Professor Endrenyi's Legacy: An Evaluation of the Regulatory Requirement "Fixed Effects, Rather Than Random Effects, Should Be Used for All Terms"

Anders Fuglsang

Fuglsang Pharma, Vejle Ø, Denmark

Corresponding author: Anders Fuglsang, Fuglsang Pharma, Beckersvej 13, 7120 Vejle Ø, Denmark; TEL: (+45) 42257533; email: a.fuglsang@ymail.com

Received, March 7, 2021; Revised, June 22, 2021; Accepted, July 27, 2021; Published, August 2, 2021

ABSTRACT -- **Purpose:** In the latest revision of the guideline for evaluation of bioequivalence (BE), European regulators introduced the requirement for using subjects as fixed factors in the underlying statistical models, even in replicate and semi-replicate studies. The implication was that estimates of within-subject variability were derived with a linear model rather than with a mixed model based on restricted maximum likelihood (REML). While REML-based methods are generally thought to give rise to less biased estimates of variance components, there have been no studies that compared the quality of REML-based estimates and estimates derived via linear models. **Methods:** A publication by Endrenyi and Tothfalusi from 1999 described simulations in a fashion that is useful for testing the European Medicines Agency's (EMA) requirement. This study defines 7 scenarios within which 10,000 individual 2-sequence, 2-treatment, 4-period trials are simulated and makes a comparison of the quality of estimates. **Results:** It is concluded that estimates based on REML are closer to the true values than estimates based on linear models, but significant differences are only shown in two of the seven scenarios tested. REML-based estimators have less variability. Both types of estimates appear negatively biased and will therefore decrease the width of the acceptance range.

INTRODUCTION

With the regulatory requirement "Fixed effects, rather than random effects, should be used for all terms" European regulators in 2010 effectively phased out the use of mixed models for evaluation of bioequivalence data (1). The topic was subject to some debate as it had been common to use mixed models with subject specified as random effects for replicate and partially replicate trials. The practice of specifying subjects as a random factor is still a prevailing principle in other areas of science, as agreed even by the regulators (2,3).

In practice it means that BE data must fit within a normal linear model using treatment, subject, period, and sequence as fixed factors (1,3); this gives effect estimates for treatment, which in turn can be used to determine the test: reference ratio (point estimate). The model can then be represented as follows (4): $Y_{ijk} = \mu + S_{ik} + P_j + F_{jk} + Q_k + e_{ijk}$ (Equation I) where, Y_{ijk} is the log-transformed metric of interest (Cmax or area under the concentration-time curve) for the i'th subject, measured in the j'th period, and who was randomized into the k'th treatment sequence, where μ is the intercept, S_{ik} is the fixed subject effect for the i't subject (who is assigned to the k'th sequence), P_j is the fixed period effect for the j'th period, F_{jk} is the fixed formulation effect for period j of the k'th sequence, and Q_k is the fixed sequence effect of the for the k'th sequence.

For the common fully-replicate trial designs where subjects are randomized into two treatment sequences ("TRTR" and "RTRT" which define the order of administration of Test (T) and Reference (R) treatments) to derive an estimate for the intra-subject variability for the reference treatment (σ^2_{WR}), data for the test treatment is removed, and a normal linear model is used on the remaining data with subject, period and sequence as fixed factors; the residual variability is taken as an estimate of σ^2_{WR}. From the effect estimates and estimate of σ^2_{WR} the confidence interval for the test:reference ratio can be derived. The acceptance range conditionally scales with the magnitude of the estimate of σ^2_{WR}. Throughout this work, sigma subscripts "w" denotes "within", subscript "b" denotes "between", "T" refers to the test treatment, "R" refers to the reference treatment. σ^2_{WR} is thus the intra- ("within") subject variance associated with the reference treatment, and so forth. See Medicines Evaluation Board (5), for an example of how it may work in practice.

EU regulators did not present arguments for their proposal, and at the time of introduction they expressed that their approach is straightforward to calculate. Laird and Ware (6) mentioned that estimates of variance components are less biased

with REML than with ordinary maximum likelihood, much accounting for the popularity mixed models with REML have later enjoyed. Yet, nothing is really known about actual observational bias, if any, associated with the bioequivalence approaches for replicated studies. Interestingly, the late Prof. Endrenyi and his colleague Laszlo Tothfalusi published a paper in 1999 (7) where a simulation approach was used to define datasets and evaluate the quality of resulting estimates arising out of REML fits. Specifically, their paper looked into estimation of the subject-by-formulation interaction. This quantity is defined as: $\sigma_D^2 = \sigma_{BR}^2 + \sigma_{BT}^2 - 2cov_{TR}$ (Equation II). The methodology used in their publication seems to provide a convenient way to simulate studies for the purpose of evaluating EMA's requirement which has never been tested. Therefore, given EMA's preference for models with all factors fixed, and given the importance of estimates of $\sigma_{WR}^2$for the assessment of bioequivalence, and given the absence of papers dealing with the quality of estimates arising from REML versus the mandatory linear model in EU, this paper tries to apply Endrenyi and Tothfalusi's simulation approach to answer the following main question: Is there an observable/quantifiable bias on estimates on σ_{WR}^2 when a normal linear model is used, and if there is, is the bias more or less pronounced than when REML is used (in which case one can specify S_{ik} of Equation I as random rather than fixed)? In addition, the paper will present evidence regarding the bias on estimates of σ_D^2. Descriptive statistics and graphing will be used. Under the European guideline (1), the acceptance limits are defined as follows: a. The (lower, upper) limits are 80.00% - 125.00% when the estimate of σ_{WR}^2 corresponds to a CV no more than 30%, and where the relationship between $\hat{\sigma}_{WR}^2$ and CV is:

$CV = \sqrt{\left(exp\left(\hat{\sigma}^2\right) - 1\right)}$ (Equation III), b. The (lower, upper) limits are 69.84%-143.19% when the estimate of σ_{WR}^2 corresponds to a CV above 50%, and c. When the estimate of σ_{WR}^2 corresponds to a CV between 30% and 50% then the limits are: $limits = exp(\pm.760\hat{\sigma_{WR}})$ (Equation IV)

In order to quantify evaluate any practical implications of the difference in estimate bias, a quantity called P'_x is defined as follows: P'_x is the probability that REML provides a better estimate of the σ_{WR}^2 than the linear model ***and*** that the difference in the two estimates translate into a difference of more than x on the upper acceptance limit, as calculated through the application of Equations II and III. P'_x will be used to discuss the extent to which the choice of estimation method is of practical relevance using levels of x being 0.01, 0.02 and 0.05. P'_x is an attempt at quantifying the practical importance of the difference between using the estimate from the linear model rather than the estimate from REML. A P of 0.01 is the probability that REML provides a better estimate which makes the confidence interval one percentage unit narrower. As this paper is the first of its kind, no alternative or better way of quantifying it exists, to the best of the author's knowledge.

MATERIALS AND METHODS

Scenarios

7 scenarios were simulated as depicted in Table 1. These were all based on dataset I which was released as supplementary material to the guideline of 2010 (1) that introduced the requirement for all fixed effects. The dataset is, in my experience, quite representative of the real-life data as it displays somewhat different variance components for Test and Reference (within as well as between). The within-subject variance for the Reference is estimated to be about 0.2025, corresponding to an intra-CV of roughly 47%. The dataset has a point estimate of about 1.157. Scenario 1 is a balanced simulation of datasets with these variance components and the original point estimate of dataset I.

Scenario 2 is a simulation of the same data but where the covariance of T and R has been adjusted so that the subject-by-formulation variance is 0.3 (=twice the critical level debated by Endrenyi and Tothfalusi (7)). In Scenario 2 ideas similar to those used by Endrenyi and Tothfalusi (7) have been used to produce a simulation with σ^2_D =0.3 which is twice the level that has been proposed as a reasonable threshold for the presence of a subject by formulation interaction.

Scenario 3 and scenario 4 are identical to scenario 1 and scenario 2 except the point estimate is 1.432 which is the maximum upper scaled acceptance limit in EU.

Scenarios 5 and 6 are identical to scenario 1 and scenario 2, except the point estimate was simulated at a value of 1.

In the paper by Endrenyi and Tothfalusi (7) a radical idea was used: in each of a series of simulations they varied but with somewhat more -subjectively- relevant values for Test and Reference. Endrenyi and Tothfalusi used a random uniform between 0.02 and 0.5 to define $\sigma_{WR}^2 = \sigma_{BR}^2 = \sigma_{WT}^2 = \sigma_{BT}^2$ and simulated a point estimate

Table 1. Scenarios for simulations in this paper.

Scenario	σ^2_{WR}, σ^2_{BR}, σ^2_{WT}, σ^2_{BT}, cov_{TR}	GMR	Remark
1	0.2025, 0.7272, 0.1175, 0.6861, 0.7067	1.157	These are the original values estimated from EMA dataset I
2	0.2025, 0.7272, 0.1175, 0.6861, 0.5567	1.157	These are the variances from EMA dataset I but with the covariance of R and T adjusted so that the subject by formulation interaction is 0.3. This corresponds to a Pearson correlation of rho ~ 0.78
3	0.2025, 0.7272, 0.1175, 0.6861, 0.7067	1.432	These are the original variance and covariance estimates from EMA dataset I but the GMR is the upper scaled limit in EU.
4	0.2025, 0.7272, 0.1175, 0.6861, 0.5567	1.432	These are the original variance estimates from EMA dataset I, where the covariance is adjusted so that the subject by formulation interaction is 0.3 and the GMR is the upper scaled limit in EU.
5	0.2025, 0.7272, 0.1175, 0.6861, 0.7067	1.000	These are the original variance and covariance estimates from EMA dataset I and where the GMR corresponds to a perfect match.
6	0.2025, 0.7272, 0.1175, 0.6861, 0.5567	1.000	These are the original variance estimates from EMA dataset I, where the covariance is adjusted so that the subject by formulation interaction is 0.3 and where the GMR corresponds to a perfect match.
7	variable, variable, variable, variable, variable	variable	Subject-by-formulation interaction maintained at 0.3. See material and methods for details on the selection of variance components.

of 1.0 every time without a subject-by-formulation interaction. It is in my experience not at all common to see estimates of within-variances of a magnitude close to estimates of between-variances. Between-variances tend to be somewhat higher than within-variances. Trying to execute simulations based on their idea, an additional scenario was defined (scenario 7), but using putatively more realistic figures for the variance components, a naïve form of rejection sampling was employed where five individual and independent variance components were generated such that: a. σ^2_{BR} was a uniform random deviate between 0.08 and 1.60, corresponding to a between-CV of between 29% and 200% (29% to about 200% corresponds to levels of between-subject variances that I have seen), and b. σ^2_{BT} was a uniform random deviate between 0.08 and 1.60, corresponding to a between-CV of between 29% and 200%, and c. σ^2_{WR} was a uniform random deviate between 0.0036 (CV =6%, the lowest I have seen) and the minimum of σ^2_{BR} and σ^2_{BT}, and d. σ^2_{WT} was a uniform random deviate between 0.0036 and the minimum of σ^2_{BR} and σ^2_{BT}, and e. cov_{TR} was derived from Equation II so that the σ^2_D was 0.3, and e. The Pearson correlation coefficient ρ was kept between -1 and 1. The relation between the correlation and the covariance is: $cov_{TR} = \rho\sigma_{BR}\sigma_{BT}$ (Equation V) , where ρ (Pearson correlation) assumes values between -1 (perfect negative correlation of test and reference) and 1 (perfect positive correlation). It is emphasized that the primary purpose of scenario 7 (or this submission in general) is *not* to make inference per se about the quality of estimates of $\sigma^2_{D\ but}$ this data is nevertheless available as a result of the simulations and are discussed. Scenario 7 provides a way to assess quality of estimates of the within-subject variability under circumstances when neither the GMR (geometric mean ratio) nor the individual variance components are fixed, and this is the purpose of the scenario.

Simulations

For each scenario 10,000 studies of the 2-treatment, 2-sequence, 4-period design were simulated, each with a balanced sample size of N=24 (a fairly common sample size). The simulations were done in the statistical language R, version 3.4.1 running under Windows 10. The choice of 10000 trial simulations per scenario reflected what is computationally feasible to accomplish in a reasonable time-frame. The Multivariate Normal and t Distributions package (mvtnorm) was used to generate pseudo-random variates from multidimensional normal distributions with the

desired levels of intra- and between-subject variances and covariances. The log-likehood function was optimized using the profile method as described by Gurka (8). The profile method was preferred as it involves inversion of much smaller matrices which may have positive implications for numerical stability. The Nelder-Mead algorithm was selected as the optimizer of choice (9) after testing of the available options in the optim function and the library of optimizers in the nlminb package. From the tested optimizers only the Nelder-Mead algorithm was able to consistently converge, so this optimizer was used throughout. A relative tolerance of 10^{-9} was applied.

The covariance matrices were defined directly from the variances and co-variances, from the estimates of which ρ and the subject-by-formulation interaction can be directly derived (an alternative, giving the same result, but parameterized differently is to use $\rho\sigma_{BT}\sigma_{BR}$ rather than the covariance explicitly, and then the covariance would be inferred from the resulting estimate of ρ) via equation V.

From the REML fit, fixed effects and random effects were extracted. The same datasets were fitted with normal linear models *ad modum* EMA and the fixed and random effects were recorded. In the following, the term *bias* denotes the difference between observed levels and the simulated (true) level. The point estimate, or PE, is the estimate of GMR.

RESULTS

Table 2 shows the general performance of the mixed model and the linear model as descriptive statistics for estimates of σ^2_{WR}, σ^2_D, and point estimates; the table is based on the 10,000 simulations of scenario 1.

First of all, Table 2 shows a tendency of the mixed model to give better estimates of σ^2_{WR} than the linear model; this can be seen by comparison of figures for medians or means and this observation holds regardless of whether there is a subject-by-formulation interaction (scenario 2, 4, and 6) or none (scenario 1, 3, and 5), and it is dependent on the magnitude of the GMR.

For scenarios 4 and 6 there is statistically significant differences in variance estimates between the two method ($p<0.05$, F-test) while for the other scenarios $p>0.05$. Figure 1 shows a histogram of $\hat{\sigma}^2_{WR,REML} - \sigma^2_{WR}$ and Figure 2 shows a corresponding histogram associated with the linear model. Both histograms are visually asymmetric, so the fair comparison of estimated bias should be based on the medians. Apart from yielding slightly better median estimates, Table 2 also shows that the mixed model is associated with a somewhat less variable estimate, cf. e.g., the span of the 10th to 90ieth percentile, or the difference between maximum and minimum.

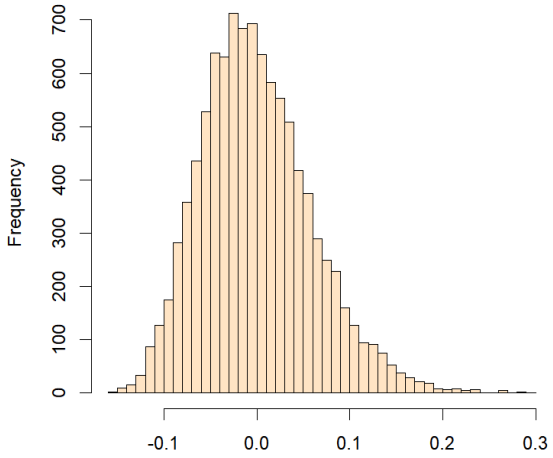

Figure 1. Histogram of $\hat{\sigma}^2_{WR,REML} - \sigma^2_{WR}$ on basis of data from scenario 1. Positive values are cases of over-estimation of the variance, negative values are cases of under-estimation. The median is 0.1965, and the mean is 0.2023 cf. Table 2.

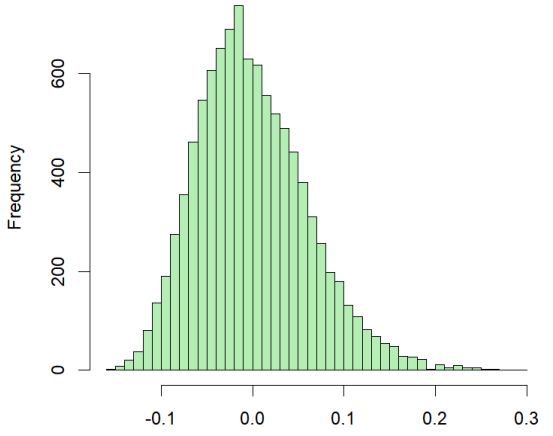

Figure 2. Histogram of $\hat{\sigma}^2_{WR,Lin.model} - \sigma^2_{WR}$ on basis of data from scenario 1. Positive values are cases of over estimation of the variance, negative values are cases of under-estimation. The median is 0.1958, and the mean is 0.2024 cf. Table 2.

Figure 3 shows a plot of $\hat{\sigma}^2_{WR,Lin.model}$ against $\hat{\sigma}^2_{WR,REML}$ (left), and the corresponding Bland-Altman plot (right). The solid black line is the line of identity; the plot on the left shows the degree of agreement between the two estimates. A linear regression gives r^2~0.96. It is noted that points that are farthest from the line of identity

Table 2. Descriptive statistics for $\hat{\sigma}^2_{\text{WR,REML}}$, $\hat{\sigma}^2_{\text{WR,Lin.model}}$, $\hat{\sigma}^2_D$ and point estimates obtained for scenarios 1-6

Metric	Scenario	Mean	Median	SD	Min	Max	Q10	Q90
$\hat{\sigma}^2_{\text{WR,REML}}$	1	0.2023	0.1965	0.0599	0.0460	0.4996	0.1308	0.2812
	2	0.2036	0.1982	0.0605	0.0471	0.5415	0.1306	0.2841
	3	0.2014	0.1949	0.0606	0.0525	0.5269	0.1286	0.2824
	4	0.2022	0.1968	0.0596	0.0490	0.4703	0.1297	0.2808
	5	0.2025	0.1961	0.0609	0.0481	0.4911	0.1297	0.2849
	6	0.2027	0.1968	0.0603	0.0468	0.4903	0.1295	0.2843
$\hat{\sigma}^2_{\text{WR,Lin.model}}$	1	0.2024	0.1958	0.0610	0.0471	0.4931	0.1295	0.2823
	2	0.2036	0.1975	0.0614	0.0476	0.5166	0.1295	0.2841
	3	0.2013	0.1946	0.0615	0.0513	0.5629	0.1272	0.2832
	4	0.2023	0.1965	0.0607	0.0485	0.4736	0.1286	0.2815
	5	0.2024	0.1959	0.0617	0.0486	0.4827	0.1292	0.2853
	6	0.2027	0.1967	0.0615	0.0481	0.5182	0.1287	0.2851
PE	1	1.1611	1.1575	0.0955	0.8195	1.5882	1.0414	1.2856
	2	1.1637	1.1556	0.1602	0.6555	1.8263	0.9656	1.3745
	3	1.4369	1.4316	0.1168	1.0482	1.9063	1.2896	1.5881
	4	1.4456	1.4319	0.2023	0.8651	2.3490	1.1980	1.7122
	5	1.0034	1.0009	0.0818	0.7215	1.4449	0.9016	1.1104
	6	1.0112	1.0018	0.1415	0.6346	1.7259	0.8351	1.1979
$\hat{\sigma}^2_D$	1	-0.0004	-0.0023	0.0592	-0.2218	0.2827	-0.0736	0.0752
	2	0.2995	0.2878	0.1420	-0.0800	0.9756	0.1248	0.4864
	3	0.0001	-0.0025	0.0594	-0.2115	0.2413	-0.0727	0.0772
	4	0.3009	0.2885	0.1434	-0.1150	1.0735	0.1283	0.4905
	5	0.0003	-0.0016	0.0597	-0.2237	0.2906	-0.0743	0.0772
	6	0.2986	0.2845	0.1435	-0.0761	0.9878	0.1258	0.4910

Point estimates are identical for REML and the linear model in all cases. Across all these scenarios the true value of σ^2_{WR} was 0.2025. See materials and methods for details on geometric mean ratios and σ^2_D. SD: Standard deviation (of a sample); Q10=10$^{\text{th}}$ percentile; Q90=90$^{\text{ieth}}$ percentile; PE=point estimate (=estimate of the test: reference ratio).

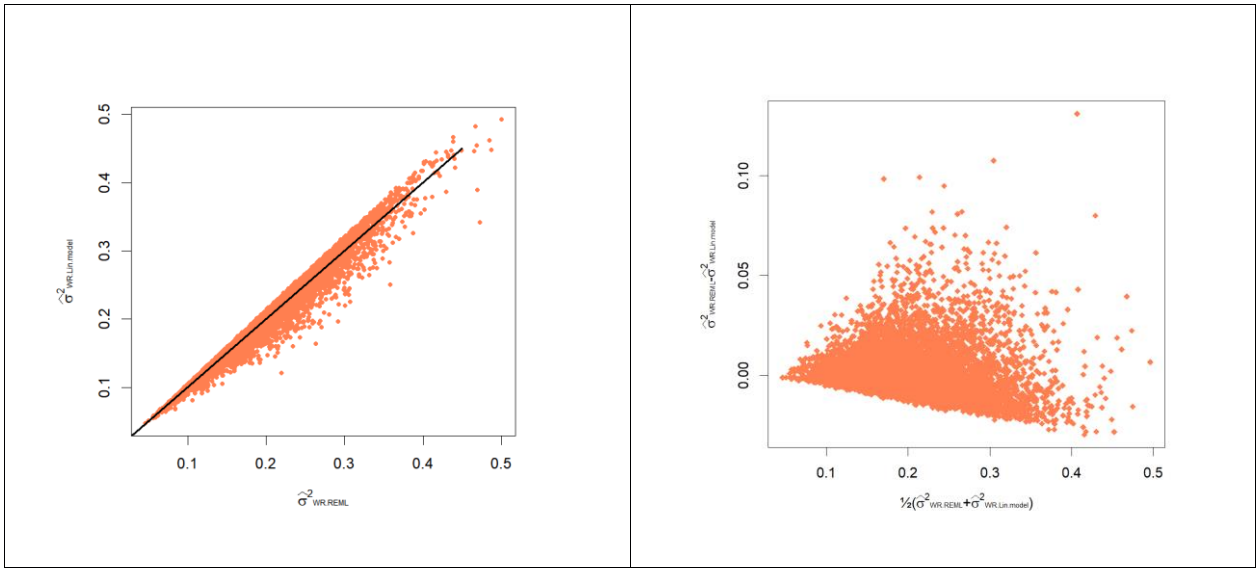

Figure 3. $\hat{\sigma}^2_{\text{WR,Lin.model}}$plotted against $\hat{\sigma}^2_{\text{WR,REML}}$ (left, line of identity shown in black) and the corresponding Bland-Altman plot of the two quantities (right). Data from scenario 1. There is a generally good correlation; differences tend to get larger as the estimates get bigger. All points stem from simulations with the true level being 0.2025, see Table 1.

fall on the lower side of it. The interpretation of the Bland-Altman plot is that differences are smaller when the estimates themselves are smaller.

Point estimates are identical for REML and the linear model in these simulations. This will be the case when data has no imbalance, but it will be a future objective to evaluate outcomes when data is imbalanced w.r.t. sequences or when data has missing periods. Figure 4 shows a histogram of PE-GMR based on scenario 1.

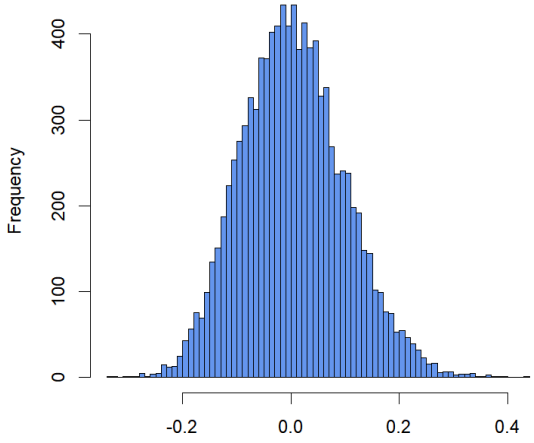

Figure 4. Histogram of PE-GMR (point estimate of GMR minus true GMR) on basis of data from scenario 1. Positive values are cases of over-estimation, negative values are cases of under-estimation. Note that since these simulations are balanced w.r.t. sequences, and have no missing values, the point estimate will be exactly the same for the linear model and for the mixed model (but the individual treatment effects will clearly not be the same).

Table 1 suggest the absence of a positive bias on the estimates of subject-by-formulation interactions derived via Equation I. Note that a few of the estimates are slightly negative; this anomaly is simply because of the unconstrained optimizer and a result of the iterative process continuing with adaptations in all five variance components individually until the convergence criterion is met. Figure 5 shows a histogram of $\hat{\sigma}_D^2 - \sigma_D^2$ from scenario 2, where σ^2_D is 0.3. It is remarkable to observe this as it is not quite in agreement with the observations of Endrenyi and Tothfalusi (7). The same is observed when scenario 7 is evaluated. Figure 6 illustrates $\hat{\sigma}_D^2 - \sigma_D^2$ from this scenario.

The median estimate of σ^2_D is 0.2858 and the mean is 0.2990. It is impossible for me, on the basis of observations made, to make a claim for σ^2_D being generally over-estimated. It may indeed be a little under-estimated, at least the results presented here suggest so.

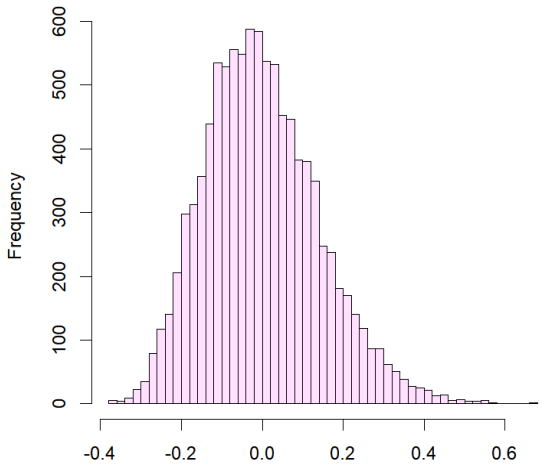

Figure 5. Histogram of $\hat{\sigma}_D^2 - \sigma_D^2$ on basis of data from scenario 2. Positive values are cases of over-estimation of the subject-by-formulation interaction, negative values are cases of under-estimation.

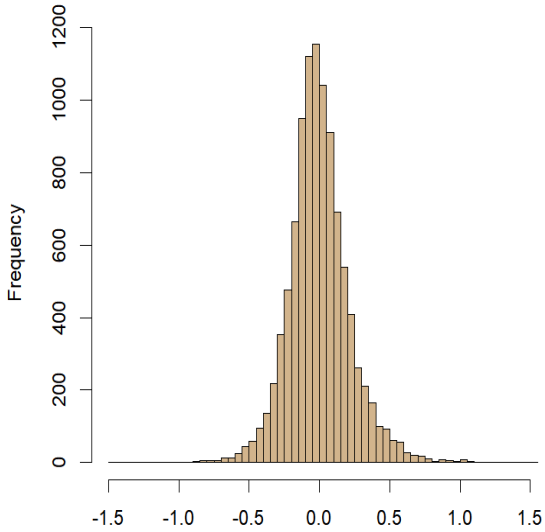

Figure 6. Histogram of $\hat{\sigma}_D^2 - \sigma_D^2$ on basis of data from scenario 7. There appears to be no visible case for claiming the presence of a bias on the estimation of the magnitude of subject-by-formulation interactions. The median is -0.0143 and the mean is -0.0010.

Table 3 shows the descriptive statistics for scenario 7. Note that since variances are variable for all trials within this scenario it does not make sense to present the estimates of the variance's components, rather the table presents how different the estimates of the variance components are from

the true values. It is observed that with variable within- and between-variances and variable point estimates (and in the presence of the subject-by-formulation interaction), REML-based estimates are still marginally better than estimates based on the linear model, and as for scenarios 1-6 the REML-based estimates have less variability than those based on the linear model.

Table 4 shows $P'_{0.01}$, $P'_{0.02}$, and $P'_{0.05}$ for the results obtained for scenario 1. The chance that REML provides a better estimate (closer to the simulated value) of σ^2_{WR} than the linear model and that the difference in terms of upper acceptance limit amount to at least 1% is 11.14%. Results for scenarios 2-6 are similar (not shown).

Table 4. Probabilities (P'_x) that REML provides a better estimate than the linear model and that the resulting difference in upper acceptance limit amounts to x=1% or more, x=2% or more or x=5% or more.

x	P'_x
0.01	0.1114
0.02	0.0365
0.05	0.0031

Figures are from data simulated under scenario 1.

Table 3. Descriptive statistics for scenario 7

Metric	Mean	Median	SD	Min	Max	Q10	Q90
$\hat{\sigma}^2_{WR,REML} - \sigma^2_{WR}$	0.0010	-0.0025	0.1261	-0.7815	1.1371	-0.1223	0.1238
$\hat{\sigma}^2_{WR,Lin.model} - \sigma^2_{WR}$	0.0008	-0.0026	0.1274	-0.8186	1.2366	-0.1220	0.1243
$\hat{\sigma}^2_D - \sigma^2_D$	-0.0010	-0.0143	0.2216	-1.4958	1.5048	-0.2481	0.2610

SD: Standard deviation (of a sample); Q10=10th percentile; Q90=90^{ieth} percentile.

DISCUSSION

The largest deviation of the estimates of σ^2_{WR} from the true value were seen in scenario 3. For products for which the true variance components are reflected by this scenario, the scaled limits would be 71.50%-139.87% when the median $\hat{\sigma}^2_{WR,lin.model}$ is plugged into equation III. The scaled limits would be 71.52%-139.83% with the median $\hat{\sigma}^2_{WR,REML}$ and they would be 71.04%-140.78% if the estimate corresponded to the true value. Therefore, the use of the linear model does decrease power for any given sample size, as compared to an alternative when we could input the true value or the REML estimate for the calculation of the scaled limits. The slight under-estimation of the width of the acceptance window will translate directly into only the sponsor's risk, not the patient's risk. The generated data cannot infer anything quantitatively about the degree by which the type I error is affected, but it would be a future relevant simulation study to examine the effect of bias on that property. It should be noted, though, that a statistically significant difference in variance estimates was only observed for scenarios 4 and 6, and only weakly so ($p<0.05$). Thus, in terms of the F-test produced by these studies we cannot say that the variances beyond a reasonable doubt differ generally.

The practical importance may be judged by the figures in Table 4; for scenario 1 there is a 11.14% chance that REML gives a better estimate of σ^2_{WR} than the linear model does and that this difference translates into a difference in the upper acceptance limit of 0.01 or more. There is a 3.65% chance that it translates into a difference of 0.02 or more. There is a 0.31% chance that it translates into a difference of 0.05 or more. Due to the numbers in table 4 and, due to the fact that variance estimates are not significantly different in 5 of 7 scenarios, and only at $p<0.05$ for two of those scenarios, I see no reason to conclude that the two approaches are meaningfully different.

The simulations undertaken by Endrenyi and Tothfalusi (7) showed over-estimation of the subject-by-formulation interaction, this phenomenon is absent here. There are methodological differences between the present approach to quantifying the phenomenon and theirs, notably apart from using equal variance levels within and between subjects for the simulations they did use a mode complex model matrix for the fixed effects as they applied a period-by-sequence-in treatment interaction (See also Hauck et al. (10) for a discussion of technical aspects of the estimation approaches and optimizer constraints). At any rate, the discussion of bias on the observed subject-by-formulation interaction is not of huge practical importance anymore. What is important in relation to subject-by-formulation interaction, however, is that Endrenyi and Tothfalusi (7) noted a relationship between the magnitude of its

estimate and the magnitude of σ^2_{WR} (and an estimate of it). A causative relationship was not investigated, in the sense that their paper did not conclude if changes in the magnitude of σ^2_{WR} causes changes in the estimate of σ^2_D or vice versa. If the observation of Endrenyi and Tothfalusi was more general, i.e., if the estimate of one variance components when fit with REML could be generally biased due to the magnitude of another variance component, then the quality of any estimator could suffer. The simulations here suggest that this isn't the apparent case, at least not to any appreciable degree.

CONCLUSIONS

The results presented give rise to the following conclusions: 1. Estimates of σ^2_{WR} obtained with the linear model correlate well with estimates obtained through REML (r^2~0.96, linear regression). 2. REML generally gives marginally better and less variable estimates of σ^2_{WR} than the linear model, but there is only a significant difference ($p<0.05$) in estimates for two of the seven scenarios. In these trials the observed levels are slightly below the (true) simulated values. 3. Estimates of σ^2_{WR} appear quite independent of the subject-by-formulation interaction, regardless of whether REML or a linear model is used. 4. The subject-by-formulation interaction σ^2_D does not appear to be over-estimated with REML.

ACKNOWLEDGMENTS

Thanks to Helmut Schütz of BEBAC, Austria for helpful discussions. And to the user PharmCat on the bebac forum (forum.bebac.at) who pointed me in the direction of the paper by Gurka and provided technical insight. Both kindly posted results generated with commercial software against which some of my scripts could be performance qualified.

CONFLICTS OF INTEREST

This paper does not make any inference about any product or any company. The author is a consultant whose present and former clients include companies, agencies, pharmacopoeias,
and the World Health Organization. The author declares no conflict of interest.

REFERENCES

1. European Medicines Agency, Committee for Human Medicinal Products. Investigation of bioequivalence. CHMP. CPMP/EWP/QWP/1401/98 Rev. 1. 2010.
2. European Medicines Agency, Committee for Human Medicinal Products. Overview of comments received on draft guideline on the investigation of bioequivalence cpmp/ewp/qwp/1401/98 rev. 1. 2010.
3. European Medicines Agency. Clinical pharmacology and pharmacokinetics: questions and answers. Undated or 2011.
4. Chow S.-C., Liu J.-P.Design and Analysis of Bioavailability and Bioequivalence Studies, 3rd edition, Chapman & Hall, 2009.
5. Medicines Evaluation Board, 2017: Public Assessment Report, NL/H/3389/001-003/DC. https://db.cbg-meb.nl/Pars/h116789.pdf
6. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 38(4) 1982: 963-974.
7. Endrenyi L, Tothfalusi L. Subject-by-formulation interaction in determinations of individual bioequivalence: bias and prevalence. Pharm. Res. 16(2) 1999:186-190.
8. Gurka M.J. Selecting the Best Linear Mixed Model Under REML. The American Statistician 60(1) 2006: 19-26
9. Nelder J, Mead R. A simplex method for function minimization". Computer Journal. 7 (4) 1965: 308–313.
10. Hauck W.W., Hyslop T, Chen M.L., Patnaik R., Williams R.L. Subject-by-formulation interaction in bioequivalence: conceptual and statistical issues. FDA Population/Individual Bioequivalence Working Group. Food and Drug Administration. Pharm. Res. 17(4) 2000: 375-380.