationship between  $S$  and  $T$  after adjusting for treatment effect<sup>1,5,9</sup> at the individual level. It can also be viewed as how accurately  $S$  can predict  $T$ , accounting for  $Z$ <sup>35</sup>. One can construct a confidence interval for  $\rho_{S,T|Z}$  based on the Fisher  $Z$  transformation<sup>1</sup> or fit the following reduced form of (7)

$$\begin{pmatrix} S_{ij} \\ T_{ij} \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu_S + \alpha_j X_i \\ \mu_T + \beta X_i \end{pmatrix}, \Sigma_j \right], \quad \Sigma_j = \begin{pmatrix} \sigma_{SS} & 0 \\ 0 & \sigma_{TT} \end{pmatrix} \quad (9)$$

to test the hypothesis  $H_0 : \rho_{S,T|Z} = 0$ . This can be achieved via the comparison of (7) and (9) using a likelihood ratio test<sup>1</sup>. Outside the realm of surrogate endpoint evaluation,  $\rho_{S,T|Z}$  has found other applications. Perualia-Tan et al<sup>25</sup> used it in computing the association between a gene expression and the response of interest, after adjusting for fingerprint effect in the QSTAR modelling framework<sup>1</sup>.

The adjusted association as a measure of surrogacy has some quite interesting features; it remains within the unit interval and is precisely computed in the case of large individual level replicates. It however does not have a direct extension to where either  $S$  or  $T$  or both are not continuous<sup>1,5,9</sup>. Moreover,  $\rho_{S,T|Z}$  does not really depict the causal relationship between  $Z, T$  and  $Z, S$  which is of main interest in evaluating a surrogate<sup>32</sup>. This can actually be avoided with use of strong unverifiable assumptions but, this points to need for a framework that allows the association of interest to be measured, hence the causal inference approach<sup>35</sup>.

### 2.3 Causal Inference Approach (CIA)

As indicated in the last paragraph of the JMA subsection, there is need to establish a causal association between the treatment effect on true endpoint ( $Z, T$ ) and treatment effect on surrogate ( $Z, S$ ), especially if the data is arising from a single clinical trial<sup>32</sup>. Furthermore, the precision of the adjusted association largely depends on the individual level replications(sample size), hence establishing surrogacy with this metric using data from a single trial with few patients might not be ideal<sup>32,35</sup>.

Assuming that there are only two treatments being evaluated(say a standard and experimental treatment in our setting) and adopting Rubin's causal inference<sup>30</sup> ideology, each patient in a trial has four potential outcomes;  $T_{0j}, T_{1j}, S_{0j}, S_{ij}$ , assuming  $Z = 0$  and  $Z = 1$  for standard and experimental treatment respectively,  $j$  here is a counter for patients and not surrogates as earlier used. One could compute Individual Causal Effects(ICEs)<sup>35</sup> as;

$$\Delta T = T_{1j} - T_j \quad (10)$$

$$\Delta S = S_{1j} - S_{0j} \quad (11)$$

Averaging ICEs across all patients in the trial gives the Expected Causal Effects (ECEs) in the trial<sup>35</sup>. ICEs are however not identifiable from the data, since patients are randomized to take a certain treatment and not the two treatments, hence only two of the four potential outcomes are observed<sup>1,35</sup>. However, if it is assumed that the observations from a patient is independent of other patients (Stable Unit Treatment Value Assumption, SUTVA), with  $T_j$  and  $S_j$  the observed true and surrogate endpoint values for patient  $j$ , then it follows that;

$$T_j = Z_j T_{1j} + (1 - Z_j) T_{0j} \quad (12)$$

$$S_j = S_j T_{1j} + (1 - Z_j) S_{0j} \quad (13)$$

where,  $Z_j$  is the treatment group the patient  $j$  is randomized to. If it is further assumed that treatment group is independent of potential outcomes (this assumption is always guaranteed in clinical trials because of randomization), i.e.  $Z \perp (T_{0j}, T_{1j})$  and  $Z \perp (S_{0j}, S_{1j})$ , then;

$$\beta = E(T_j | Z_j = 1) - E(T_j | Z_j = 0) \quad (14)$$

$$\alpha = E(S_j | Z_j = 1) - E(S_j | Z_j = 0) \quad (15)$$

where  $\alpha$  and  $\beta$  retain the same meaning as in (1) and (2) because of the for-mentioned assumptions<sup>32,35</sup>. Furthermore, these two quantities are the expected causal effects of  $Z$  on  $S$  and  $T$ <sup>1</sup> and are estimable from the observed data.

Let's assume that the set of potential outcomes is normally distributed, i.e.  $Y = (T_{0j}, T_{1j}, S_{0j}, S_{1j}) \sim$

$N(\mu, \Sigma)$ ,  $\mu = (\mu_{0j}, \mu_{1j}, \mu_{0j}, \mu_{1j})$  and  $\Sigma = \begin{pmatrix} \sigma_{T_0 T_0} & \sigma_{T_0 T_1} & \sigma_{T_0 S_0} & \sigma_{T_0 S_1} \\ \sigma_{T_0 T_1} & \sigma_{T_1 T_1} & \sigma_{T_1 S_0} & \sigma_{T_1 S_1} \\ \sigma_{T_0 S_0} & \sigma_{T_1 S_0} & \sigma_{S_0 S_0} & \sigma_{S_0 S_1} \\ \sigma_{T_0 S_1} & \sigma_{T_1 S_1} & \sigma_{S_0 S_1} & \sigma_{S_1 S_1} \end{pmatrix}$ , it then follows

that the vector of ICEs can be written as  $AY$ , where  $A = (1, -1, 1, -1)$ . It further follows that  $AY \sim N(A\mu, A\Sigma A')$ , where  $A\mu = (\beta, \alpha)$ <sup>1</sup>.

This implies that a confidence interval cannot be estimated for  $\rho_{ICA}$  since the quantity in itself is estimated via several guess of plausible values.

## 2.4 Evaluating Use of Multiple Surrogates

## 2.5 Softwares and Tools

The packages *Surrogate*<sup>34</sup> and *nlme*<sup>26</sup> in R 3.4.0<sup>29</sup> are the major software tools employed in analyzing the data used for this thesis. Also, the packages *exactRankTests*<sup>18</sup> and *multtest*<sup>27</sup> were used for non-parametric testing of treatment effect and correcting for multiplicity respectively.

### 3 Data Analysis

This section presents the results obtained by applying the Joint Modelling and Causal Inference surrogate endpoint evaluation approach to the dataset described in the first section of this report. First, an exploration of the association between each surrogate and the true endpoint is presented, then the treatment effect on insulin sensitivity as well as the surrogates is carried out using both parametric (paired t-test<sup>2</sup>) as well as non-parametric test (Wilcoxon rank sum test<sup>2</sup>). The non-parametric alternative to the paired t-test was considered because of concerns about the number of individuals involved in the study. Also, focus will be on log of the surrogates since they are primarily difference in counts.

#### 3.1 Exploratory Analysis

A scatter plot of insulin sensitivity ( $T$ ) against bacteria counts ( $S$ ), with each point coloured based on treatment received ( $Z$ ) is used to explore the association between these variables. This is done using both the raw counts and log of counts for each bacteria cell, however explanations will be focused on log counts rather than the counts themselves.

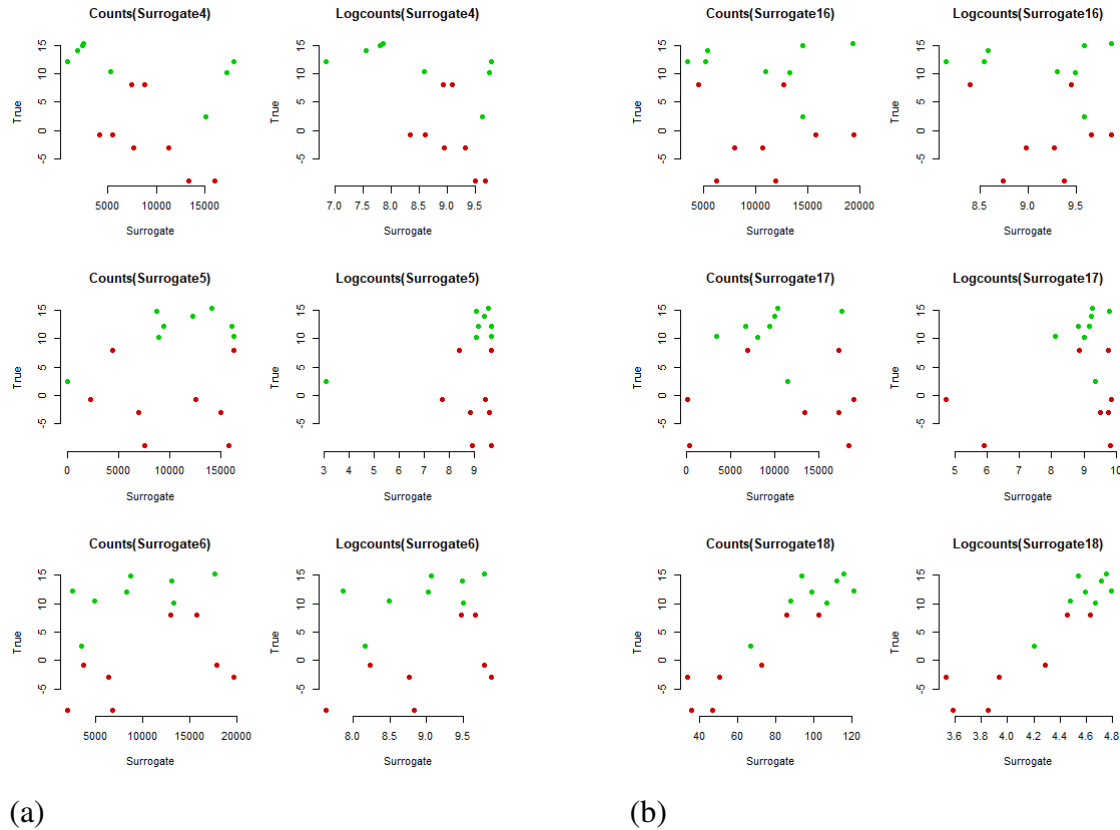


Figure 2: Scatterplots of insulin sensitivity against each surrogate coloured by treatment received. Red=autologous gut microbiota infusion (control treatment), green=allogenic gut microbiota infusion (experimental treatment)

A look at the grouping of colours from the y-axis shows the treatment effect on insulin sensitivity, the colour grouping along the x-axis along the would indicate treatment effect on the surrogate, while the arrangement of the points across and within the colour groups gives an idea of the association between insulin sensitivity and the surrogates.

Generally in all the figures (2,3 and the rest in appendix), one could observe that the green points (experimental treatment) is above the red ones (control treatment) indicating a treatment effect on insulin sensitivity. In terms of the surrogates, the plots suggests a treatment effect on surrogates; 18 (figure 2b), 29 and 13 (figure 3a and b).

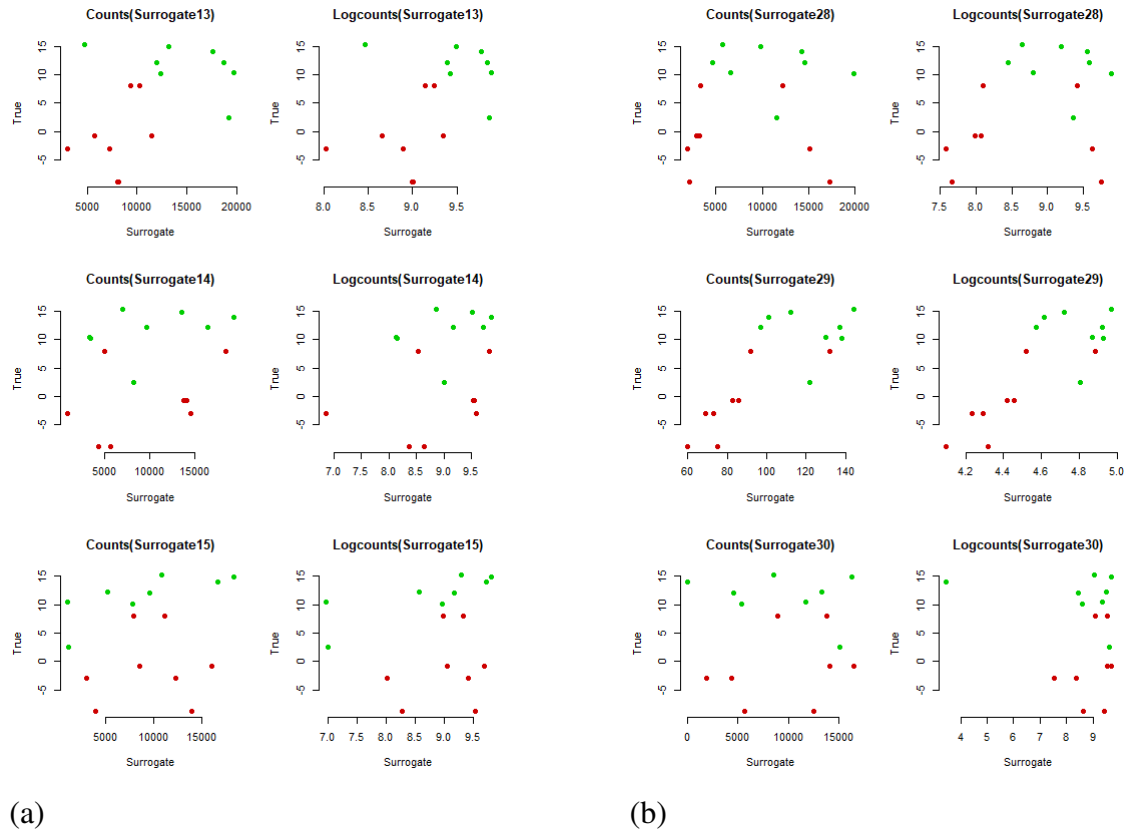


Figure 3: Scatterplots of insulin sensitivity against each surrogate coloured by treatment received. Red=autologous gut microbiota infusion (control treatment), green=allogenic gut microbiota infusion (experimental treatment)

In terms of the association, surrogates 18 appears interesting since a linear association is visible between both the red (control) and green (experimental) points. This indicates the likelihood of a high adjusted association between this surrogate and the true endpoint. This observation is also consistent for surrogate 29, however the association appears stronger within the control group than within the experimental group. Furthermore, surrogate 4 appears to have a negative association with insulin sensitivity in both treatment groups. These observations remain consistent with those that can be deduced from table 1. A careful look at surrogate 29 in table 1 reveals high  $\rho_{UN}$

(0.82) but this mainly driven by the high  $\rho_0$ . Furthermore, it is also important to note that  $\rho_0$  and  $\rho_1$  are in different directions.

| $S$ | $\rho_0$ | $\rho_1$ | $S$ | $\rho_0$ | $\rho_1$ | $S$ | $\rho_{UN}$ | $S$ | $\rho_{UN}$ |
|-----|----------|----------|-----|----------|----------|-----|-------------|-----|-------------|
| 1   | 0.01     | -0.24    | 16  | -0.10    | -0.15    | 1   | -0.23       | 16  | -0.13       |
| 2   | 0.32     | -0.01    | 17  | 0.21     | 0.14     | 2   | -0.34       | 17  | 0.27        |
| 3   | 0.20     | -0.38    | 18  | 0.84     | 0.81     | 3   | -0.23       | 18  | 0.90        |
| 4   | -0.49    | -0.59    | 19  | 0.00     | 0.21     | 4   | -0.52       | 19  | -0.22       |
| 5   | -0.16    | 0.88     | 20  | -0.51    | 0.73     | 5   | 0.14        | 20  | -0.14       |
| 6   | 0.57     | 0.56     | 21  | 0.12     | -0.47    | 6   | 0.29        | 21  | -0.03       |
| 7   | 0.06     | -0.32    | 22  | 0.43     | -0.10    | 7   | -0.29       | 22  | 0.29        |
| 8   | -0.52    | 0.37     | 23  | 0.06     | 0.30     | 8   | 0.22        | 23  | 0.17        |
| 9   | 0.19     | 0.01     | 24  | -0.10    | 0.81     | 9   | 0.33        | 24  | -0.18       |
| 10  | -0.48    | 0.18     | 25  | 0.18     | -0.47    | 10  | -0.20       | 25  | -0.01       |
| 11  | -0.57    | 0.70     | 26  | 0.39     | -0.19    | 11  | -0.05       | 26  | 0.31        |
| 12  | -0.27    | 0.09     | 27  | -0.44    | 0.16     | 12  | -0.10       | 27  | 0.10        |
| 13  | 0.29     | -0.50    | 28  | 0.02     | -0.23    | 13  | 0.44        | 28  | 0.31        |
| 14  | 0.32     | 0.32     | 29  | 0.85     | -0.11    | 14  | 0.29        | 29  | 0.82        |
| 15  | 0.20     | 0.79     | 30  | 0.28     | -0.28    | 15  | 0.12        | 30  | -0.18       |

Table 1: *Correlations between each surrogate( $S$ ) and insulin sensitivity( $T$ ).  $S$ =Surrogates,  $\rho_0$ =correlation between  $S$  and  $T$  in the control group,  $\rho_1$ =correlation between  $S$  and  $T$  in the experimental group and  $\rho_{UN}$ =correlation between  $S$  and  $T$  without paying attention to treatment group.*

This implies that the association between insulin sensitivity and surrogate 29 is different based on the treatment group. The association between surrogate 4 and insulin sensitivity is consistent in both treatment groups (see  $\rho_0$  and  $\rho_1$  in table, 1), this observation is also consistent for surrogate 18 and 6.

### 3.2 Estimation of Treatment Effects

In estimating treatment effects, the following null hypothesis is of interest for insulin sensitivity;

$$H_0 : \mu_0 = \mu_1$$

where  $\mu_0$  is the average insulin sensitivity value for the patients administered the control treatment and  $\mu_1$  the average insulin sensitivity value for patients administered the experimental treatment. The same hypothesis is of interest for each surrogate but because there are several surrogates, the index  $j$  will be added to the subscript 0 and 1 in the previous hypothesis. Hence the null;

$$H_0 : \mu_{0j} = \mu_{1j}, \quad j = 1, 2, \dots, 30$$

is of interest. Here,  $\mu_{0j}$  is the average difference in count for surrogate  $j$  from patients administered the control treatment, while  $\mu_{1j}$  is the average difference in count for surrogate  $j$  from patients administered the experimental treatment. The second hypothesis testing involves 30 different hypothesis, hence the need for multiplicity correction. The Benjamini-Hockberg (BH) approach<sup>2</sup> to correcting False Discovery Rate (FDR) was employed. Also, these null hypotheses are tested against their two-sided alternatives.

| Type | MeanDifference | LowerCI | UpperCI |
|------|----------------|---------|---------|
| t    | 12.57          | 6.68    | 18.47   |
| W    | 13.23          | 5.45    | 18.96   |

Table 2: Results of student-t ( $t$ ) and Wilcoxon ( $W$ ) test for difference in means. MeanDifference is estimated as  $\mu_1 - \mu_0$  for  $t$  and  $R_1 - R_2$  for  $W$ .  $R_1$  is sum of ranks for experimental group while  $R_0$  is the sum of ranks for the control group.

Results from table 2 showed that the experimental group had a higher average insulin sensitivity compared to the control group. The resulting confidence interval for the difference in means also did not contain 0, hence it can be concluded that there is a significant difference in means of the experimental and control group. Due to concerns about the sample size (as indicated earlier), the same hypothesis was also tested in a non-parametric fashion using the Wilcoxon ranksum test. Similar conclusions were derived from this test as well, however the confidence interval for the treatment effect was larger compared to the student-t.

In terms of treatment effect on surrogates, there is a significant treatment effect (at 5% level of significance) on surrogates 29, 18 and 13 when student-t is applied, however surrogate is added to the list when Wilcoxon ranksum test is used (see raw p-value column in table 7). When multiplicity is adjusted for, none of the surrogates have a significant treatment effect in both tests.

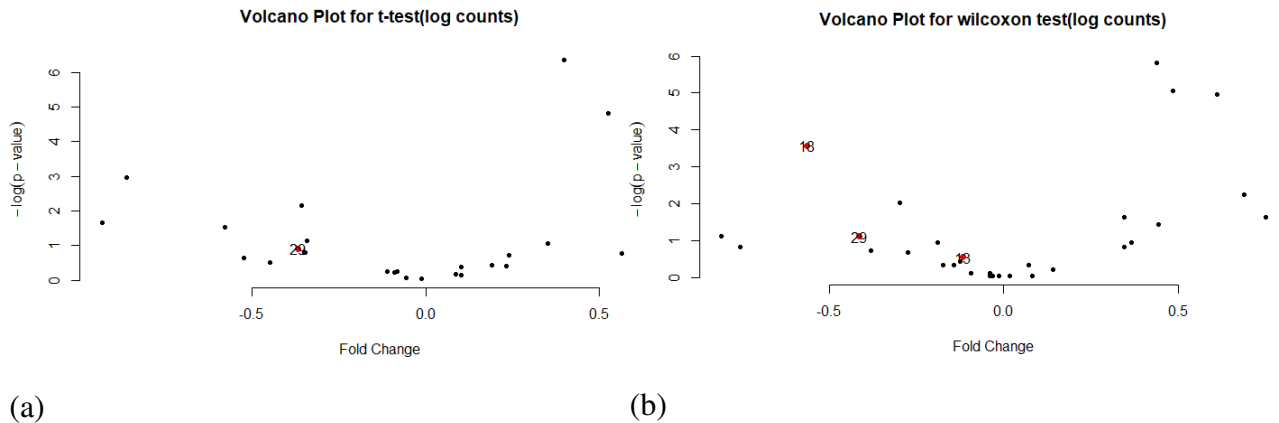


Figure 4: Volcano plots of fold change against  $-\log(\text{raw } p\text{-value})$  for both  $t$ -test and Wilcoxon. Fold change in  $t$ -test is computed as difference in means while it is computed as difference in ranks for Wilcoxon ranksum test. Surrogates with adjusted  $p\text{-value} < 0.1$  is in red with the surrogate number on it.

Figures 4 (a) and (b) also showed that surrogates 29 have an adjusted p-value  $< 10\%$  in both student-t and Wilcoxon test while surrogates 18 and 13 also have adjusted p-values  $< 10\%$ . Therefore, in terms of treatment effects, surrogates 29, 18 and 13 could be explored for treatment effects, however this is not certain because they aren't validated surrogates and a relationship has not been established between the effect of treatment on insulin sensitivity and the effect of treatment on these surrogates. This is the topic of the following subsections.

### 3.3 Univariate Surrogate Evaluation

In this part of the report, results obtained from applying the two methods discussed in the previous section to each surrogate is presented and discussed.

#### 3.3.1 Joint Modelling Results

For each surrogate, model (7) and (9) is fitted and compared using a likelihood ratio test, the adjusted association  $(\rho_{S,T|Z})$  is computed as described previously and the Fisher Z approach is used to construct a confidence interval for the adjusted association. Of interest in comparison of these 2 models is the null hypothesis,

$$H_0 : \rho_{S,T|Z} = 0$$

against its two sided alternative. Since there are 30 surrogates, it implies that a multiplicity issue arises again, hence the BH procedure to controlling FDR is also employed here. Also, treatment effect on each surrogate and insulin sensitivity is evaluated.

| S  | $\rho_{S,T Z}$ | $S_{Trt}$ | $PV_S$ | $BH_S$ | $PV_{\rho_{S,T Z}}$ | $BH_{\rho_{S,T Z}}$ | S  | $\rho_{S,T Z}$ | $S_{Trt}$ | $PV_S$ | $BH_S$ | $PV_{\rho_{S,T Z}}$ | $BH_{\rho_{S,T Z}}$ |
|----|----------------|-----------|--------|--------|---------------------|---------------------|----|----------------|-----------|--------|--------|---------------------|---------------------|
| 18 | 0.83           | 0.52      | 0.00   | 0.04   | 0.00                | 0.00                | 9  | 0.15           | 0.67      | 0.24   | 0.65   | 0.55                | 0.82                |
| 29 | 0.57           | 0.40      | 0.00   | 0.01   | 0.01                | 0.13                | 12 | -0.14          | -0.01     | 0.96   | 0.96   | 0.58                | 0.82                |
| 6  | 0.56           | -0.11     | 0.77   | 0.92   | 0.01                | 0.13                | 7  | -0.13          | -0.34     | 0.31   | 0.77   | 0.59                | 0.82                |
| 15 | 0.46           | -0.35     | 0.44   | 0.80   | 0.05                | 0.36                | 23 | 0.12           | 0.23      | 0.66   | 0.88   | 0.62                | 0.82                |
| 4  | -0.44          | -0.58     | 0.19   | 0.58   | 0.06                | 0.36                | 3  | -0.12          | -0.35     | 0.43   | 0.80   | 0.64                | 0.82                |
| 5  | 0.41           | -0.45     | 0.60   | 0.88   | 0.09                | 0.44                | 16 | -0.12          | -0.08     | 0.77   | 0.92   | 0.64                | 0.82                |
| 14 | 0.32           | 0.19      | 0.65   | 0.88   | 0.19                | 0.74                | 21 | -0.11          | 0.10      | 0.86   | 0.92   | 0.65                | 0.82                |
| 27 | -0.31          | 0.62      | 0.16   | 0.58   | 0.20                | 0.74                | 2  | 0.10           | -0.86     | 0.03   | 0.26   | 0.69                | 0.82                |
| 24 | 0.25           | -0.36     | 0.10   | 0.52   | 0.30                | 0.82                | 8  | -0.10          | 0.60      | 0.16   | 0.58   | 0.70                | 0.82                |
| 10 | -0.24          | -0.09     | 0.80   | 0.92   | 0.34                | 0.82                | 19 | 0.10           | -0.93     | 0.17   | 0.58   | 0.70                | 0.82                |
| 22 | 0.23           | 0.24      | 0.48   | 0.80   | 0.35                | 0.82                | 1  | -0.09          | -0.37     | 0.40   | 0.80   | 0.71                | 0.82                |
| 11 | -0.22          | 0.10      | 0.67   | 0.88   | 0.38                | 0.82                | 25 | -0.08          | 0.09      | 0.85   | 0.92   | 0.75                | 0.84                |
| 17 | 0.19           | 0.56      | 0.45   | 0.80   | 0.44                | 0.82                | 30 | -0.06          | -0.52     | 0.51   | 0.80   | 0.80                | 0.86                |
| 20 | -0.19          | -0.06     | 0.93   | 0.96   | 0.45                | 0.82                | 28 | -0.05          | 0.66      | 0.08   | 0.51   | 0.84                | 0.87                |
| 26 | 0.19           | 0.35      | 0.34   | 0.78   | 0.45                | 0.82                | 13 | -0.04          | 0.61      | 0.01   | 0.11   | 0.88                | 0.88                |

Table 3: Table showing  $\rho_{S,T|Z}$ , treatment effect on surrogate as well as raw and adjusted p-values for these quantities, ordered by absolute value of  $\rho_{S,T|Z}$ .  $S_{trt}$ =treatment effect on surrogate,  $PV_S$ ,  $PV_{\rho_{S,T|Z}}$ =raw p-values for surrogate treatment effect and  $\rho_{S,T|Z}$ ,  $BH_S$ ,  $BH_{\rho_{S,T|Z}}$ =adjusted p-values for surrogate treatment effect and  $\rho_{S,T|Z}$ .

The estimated treatment effect on insulin sensitivity ( $\beta$  from the model framework) is 12.57, same as the difference in means for the two treatment groups, however there is a significant effect of treatment only on surrogate 18 and 29 after adjusting for multiplicity. In terms of adjusted association, all computed  $\rho_{S,T|Z}$  is lesser than  $\rho_{UN}$  since the effect of treatment has been removed before computing  $\rho_{S,T|Z}$ . A closer look at surrogate 29 revealed that the value of 0.82 for  $\rho_{UN}$  reduced to 0.57 after adjusting for treatment, this association is even not significantly different from 0. Only surrogate 18 had an adjusted association value significantly different from 0 after adjusting for multiplicity. This implies that surrogate 18 is significantly correlated with insulin sensitivity regardless of treatment group, which makes surrogate 18 a good surrogate if conclusions were drawn based on this approach.

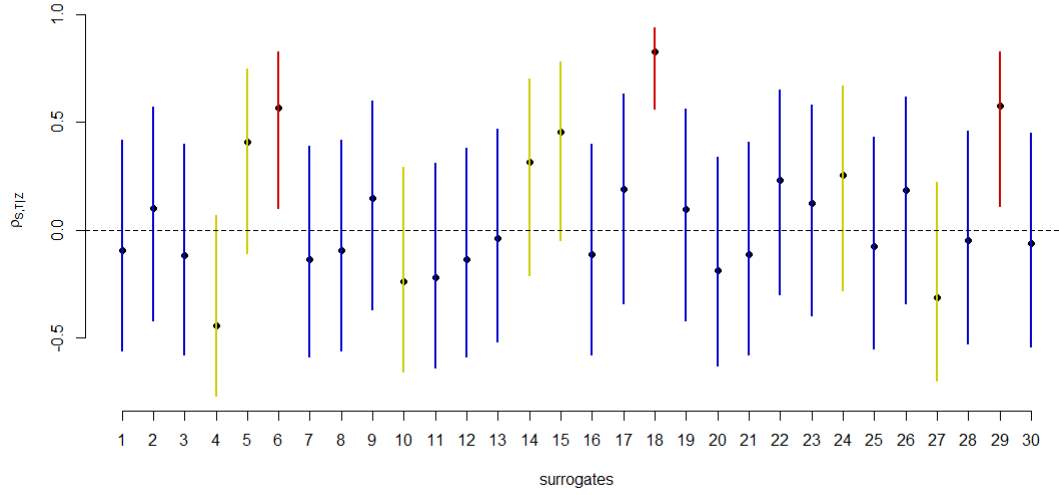


Figure 5:  $\rho_{S,T|Z}$  (black points) for each surrogate alongside their confidence intervals. Confidence interval is coloured by order in table 3. Top 3 surrogates in red, next 7 in yellow and others in blue.

The confidence interval represents a measure of uncertainty about the estimated adjusted association. It is important to examine this in order to have an idea of the precision of the estimated adjusted association for each surrogate. A look at figure 5 showed that  $\rho_{S,T|Z}$  for surrogate 18 is estimated with high precision while  $\rho_{S,T|Z}$  surrogates 6 and 29 were also estimated fairly precisely. Regarding information contained in these surrogates in relation to insulin sensitivity, surrogates 4, 10 and 27 contain somewhat different information compared to other top 10 surrogates, since the confidence interval for their  $\rho_{S,T|Z}$  is in the opposite direction of others. In other words, while surrogates 4, 10 and 27 are negatively associated with insulin sensitivity after adjusting for treatment, other surrogates among the top 10 surrogates are positively associated with insulin sensitivity.



### 3.3.2 Causal Inference Results

For each surrogate,  $\rho_{\Delta}$ , the Individual Causal Association (ICA), its distribution and  $\delta$ , the Prediction Mean Square Error (PMSE) are computed. This is done by estimating the identifiable part of  $\Sigma$ , convert it to a correlation matrix and filling other unidentifiable parts with plausible values between -1 and 1 as described in the methods section. The range of  $\rho_{\Delta}$  values obtained from its distribution is used as a measure of uncertainty about the computed ICA values, since a confidence interval cannot be computed for the ICA values. The distribution of  $\rho_{\Delta}$  is also examined to see the general direction of computed individual causal association between each surrogate and insulin sensitivity.

| S  | $\rho_{\Delta}$ | LR    | UR   | S  | $\rho_{\Delta}$ | LR    | UR   |
|----|-----------------|-------|------|----|-----------------|-------|------|
| 18 | 0.91            | -0.76 | 1.00 | 26 | 0.51            | -0.95 | 0.95 |
| 6  | 0.84            | -0.96 | 1.00 | 11 | 0.48            | -0.75 | 0.74 |
| 15 | 0.83            | -0.91 | 0.93 | 9  | 0.48            | -0.99 | 0.99 |
| 4  | -0.79           | -1.00 | 0.97 | 3  | 0.34            | -0.94 | 0.94 |
| 17 | 0.74            | -1.00 | 1.00 | 30 | 0.33            | -0.95 | 0.95 |
| 19 | 0.70            | -0.99 | 0.99 | 28 | 0.31            | -0.99 | 0.99 |
| 24 | 0.69            | -0.85 | 0.86 | 1  | 0.30            | -0.99 | 0.98 |
| 29 | 0.67            | -0.81 | 0.84 | 12 | 0.28            | -0.97 | 0.98 |
| 14 | 0.63            | -0.99 | 1.00 | 13 | 0.17            | -0.91 | 0.90 |
| 5  | 0.60            | -0.73 | 0.80 | 7  | 0.14            | -0.98 | 0.97 |
| 22 | 0.58            | -0.96 | 0.95 | 16 | 0.13            | -0.99 | 0.99 |
| 2  | 0.57            | -0.98 | 0.98 | 25 | 0.05            | -0.93 | 0.93 |
| 20 | 0.53            | -0.75 | 0.77 | 27 | 0.05            | -0.95 | 0.95 |
| 23 | 0.52            | -0.99 | 0.98 | 10 | 0.05            | -0.93 | 0.94 |
| 8  | 0.52            | -0.87 | 0.87 | 21 | 0.03            | -0.95 | 0.94 |

Table 4: Estimates  $\rho_{\Delta}$  for each surrogates including their range ranked by absolute value of  $\rho_{\Delta}$ .  $\rho_{\Delta}$ =ICA, LR=lower range, UR=upper range

Results from table 4 appear to show surrogate 18 as a very good surrogate for insulin sensitivity, but the uncertainty about the computed  $\rho_{\Delta}$  raises concern since its range goes from -0.76 to 1. Although the top 10 surrogates have fairly high individual causal association with insulin sensitivity, their range of computed ICA is wide enough to cast doubts about the use of these bacterias as surrogates. To really decipher the potentially good surrogates among the 30, their  $\rho_{\Delta}$  distribution and the percentage of values greater than 0.5 in absolute term, out of all computed  $\rho_{\Delta}$  values is examined closely.

A look at the distribution of ICA values for surrogate 18 revealed that there is a high chance of obtaining ICA values higher than 0.5 (see figure 7 a and figure 8), specifically this chance was

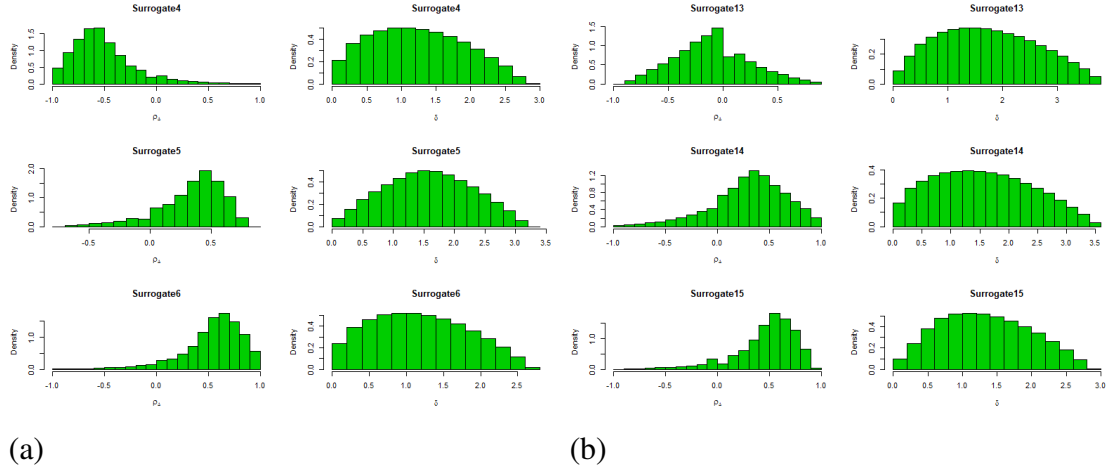


Figure 6: Distribution of  $\rho_\Delta$  and  $\delta$

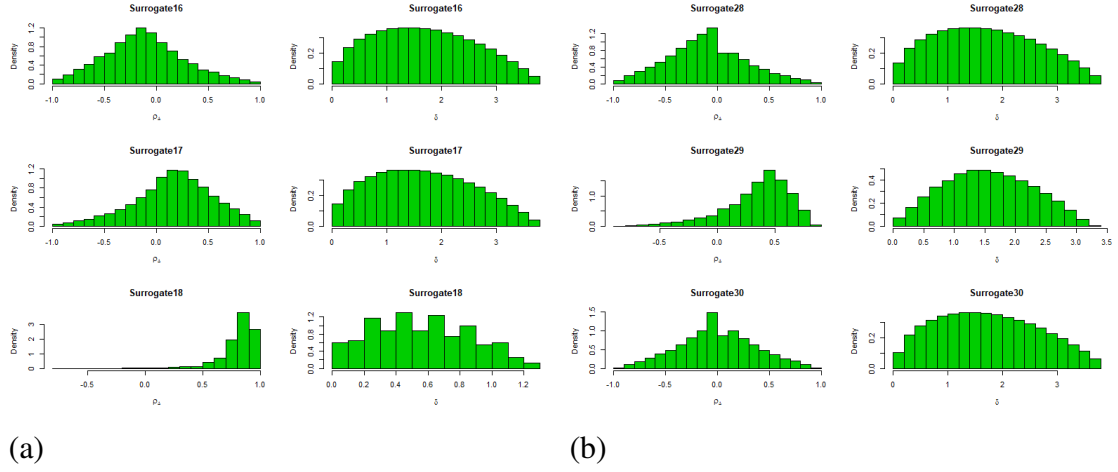


Figure 7: Distribution of  $\rho_\Delta$  and  $\delta$

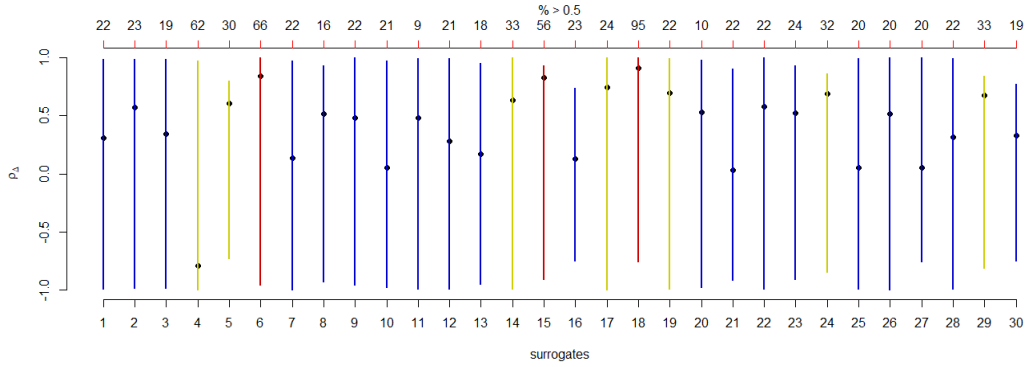


Figure 8:  $\rho_\Delta$  and their range including percentage greater than 0.5 in absolute value

computed to be 95%, while this chance is 56% and 66% for surrogates 6 and 15. Furthermore, the skewness of the distribution of  $\rho_{\Delta}$  values for these surrogates identify them as good candidates and their individual causal effects can predict the individual causal effects on insulin sensitivity with prediction error ranging from 0.5-2.5 for surrogate 15, 0.2-1.2 for surrogate 18 and 0-2.5 for surrogate 6 (see distribution of  $\delta$  in figures 6 and 7). For surrogate 29 which appears to be a promising surrogate judging from previous results, the chance of obtaining a  $\rho_{\Delta} > 0.5$  is estimated to be 33% while its individual causal effects will predict the individual causal effects on insulin sensitivity with errors ranging from 0-3.5.

In terms of agreement of results obtained from joint modelling and the ones obtained here, it was observed that both methods agree on surrogate 18 been the best in terms of predicting treatment effect on insulin sensitivity from treatment effect on surrogates. Furthermore, two surrogates, 17 and 19 are found amongst the top 10 surrogates from the causal inference results and not among the top 10 from the joint modelling results, while surrogates 27 and 10 are amongst the top 10 surrogates from the joint modelling results and not among the causal inference results.

To summarize, results obtained from surrogate by surrogate analysis, it was observed that surrogate 18 (*Akkermansia*) was judged the best surrogate in terms of predicting insulin sensitivity as well as treatment effect on this endpoint. While the adjusted association showed that the association between insulin sensitivity and this surrogate is precisely estimated, the range of its individual causal association covers almost the entirety of the parameter space of a correlation (-1,1). On closer inspection of the computed individual causal association, it was observed that there is a 95% chance of obtaining individual causal association values above 0.5. It was also observed that individual causal effect on this surrogate can predict individual causal effect on insulin sensitivity with little prediction errors ranging between 0.2-1.2, this range of prediction error is the smallest when compared to other surrogates.

Furthermore, it was observed that some surrogates contain information that is not captured by other surrogates, these surrogates are 4, 10 and 27. Both their adjusted association confidence interval and distribution of individual causal association values are in the opposite direction of other surrogates. Of these three surrogates, surrogate 4 appear to be most promising since its estimated individual causal association and adjusted association are negative.

### 3.4 Multiple Surrogates Evaluation

In light of the fact that some surrogates contain information not captured by other surrogates and biological evidence<sup>3,13,37</sup> that demonstrates that several element of the gut microbiota are involved in human metabolism, it is only natural to assume that a combination treatment effect on these elements would better predict treatment effect on insulin. Furthermore, knowing that the composition of the gut microbiota vary from individual to individual but certain genes and biological pathway are maintained makes it imperative to look beyond using single elements of

the microbiome as surrogates for insulin sensitivity.

Building on results from the univariate analysis, focus will be primarily on the top 10 surrogates from both approach including the 4 surrogates not agreed on by the two methods employed. Also, since surrogate 18 is the most promising surrogate, focus will be primarily on improving this surrogate by combining it with other surrogates in the top 10. For each combination, the multivariate individual causal association and multivariate adjusted association was computed as discussed in the methods section.

| S     | $R_H^2$ | LR   | UR   | $AR_H^2$ | S     | $\gamma_\Delta^2$ | LCL  | UCL  | $A\gamma_\Delta^2$ |
|-------|---------|------|------|----------|-------|-------------------|------|------|--------------------|
| 18,4  | 0.80    | 0.22 | 0.97 | 0.77     | 18,4  | 0.75              | 0.52 | 0.99 | 0.71               |
| 18,5  | 0.80    | 0.19 | 0.98 | 0.77     | 18,6  | 0.75              | 0.52 | 0.99 | 0.71               |
| 18,6  | 0.80    | 0.06 | 0.98 | 0.77     | 18,10 | 0.73              | 0.47 | 0.98 | 0.68               |
| 18,17 | 0.77    | 0.06 | 0.97 | 0.73     | 18,5  | 0.71              | 0.45 | 0.98 | 0.67               |
| 18,10 | 0.76    | 0.02 | 0.99 | 0.73     | 18,19 | 0.71              | 0.44 | 0.98 | 0.66               |
| 18,15 | 0.76    | 0.04 | 0.99 | 0.73     | 18,14 | 0.70              | 0.43 | 0.97 | 0.66               |
| 18,29 | 0.76    | 0.22 | 0.97 | 0.72     | 18,29 | 0.70              | 0.43 | 0.97 | 0.66               |
| 18,14 | 0.76    | 0.14 | 0.98 | 0.72     | 18,17 | 0.70              | 0.43 | 0.97 | 0.66               |
| 18,19 | 0.75    | 0.14 | 0.99 | 0.71     | 18,27 | 0.69              | 0.41 | 0.97 | 0.65               |
| 18,27 | 0.75    | 0.05 | 0.97 | 0.71     | 18,15 | 0.69              | 0.40 | 0.97 | 0.64               |
| 18,24 | 0.73    | 0.07 | 0.99 | 0.68     | 18,24 | 0.69              | 0.40 | 0.97 | 0.64               |

(a)
(b)

Table 5: *Multivariate Individual Causal Association ( $R_H^2$ ), its range, Adjusted Multivariate Individual Causal Association ( $AR_H^2$ ), Multivariate Adjusted Association ( $\gamma_\Delta^2$ ), its confidence interval and Adjusted Multivariate Adjusted Association ( $A\gamma_\Delta^2$ ). Both tables are ordered by  $R_H^2$  and  $\gamma_\Delta^2$*

Results of the exercise is presented in table 5 and figure 9. The (a) part of both figure and table contain results obtained from the causal inference approach, while the (b) part contain results from the joint modelling approach.

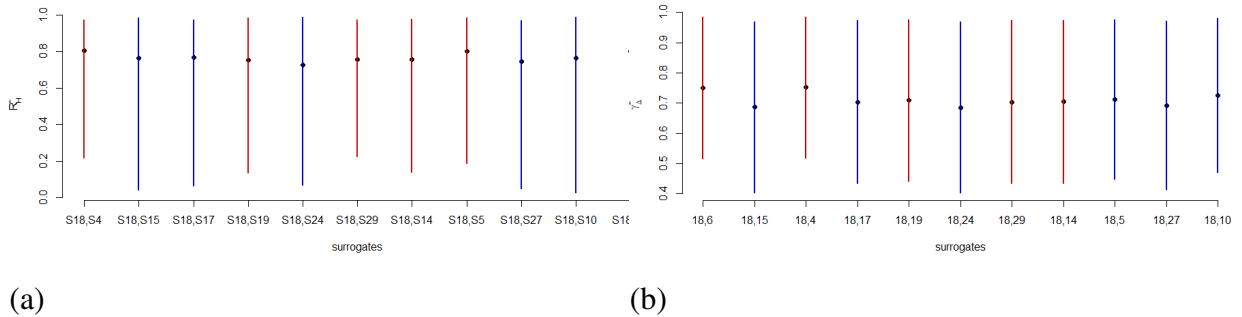


Figure 9: *Computed  $R_H^2$ , its range and  $\gamma_\Delta^2$  and its confidence interval with top surrogates in red.*

To properly digest these results, it is important to see if any of these surrogates improves surrogate 18, and more importantly is to check if any of the surrogates add some information to that already contained in surrogate 18. This evaluation is done by comparing results on surrogate 18 alongside those obtained here. To enable comparison on the same footing,  $\rho_{S,T|Z}^2$  for surrogate 18 is compared with  $\rho_{\Delta}$  obtained from the bivariate results. Also, the length of the range from the univariate and bivariate result are compared after the Fisher Z and logit transformation has been applied respectively. These transformations will move the range of values from  $[-1, 1]$  and  $[0, 1]$  to  $[-\infty, \infty]$ , thereby allowing the comparison of the range.

The individual causal association for surrogate 18 was computed to be

## 4 Discussion & Conclusion

### References

- [1] Alonso, A., Bigirimurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N. J., Shkedy, Z., and Van der Elst, W. (2017). *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. CRC Press, Florida.
- [2] Amaratunga, D., Cabrera, J., and Shkedy, Z. (2014). *Exploration and analysis of DNA microarray and other high-dimensional data*, volume 2014: 10. Wiley-Interscience, Hoboken, N.J, 2;second;2;2nd; edition.
- [3] Bäckhed, F., Fraser, C. M., Ringel, Y., Sanders, M. E., Sartor, R. B., Sherman, P. M., Versalovic, J., Young, V., and Finlay, B. B. (2012). Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. *Cell host & microbe*, 12(5):611–622.
- [4] Bull, M. J. and Plummer, N. T. (2014). Part 1: The human gut microbiome in health and disease. *Integrative Medicine: A Clinician’s Journal*, 13(6):17–22.
- [5] Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The evaluation of surrogate endpoints*. Springer, New York, NY.
- [6] Buyse, M., Burzykowski, T., Carroll, K., Michiels, S., Sargent, D. J., and Miller, L. L. (2007). Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology*, 25(33):5218–5224.
- [7] Buyse, M., Michiels, S., Squifflet, P., Lucchesi, K. J., Hellstrand, K., Brune, M. L., and Rowe, J. M. (2011). Leukemia-free survival as a surrogate end point for overall survival in the evaluation of maintenance therapy for patients with acute myeloid leukemia in complete remission. *Haematologica*, 96(8):1106–1112.

- [8] Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 54:1014–1029.
- [9] Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., Alonso, A. 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.
- [10] DeGruttola, V., Fleming, T. R., Lin, D. Y., and Coombs, R. (1997). Validating surrogate markers—are we being naive? *Journal of Infectious Diseases*, 175:237–246.
- [11] DeGruttola, V. and Tu, X. M. (1994). Modelling progression of cd-4 lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014.
- [12] Dumas, M., Barton, R. H., Toye, A., Cloarec, O., Blancher, C., Rothwell, A., Fearnside, J., Tatoud, R., Blanc, V., Lindon, J. C., Mitchell, S. C., Holmes, E., McCarthy, M. I., Scott, J., Gauguier, D., and Nicholson, J. K. (2006). Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. *Proceedings of the National Academy of Sciences of the United States of America*, 103(33):12511–12516.
- [13] Enviromedica (2012). Microbiome 101: Understanding gut microbiota. <http://www.prescript-assist.com/intestinal-health/gut-microbiome/>. Online, accessed 21 July 2017.
- [14] Fleming, T. R. (1994). Surrogate markers in aids and cancer trials. *Statistics in Medicine*, 13:1423–1435.
- [15] Fleming, T. R. and DeMets, D. L. (1996). Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine*, 125:605–613.
- [16] Fleming, T. R. and Powers, J. H. (2012). Biomarkers and surrogate endpoints in clinical trials. *Statistics in Medicine*, 31(25):2973–2984.
- [17] Freedman, L. S., Graubard, B. I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11:167–178.
- [18] Hothorn, T. and Hornik, K. (2017). *exactRankTests: Exact Distributions for Rank and Permutation Tests*. R package version 0.8-29.
- [19] Investigators, C. A. S. T. (1989). Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infraction. *New England Journal of Medicine* 321, 321:406–412.
- [20] Lagakos, S. W. and Hoth, D. F. (1992). Surrogate markers in aids: where are we? where are we going? *Annals of Internal Medicine*, 116:599–601.

- [21] Ley, R. E., Gordon, J. I., Fraser-Liggett, C. M., Knight, R., Hamady, M., and Turnbaugh, P. J. (2007). The human microbiome project. *Nature*, 449(7164):804–810.
- [22] Mahowald, M. A., Magrini, V., Turnbaugh, P. J., Gordon, J. I., Ley, R. E., and Mardis, E. R. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–131.
- [23] Molenberghs, G., Abad, A. A., Van der Elst, W., Burzykowski, T., and Buyse, M. (2013). Surrogate endpoints: when should they be used? *Clinical Investigation*, 3:1147–1155.
- [24] Molenberghs, G., Buyse, M., Geys, H., and Burzykowski, T. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*, 23:607–625.
- [25] Perualila-Tan, N., Kasim, A., Talloen, W., Verbist, B., Ghoelmann, H., Consortium, Q., Shkedy, Z., Khamiakova, T., and Otava, M. (2016). A joint modeling approach for uncovering associations between gene expression, bioactivity and chemical structure in early drug discovery to guide lead selection and genomic biomarker development. *Statistical Applications in Genetics and Molecular Biology*, 15(4):291–304.
- [26] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.
- [27] Pollard, K. S., Dudoit, S., and van der Laan, M. J. (2005). *Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer.
- [28] Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, 8:431–440.
- [29] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [30] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.
- [31] Ursell, L. K., Metcalf, J. L., Parfrey, L. W., and Knight, R. (2012). Defining the human microbiome. *Nutrition Reviews*, 70(Suppl 1):S38–S44.
- [32] Van der Elst, W. (2016). *Statistical Evaluation Methodology for Surrogate Endpoints in Clinical Studies*. PhD thesis, Universiteit Hasselt.
- [33] Van der Elst, W., Abad, A. A., Geys, H., Meyvisch, P., Bijmens, L., and Molenberghs, G. (2017). Univariate versus multivariate surrogate endpoints. *Submitted for consideration for publication in Biometrics*.

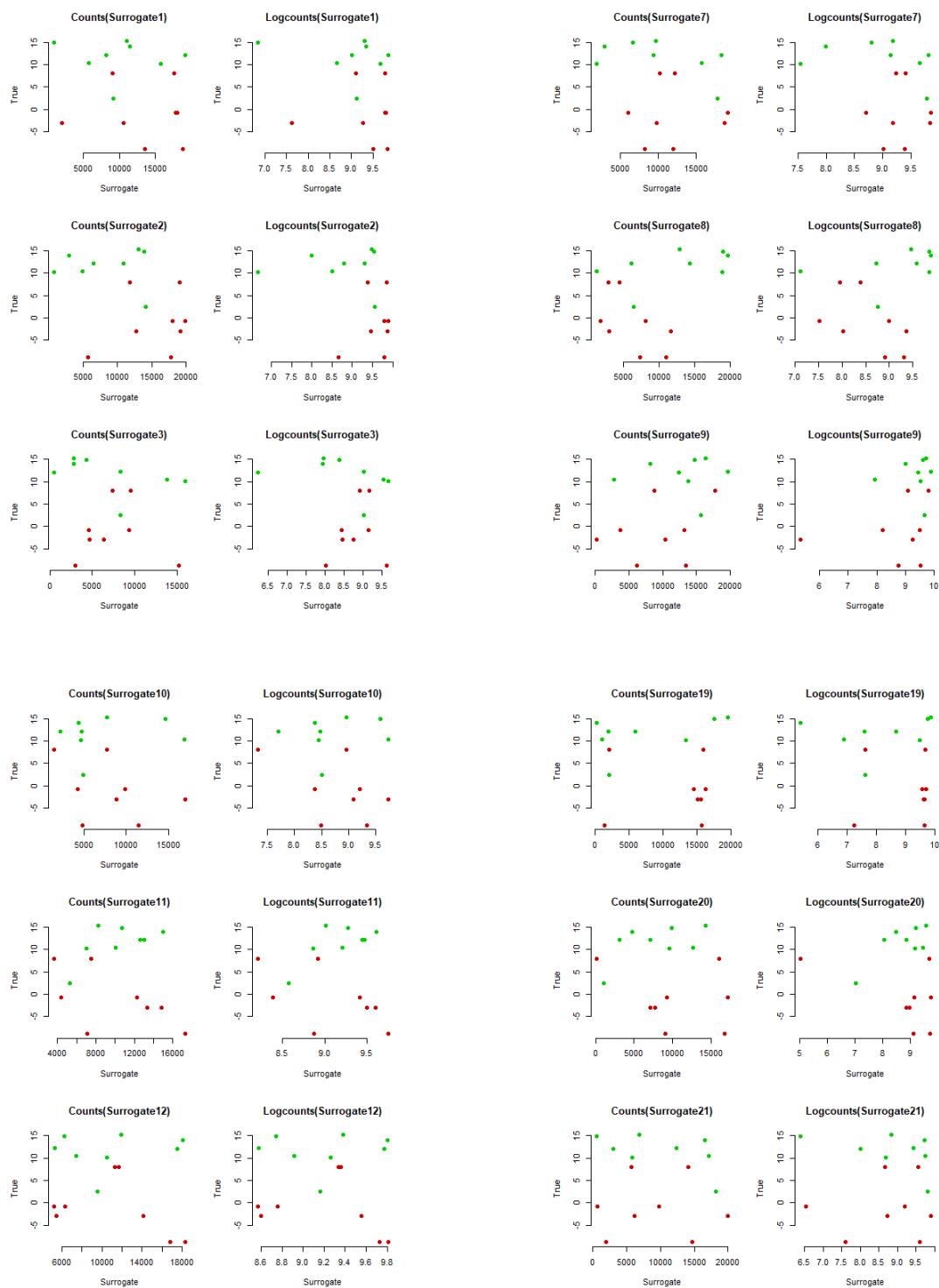
- [34] Van der Elst, W., Meyvisch, P., Alonso, A., Ensor, H. M., Weir, C. J., and Molenberghs, G. (2017). *Surrogate: Evaluation of Surrogate Endpoints in Clinical Trials*. R package version 0.2.
- [35] Van der Elst, W., Molenberghs, G., and Alonso, A. A. (2016). Exploring the relationship between the causal-inference and meta-analytic paradigms for the evaluation of surrogate endpoints. *STATISTICS IN MEDICINE*, 35(8):1281–1298.
- [36] Volberding, P. A., Lagakos, S. W., Koch, M. A., Pettinelli, C., Myers, M. W., Booth, D. K., Balfour Jr, H. H., Reichman, R. C., Bartlett, J. A., Hirsch, M. S., Murphy, R. L., Hardy, W. D., Soeiro, R., Fischl, M. A., Bartlett, J. G., Merigan, T. C., Hyslop, N. E., Richman, D. D., Valentine, F. T., Corey, L., and the AIDS Clinical Trials Group of the National Institute of Allergy & Infectious Diseases (1990). Zidovudine in asymptomatic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 cd4-positive cells per cubic millimeter. *New England Journal of Medicine*, 322:941–949.
- [37] Vrieze, A., Van Nood, E., Holleman, F., Salojärvi, J., Kootte, R. S., Bartelsman, J. F., Dallinga-Thie, G. M., Ackermans, M. T., Serlie, M. J., Oozeer, R., Derrien, M., Druesne, A., Van Hylckama Vlieg, J. E., Bloks, V. W., Groen, A. K., Heilig, H. G., Zoetendal, E. G., Stroes, E. S., de Vos, W. M., Hoekstra, J. B., and Nieuwdorp, M. (2012). Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology*, 43(4):913–916.



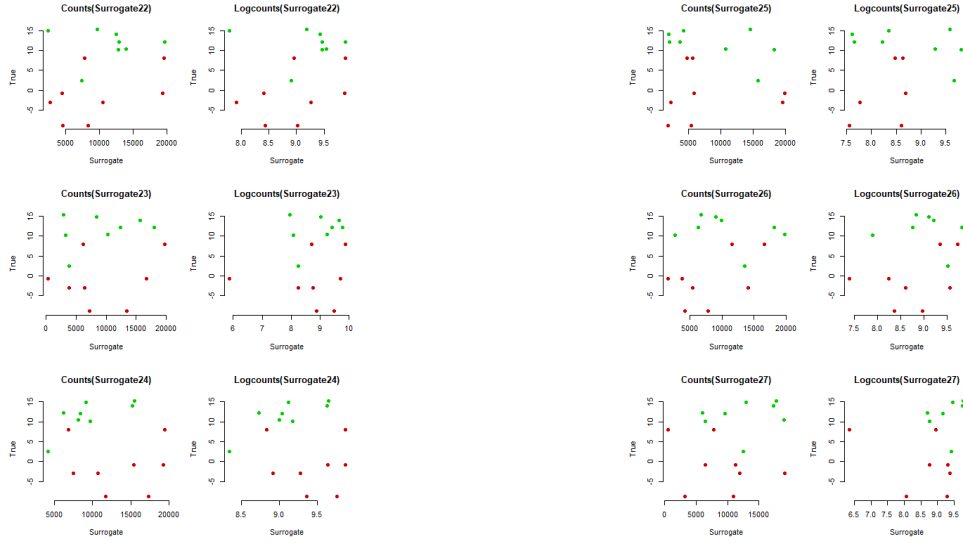
# Appendix

## Extra Plots

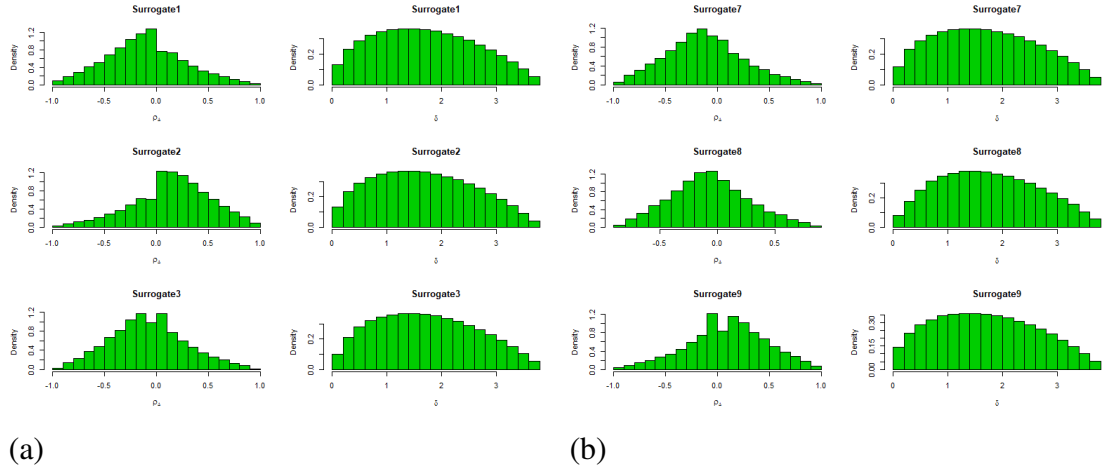
### Exploratory Plots



# Evaluation of Surrogate Endpoints in Human Microbiome

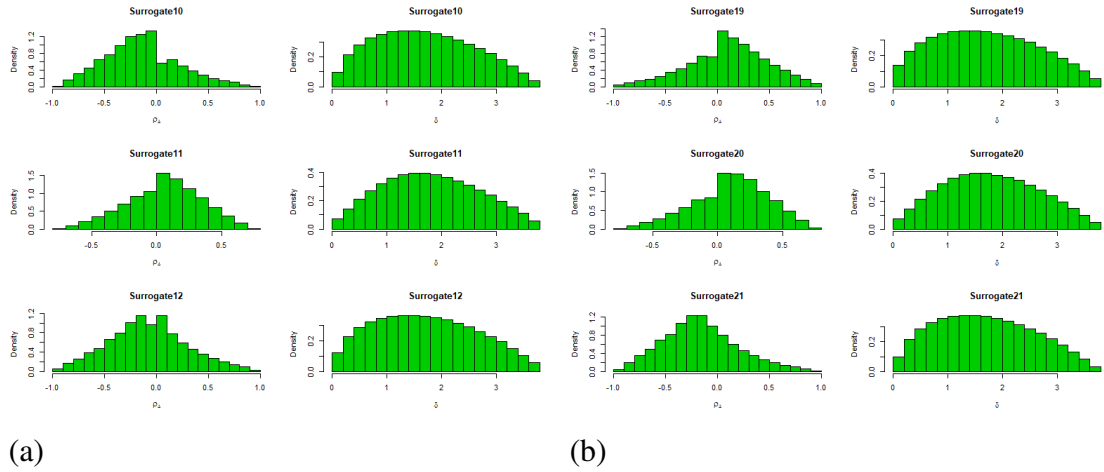


## $\rho_\Delta$ and $\delta$ Distribution



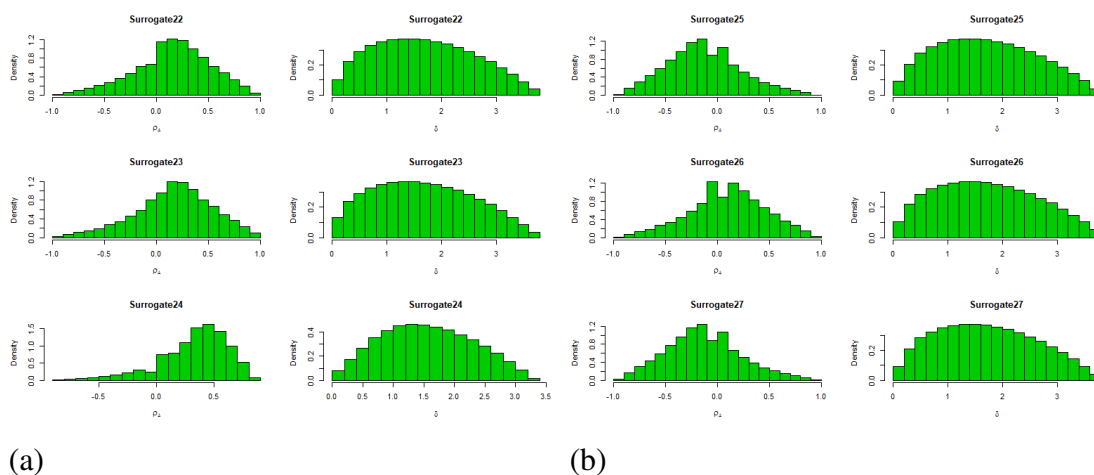
(a)

(b)



(a)

(b)



## Extra Tables

### Surrogates

| S  | Full Names                          | S  | Full Names                           |
|----|-------------------------------------|----|--------------------------------------|
| 1  | Brachyspira                         | 16 | Klebsiella.pneumoniae.et.rel.        |
| 2  | Lachnospira.pectinoschiza.et.rel.   | 17 | Bacteroides.ovatus.et.rel.           |
| 3  | Clostridium.felsineum.et.rel.       | 18 | Akkermansia                          |
| 4  | Megamonas.hypermegale.et.rel.       | 19 | Bifidobacterium                      |
| 5  | Mitsuokella.mutiacidia.et.rel.      | 20 | Lactococcus                          |
| 6  | Alcaligenes.faecalis.et.rel.        | 21 | Faecalibacterium.prausnitzii.et.rel. |
| 7  | Clostridium.thermocellum.et.rel.    | 22 | Corynebacterium                      |
| 8  | Lactobacillus.salivarius.et.rel.    | 23 | Micrococcaceae                       |
| 9  | Collinsella                         | 24 | Helicobacter                         |
| 10 | Clostridium.leptum.et.rel.          | 25 | Lactobacillus.cateniformis.et.rel.   |
| 11 | Oxalobacter.formigenes.et.rel.      | 26 | Coproccoccus.eutactus.et.rel.        |
| 12 | Papillibacter.cinnamivorans.et.rel. | 27 | Clostridium.difficile.et.rel.        |
| 13 | Enterobacter.aerogenes.et.rel.      | 28 | Clostridium.nexile.et.rel.           |
| 14 | Oceanospirillum                     | 29 | Novosphingobium                      |
| 15 | Anaerostipes.caccae.et.rel.         | 30 | Anaerotruncus.colihominis.et.rel.    |

Table 6: Full name of each Surrogate, i.e. full names of bacteria that makes up the Gut microbiota

### Raw and Adjusted p-values for Surrogates

| S  | PV(t) | BH   | S  | PV(t) | BH   | S  | PV(W) | BH   | S  | PV(W) | BH   |
|----|-------|------|----|-------|------|----|-------|------|----|-------|------|
| 29 | 0.00  | 0.05 | 15 | 0.45  | 0.81 | 29 | 0.00  | 0.07 | 17 | 0.49  | 0.89 |
| 18 | 0.01  | 0.12 | 17 | 0.47  | 0.81 | 18 | 0.01  | 0.07 | 20 | 0.51  | 0.89 |
| 13 | 0.02  | 0.17 | 22 | 0.49  | 0.81 | 13 | 0.01  | 0.07 | 3  | 0.57  | 0.96 |
| 2  | 0.05  | 0.39 | 30 | 0.52  | 0.82 | 2  | 0.03  | 0.21 | 10 | 0.65  | 0.96 |
| 28 | 0.10  | 0.57 | 5  | 0.61  | 0.88 | 8  | 0.10  | 0.63 | 5  | 0.72  | 0.96 |
| 24 | 0.11  | 0.57 | 14 | 0.66  | 0.88 | 24 | 0.13  | 0.65 | 6  | 0.72  | 0.96 |
| 8  | 0.17  | 0.63 | 23 | 0.66  | 0.88 | 9  | 0.19  | 0.73 | 15 | 0.72  | 0.96 |
| 27 | 0.18  | 0.63 | 11 | 0.68  | 0.88 | 28 | 0.19  | 0.73 | 21 | 0.80  | 0.96 |
| 19 | 0.19  | 0.63 | 6  | 0.77  | 0.92 | 27 | 0.23  | 0.78 | 25 | 0.88  | 0.96 |
| 4  | 0.21  | 0.64 | 16 | 0.78  | 0.92 | 1  | 0.33  | 0.88 | 30 | 0.88  | 0.96 |
| 9  | 0.26  | 0.70 | 10 | 0.80  | 0.92 | 19 | 0.33  | 0.88 | 11 | 0.96  | 0.96 |
| 7  | 0.32  | 0.80 | 25 | 0.85  | 0.92 | 7  | 0.38  | 0.88 | 12 | 0.96  | 0.96 |
| 26 | 0.34  | 0.80 | 21 | 0.86  | 0.92 | 26 | 0.38  | 0.88 | 14 | 0.96  | 0.96 |
| 1  | 0.40  | 0.81 | 20 | 0.93  | 0.96 | 4  | 0.44  | 0.88 | 16 | 0.96  | 0.96 |
| 3  | 0.45  | 0.81 | 12 | 0.96  | 0.96 | 22 | 0.44  | 0.88 | 23 | 0.96  | 0.96 |

Table 7: Raw and Adjusted P-values for Treatment Effect on Surrogates. PV(t)=raw pvalue from t-test and BH=Benjamini Hockberg corrected p-values

### R Codes