Fraud Detection in Bioequivalence Studies

Johanna Gerstmeyer, Jonathan Mißler
Collaboration partners: Norbert Benda, Tim Friede

Final Report: Practical Statistical Training
University of Göttingen

March 14, 2022

Contents

1 Problem Outline **1**

2 Data **2**
2.1 Provided Data Sets 2
2.2 Simulated Data Sets 3

3 Literature Review and Replication **5**
3.1 Buster 5
3.1.1 Overview 5
3.1.2 Details 6
3.2 SaToWIB 7
3.3 Replication 8
3.3.1 Fuglsang 8
3.3.2 Canada Data 11

4 Statistical Background **13**
4.1 Time Series 13
4.2 The Augmented Dickey-Fuller Test 14
4.3 The Kwiatkowski–Phillips–Schmidt–Shin Test 15

5 Testing Process **15**
5.1 KPSS Test 16
5.2 ADF Test 16
5.3 Confidence Interval Width Rule 17
5.4 KPSS Test Reconsideration 17
5.5 ADF Test Reconsideration 18
5.6 Overview 19

6 Discussion **20**
6.1 Test Procedures 20
6.2 Practical Use 21
6.3 Considerations 22

7 Conclusion **23**

List of Figures

1 Buster plot 1: Replication vs. Fuglsang [2021] 8
2 Buster plot 2: Replication vs. Fuglsang [2021] 9
3 Buster plot 3: Replication vs. Fuglsang [2021] 9
4 Buster plot 4: Replication vs. Fuglsang [2021] 10
5 Buster plot 5: Replication vs. Fuglsang [2021] 10
6 Buster plot 1: Buster on the Health Canada [2014] data 11
7 Buster plot 2: Buster on the Health Canada [2014] data 11
8 Buster plot 3: Buster on the Health Canada [2014] data 12
9 Buster plot 4: Buster on the Health Canada [2014] data 12
10 Buster plot 5: Buster on the Health Canada [2014] data 13
11 Plot of the point estimates without transformation (above) and with transformation (below) 19

List of Tables

1 Excerpt of a simulated dataset 5
2 SaToWIB Replication 10
3 SaToWIB: Original 10
4 Overview of test performances 20
5 Prevalence vs. Positive Predictive Value 22

1 Problem Outline

This project is motivated by a scientific paper of Fuglsang [2021]. In the paper the author describes one kind of data manipulation that they encountered in the field of bioequivalence trials. Typically bioequivalence trials are conducted to compare certain properties of two formulations, a new Test formulation and an already established Reference formulation. The trials examine if the rate and extent of absorption into the blood stream is equivalent for both formulations. A successful bioequivalence trial often presents a faster and cheaper approach to facilitating the approval of a new formulation for public usage.

The trials are realized as cross-over trials where a subject gets each formulation administered separately over two time periods with a sufficient wash-out period in between. The sequence, i.e. the order of Test and Reference administration periods is also factored in. For each period the concentration of the formulation in the blood is measured at fixed time points. This results in two concentration profiles for each subject, one for the period where the Test formulation was administered and one for the Reference formulation. The largest observed value of a concentration profile is denoted as the Cmax value. A trial fails if the 90%-confidence interval for the ratio of the respective mean Cmax of both formulations, i.e. the exponentiated treatment effect (see equation 4), reveals too high a discrepancy between the two treatments. Such is the case, if the interval doesn't fall within a certain acceptance range, typically 80% to 125% (World Health Organization). A detailed formulation of the underlying linear model will be presented in the Methods section.

$$\log(Cmax_R) = \mathbf{x_i'}\hat{\boldsymbol{\beta}} + 0 \cdot \hat{\beta}_{treatment} \tag{1}$$

$$\log(Cmax_T) = \mathbf{x_i'}\hat{\boldsymbol{\beta}} + 1 \cdot \hat{\beta}_{treatment} \tag{2}$$

$$\implies \log\left(\frac{Cmax_T}{Cmax_R}\right) = \hat{\beta}_{treatment} \tag{3}$$

$$\frac{Cmax_T}{Cmax_R} = \exp(\hat{\beta}_{treatment}) \tag{4}$$

Fuglsang [2021] outlines one method of manipulating these bioequivalence trials in favor of the approval of the new Test formulation: while conducting the trial,

an undocumented statistical analysis is performed partway through. This analysis evaluates the ratios between the Cmax of the concentration profiles that were obtained up to this point. If the initial point estimate of the ratio between Test and Reference formulations suggests that the trial is failing, for instance a value above 1.25 or below 0.8, an act of manipulation is performed.

This act entails to first detect those subjects among the initially analysed that exhibit the most adverse behavior, i.e. the ones with the lowest ratios if the initial point estimate is too low or with the highest ratios if the initial point estimate is too high. For these subjects, Test and Reference values are then swapped and they are brought back into the trial under the identities of yet-to-be-analysed subjects. For example a subject with a ratio of 0.6 would be re-injected as a subject with a ratio of $\frac{1}{0.6} \approx 1.67$. This act of manipulation evidently skews the overall point estimate of the trial towards the desired direction. Alternatively or additionally to switching, samples can be diluted. For example, a ratio of 0.6 with a two-fold Reference dilution results in a ratio of $\frac{0.6}{0.5} = 1.2$.

Per Fuglsang [2021], these acts of data manipulation are not traceable by the commonly used methods of audit trials. Consequently, this opens up the opportunity of contemplating and developing techniques in order to reliably detect this form of fraud in the field of bioequivalence trials.

In the following we will first outline the data that was provided as well as our own simulated data. Then, we will summarize the methods for fraud detection developed by Fuglsang [2021] and showcase our replication of his findings as well as another application of the methods. Next, we will derive our own method for detecting the manipulation and outline the underlying statistical theory. Then, we will describe the process of applying those methods while also assessing their performance. Finally, we will contemplate our findings and reflect on their meaningfulness.

2 Data

2.1 Provided Data Sets

Two data sets were provided to us at the beginning. The first data set is one created by Fuglsang [2021] and used for showcasing his findings in the paper.

It is simulated and contains 72 concentration profiles obtained from a standard 2-treatment, 2-period, 2-sequence bioequivalence trial of 36 subjects. For each observation the subject and period numbers as well as the respective concentration levels for 19 defined time points are given. From the profiles the Cmax values can be obtained trivially. An act of manipulation as described in the previous section has been performed on the data: among the first 24 subjects, the 12 subjects with the lowest Cmax ratios have been re-injected with Test and Reference switched. The two subjects with the 11th and 12th lowest ratio additionally had the Reference profile diluted by a factor 2.

The other data set is disclosed by Health Canada for non-commercial purposes and contains the data from an actual bioequivalence study conducted by Apotex Research Pvt. Ltd.. For this study the test product was Tenofovir Disoproxil Fumarate Tablets 300 mg and the reference product was Viread Tablets 300 mg. The data was only available through a PDF document of the final study report. Thus, some pre-processing was required at first to transform the data into a readable input format. The data set contains again concentration profiles obtained from a standard 2-treatment, 2-period, 2-sequence bioequivalence trial. In this case, it was comprised of 21 subjects, resulting in 42 concentration profiles measured at 24 time points. Again, the Cmax values could be obtained trivially. For this data set we were to assume no presence of the type of data manipulation that is relevant to this project.

2.2 Simulated Data Sets

To allow for a more thorough examination of the performance of various detection methods down the line, we decided to write a function in `R` that produces simulated datasets containing the results of bioequivalence trials. This way we would have access to as many data sets as needed while also knowing their true status. For our purposes, only the simulation of the Cmax values was required. The function is designed as follows:

```
> sim_data(n, mean, sd, treat_effect, seed, manipulation = TRUE)
```

The input `n` controls the size of the data set, i.e. how many subjects it contains, and the `seed` enables reproducibility. There is also the option of performing data manipulation on the data set. The input parameters `mean`, `sd` and `treat_effect`

are to be given on a log-scale. The function then takes the `mean` input and adds a random normally distributed error term to arrive at the `n` simulated Reference Cmax. For the Test Cmax it additionally adds the input `treat_effect`. The error term follows a normal distribution with distribution parameters $\mu = 0$ and $\sigma =$ `sd`. This approach is in accordance with the assumption stated by Fuglsang [2021], that the logarithmised Cmax are normally distributed. The function then exponentiates the simulated Cmax values and sorts both Test and Reference values by increasing order. This way, we are able to introduce a subject effect, since one subject gets attributed the highest Cmax value of Test and Reference each, another subject the second highest of each and so on. Then, the data set get shuffled pairwise.

If the option of manipulation is selected, the manipulation procedure is performed at this point. For this, we assume an initial analysis is conducted after two-thirds of the samples have been obtained, meaning from two-thirds of the subjects the ratios are computed and sorted. If the input `treat_effect` is negative, the half with the lowest ratios are selected as manipulation candidates. For a positive input the half with the highest ratios become candidates. For these manipulation candidates the Test and Reference values get swapped. Additionally, for a quarter of the candidates a dilution by factor 2 is applied to the appropriate side. This concludes the manipulation procedure. The general process as well as the proportions used are based on the manipulated Fuglsang data set.

Subsequently, all subjects get a sequence 'RT' or 'TR' randomly allocated and each period value gets the appropriate corresponding period number denoted. As this is a random allocation, the sequence effect is assumed to be non-significant. Finally, the data set is sorted by subject. An excerpt of an exemplary data set produced by our function can be seen in table 1.

For the data sets to be used in our simulation study, we chose the data-specific input parameters by reference to the non-manipulated part of the Fuglsang data set. Based on this we arrived at the following values: `mean = 6.49`, `sd = 0.2`, `treat_effect = -0.311`.

subject	cmax	treat	sequence	period
1	669.3158	R	RT	1
1	514.5085	T	RT	2
2	766.6055	R	RT	1
2	611.8824	T	RT	2
⋮	⋮	⋮	⋮	⋮

Table 1: Excerpt of a simulated dataset

3 Literature Review and Replication

Fuglsang [2021] establishes two methods for detecting and analyzing the specific type of fraud in bioequivalence studies explained in the problem outline.

3.1 Buster

3.1.1 Overview

The *Buster routine* visualizes the data used in the trial to detect patterns or anomalies that indicate manipulation. Fuglsang [2021] introduces five different kinds of plots:

1. *Buster* plot 1: A bar plot that visualizes the difference of the logarithmized maximal concentrations in the test group to their mean, i.e.

$$\log Cmax_{T_i} - \overline{\log Cmax_T}.$$

2. *Buster* plot 2: The same plot as described before, only with observations from the reference group instead of test group, i.e.

$$\log Cmax_{R_i} - \overline{\log Cmax_R}.$$

3. *Buster* plot 3: A plot visualizing the linear model used to examine the success of the trial; the details of this procedure will be described later. The plot shows point estimates of the treatment effect and their 90%-confidence interval.

4. *Buster* plot 4: A plot visualizing the mean squared errors (MSE) of the models examined in plot 3. Here, the MSE is defined as the mean squared residuals, i.e.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

where y_i are the true values for the maximal blood concentrations obtained in the study and $\hat{y}_i$ are the fitted values estimated by the model.

5. *Buster* plot 5: A bar plot of the residuals of the test group obtained from the model mentioned before. The residuals are calculated as

$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

3.1.2 Details

The linear model that is used to determine whether or not a formulation passes the trial takes the form of

$$\log Cmax_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma^2),$$

with β_1 as the treatment effect, β_2 as the subject effect, β_3 as the period effect and β_4 as the sequence effect. In order to arrive at plot 3 described before, this model is fit iteratively with 5,6,...,N subjects in the analysis. This simulates the analysis conducted partway through described in the problem outline. The treatment effects and their corresponding 90%-confidence intervals of each of these models are exponentiated and plotted versus the given values 0.8 and 1.25 that span the interval in which the 90%-confidence intervals must lie in order to pass the trial.

In general, all of the *Buster* plots reveal manipulation by showing a sudden break of a pattern in the data. Buster plot 1 and 2 indicate fraud when there is a change in the sign. Buster plot 3 indicates fraud when the point estimates suddenly move away from the value they were converging to before and the confidence intervals suddenly widen although there are more subjects injected into the analysis. The same behavior is shown in plot 4 when the MSE starts

to increase at a specific point. Examples of these plots can be found in Figures 1-5.

3.2 SaToWIB

SaToWIB is a method of comparing blood concentration profiles and measuring their similarity. It is the second method introduced by Fuglsang [2021]. Once *Buster* has shown strong signs of data manipulation, *SaToWIB* can be used to determine which observations were most likely included in the fraudulent manipulation, however, it cannot be used to detect manipulation alone. *SaToWIB* makes use of scores to calculate the similarity of two blood concentration profiles. These scores have to abide by certain rules:

1. A score of zero must indicate a perfect match, i.e. the two concentration profiles are identical
2. The score must only take non negative values and any value above zero indicates a difference in the profiles.
3. The score must be symmetric so it does not matter whether one calculates the score of profile A and profile B or profile B and profile A.
4. If profile A is more similar to profile B than to profile C, it must hold that $S_{AB} < S_{AC}$, where S_{ij} denotes the similarity score of profile i and profile j.

Fuglsang [2021] proposes the following metric:

$$Score = \frac{1}{N} \sum_{i=1}^{N} \frac{|(v_{1i} - v_{2i})|}{\frac{1}{2}(v_{1i} + v_{2i})},$$

where v_{1i} is the i'th concentration of concentration profile 1 and v_{2i} is the i'th concentration of concentration profile 2. A problem arises when $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ have a zero on the same position in the vector. The term $\frac{1}{2}(v_{1i} + v_{2i})$ would imply a division by zero. To tackle this issue, we add `.Machine\$double.xmin`, which equals $2.225074e-308$ to the denominator. This value should be small enough to not significantly alter results.

The obvious weakness of this method is manipulation by dilution. When comparing two profiles of the same subject, the score can be artificially increased

when one profile is diluted. In order to tackle this problem, Fuglsang [2021] introduces "method 32". It consists of dividing every concentration in a profile by the sum of the three largest concentrations in that profile. This scales all profiles to a similar level.

3.3 Replication

We implemented the methods described above in `R` and replicated the results of Fuglsang [2021]. We also applied these methods to the Canada data.

3.3.1 Fuglsang

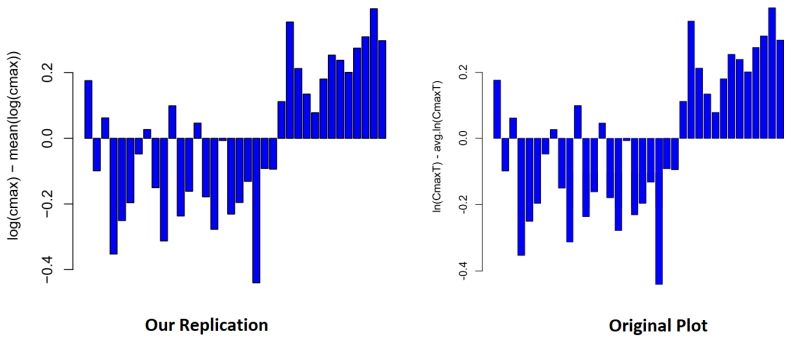


Figure 1: Buster plot 1: Replication vs. Fuglsang [2021]

As one can see from Figure 1, our replication looks the exact same as Buster plot 1 in Fuglsang [2021] except for minor aesthetic differences. The same can be said for Figure 2. Figure 3 is the first plot to show slight differences between our replication and the original plot by Fuglsang [2021]. They can be seen in the width of the confidence intervals. The most likely reason for these differences are variance estimators. Fuglsang [2021] does not specify the estimation of the underlying linear models. Qualitatively, however, the plots are equivalent. The same points about the variance mentioned before can be seen in Figure 4. In Figure 5 our replication looks largely the same as the original plot by Fuglsang [2021] except for minor differences that again occur most likely due to differences in model estimation.

Overall, all of these plots indicate the manipulation explained in the problem

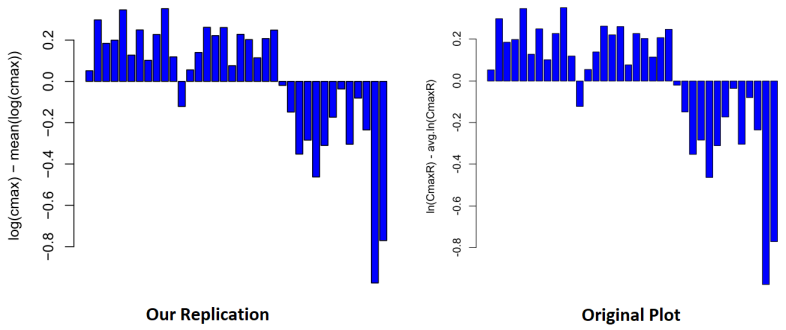


Figure 2: Buster plot 2: Replication vs. Fuglsang [2021]

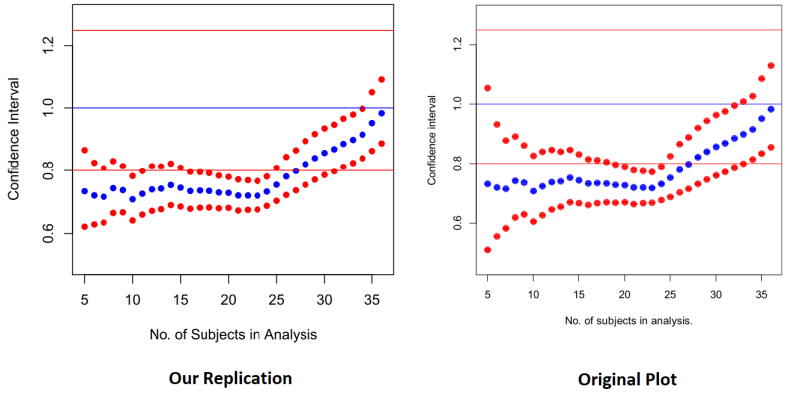


Figure 3: Buster plot 3: Replication vs. Fuglsang [2021]

outline. It is clearly visible that after the first 24 subjects there is a sudden and significant break in the data. The manipulation changed the underlying distribution of the data fundamentally. When looking at Figure 3 one can see that aside from the point estimates suddenly moving upwards after injecting subject 24 into the analysis, the confidence intervals also get wider. This can be explained by the added uncertainty that comes with changing the underlying distribution during the analysis.

Tables 2 and 3 show an excerpt of the results of the *SaToWIB* routine from Fuglsang [2021] and our replication. Since there are 2556 profile pairs, we only show the five lowest scores. Most of the scores we obtained are identical to those

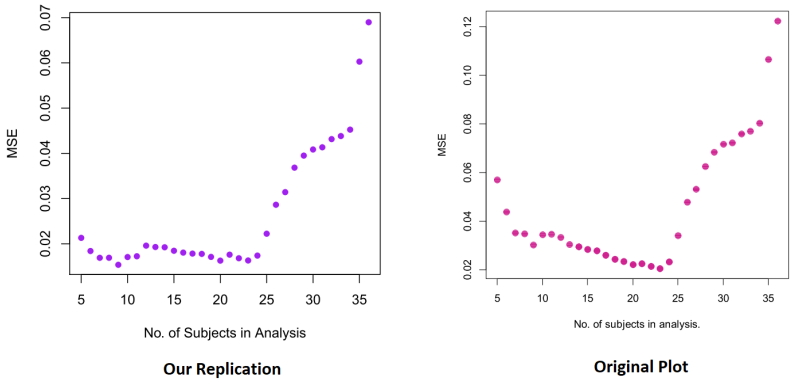


Figure 4: Buster plot 4: Replication vs. Fuglsang [2021]

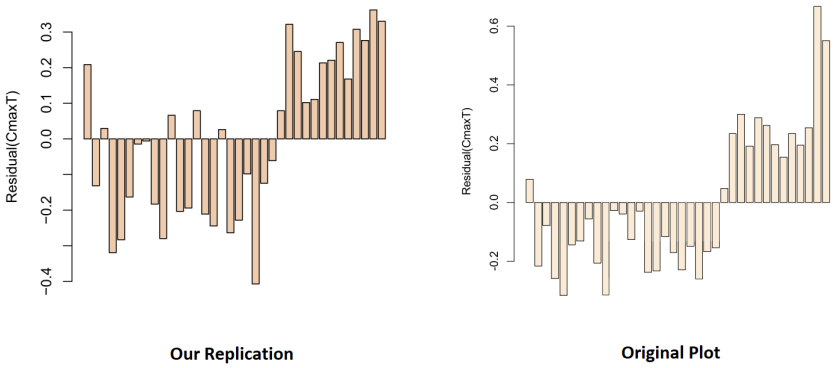


Figure 5: Buster plot 5: Replication vs. Fuglsang [2021]

Profile 1	Profile 2	Score	Ratio
S18P1	S27P2	0.03891	0.9561
S21P1	S28P2	0.03990	1.0051
S4P1	S26P2	0.04003	1.0189
S5P1	S25P2	0.04041	0.9901
S2P2	S33P2	0.04182	0.9639

Table 2: SaToWIB Replication

Profile 1	Profile 2	Score	Ratio
S18P1	S27P2	0.03891	0.9561
S21P1	S28P2	0.03990	1.0051
S5P1	S25P2	0.04041	0.9901
S2P2	S33P2	0.04182	0.9639
S9P2	S30P1	0.04187	1.0264

Table 3: SaToWIB: Original

in Fuglsang [2021], however, at some places the order of profile pairs is different. Overall, of the 18 profile pairs with the lowest scores, 17 are manipulated, both in the original results and our replication.

3.3.2 Canada Data

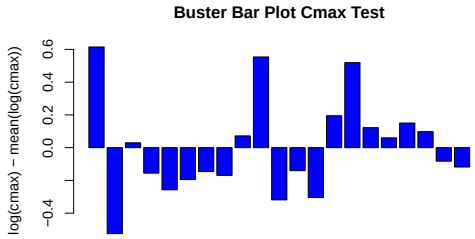


Figure 6: Buster plot 1: Buster on the Health Canada [2014] data

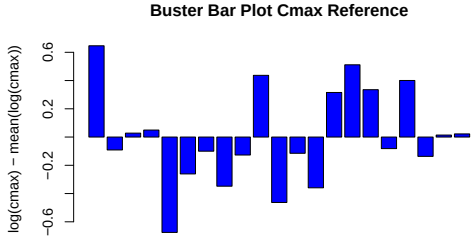


Figure 7: Buster plot 2: Buster on the Health Canada [2014] data

Figures 6-10 show the *Buster* routine applied to the data of Health Canada [2014]. The plots show no signs of fraudulent manipulation. Figures 6, 7 and 10 exhibit multiple changes of sign and no significant deviance from the mean for an extended series of observations. In Figure 8 we can observe little changes in the point estimates for the treatment effect. After initial uncertainty, they seem to converge to a value slightly above one. Furthermore, the confidence intervals become narrower the more subjects are injected into the analysis. Figure 9 also displays the behavior one would expect from non-manipulated data. The MSE

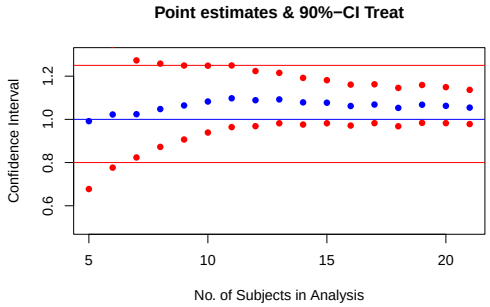


Figure 8: Buster plot 3: Buster on the Health Canada [2014] data

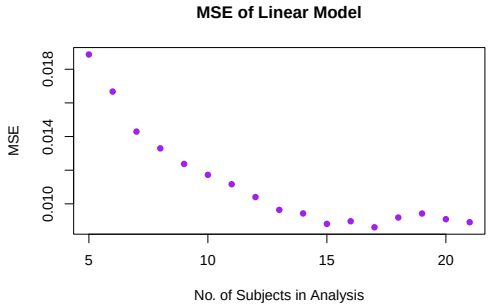


Figure 9: Buster plot 4: Buster on the Health Canada [2014] data

decreases the more subjects are injected into the analysis up to a certain point. It fluctuates for the last few observations.

Overall, the Health Canada [2014] data does not seem to be manipulated, at least not in a way the *Buster* routine would be able to detect. Since there are no signs of manipulation, there is no reason to apply the *SaToWIB* routine to the data.

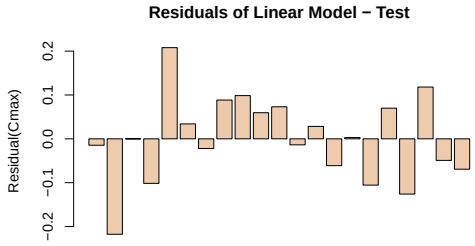


Figure 10: Buster plot 5: Buster on the Health Canada [2014] data

4 Statistical Background

The previous section exhibited and applied the methods introduced by Fuglsang [2021]. In this section we will try to extend these methods and establish statistical tests to detect fraud in bioequivalence studies. More specifically, we want to quantify the visual assessment of the *Buster* routine. In order to create these methods, we will first introduce the statistical background necessary to conduct these tests. The main focus of this section lies on time series analysis.

4.1 Time Series

From *Buster* plot 3 we obtain a series of point estimates. These point estimates are not independent of one another, since every point estimate depends on the observations also used to obtain the previous point estimates. We can interpret this property as a time series. More specifically, we can model the series of point estimates as an autoregressive (AR) stochastic process. Hackl [2013] defines an AR(p)-process as

$$T_t = \alpha_1 T_{t-1} + \alpha_2 T_{t-2} + ... + \alpha_p T_{t-p} + u_t, \quad u_t \sim i.i.d.(0, \sigma^2),$$

with treatment effect T_t at point t, regression coefficients α_j and error term u_t.

When denoting the linear model that is used to obtain the point estimates in matrix notation, i.e.

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

one can easily show that the OLS estimator $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ is an unbiased estimator of the true parameter $\boldsymbol{\beta}$. i.e.

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

This holds true for all estimated treatment effects obtained from arriving at *Buster* plot 3, which means that the mean of the time series of point estimates is time invariant. A time series with time invariant mean and time invariant variance is called stationary.

The condition for stationarity can be denoted as the following:
For all roots $z_i, i = 1, ..., p$ of the characteristic polynomial

$$\phi(z) = 1 - \alpha_1 z - ... - \alpha_p z^p$$

it holds that $|z_i| > 1$. If one of these roots $z_i = 1$, the process is non-stationary. From all of this we conclude that a non-manipulated series of treatment effect estimates should resemble a stationary time series and that a unit root test should confirm that suspicion.

4.2 The Augmented Dickey-Fuller Test

The Augmented Dickey-Fuller (ADF) Test is a unit root test developed by Dickey and Fuller [1979]. The following short outline is based on Kirchgässner et al. [2013].

The ADF Test tests the null hypothesis of the presence of a unit root for an AR(p) process:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} + u_t$$

with its reparameterisation:

$$y_t = \rho\, y_{t-1} + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + ... + \theta_{p-1} \Delta y_{t-p+1} + u_t$$

where

$$\rho = \theta_0 = \sum_{j=1}^{p} \alpha_j, \quad \theta_i = -\sum_{j=i+1}^{p} \alpha_j, \quad i = 1, 2, 3, ..., p-1.$$

The null hypothesis is represented by $\rho = 1$. If testing against a stationary process, the alternative hypothesis is represented by $|\rho| < 1$, whereas for testing against an explosive process it is represented by $\rho > 1$.

The test statistic $DF = \dfrac{\hat{\rho} - 1}{SE_{\hat{\rho}}}$ follows a distribution derived from simulations.

4.3 The Kwiatkowski–Phillips–Schmidt–Shin Test

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) Test was developed by Kwiatkowski et al. [1992]. It tests the null hypothesis of stationarity around either a mean or a linear trend. The following outline is again based on Kirchgässner et al. [2013]

For a decomposition of a time series

$$y_t = \alpha_t + \beta t + u_t$$

which contains a random walk $\alpha_t = \alpha_{t-1} + \varepsilon_t$ with $\varepsilon_t \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$, the null hypothesis for trend stationarity is represented by $\sigma_\varepsilon^2 = 0$. The null hypothesis for level stationarity is equivalent, except setting $\beta = 0$ takes out the linear trend part of the series, resulting in the residuals $\hat{u}_\mu$ instead of $\hat{u}_{Tr}$.
The test statistic is

$$\hat{\eta}_j = \frac{1}{T^2} \frac{\sum_{t=1}^{T} (S_{t,j})^2}{\hat{\sigma}_u^2}$$

where $S_{t,j} = \sum_{i=1}^{T} \hat{u}_{i,j}$ and $j = \mu$ when testing for level stationarity and $j = Tr$ when testing for trend stationarity. The distribution for this test statistic is again derived from simulations.

5 Testing Process

In the follwing section we will outline our considerations of various test procedures for fraud detection and an assessment of their respective performances. At the end we will present an overview of our findings in terms of sensitivity, specificity and overall accuracy of the testing methods.

These assessments are based on treating the series of point estimates for the treatments effect as a time series as described in the previous section. The

point estimates arise from simulated data sets with the specifications as noted in the corresponding section. We looked at 50 data sets with no manipulation and 50 data sets with manipulation. The sample size for each data set was 36 subjects.

5.1 KPSS Test

Our first approach was to establish the stationarity of a series, and consequently no indication of manipulation, through the means of a statistical test. For this purpose, the KPSS Test with a null hypothesis of level stationarity seemed appropriate. Our chosen level of significance was 5%. We used the test implementation from the `tseries` [Trapletti and Hornik, 2021] package in `R`.

Out of the 50 manipulated data sets 50 were recognized as such. On the other hand, the test also classified 29 of the non-manipulated data sets as manipulated. This is obviously not a satisfactory performance with respect to specificity. It also reveals unexpected behavior, since for a significance level of 5%, one would expect around 3 false positive classifications for this sample size. The KPSS Test with a null hypothesis of trend stationarity achieved the same sensitivity and performed slightly worse in terms of specificity.

5.2 ADF Test

Next, we looked at the ADF Test with the null hypothesis of a unit root against the hypothesis of a stationary process. Again, the chosen significance level was 5% and the test implementation from the `tseries` [Trapletti and Hornik, 2021] package was employed. This test was also able to detect all 50 of the manipulated cases. However, in terms of specificity it performed even worse than the KPSS test and classified 43 of the non-manipulated data sets as manipulated. Consequently, this version of the test offers no meaningful insight for our purposes.

We also considered the ADF Test with an alternative hypothesis of an explosive process. The applied significance level was again 5%. Out of the 50 manipulated data sets this test wasn't able to detect any as such. On the other hand, one of the 50 non-manipulated data sets was classified as manipulated. This results in a high specificity, but evidently no utility in regard to sensitivity.

5.3 Confidence Interval Width Rule

Looking at the poor performance of the statistical tests so far, we decided to try a different approach. When reviewing the Buster plot 3, one can observe that the confidence intervals for the point estimates generally get wider after the point where the manipulation happens. This is sensible since the manipulation procedure introduces new data points of a differing distribution and consequently increases the variance of the data overall. This added uncertainty is reflected in the width of the confidence intervals. For non-manipulated data on the other hand, the confidence intervals are generally expected to get smaller, since more data points are included in the underlying model. Thus, we decided to define and implement a rule that classifies a data set as manipulated, if the CI-width increases a certain number of times consecutively. We wrote a corresponding function in `R` where the required number of consecutive increases could be set as an input parameter. For our data sets of 36 subjects we decided on a parameter of 3.

The CI-Width test was able to detect all of the 50 manipulated cases. It also displayed a respectable performance in terms of specificity. Out of the 50 non-manipulated cases it classified 6 as manipulated. While there is still some room for improvement, this method presents an acceptable option for this type of fraud detection. On the downside, one loses the advantageous properties that a statistical test offers, such as defining an acceptable α-error.

5.4 KPSS Test Reconsideration

We decided to take another look at the statistical tests and to try to apply some alterations that might make them viable for our purposes. For the KPSS-Test, its performance was clearly lacking in terms of specificity. Meaning, data sets that should have been classified as stationary were not recognized as such by the test. As a consequence, we tried to apply a transformation to the data that would make the non-manipulated data sets appear more stationary while also not altering the manipulated ones to the point that they would be classified as stationary. We decided on a transformation where we would subtract a form of a cumulative mean from each data point. This form of cumulative mean is computed in a way where after a certain point, a number of the first observations are no longer included. Moreover, the last few observations are left out of the computation as well. The number of data points to be left out from each side

is somewhat arbitrary, for our purposes we decided on a value of 7 for each.

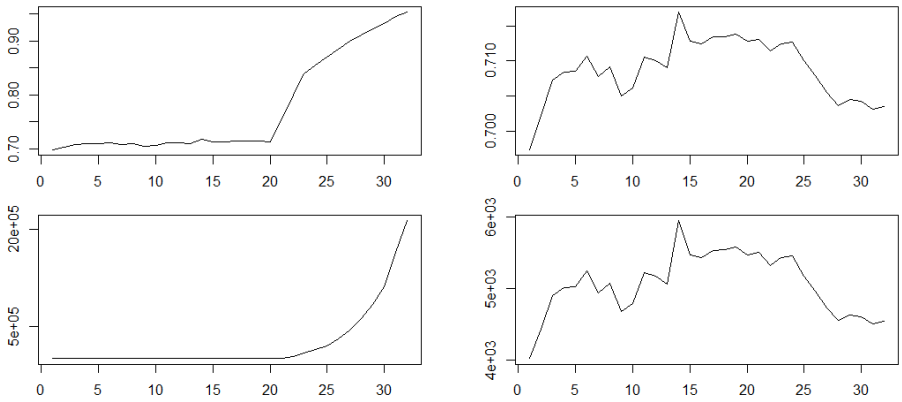
Applying this transformation to the data and then deploying the KPSS-test for level stationarity with $\alpha = 5\%$, we were able to achieve an improved performance with respect to specificity in comparison to using non-transformed data. 10 of the 50 non-manipulated cases were classified as manipulated. The sensitivity was again at 100%. This is an improvement in terms of performance and also retains the advantages of using a statistical test. On the other hand, as mentioned before, the choice of values for the computation of the cumulative mean is somewhat arbitrary and might result in unreliable behavior. Please note that in table 4 this method is labeled as KPSS (H_0 : *level-stat.*) + CM.

5.5 ADF Test Reconsideration

We also considered data transformations that might improve the performance of the ADF-Test. Here we looked at the ADF-Test with the alternative hypothesis of an explosive process. While performing well in terms of specificity, the issue with this test was, that it wasn't able to recognise the explosive behavior in the manipulated data and consequently didn't classify them as manipulated. Based on this, our approach for the data transformation here was to amplify the explosive behavior in the manipulated data. Thus, we were looking for transformations that would amplify larger values and simultaneously scale the data in such a way that for large values the slope would increase faster. This resulted in the following transformation formula, where b_i denotes the observation before and $\tilde{b}_i$ the observation after applying the transformation:

$$\tilde{b}_i = \exp((b_i + 1)^4)$$

Adding 1 to b_i ensures that all $b_i > 1$, so that the following operations would work appropriately. Applying the fourth power causes the amplification of larger values while the exponentiation induces exponential growth in the data, increasing the intervals between values of data points. This exaggerates any explosive trends already inherent to the data and potentially makes explosiveness more detectable for statistical tests. Figure 11a shows an example of the effect of applying the transformation to a manipulated data set. We can see the desired effect has been achieved and the explosive behavior is amplified in terms of scale and slope.



(a) Manipulated data set (b) Non-manipulated data set

Figure 11: Plot of the point estimates without transformation (above) and with transformation (below)

Obviously, this transformation is only useful, if it does not affect the behavior of the non-manipulated data in any crucial way. Figure 11b shows that outside of the magnitude of the values the non-manipulated data remains unaffected.

We deployed the ADF-Test with the alternative hypothesis of an explosive process and an $\alpha = 5\%$ on the simulated data sets with the aforementioned transformation applied. Out of the 50 manipulated cases the test was now able to detect all 50. This is a definite improvement to the test when using the non-transformed data. Moreover, 48 of the 50 non-manipulated cases were correctly classified. Thus, the tests achieves a sensitivity of 100% and a specificity of 96%. Since only 2 of the non-manipulated data sets were classified as manipulated, this amounts to a false positive rate of 4%, which is in accordance with our chosen significance level. Consequently, this methods not only accomplishes satisfactory results in terms of performance, but also displays the expected statistical properties. Please note that in table 4 this method is labeled as ADF $(H_1 : explosive)$ + TF.

5.6 Overview

Table 4 presents a summary of the various test performances. In terms of overall accuracy the ADF $(H_1 : explosive)$ + TF method achieves the best performance. It also displays excellent results in respect to both sensitivity and specificity.

Method	Sensitivity	Specificity	Accuracy
KPSS (H_0 : *level-stationary*)	100%	44%	72%
ADF (H_1 : *stationary*)	100%	10%	55%
ADF (H_1 : *explosive*)	0%	98%	48%
KPSS (H_0 : *level-stat.*) + CM	100%	80%	90%
ADF (H_1 : *explosive*) + TF	100%	96%	98%
CI-Width Rule	100%	88%	94%

Table 4: Overview of test performances

Apart from that, only the KPSS (H_0 : *level-stat.*) + CM and the CI-Width Rule methods can achieve adequate overall results. A contemplation of those three methods in particular as well as their respective usages and limitations in practice will follow in the discussion section.

6 Discussion

6.1 Test Procedures

Based on our findings, the ADF (H_1 : *explosive*) + TF method clearly showed the best performance and seems like the most suitable candidate for fraud detection. Since it is a statistical test, one can make use of its known properties. On the other hand, it is not certain, that the transformation that was used here is generally applicable. It might need some adjustments depending on specific situations. However, the idea of amplifying the abnormal behavior of the manipulated data should still hold and allow for a comparable performance using this method. Moreover, for smaller sample sizes, which are not too unusual in the field of bioequivalence studies, we found that the test power slightly declines, which seems somewhat inevitable.

The KPSS (H_0 : *level-stat.*) + CM method performs well in terms of sensitivity. For situations where a high specificity is not required, it could be a

viable option, if the ADF ($H_1 : explosive$) + TF method is not employable, but a statistical test is still desired. However, as mentioned before, the inputs for computing the cumulative mean are somewhat arbitrary and non-transparent.

If one can forgo the usage of a statistical test, the CI-Width Rule method represents a serviceable alternative. It offers a reliable performance in regard to sensitivity and an adequate rate of specificity. By adjusting the input parameter of the number of required increases in CI-width, one can also further improve the specificity, albeit potentially at the cost of sensitivity. This would depend on the objective of the detection analysis. A smaller sample size can be compensated for similarly.

While all our test methods were developed and assessed based on the type of manipulation that was described by Fuglsang [2021], they should also provide some general usefulness in detecting at least unusual or unexpected behavior in the data of bioequivalence trials. The tests based around stationarity detect, if a noticeable trend is introduced to the data at a certain point in the evaluation. The CI-Width Rule is even more general in its detection since it is only based on discerning an odd increase of variation in the data.

6.2 Practical Use

Based on the properties of the statistical tests described before we can now consider potential use cases. The practical usefulness of these tests depends on the prevalence of manipulation in the real world. To demonstrate this point, consider the Bayes theorem (Bayes [1763]) in the form of

$$PPV = P(M^+|T^+) = \frac{P(M^+)P(T^+|M^+)}{P(M^+)P(T^+|M^+) + (1 - P(M^+))(1 - P(T^-|M^-))}.$$

with prevalence of manipulation $P(M^+)$, test sensitivity $P(T^+|M^+)$ and test specificity $P(T^-|M^-)$. This equation gives the positive predictive value PPV; it is the probability that we actually observe manipulation given that the test suggested that there is manipulation. We can use this metric to determine how useful our tests actually are given the prevalence of manipulation. For this purpose we calculate the PPV for a few different prevalences. This illustration will only be done on the test with the best performance, i.e. ADF ($H_1 : explosive$) + TF, since its implications hold true for all other tests. As can be seen in table

$P(M^+)$	PPV
1/3	92.59%
0.1	73.53%
0.01	20.16%
0.001	2.44%

Table 5: Prevalence vs. Positive Predictive Value

4 the test has a sensitivity of 100% and a specificity of 96%.

Table 5 shows the $PPVs$ of ADF ($H_1 : explosive$) + TF for several different prevalences. It becomes clear that the test's efficiency depends on how often manipulation actually happens. If only one in 1000 studies are manipulated, the probability of observing manipulation given a positive test is 2.44%. This means that 97.56% of positive tests are false positives. This clearly restricts the practical use of statistical tests for fraud detection. However, these tests can still be used as screening tools. If a study shows a positive test result, it should be further investigated. In that case other tests mentioned in this work can additionally be used on these suspicious studies to determine which of them are likely to contain manipulated data.

This also raises the question which metric one should focus on when deciding which test to use. Is sensitivity more important than specificity or the other way around? This highly depends on the use case. Using the CI-width rule test one can create highly specific tests by increasing the input parameter of the test. However, this would lead to a loss in sensitivity. If a test is supposed to find as many manipulated studies as possible, one should use a test with the highest sensitivity. If, on the other hand, one wants to be sure that a positive test result actually implies manipulation, e.g. because time and money are limiting factors, one should use a highly specific test. Since we recommend these tests as screening tools and not as definitive evidence of manipulation, sensitivity is the deciding metric to be considered.

6.3 Considerations

We also want to investigate how generalizable these methods are. One might be tempted to think that the tests introduced in this work are only applicable to the specific method of manipulation described by Fuglsang [2021]. However,

this is not entirely true. The tests do not depend on switching reference and test samples or diluting them. They only depend on the partway through analysis and the following manipulation of the underlying distribution of the data. It does not matter how that change in the underlying distribution has occurred. This means that our tests should be able to detect every sort of manipulation as long as it happens during the analysis. Especially the CI-width rule test should capture every sort of manipulation that increases estimation uncertainty from a certain point onward. Should the size of confidence intervals increase during the analysis, there has to be something that causes this additional uncertainty. How that additional uncertainty came about is not relevant to the test.

Overall, the choice of methods to detect fraud is limited by the sample sizes common in bioequivalence studies. Methods like clustering or machine learning approaches are not viable as they usually require a larger amount of observations. More heuristic approaches like visualization and the tests introduced in this paper are more suitable for the sample sizes one usually encounters in bioequivalence trials. Still, even these approaches suffer from a loss of accuracy in the lower range of sample sizes.

7 Conclusion

This project addressed the problem of fraud detection in bioequivalence trials. It was motivated by a research paper by Fuglsang [2021] where the author outlined a specific form of data manipulation as well as methods, the Buster routine in particular, that can be utilized to detect this manipulation.

First, we successfully implemented the methods developed in the paper, which resulted in an effective replication of Fuglsang's findings. We also applied the method to the provided data from an actual bioequivalence trial and were able to affirm the assumption of no manipulation in this data.

Subsequently, we derived our own methods based on the idea of quantifying the visual assessments from Fuglsang's methods by treating the point estimates of the treatment effect as a time series. For this purpose, various testing procedures were developed and applied and their performance was assessed using simulated data. We were able to devise a method of using the ADF Test, which

tests against an explosive process, on the transformed point estimates. The transformation serves the purpose of amplifying the behavior caused by the manipulation to facilitate a successful detection by the statistical test. This method achieved satisfactory results both in terms of sensitivity as well as specificity.

Finally, we reflected on our findings and their limitations and applications in practice. While the utility of our findings is influenced by various factors, such as the prevalence and type of manipulation as well as the objective of the detection analysis, the methods should still be able to offer some additional insight for investigating data manipulation in bioequivalence trials.

References

Anders Fuglsang. Detection of data manipulation in bioequivalence trials. *European Journal of Pharmaceutical Sciences*, 156:105595, 2021. ISSN 0928-0987. doi: http://doi.org/10.1016/j.ejps.2020.105595.

Health Canada. Available information for APO-TENOFOVIR - submission control number 172688. `http://clinical-information.canada.ca/ci-rc/item/172688`, 2014. Accessed: 17.11.2021.

World Health Organization. Annex 7: Multisource (generic) pharmaceutical products: guidelines on registration requirements to establish interchangeability. `http://www.who.int/medicines/areas/quality_safety/quality_assurance/Annex7-TRS992.pdf`, WHO Technical Report Series No. 992, 2015. Accessed: 09.03.2022.

Peter Hackl. *Einführung in die Ökonometrie*. Pearson, 2013.

D. Dickey and Wayne Fuller. Distribution of the estimators for autoregressive time series with a unit root. *JASA. Journal of the American Statistical Association*, 74, 06 1979. doi: 10.2307/2286348.

Gebhard Kirchgässner, Jürgen Wolters, and Uwe Hassler. *Introduction to Modern Time Series Analysis*. Springer, February 2013. ISBN 978-3-642-33436-8.

Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178, 1992. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(92)90104-Y.

Adrian Trapletti and Kurt Hornik. *tseries: Time Series Analysis and Computational Finance*, 2021. URL `https://CRAN.R-project.org/package=tseries`. R package version 0.10-49.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. of the Royal Soc. of London*, 53:370–418, 1763.