European Journal of Pharmaceutical Sciences 156 (2021) 105595

Contents lists available at ScienceDirect

European Journal of Pharmaceutical Sciences

journal homepage: www.elsevier.com/locate/ejps

Detection of data manipulation in bioequivalence trials

Anders Fuglsang

Hiort Lorenzens Vej 6c, DK6100 Haderslev, Denmark

ARTICLE INFO

Keywords:
Fraud
Bioequivalence
Residuals
Similarity
Re-analysis

ABSTRACT

In recent years regulators have documented how pharmaceutical companies or clinical research organisation can manipulate bioequivalence trial data for non-approvable formulations by performing an interim analysis followed by re-analysis of pharmacokinetic profiles under new subject aliases, with a switch of Test and Reference and/or dilutions. The net effect is that point estimates for failing products will be forced artifically towards 1 and that trials will pass the test for bioequivalence. This is not detectable by any pharmacopoeial method, and is not addressed by common assessment practices at agencies. This paper aims at demonstrating how the signals of such fraudulent study conduct can be detected. The approaches presented are called ”Buster” and ”SaToWIB” routines; these are computer programs that have been used extensively by regulators to detect signals of fraud but they have not been described in the public domain.

The Buster routines visualize trends in the form of partial statistics, residual plots, cumulative confidence intervals, cumulative mean squared errors, and more. Runs tests on the sign of the residuals may constitute a potential test for the manipulation. It is noteworthy that in 2020, regulators in the European Union have publicly begun questioning trial validity on basis of PK profile similarity. The SaToWIB routines rank profile pairs according to numerical similarity on basis of an objective function. It is shown that the rank (as determined by score) is an indicator of fraud in that the actual fraud cases will have higher rank than if there were no relationship between rank and score.

The paper also comments on the use of multivariate statistics and discusses the need for development of formal tests for manipulation in view of e.g. multiplicity.

1. Introduction

Bioequivalence trials are typically comparative pharmacokinetic studies in which healthy human volunteers are exposed to two or more formulations and where the rate and extent of absorption into the blood stream is compared via quantification in plasma or other matrix. The outcome of a bioequivalence trial is in most cases presented as 90% confidence intervals for the ratios of Cmax (maximum observed concentration) and for AUC's (area under the concentration time curve) in one or more varieties. A trial is declared failing if the confidence intervals are not contained within the acceptance range. For a typical trial, the acceptance range is 80.00% - 125.00% (European Medicines Agency 2010; United States Food and Drug Administration 2001; United States Food and Drug Administration 2013; World Health Organization 2015).

Trials may fail by chance, or because the variability of the drugs is too large (for the given sample size), or because the rate or extent of absorption simple is not similar. The latter is a formulation issue.

In recent years, some CROs have employed a simple yet very powerful method to make studies pass by manipulation of the trial outcome. There is not a single publication about it in the scientific literature, but the phenomenon is mentioned and described in one Notice of Concern from the World Health Organization (2016), an Untitled Letter and form 483 from United States Food and Drug Administration (2015, 2016). As an auditor for private companies and as a consultant for authorities I have been involved in such cases, and during the past 6–7 years I have been writing software for the detection. The need for such software became particularly obvious to me after I was invited to present a prototype of the Buster software at a meeting at EMA in 2016, where inspectors from Europe, United States FDA, WHO and elsewhere were invited. At that meeting it was obvious that the need for detection opportunities was pronounced since:

1 Neither assessors nor inspectors apparently had reliable and well-defined ways of detecting the issues in the absence of either proof or whistle-blowers.
2 The fraud was suspected to be happening at several different CROs according to a common and simple recipe.

Now, as of 2020, the software (Buster and SaToWIB routines) has been used to produce analyses for signal detection requested by European authorities, and this has resulted in more than 1300 files being handed in. EU regulators have informed me that the output of the software has been used to trigger a “potential serious risk to public

https://doi.org/10.1016/j.ejps.2020.105595
Received 6 July 2020; Received in revised form 7 October 2020; Accepted 8 October 2020
Available online 11 October 2020

health" towards a (to me) unknown applicant and that at least one referral in Europe on that basis was triggered. The software have not been shared with private companies so far.

The purpose of this paper is to put into the public domain information that allows any stakeholder (CRO, Sponsor, Authority, or any party doing due diligence on dossiers) with access to data to review it for such signals. To fulfil this purpose a dataset is put forward which displays most of the features that are detectable with statistical analysis, and to describe in detail what the Buster and SaToWIB routines do, and thereby to illustrate how the signs of the manipulation can be detected, which will enable stakeholders, namely sponsors or regulators, to create tools for detection of the phenomenon.

It seems the right time to publish this paper and the details of Buster and SaToWIB since the fraud is prevalent and its nature is described in some detail by regulatory bodies. In order to make sure this manuscript does not constitute a recipe for fraud this submission has been held off until regulators have made public how this fraud actually takes place and that it is not confined to a single company; they have done already (World Health Organization 2016; European Medicines Agency 2020). The confirmation in the public domain that detection of signs alone (such as similarity of PK-profiles), even in the absence of solid proof, would be grounds enough for questioning a study's validity was published in 2020 (European Medicines Agency 2020).

The manipulation is brought about by performing an undocumented interim statistical analysis after a portion of the subject PK-data has become available. Conceptually, if the point estimate after the first M subjects is particularly low or high and suggests the trial is failing (some of) the initially analysed samples (which have a known concentration at this point) can be re-analysed under the new subject identity, typically with Test and Reference switched in order to bring the point estimate close to one. More formally, on basis of the PK-data from the first M subjects in a trial with N subjects ($N > M$) it is possible to check what the point estimate is, and to re-analyse PK-profiles from the first M subjects under new aliases of the remaining (N-M) subjects but with Test and Reference switched.

Example: if we have a trial with 50 subjects an interim analysis can be done after the first 25 subjects. If the point estimate or the confidence interval at this point indicates a formulation issue (like GMR~75%, which is clearly outside the acceptance range) then one can identify the subjects amongst those first 25 subjects who have a particularly low Test/Reference ratio and use these to manipulate the remainder of the samples. If for example subject 10 has a distinctly low Test/Reference ratio like 0.7 we can re-inject (re-analyse) this subject as subject 26 but with Test and Reference switched. Subject 26 will now have a T/R of 1/0.7 (± analytical variation), and will thus drag the GMR upwards. An alternative option is to dilute Test or Reference; if a subject displays a Test:Reference ratio of 0.7 we can e.g. dilute the Reference 2-fold and re-inject under a new subject number without Test and Reference switched, and the new subject will have a value of 0.7/0.5 = 1.4 (± analytical variation). Dilution and sample swap may be combined in which case the result would be 1/(0.5 × 0.7) ~2.9 (± analytical variation). There will be no trace of it in audit trails on e.g. chromatographic software.

2. Materials and methods

2.1. Dataset

The main dataset files can be downloaded as supplementary material:

SWa627311.csv: Contains simulated pharmacokinetic profiles (concentration levels for every defined time point) in a dataset corresponding to a standard 2-treatment, 2-period, 2-sequence bioequivalence trial (222BE trial) with $N = 36$ subjects, and thus 72 periods. The first column being time (implicitly in hours, but the units are not of importance here) and the rest of the columns being the corresponding PK-profiles from the 36 subjects each in 2 periods. Columns are named e.g. "S10P1", meaning subject 10, period 12 and so forth. *Re.* missing values, samples being blow limits of quantification ("BLQ") etc.: For the purpose of this paper itself such phenomena are of no particular importance, but they do need to be dealt with in practice. This is discussed later.

Table 1
The manipulation used for the simulation of the dataset 627,311. In all cases Test and Reference are switched.
Example: Subject 25′s original PK-profiles were discarded and replaced by those of subject 5, and with Test and Reference switched. For both treatments, no dilution (dilutation factor 1) applied. For subject 35 and 36 a dilution for the Reference by a factor 2 was applied to the re-injected profiles defined as the Reference.

Subject ID	Original ID	Dilution, Test	Dilution, Reference
25	5	1	1
26	4	1	1
27	18	1	1
28	21	1	1
29	19	1	1
30	9	1	1
31	7	1	1
32	16	1	1
33	2	1	1
34	15	1	1
35	10	1	2
36	23	1	2

In this simulated set, the simulated GMR is around 0.8, and the samples have been manipulated as depicted in table 1; From the first 24 subjects, the 10 subjects with the lowest T/R Cmax ratios have been re-injected (re-used) under a new ID, and two subjects corresponding to 11th and 12th lowest T/R Cmax ratio amongst the first 24 subjects have been re-injected with the reference profile diluted by a factor 2. All all 12 cases of manipulation, Test and Reference have been switched.

Note that the dataset does not contain identical profiles because the simulation includes addition of analytical scatter. The simulation details are immaterial to this publication, since the manuscript solely aims to put a dataset into the public domain which has features in common with actual fraud cases and to highlight the indicators of fraud.

BUa627311.csv: Contains the extracted Cmax values from the profiles in the file SWa627311.csv in the "Var" column (the PK variable), along with subject number, period, sequence (i.e. the randomisation), and treatment.

2.2. Buster routines

These routines presents graphs and tables based on the files having the structure like in file BUa627311.csv. The graphs presented are:

1 A plot of ($\ln(\mathrm{Cmax}_T)$-average $\ln(\mathrm{Cmax}_T)$) by subject in a bar plot with one bar representing one subject.
2 A plot of ($\ln(\mathrm{Cmax}_R)$-average $\ln(\mathrm{Cmax}_R)$) by subject in a bar plot with one bar representing one subject.
3 A plot of confidence interval as function of the number of analysed subjects in the statistical analysis. I.e. the confidence interval at $N = 10$ contains the PK-outcome for a pool of data corresponding to the first 10 subjects only. And so forth. Note that chronology of bioanalytical work is of utmost importance in practice, see discussion. The model used for the fit is the standard normal linear model on logarithmised Cmax with factors being subject (nested in sequence), treatment, period, and sequence. Where applicable, one could specify any other factor deemed relevant (like a group factor if subjects are dosed in groups, or covariates like body weight which are occasionally used for e.g. biosimilar equivalence trials).
4 A plot of mean squared error (mean squared residual, error

variance) from the normal linear model as function of the number of analysed subjects in the statistical analysis, i.e. this plot is closely related to the plot described under point 3, and can be derived from, the analysis mentioned in the previous point. The mean squared error is the overall sums of squares of the normal linear model divided by the residual degrees of freedom. When N subjects are analysed the degrees of freedom is N-2 for a standard bioequivalence evaluation with factors subject (nested in sequence), treatment, period, and sequence.

5 A barplot of residuals for the Test treatment from the overall Cmax analysis (the analysis with all $N = 36$ subjects). The residuals are extracted directly from the model mentioned under point 3. One bar corresponds to one subject.

Note that by nature of the 222BE design, residuals for the Reference treatment will have the same values but with opposite sign, so there is no additional information associated with presenting a plot of those residuals.

Similar Buster plots can be made for any other relevant metric, like AUCt, AUC extrapolated to infinity etc.

2.3. SaToWIB routines

These routines make an all-against-all comparison of pharmacokinetic profiles, for the purpose of identifying the profile pairs that have closest resemblance as defined by an objective function. This is done by calculating a “similarity score” for every pair of PK-profiles. In a dataset with N PK-profiles there will be a maximum of N(2N-1) profiles comparisons since profiles are only compared to each other once and never compared to self. The actual number will, for realistic datasets, occasionally be slightly lower since BLQ's or missing values may make specific comparisons impossible. BLQ's and missing values are discussed later.

When the SaToWIB routines were written, some concepts from previous studies on codon usage in bacteria, which relied on similarity between vectors (Fuglsang 2006) . For SaToWIB, I defined similarity score rules as follows:

1 A score of zero is a perfect match (=two PK profiles are identical)
2 Scores can not be negative.
3 A score above 0 means the two profiles are not identical and there is no requirement for a defined upper boundary for the range of scores.
4 A score computed for profile A versus profile B must be identical to the score calculated from comparing profile B to profile A, see discussion.
5 If a profile A has a score versus profile B which is S_{AB}, and the score of profile A versus profile C is S_{AC}, then A is considered more similar to B if $S_{AB} < S_{AC}$ and vice versa.

One could for example do ordinary linear regression on the profiles (with or without weights, with or without a forced intercept through (0,0)). In that case 1-r^2 (where r is the correlation coefficient) would meet the rules for being a valid score as defined above and thus a valid indicator of profile similarity. SaToWIB is implemented with scores based on this principle but I have empirically noticed that other methods may have better performance. One is presented in the following.

Olivier Le Blaye of the French Medicines Agency (ANSM) suggested me some years back to look at the opportunity for using a score based on the way incurred sample reproducibility is quantified, see for example section 6 of the current EU guideline on bioanalysis (European Medicines Agency 2009).

The implementation was initially as follows: When two profiles have N time points in common for which a concentration is quantifiable (not BLQ or missing), and when there are N data points for which the sum of the concentration pair is not zero, we can derive:

$$Score = \left(\frac{1}{N}\right) \sum_{i=1}^{N} \frac{|(v_{1i} - v_{2i})|}{\frac{1}{2}(v_{1i} + v_{2i})} \qquad \text{(Eq. I)}$$

where v_{1i} is the i'th concentration from the first profile, and v_{2i} is the i'th concentration in the other profile.

This score obeys all rules as defined above. The score can be seen as the ‘average absolute relative difference’. Empirically, it performs well for undiluted profiles, but obviously perform less well when some profiles are diluted in the sense that scores will be high. To mitigate it, a version (“method 32″) of the score was implemented where all profiles are normalised by dividing all concentrations in any given profile by the sum of the three largest concentrations in the same profiles. The ratio of those normalising sums is then an indicator of the dilution, *if one profile is a dilution of the other,* and the concentrations from the resulting normalised profiles can be plugged into equation I.

Example: We wish to compare the two profiles S10P1 and S35P1 which are in the dataset SWa627311.csv. The three highest values in S10P1 sum to 1395.34 and the three highest values in S35P1 sum to 694.65. We divide all concentration in profile S10P1 by 1395.34 and we divide all concentrations in profile S35P1 by 694.65. From the resulting normalisation, and plugging the normalised profiles into equation I, we get a similarity score of 0.0467 and the dilution estimate is 1395.34/694.65 = 2.0.

It is very important at this point to emphasize that a similarity score of 0.0467 is in itself not informative in the sense that the magnitude per se does not translate into a proof of falsification. For a discussion of the opportunity and necessity to establish formal testing or formal thresholds below which re-injection is believed to have taken place, see later.

The SaToWIB will compare all profiles against each other, calculate the score and the putative dilutions. It will then rank all these comparisons on basis of the scores. The profile pairs displaying lowest scores are ideally the candidates of re-injection. This also implies that SaToWIB does not per se detect those samples subjected to fraud, rather it tries to put on top of the ranking the most probable profile pairs.

2.4. Calculations

All calculations for this manuscript were done in R 3.4.0 running under Windows 10.

3. Results

Fig. 1 is a bar plot of Bar plot of ln(Cmax)-avg.ln(Cmax) by subject

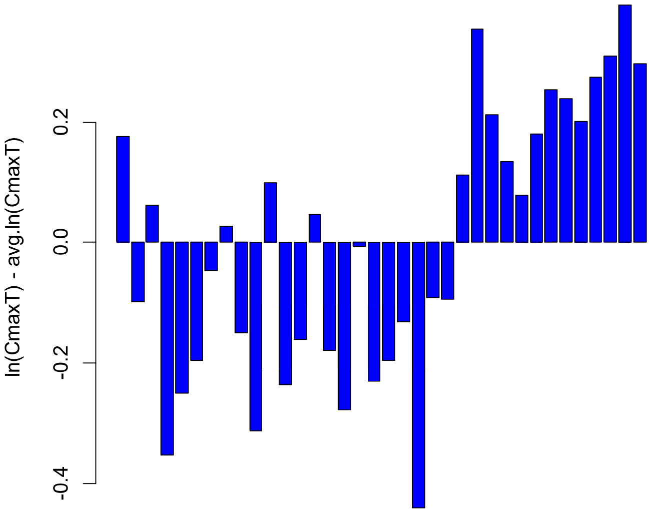

Fig. 1. Bar plot of ln(Cmax)-avg.ln(Cmax) for the test formulation. One bar represents on subject. A trend is clearly visible, as the magnitude switches from mostly negative to most positive in the last 12 subjects.

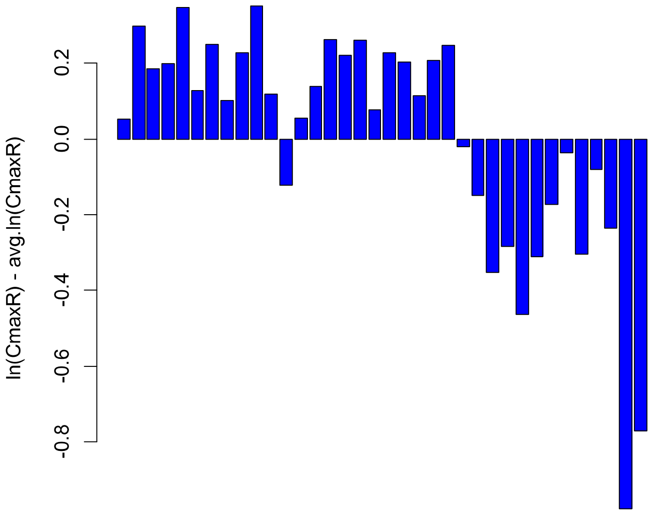

Fig. 2. Bar plot of ln(Cmax)-avg.ln(Cmax) for the test formulation. One bar represents on subject. A trend is clearly visible, as the magnitude switches from mostly positive to mostly negative in the last 12 subjects.

for the test formulation. The manipulation reveals as a trend visible after the first 24 subjects when the bars change direction (sign). If we just plot ln(Cmax) by subject without subtracting the overall mean, then it is more visually difficult to see this trend (data not shown). Fig. 2 is similar to figure but for the Reference and thus bars that are positive for a subject in Fig. 1 are often negative in Fig. 2 and vice versa.

Fig. 3 displays the confidence interval and point estimates as function of the number of subjects included in the analysis. The point estimate stabilises close to 0.8 but is brought upwards after the first 24 subjects by the manipulation and the final confidence interval is comfortably within the acceptance range as marked by the horizontal lines. Note also that the confidence intervals get a bit wider after the manipulation is done at subject 25 and onwards. The width of confidence intervals is determined by sample sizes and by the mean squared error (or mean squared residual / error variance). Fig. 4 is a plot of the mean squared error extracted from the fits done to produce Fig. 3. The trend after subject 24 is striking in that the MSE jumps radically upwards.

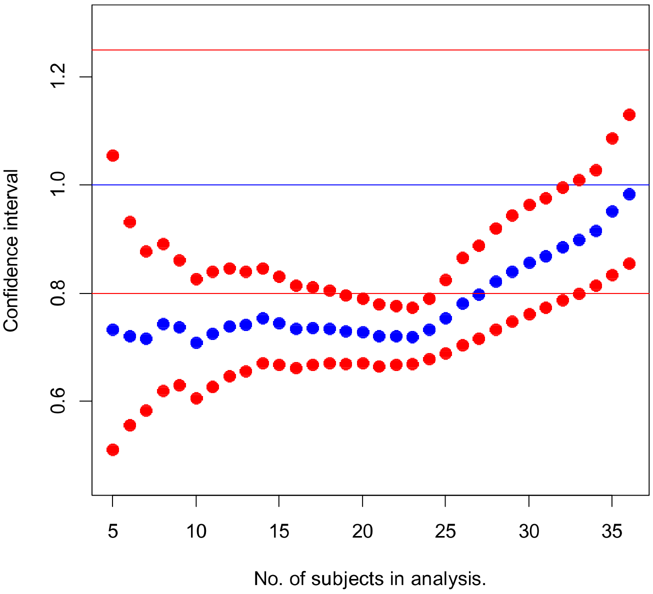

Fig. 3. Confidence interval as function of number of subjects in the analysis. The red dots indicate the confidence limits, the blue dots are point estimates. The vertial lines in red indicate the classical acceptance limits 0.8000 and 1.2500 (80.00%-125.00%). An increasing trend is seen in the last third of the plot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

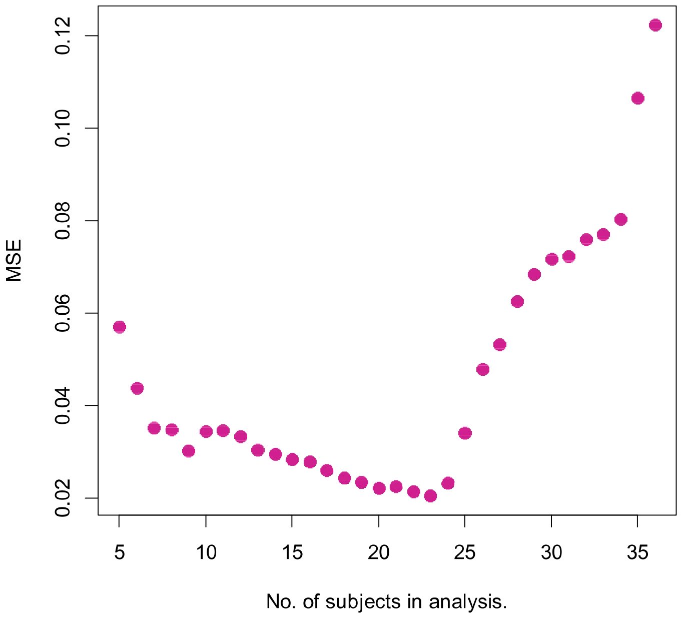

Fig. 4. A plot of the mean squared residual versus the number of subjects analysed. Data from the normal linear model used to fit the confidence intervals presented in Fig. 3. The mean squared residual shows a radical increase in the last third of the plot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

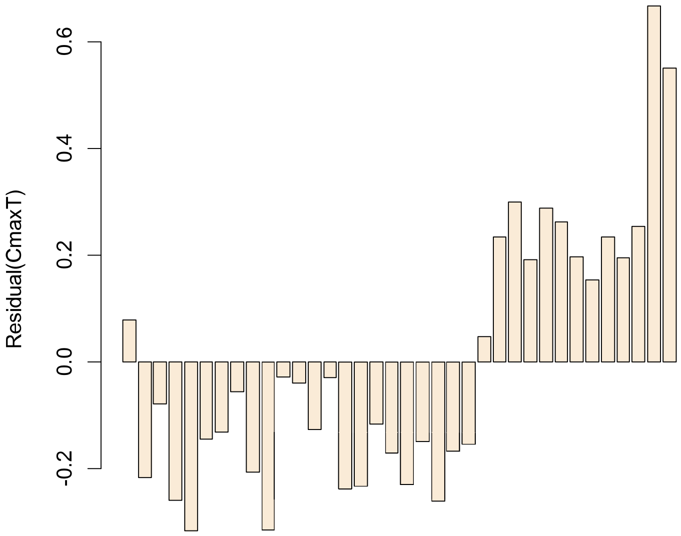

Fig. 5. Plot of residuals associated with the Test treatment from overall analysis. One bar represents on subject. A runs test onm the sign on the bar magnitude (positive or negative) is highly significant ($p < 0.0001$) as there are 36 bars and just three runs or (two changes of sign). The sign of the bar change exactly where the manipulation begins. There is similarity between this plot and Fig. 1.

Finally, Fig. 5 shows the residuals associated with the Test treatment from the model used to fit the final dataset. The plot has natural resemblance to Fig. 1, and displays the same trend. If we do a runs test on the sign on the residuals, it is highly significant ($\boldsymbol{p} < \mathbf{0.0001}$) as there are 36 bars and only two changes in the sign of the magnitude of the residuals.

Figs. 1-5 are all examples of the way trends arising out of sample manipulation can be visualized.

Table 2 shows the 30 first profile pairs identified by SatoWIB

Table 2
The 30 best matches from SaToWIB method 32. The first 17 on the list are pairs with a simulated manipulation cf table 1 and the randomisation in the PK data file. The first non-match is at rank 18.

Profile1	Profile2	Score	Rank	Ratio
S18P1	S27P2	0.03891	1	0.9561
S21P1	S28P2	0.03990	2	1.0051
S5P1	S25P2	0.04041	3	0.9901
S2P2	S33P2	0.04182	4	0.9639
S9P2	S30P1	0.04187	5	1.0264
S4P1	S26P2	0.04239	6	1.0189
S2P1	S33P1	0.04487	7	1.0273
S15P2	S34P1	0.04557	8	1.0024
S7P2	S31P1	0.04609	9	1.0282
S10P1	S35P1	0.04671	10	2.0087
S23P2	S36P1	0.04716	11	0.9593
S19P2	S29P2	0.04827	12	1.0492
S21P2	S28P1	0.04924	13	1.0083
S19P1	S29P1	0.04966	14	1.0107
S7P1	S31P2	0.05015	15	0.9962
S16P1	S32P1	0.05158	16	0.9945
S9P1	S30P2	0.05254	17	1.0306
S3P2	S31P2	0.05284	18	1.1041
S4P2	S26P1	0.05614	19	1.0307
S10P2	S35P2	0.05625	20	0.9636
S16P2	S32P2	0.05971	21	1.0389
S15P1	S34P2	0.05973	22	0.9991
S18P2	S27P1	0.06248	23	1.0299
S1P1	S11P2	0.06388	24	1.1132
S5P1	S10P2	0.06389	25	0.9801
S3P2	S7P1	0.06556	26	1.1083
S20P1	S34P1	0.06759	27	1.132
S5P2	S25P1	0.07000	28	0.919
S6P2	S8P1	0.07123	29	0.9892
S8P2	S25P1	0.07207	30	1.1541

"method 32″ and ranked according to their degree of match (best scores on top). As can be seen from this table the first 17 entries in the table are actual cases of manipulation (see table 1 for subjects affected by the switch operation, work out the periods from the file BUa627311.csv (supplementary)). The ratio column is the putative dilution if two profiles were to be assumed identical within the meaning of analytical variation; the ratio tells with reasonable accuracy the true dilutions cf. table 1. The table has 2556 rows arising from 72 profiles all being compared to each other once and not to their identity. The first non-match has rank 18, and the last true match is at rank 38 (not shown in table 2). If SaToWIB does not in any way rank profile pairs relevantly then the true matches would be randomly scattered across ranks 1–2556. With "method 32″ the true matches all appeared with the first 38 ranks. A rank sum test on the scores of manipulated pairs versus non-manipulated is associated with $\boldsymbol{p} < \mathbf{0.0001}$. Thus SaToWIB in this case has done a proper job identifying candidates for the manipulation. I'd like to stress that I am not referring to rank 18 and so forth as "false positives" - ranks truly aren't neither positives nor negatives. SaToWIB at this point only ranks profile pairs according to a numerically derived match score. The p-value indicates very strongly that this ranking is associated with the data manipulation.

Outcomes of SaToWIB can be easily graphed, see e.g. Fig. 6 where two profiles ranking high on SaToWIB's list (S18P1 and S27P2 are the pair at rank 1 in table 2) are shown along with a profile for which there is no particularly good match, S21P1, which in turn displays numerical similarity to another profile. The profiles graphed in two ways with concentration versus time (left) or concentration versus sample number (numbered 1–19, right). It is easily recognisable for the human eye that S18P1 and S27P2 are much more similar to each other than S21P1 is to either of them.

Table 3 presents an analysis of variance done on $\ln(\mathrm{Cmax}_T/\mathrm{Cmax}_R) = \ln(\mathrm{Cmax}_T)-\ln(\mathrm{Cmax}_R)$ in three groups corresponding to subjects 1–12, subjects 13–24, and subjects 25–36, and with fixed factors being group and sequence. Group comes out highly significant. Along such lines, in a letter to a CRO FDA presented confidence intervals for the data in the first third, middle third and last third of. Similar results are presented in table 4 for the groups defined above. There is an evidently distinct difference in confidence intervals in the thee groups. Tables 3 + 4 thus provide a tabulation of the way the trends in Figs. 1-5 manifest themselves in the comparative statistics. If we split the subjects into two halves rather than three thirds then the group factor is still significant (data not shown).

4. Discussion and conclusion

This manuscript illustrates how the manipulation reveals itself as trends (detectable via the Buster routines) or as profile similarities (the SaToWIB routines), and provides some ways to tabulate the way such trends show themselves on outcome statistics. The basis here is a dataset corresponding to a 2-treatment, 2-period, 2-sequence bioequivalence trial, but the concepts are easily applicable to other common BE designs, like replicated or semi-replicacted crossover trials, parallel trials, possibly even to other types of time series data. It is emphasized that the approach here is not a general method to detect fraud, but only a method to detect fraud arising out of the particular manipulation scheme that arises out of interim analysis and subsequent manipulation of samples if/when the point estimate after interim analysis appears to be too different from 1.00.

Interim analysis, done on the first M out of N subjects ($N > M$), may necessitate re-numbering of subjects according to the time of analysis, in case the M subjects were not analysed in a chronology corresponding to their study IDs. In this case the time of analysis determines the time by which data became available for interim analysis, and may thus determine the provisional ordering (renumbering) of subjects.

The scores presented here are for SaToWIB "method 32″ but other types of similarity can certainly be used. Linear regression where the score is 1-r^2 is handy and also easy to implement (note that this caps the maximum scores at 1, but this is not a violation of rule #3); it performs less well than "method 32″, empirically. For example an objective rule is to check where the first true non-match occurs in the ranking, or where the last true manipulated pair occurs. Handily, the slope of the linear regression will indicate the putative dilution. Using Theil-Sen regression (Sen, 1968) may often offer improved detection as compared to ordinary linear regression. The Theil-Sen regression approach is considered somewhat resistant to outliers, i.e. residuals which in magnitude may be suggestive of departure from the underlying assumption. The better performance of Theil-Sen regression over ordinary regression is again an empirical observation; it is nearly impossible to study residual distributions since PK datasets in bioequivalence generally do not contain more than roughly 15–25 samples per period, and because the number of datasets with known mode of fraud that I have access to is around 20.

PK-profile similarity could also be calculated on basis of ideas borrowed from dissolution. The classical measure for that is f2, but it will in its native form violate SaToWIB rule #4, but it otherwise performs quite well. Regarding f2 which traditionally is a quantification of dissolution similarity of Test against reference, one should note that when the fraud involves dilution without swapping of Test and Reference, then f2 will be comparing Reference-to-Reference or Test-to-Test for such profile pairs; therefore identifying matching profile pairs is not per se only about comparing profiles associated with Test against profiles associated with Reference and that is why rule #4 to has been of some importance.

Samples that are below the limit of quantification ("BLQ") or missing are common, and they need to be dealt with for the purposes of the implementation of such matching routines. For example, if a time point for one subject's profile has a quantifiable concentration and the other subject's concentration is missing then the time point in question may not contribute to our understanding or calculation of similarity.

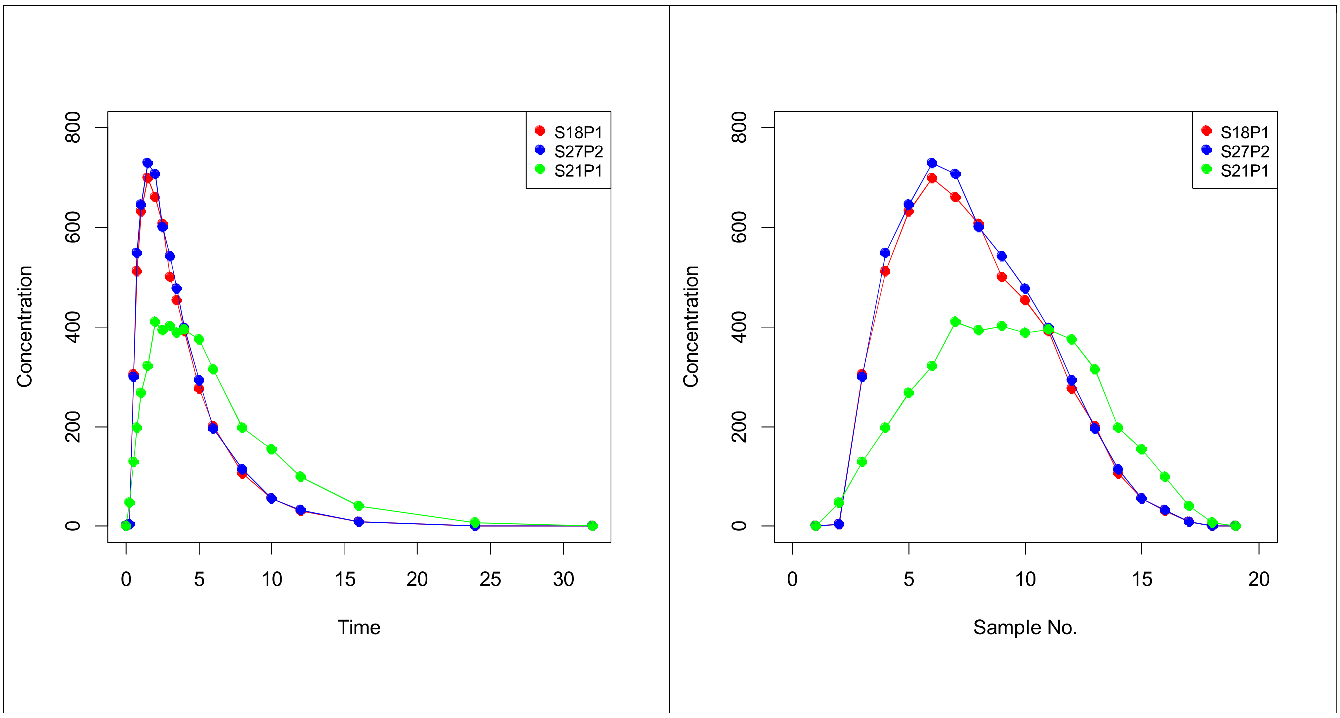

Fig. 6. Three PK profiles from the dataset. S18P1 and S27P2 was identified by SatoWIIB method 32 as the closest match (see table 2). Also shown for comparison is the profile S21P1 to which neither S18P1 nor S27P2 has much resemblance. To the left is concentration versus time and to the right is concentration versus sample number.

Table 3
An analysis of variance done on ln($Cmax_T/Cmax_R$) = ln($Cmax_T$)-ln($Cmax_R$) in three groups corresponding to subjects 1–12, subjects 13–24, and subjects 25–36, and with fixed factors being group and sequence. Group comes out very significant.

Source	Df	Sum Sq	Mean Sq	F value	Pr(>*F*)
Group	2	6.2380	3.11901	47.8490	2.418e-10
Seq	1	0.0433	0.04330	0.6642	4.211e-01
Residuals	32	2.0859	0.06518		

Table 4
Confidence limits for the geometric mean ratio of $Cmax_T$ to $Cmax_R$ in three groups.

Legend: CI.Lo = 90% confidence limit; CI.Hi = upper 90% confidence limit; GMR = geometric mean ratio; MSE = mean squared error; CV = coefficient of variation.

Dataset	CI.Lo	CI.Hi	GMR	MSE	CV
Subject 1–12	0.6461	0.8468	0.7397	0.0334	0.1843
Subject 13–24	0.6575	0.8008	0.7256	0.0177	0.1338
Subject 25–36	1.4940	2.1002	1.7713	0.0530	0.2332
All	0.8553	1.1304	0.9833	0.1224	0.3609

BLQs and missing values may in very extreme cases render two profile uncomparable in terms of the score. Method 32 is forgiving in this regard, as it only requires that three time points exist in the two profiles under comparison for which the concentrations are quantified and add to a value larger than zero.

Time point deviations, just like BLQ's and missing values, occur in almost all trials, and they can affect the calculation of AUC levels, but I do not see an obvious way or a need to take the info for such deviations into relevant consideration, though I am eager to hear opinions to the contrary. During the development of SaToWIB methods I have so far reviewed about 75 approaches towards quantification of similarity, some of which having similarity scores based on comparison of AUCs. A reported time point deviation would affect the magnitude of AUC, but if it falsified anyway then the time point deviation is not applicable and the methods I reviewed based on AUC ratios as basis for similarity scores were never performing anywhere near as well as method 32 for SatoWIB which is what this paper presents. That said, method 32 as presented here is one of the better performers, but it is not consistently the best amongst the actual cases of fraud that I have been involved in.

One ultimate goal, and one for which I am asking the scientific community to contribute, is the development of a formal test for irregular trends or fraud. At the moment I believe a runs tests on the sign of the residuals for either Test or Reference, as presented in Fig. 5, is a good opportunity and it is based on the same normal linear model that is used to establish BE as per guidelines. Therefore, it is not a test that introduces assumptions violating common evaluation and assessment principles. However, I have seen cases where the number of re-injected profiles was less than 8 (i.e. less than 4 subjects in a 222BE trial) and where the runs tests was not significant and where I could not see any trend in the plots produced by Buster routines, and therefore in as much as the runs tests looks very promising when the p-level is low, I think it will be associated with false negatives if employed on data sets with such low re-injection figures.

What also needs to be borne in mind is that the scatter or variation introduced by the analytical method and while there are formal requirements for such variation, all methods are inherently associated with a unique analytical variation. For that reason a score of e.g. 0.04 with “method 32″ (or any other variant) in one study may not be comparable to a score of 0.04 in another dataset. I believe at some point the analytical precision and accuracy as measured during the bioanalytical method validation (which accompanies every submission to regulatory authorities) could be taken into consideration if or when a formal test for profile similarity is being invented.

Occasionally, I have discussed with regulators the opportunity to use e.g. a polymerase chain reaction (PCR)-based method to create relative or absolute identity of plasma samples. While red blood cells are not nucleated (void of genomic DNA), plasma generally contains traces of nucleic acids (either extracellularly or e.g. in the minute

Table 5
Result of kmeans clustering for 2 arbitrary clusters: Numbers assigned to clusters and whether they are manipulated or not. amongst the 2556 profile pairs the kmeans algorithm assigned all 24 manipulated pairs to one cluster. Yet, that cluster also includes most of the unmanipulated pairs, so in its present implementation is isn't clear how the clustering is useful.

	Not Manipulated	Manipulated
Cluster 1	1319	0
Cluster 2	1213	24

volume of white blood cells that still exists after siphooning the red blood cells off of a centrifuged blood sample) that can be used for DNA amplification. PCR methods are widely accepted in forensic science (see a general review by Dumache et al., 2016, see O'Driscoll 2007 and references therein for info on methodologies to detect nucleic acids in fluids void of nuclei) and there is little doubt that something along such lines could be done in practice but it is not at all a straightforward proposition. It is not clear who should bear the cost of such analysis, and it does create potential ethical/GCP-related issues if one uses PCR to establish the absolute or relative identity of the subjects that gave specific samples. Some countries do not readily allow export of biological fluids, so plasma samples can not be collected like retention samples of investigational products. Plasma separation is generally not done under sterile conditions, so whatever RNA/DNA is subjected to amplification may be contaminated. It may require a validation trial on its own to establish that the approach works. Thus, the opportunity to investigate this type of fraud by means of molecular identification technqiues may not be ready for prime time despite its othjerwise well-established applications in other fields of science.

Further, if a test in the future becomes established and agreed upon amongst regulators for rejection of data, it will be necessary for sponsors to factor this testing into the power and sample size calculation and this could somewhat increase the number of subjects necessary to achieve the desired level of power, where power is the chance of showing BE at the given sample size while at the same time not having a significant flag for falsification. Multiplicity would become a serious issue if a battery of formal tests would be applied at e.g. the 5% and if each of them were potential grounds for refusal.

SaToWIB provides an attempt at ranking profiles on basis of a univariate (unidimensional) score. If we think of the difference between any two profiles (one vector subtracted from the other) as a multidimensional indicator of similarity then there are opportunities for exploring the potential of clustering algorithms or principal component analysis to separate the manipulated pairs from the nonmanipulated pairs. Table 5 illustrates how the kmeans algorithm (McQueen 1967) with two arbitrary target clusters provides separation of the truly manipulated pairs into one of them, which looks promising. Yet, many of the unmanipulated pairs end up in the same cluster. A similar effects is seen if applying e.g. 10 clusters in stead of 2. So, although the kmeans approach is very effective in clustering the manipulated cases, a lot of unmanipulated pairs end up in the same cluster as the manipulated pairs. I have tried other transformations of the vector difference (e.g. absolute, squared) but the outcome is much the same (not shown). More work is needed in order to make multivariate statistics useful for purposes of identifying potentially manipulated pairs; it is an area of future work.

The main conclusions are:

1. Buster routines may reveal trends in data which signify re-injection (re-analysis) of subject data.
2. Profile similarity in itself is ground enough for questioning a study's validity from regulator's point of view (8).
3. SaToWIB evaluates profile similarity and ranks PK-profile pairs according to an objective function. If manipulation has taken place then the profiles pairs that constitute the manipulation are likely to be near the top of the ranking.
4. Manipulation is detectable as evident from pt. 1 + 3 but in their presents forms neither Buster nor SaToWIB offers an objective test for the presence of manipulation; it would be highly desirable to expand on the methodology along these lines.
5. The work on Buster and SaToWIB routines is by no means a final or finished work, and there is plenty of room for improvement.

Along these lines, comments on this paper are invited from sponsors, agencies, and CROs, in order to explore better and alternative ways to detect similarities or trends arising out of fraud. At the end of the day it is in everyone's interest that medicines are safe and efficaceous. Since I have no reason to believe fraud in BE is generally on the decline, it is my opinion that it is of quite some importance that formal testing for fraud as part of the evaluation be done by the parties involved, so the need for an objective testing method is present.

ICH E6 in its most recent iteration introduced risk-based oversight of trials. In the context of BE, and given the recent experience with CROs subjected to regulatory scrutiny, interim analysis and the signs of sample manipulation may be highly relevant aspects associated with risks which sponsors would take into consideration.

Credit author statement

Anders Fuglsang:

Single author, has done all underlying planning, execution, analysis and manuscript writing.

Declaration of Competing Interest

The authors a consultant. Present and former clients or requestors of service include agencies, regulatory bodies, pharmacopoeias and private companies. This paper does not express any opinion about any particular company, any particular product or any specific agency/body, and I am not declaring the existence of any conflict of interest.

Acknowledgments

I would like to thank everyone participating at the regulatory inspector's meeting held in October 2016 at the European Medicines Agency in London, UK, for the fruitful discussion and for the encouraging comments and questions attendees expressed during and after the meeting, in relation to my development of software for fraud detection. In particular I would like to thank Mr. Olivier Le Blaye, Agence Nationale de Sécurité du Médicament et des Produits de Santé (ANSM, France), for his supporting comments and suggestions and not least for the idea to look towards incurred sample re-analysis when developing similarity scores for SaToWIB. This research has been undertaken independently by the author only, and the development of Buster/SaToWIB, did not receive any funding.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejps.2020.105595.

References

Dumache, R., Ciocan, V., Muresan, C., Enache, A., 2016. Molecular DNA analysis in forensic identification. Clin. Lab. 62, 245–248.

European Medicines Agency, Referral Notification, February 19, 2020. https://www.ema.europa.eu/en/documents/referral/panexcell-article-31-referral-notification_en.pdf, Accessed June 20, 2020.

European Medicines Agency, Committee for human medicinal products. guideline on bionanalytical method validation. EMEA/CHMP/EWP/192217/2009 Rev.1. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-bioanalytical-method-validation_en.pdf, Accessed June 20, 2020.

European Medicines Agency, Committee for Human Medicinal Products. Investigation of bioequivalence. CHMP CPMP/EWP/QWP/1401/98 Rev. 1. 2010. http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500070039. Accessed June 20, 2020.

Fuglsang, A., 2006. Estimating the “effective number of codons”: the Wright way of determining codon homozygosity leads to superior estimates. Genetics 172 (2), 1301–1307.

MacQueen, J.B., 1967. Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.

O'Driscoll, L., 2007. Extracellular nucleic acids and their potential as diagnostic, prognostic and predictive biomarkers. Anticancer Res. 27, 1257–1265.

Sen, P.K., 1968. "Estimates of the regression coefficient based on Kendall’s tau. J. Am. Stat. Assoc. 63 (1968), 1379–1389.

United States Food and Drug Administration, 2013. Center for Drug Evaluation and Research. Bioequivalence. Studies With Pharmacokinetic Endpoints for Drugs Submitted Under an Abbreviated New Drug Application. . https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioequivalence-studies-pharmacokinetic-endpoints-drugs-submitted-under-abbreviated-new-drug Accessed June 20, 2020.

United States Food and Drug Administration, 2001. Center for Drug Evaluation and Research. Statistical Approaches to Establishing Bioequivalence. Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. . https://www.fda.gov/media/70958/download Accessed June 20, 2020.

United States Food and Drug Administration. Untitled letter issued on April 19, 2016, to a company in Bangalore, India. https://www.fda.gov/media/97413/download. Accessed June 20, 2020.

United States Food and Drug Administration. Form 483 issued on October 9, 2015, to a company in Bangalore, India. https://www.fda.gov/media/97413/download. Accessed June 20, 2020.

World Health Organization. Annex 7, Multisource (generic) pharmaceutical products: guidelines on registration requirements to establish interchangeability. WHO Technical Report Series No. 992, 2015. https://www.who.int/medicines/areas/quality_safety/quality_assurance/Annex7-TRS992.pdf. Accessed June 20, 2020.

World Health Organization. Notice of Concern, issued on February 12, 2016, to a company in Bangalore, India.The Notice of Concern is in the public domain but no copy could be found on WHO's website as of the date of this submission. A copy is available from the author upon request.