Longitudinal assessment of the impact of multiple switches between a biosimilar and its reference product on efficacy parameters

Johanna Mielke¹ | Heike Woehling² | Byron Jones¹

¹ Novartis Pharma AG, Basel, Switzerland
² Sandoz Biopharmaceuticals, Holzkirchen, Germany

Correspondence
Johanna Mielke, Novartis Pharma AG, Basel, Switzerland.
Email: johanna.mielke@udo.edu

Funding information
Marie Sklodowska-Curie, Grant/Award Number: 633567; Swiss State Secretariat for Education, Research, and Innovation (SERI), Grant/Award Number: 999754557

Patients, physicians, and health care providers in Europe have more than 10 years of experience with biosimilars. However, there are still debates if switching between a biosimilar and its reference product influences the efficacy of the treatment. In this paper, we address this uncertainty by developing a formal statistical test that can be used for showing that switching has no negative impact on the efficacy of biosimilars. For that, we first introduce a linear mixed-effects model that is used for defining the null hypothesis (switching influences the efficacy) and the alternative hypothesis (switching has no influence on the efficacy). Using this as the foundation of our work, we propose several approaches for testing for changes in the efficacy of the treatment due to switching and discuss the properties of these tests in an extensive simulation study. It is shown that all these methods have advantages and disadvantages and the decision regarding which method is preferred depends on the expectation of a switching assessment. To demonstrate the applicability of the methods in practice, the approaches were applied to the data of the EGALITY study, which compares the reference product Enbrel® (Amgen) with the approved biosimilar Erelzi® (Sandoz).

KEYWORDS
biosimilarity, crossover design, Enbrel, equivalence testing, switching

1 | INTRODUCTION

Due to the high costs and the expiration of patents of biologics in the last few years, the development and approval pathway of biosimilars¹ has recently gained much attention. Both the European Medicines Agency and the Food and Drug Administration have already approved several biosimilars, and considering the high number of applications that are currently under investigation by these agencies, it can be expected that the number of approved biosimilars will greatly increase in the future.

Nonetheless, patients, physicians, and health care providers are still debating if patients who are already taking the originator product should be switched to the biosimilar and if switching between the biosimilar and its reference product should be allowed, eg, by automatic substitution at the pharmacy level, as it is often the case for generics. The regulatory authorities in different regions handle this differently²: the European Medicines Agency states that “the Agency’s evaluations do not include recommendations on whether a biosimilar should be used interchangeably with its reference medicine” and recommends that “for questions related to switching from one biological medicine to another, patients...
should speak to their doctor or pharmacist". Within the member states of the European Union, the viewpoints and handling of switching are diverse: the Finnish Medicines Agency, for example, published a position paper stating that "switches between biological products are common and usually not problematic", whereas the Health Products Regulatory Agency in Ireland, for example, explicitly "does not recommend that patients are switched back and forth between a biosimilar and the reference medicinal product". The European Biopharmaceutical Enterprises, the European Federation of Pharmaceutical Industries and Associations, and the International Federation of Pharmaceutical Manufacturers and Associations also recently published a joint position paper on switching, advocating that the decision as to whether a patient can be switched to another biological product should be made by the physicians individually for each patient and physicians should not be pushed into switching patients. In the United States, the Food and Drug Administration has the legal option to approve a biosimilar as an interchangeable biosimilar. This status means that patients can be switched between the biosimilar and its reference product without the knowledge or approval of the prescribing doctor. So far, there are no approved interchangeable biosimilars in the United States. It is important to point out that there is a clear hierarchy between showing that a product is biosimilar and that switching between the test and reference products does not impact the efficacy and safety of the treatment. Biosimilarity does not imply switchability, and if a product is not biosimilar, it is not reasonable to debate whether patients can be switched to the biosimilar. Therefore, we assume that biosimilarity has already been proven beforehand. Hence, the focus is solely on showing that switching between the biosimilar and the reference product does not negatively impact efficacy.

There are several published results of clinical studies focussing on the effect of switching between a biosimilar and its reference product, eg, on switching between drugs with the active substance somatropin and on switching between products that use the active substance infliximab. None of these studies found any evidence that switching has a negative impact, but no multiple switches were considered, and the authors often did not use any formal statistical testing procedure for establishing switchability and merely relied on descriptive and graphical comparisons. Faccin et al and Chow et al discussed possible study designs for testing switchability, and there are also a few publications focussing on the statistical methodology for assessing switchability: Chow et al proposed a Switching and Alternating Index and a Scaled Criterion for Drug Interchangeability, Zheng et al discussed safety margins and Belleli et al presented a model-based approach. All 4 methods were not specifically developed for assessing the impact of multiple switches in comparison with continuously receiving treatment with the test or reference product and do not offer a statistical testing procedure. In this paper, we propose several longitudinal statistical testing procedures that are applicable to study designs with multiple switches for normally distributed endpoints.

The rest of this paper is such that in Section 2, data situations under the null hypothesis (switching influences the efficacy) and the alternative hypothesis (switching has no influence on the efficacy) are described. Building on these ideas, we define in Section 3 the assumed statistical model and formally introduce the hypotheses that are to be tested. The proposed testing procedures are introduced in Section 4 and compared in a simulation study using various scenarios in Section 5. A sensitivity analysis is conducted in Section 6. In Section 7, the proposed methods are applied to the data of the EGALITY study. A conclusion and additional remarks are given in Section 8.

2 | THE CONCEPT OF SWITCHABILITY

There are, to the best of our knowledge, no proposals yet published for the definition of the null and alternative hypotheses in an assessment of multiple switches in a longitudinal setting. As the goal is to show that switching has no impact on the efficacy of the treatment, any considered test will need to be an equivalence or non-inferiority type test. Therefore, the definition of the null hypothesis is crucial for the test decision. In addition, the test can be based on several characteristics of the data (eg, difference in mean value or difference in distribution) and so is not unique. The few clinical trials that have focussed on switching all claim that the trial data confirm switchability. Thus, it is unclear which pattern in the data can be expected to be seen if switching decreases efficacy and against which kind of deviation from the setting in which switching between test (T) and reference (R) does not impact efficacy one would like to be assured, ie, how should the null hypothesis be formulated. One could, for example, wish to have a high chance not to reject the null hypothesis and not to conclude switchability if the mean or variability of the response changes because of switching. Two examples of such situations are shown in Figure 1. The Figure shows (simulated) individual subject trajectories over 6 periods. A continuous reference treatment is given to subjects in red (dashed lines, eg, treatment sequence RRRRRR). The subjects in black (solid lines) are in a switching sequence starting with the test treatment and switching between T and R after each period (treatment sequence TRTRTR). In Figure 1A, there does not seem to be a systematic difference between the
red and black lines; switching between the 2 products does not seem to have an impact on the response. In Figure 1B, the mean response decreases after the subjects start to switch. This is clearly not desirable, and one would not recommend patients to switch back and forth if this would have been observed in a study. Figure 1C shows an example in which the variability of the subjects who are switching is higher and allowing switches between T and R might be questionable.

As the null and alternative hypotheses for biosimilarity are also only defined and assessed in terms of difference of the mean values and concepts that also consider differences in variability like individual bioequivalence have not yet been introduced for biosimilarity development, we will put our focus on detecting changes in mean, which is the situation in Figure 1B. Nonetheless, in Section 6, we will also discuss another setting in which the switchability might be questionable (ie, the increase of the variability after switching, Figure 1C) and analyse the performance of the proposed methods in detecting this type of deviation.

### 3 THE STATISTICAL MODEL AND THE HYPOTHESES

In this section, we formally introduce the statistical model and the null and alternative hypotheses. For that, we will assume that T and R will be compared using a design with q sequences and p periods. For defining the null hypothesis (switching influences the efficacy), we assume a linear mixed-effects model, which is given by

\[ y_{i,j,k} = p_k + I_{at(i,j,k)} + I_{at(i,j,k−1),at(i,j,k)} + \kappa_{i,j,k} + \epsilon_{i,j,k}. \]

The response of the ith subject \((i = 1, \ldots, n)\) in period \(k (k = 1, \ldots, p)\) and sequence \(j (j = 1, \ldots, q)\) is denoted by \(y_{i,j,k}\). The term \(p_k\) is the period effect, \(I_{at(i,j,k)}\) is the effect of the treatment administered in period \(k\) of sequence \(j\) with \(a(i,j,k) = T\) if treatment T was given to subject \(i\) in sequence \(j\) and period \(k\) and \(a(i,j,k) = R\) is defined analogously. The term \(\kappa_{i,j,k}\) is the random effect of subject \(i\) in sequence \(j\) with mean 0 and variance \(\sigma^2\), and \(\epsilon_{i,j,k}\) are the independent residuals with mean 0 and variance \(\sigma^2\). We assume that the random subject effects \(\kappa_{i,j,k}\) are mutually independent and also independent of the residuals \(\epsilon_{i,j,k}\). \(I_{at(i,j,k−1),at(i,j,k)}\) is the effect of switching from the treatment in period \(k−1\) to the treatment in period \(k\) for subject \(i\) in sequence \(j\). In addition, it is assumed that \(I_{at(i,0),at(i,1)} = 0\). We assume that the switching effect depends only on the treatment in the current period and the treatment in the previous period and that it is a fixed effect whereby the switch from T to R leads to the same switching effect \(I_{TR}\) as a switch from R to T, which is denoted by \(I_{RT}\). This appears reasonable because biosimilarity is assumed to be proven beforehand. The switching effect is 0 if the subjects do not switch \(I_{RR} = I_{TT} = 0\), ie, when the treatment in period \(k\) is the same as the treatment in period \(k−1\).

It is reasonable to claim that switching between T and R does not influence the efficacy (the alternative hypothesis) if \(|I_{RT}| = |I_{TR}|\) is “small”. As in all applications of equivalence testing, it is necessary to decide what “small” means, based on the clinical context. If log-transformed PK parameters are analysed, \(\Delta = \log(1.25)\) is commonly used for the equivalence margins. If an efficacy endpoint is assessed, the difficulty of this choice is comparable with the choice of the margins for non-inferiority studies and beyond the scope of this paper. Here, we will assume that there is a margin \(\Delta\) related to the effect of switching such that if the effect is smaller than \(\Delta\), the impact of switching is not clinically relevant. Then, we can introduce the null and alternative hypotheses more formally as

\[ \begin{align*}
H_0: & \quad \Delta \geq 0, \\
H_A: & \quad \Delta < 0.
\end{align*} \]

### FIGURE 1

Example of trajectories of subjects who were followed over 6 periods. A, Switching does not influence the response. B, Switching decreases the response. C, Variability is higher for subjects who are switching.
4 | PROPOSED TESTING PROCEDURES

In this section, we propose 3 different tests for normally distributed endpoints that follow the ideas of Belleli et al.\(^{18}\): one is based on directly estimating the effect of switching (Section 4.1) and 2 are based on predictions from a linear model (Section 4.2). All methods have advantages and disadvantages that are discussed in Sections 5 and 6. It is assumed that the study design consists of sequences that can be split into “switching” (subjects are alternating between test and reference products) and “non-switching” (continuous treatment with test or reference) sequences. An example for such a study design is given in Table 1: the subjects are followed over 6 periods. While subjects in the non-switching sequences (sequences 1 and 2) stay on T or R during the study, subjects in the switching sequences (sequences 3 and 4) switch the treatment after each period. This study design is also one of the designs discussed by Chow et al.\(^{13}\) for the assessment of interchangeability.

The distribution of the proposed methods under the null hypothesis depends on the variance of the subject effect and the variance of the residual error. In practice, however, these are not known and need to be estimated. For that, the linear mixed-effects model

\[
y_{i,j,k} = p_k + I_{a(i,j,k)} + \kappa_{i,j} + \epsilon_{i,j,k},
\]

which uses the same notation as the model in Section 3, is fitted to the observations in the non-switching sequences, and the variances are extracted. If the restricted maximum likelihood approach is used, the estimates of the variances are unbiased.\(^{22}\) We note that for the test statistics developed in the next subsections, the additional uncertainty of the variance estimators is not incorporated in the distribution of the test statistics. However, in Section 5, we will show that the tests nonetheless have good properties, the tests keep the nominal significance level, and the impact on power is negligible.

4.1 | Estimation method

The method that is proposed in this subsection directly estimates the effect of switching by fitting a linear mixed-effects model to the dataset that is similar to the one already introduced in Section 3:

\[
y_{i,j,k} = p_k + I_{a(i,j,k)} + c_{(j,k,k-1)} + \kappa_{i,j} + \epsilon_{i,j,k}.
\]

The only difference compared with the model in Section 3 is that the switching effect \(I_{a(i,j,k-1),a(i,j,k)}\) is replaced by the carryover effect \(c_{(j,k,k-1)}\), which is a categorical factor with 3 levels \((c_0, c_1, and c_2)\): the first level \(c_0\) is used for all observations in the first period in which no switching effect is expected and for observations for which the treatment in period \(k\) is the same as in period \(k-1\). For a switch from T to R, the second level \(c_1\) will be used, and for a switch from R to T, the third level \(c_2\) is used. Therefore, the carryover effect distinguishes between “no switch,” “switch R to T,” and “switch T to R.” It should be noted that this approach is more general than the assumed model, because the assumption that the effect due to a switch from T to R is the same as the effect due to a switch from R to T is not used. The level “no switch” \((c_0)\) will be

| Sequence | Period | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|--------|---|---|---|---|---|---|
| 1        |       | T | T | T | T | T | T |
| 2        |       | R | R | R | R | R | R |
| 3        |       | R | T | R | T | R | T |
| 4        |       | T | R | T | R | T | R |

\(T\) is the test treatment (the biosimilar), and \(R\) is the reference treatment.
used as the reference level so that the relative effects of “T to R” \((c_1)\) and “R to T” \((c_2)\) can be directly estimated \((\hat{c}_1, \hat{c}_2)\).

The null hypothesis should be rejected if the absolute values of \(c_1\) and \(c_2\) are smaller than a prespecified value. To avoid dealing with the absolute value, we consider the equivalent condition that \(c_1, c_2, -c_1,\) and \(-c_2\) are smaller than this value. Therefore, the test decision can be based on a multivariate normal distribution because, as shown by Jiang,\(^2^3\) the vector of the estimates

\[ \begin{pmatrix} \hat{c}_1, \hat{c}_2, -\hat{c}_1, -\hat{c}_2 \end{pmatrix}^T \]

asymptotically follows a multivariate normal distribution with mean

\[ \mu_1 = (c_1, c_2, -c_1, -c_2)^T \]

and variance-covariance matrix \(\Sigma_1\), which is derived in Appendix A. The null hypothesis should be rejected if the absolute value of both switching effects are smaller than a critical value \(x\). For controlling the type 1 error rate, it must hold true under the null hypothesis that

\[ P(|\hat{c}_1| \leq x, |\hat{c}_2| \leq x) = P(\hat{c}_1 \leq x, \hat{c}_2 \leq x, -\hat{c}_1 \leq x, -\hat{c}_2 \leq x) = F_{\mu_1, \Sigma_1}(x) = \alpha, \]

where \(F\) denotes the joint distribution of \((\hat{c}_1, \hat{c}_2, -\hat{c}_1, -\hat{c}_2)\) under the null hypothesis. Therefore, the \(a\) quantile of this multivariate normal distribution can be used as the critical value. The distribution under the null hypothesis is based on the variance-covariance matrix \(\Sigma_1\) (see Appendix A) that depends on the variance of the subject effects and the residual errors and the mean value that corresponds to the acceptable difference in mean: \(\mu_1 = (\Delta, \Delta, -\Delta, -\Delta)^T\). It should be noted that the variance of the subject effects and the residual errors that are incorporated into the variance-covariance matrix \(\Sigma_1\) must be estimated as described previously.

### 4.2 Prediction methods

The idea of these methods is based on prediction with a linear model: one observation from each subject which is free of any switching effect (the modelling dataset) is used for fitting a linear model with treatment and period as fixed-effects. The rest of the data (the evaluation dataset) are used for comparing the observed responses with the predictions obtained from the linear model that was fitted in the first step. If switching has an impact on the efficacy, it can be expected that the difference between predictions and observations will be much larger for the switching sequences than for the non-switching sequences. For simplification of the notation, the methods are described for the sample study design that was given in Table 1 with 4 sequences and 6 periods, whereby the first 2 sequences are non-switching sequences and the last 2 are switching sequences with equal sample sizes in all sequences. We assume that the data within the 2 switching arms and within the 2 non-switching arms can be pooled as biosimilarity has already been proven beforehand.

First, the modelling dataset needs to be defined. It consists of one observation per subject: from the switching sequences, always the observation from the first period is used because no switching effect can be expected in the first period (assuming treatment-naive patients or a sufficiently long washout period prior to the start of the study). From the non-switching sequences, one observation is used per person from periods 2 to 6. More concretely, from the first subject the response from the second period, from the second subject the response from the third period, and so on are used. With this strategy, observations from all periods are included, allowing for a reliable estimate of a period effect. With the notation that was previously introduced \((y_{i,j,k} \text{ is the observation from subject } i \text{ in sequence } j \text{ and period } k)\) and assuming \(n\) observations per sequence, the modelling dataset is defined as

\[ M := \{ y_{1,1,2}, y_{1,1,3}, y_{3,1,4}, y_{4,1,5}, y_{5,1,6}, \ldots, y_{1,2,2}, y_{2,2,3}, \ldots, y_{1,3,1}, \ldots, y_{n,3,1}, y_{1,4,1}, \ldots, y_{n,4,1} \}. \]

As only 1 observation from each person is used, the observations in \(M\) are independent. The evaluation dataset \((E)\) consists of all other observations.

It is important to note that there are several options for splitting the data into the evaluation and modelling datasets. For example, one could also use all observations from the non-switching sequences as the modelling dataset and the data from the switching sequences as the evaluation dataset. Using data from all subjects and comparing the switching and non-switching sequences tend to be more robust against model misspecification, and therefore, we prefer this split. However, the proposed methodology can be easily adjusted also for alternative splits.

In the first step of the proposed methodology, a linear model is fitted to the modelling dataset \(M\):

\[ y_{i,j,k} = p_k + t_{a(i,j,k)} + e_{i,j,k}, \]
where \( l_{a(i,j,k)} \) is the effect of the treatment which is given to subject \( i \) of sequence \( j \) in period \( k \) and \( p_k \) is the period effect. A linear model is used instead of the linear mixed-effects model that was mentioned earlier because only 1 observation per subject is used and the observations are therefore independent. Afterwards, with this model, the predictions for the observations in the evaluation dataset \( E \) are calculated and are denoted by \( \hat{y}_{i,j,k} \). The difference between an observation and its prediction is defined as

\[
\hat{y}_{i,j,k} = y_{i,j,k} - \hat{y}_{i,j,k}, \quad (i,j,k) \in E.
\]

The test decision is based on comparing the prediction error of the switching and non-switching sequences. For that, 2 different tests are proposed: the first is based on the difference of the mean squared differences (MSDs) of observations and predictions between the switching and non-switching sequences (quadratic prediction method), and the second uses the difference between the distributions of \( \hat{y}_{i,j,k} \) (distribution prediction method). For both methods, the distribution of \( \hat{y} \) must be known. It can be shown (see Appendix B) that, assuming that the variance of the subject effect and the variance between predictions and observations, respectively, as

\[
\begin{align*}
\nu^2 & = \text{variance of the subject effect}, \\
\Sigma^2 & = \text{variance-covariance matrix}. 
\end{align*}
\]

is distributed as a random variable that follows a generalised \( \chi^2 \) distribution, and there are numerical procedures that can be used for calculating the distribution function of \( Y \).24 In relation to our test statistics, \( X \) is the vector of differences between predictions and observations, \( \hat{y} \), and it can then be easily seen that the matrix \( A \) is a diagonal matrix with entries

\[
a_{l,l} = \begin{cases} 
-1 & l \in \{1, \ldots, 10n\}; \\
1 & l \in \{10n + 1, \ldots, 20n\}.
\end{cases}
\]

With all the information put together, the \( \alpha \) quantile of the test statistic under the null hypothesis can be calculated and used as the critical value. Under the null hypothesis, the mean value of the multivariate normal distribution of \( \hat{y} \) depends on the equivalence margin \( \Delta \), and the variance-covariance matrix is specified by the variance of the subject effects and the residual errors. This leads to a variance-covariance matrix \( \Sigma^2 \) (see Appendix B) and the mean vector \( \mu_2 \) under the null hypothesis, which is given by a vector that consists of \( 10n \) entries with the value 0 and \( 10n \) entries with the value \( \Delta \), where \( \Delta \) is the equivalence margin:

\[
\mu^T = (0, \ldots, 0, \Delta, \ldots, \Delta). 
\]

The test rejects the null hypothesis and claims switchability if \( t_{\text{MSD}} \), the observed value of \( T_{\text{MSD}} \), is smaller than this critical value.

The distribution prediction method is based on the idea of comparing the distributions of the vector of the difference between prediction and observation in the switching and non-switching sequences using the Kolmogorov-Smirnov distance \( D \), which is defined as the maximum absolute distance between the 2 empirical distribution functions. For 2 empirical distribution functions \( F_n^{(1)} \) and \( F_n^{(2)} \), the Kolmogorov-Smirnov distance \( D \) is given by
\[ D = \max_{z \in \mathbb{R}} |F_n^{(1)}(z) - F_n^{(2)}(z)|, \]

where for a vector \( x = (x_1, \ldots, x_n) \) the empirical distribution function is defined as

\[ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}, \]

where \( \mathbb{1}_{x_i \leq x} \) is the indicator function and is 1 if \( x_i \leq x \) and 0 otherwise. Following this formula, the function \( F_n \) is estimated for the switching and non-switching sequences using the differences between prediction and observation, \( \tilde{y} \), as vector \( x \). It should be noted that the estimated functions are not ordinary empirical distribution functions because the observations that are used for their estimation are not independent, so we will refer to them as estimated functions. However, as the critical value is approximated by simulation (see below), this is not critical and the concept of comparing distributions is applicable here. Assuming that the estimated function in the switching arms is denoted by \( F_n^{(s)} \) and the estimated function for the non-switching arms is denoted by \( F_n^{(ns)} \), the test statistic is defined as

\[ T_{DP} = \max_{z \in \mathbb{R}} |F_n^{(s)}(z) - F_n^{(ns)}(z)|. \]

If switching has no impact on the response, there should be no difference between the 2 estimated functions, and therefore, the realisation of the test statistic should be close to 0. Thus, the test rejects the null hypothesis (not switchable) if the observed value \( t_{DP} \) of the test statistic \( T_{DP} \) is smaller than the \( \alpha \) quantile of the distribution of the test statistic under the null hypothesis.

Because of the dependencies between the elements of \( \hat{y} \), the distribution of the test statistic under the null hypothesis cannot be easily derived analytically and will be approximated by simulations in this paper. For that, the derived distribution of \( \hat{y} \) is used: 10 000 realisations of the vector \( \hat{y} \) are generated using the mean value under the null hypothesis \( \mu^*_1 = (0, \ldots, 0, \Delta, \ldots, \Delta) \) and the derived variance-covariance matrix \( \Sigma_2 \) (see Appendix B). The test statistic is calculated for each simulated dataset, and the empirical \( \alpha \) quantile of these observed test statistics is used as the critical value.

In a way similar to the estimation method (Section 4.1), the variance of the subject effects and the variance of the residual errors are used in the variance-covariance matrix \( \Sigma_2 \) of the difference between prediction and observation in both methods. As described previously, they are estimated and plugged in, and the resulting distribution is not adjusted for the additional uncertainty.

5 | SIMULATION STUDY

5.1 | Set-up

The datasets for the simulation study are generated using the linear mixed-effects model

\[ y_{i,j,k} = p_k + t_{a(i,j,k)} + I_d(a(i,j,k-1),a(i,j,k)) + \kappa_{i,j} + \epsilon_{i,j,k} \]

that was already introduced in Section 3 and where \( y_{i,j,k} \) is the observed response from subject \( i \) in sequence \( j \) and period \( k \). It is assumed that the subject effect, \( \kappa_{i,j} \), follows a normal distribution with mean 0 and variance \( \sigma_k^2 \) and is constant within each subject; the residual error, \( \epsilon_{i,j,k} \), follows a normal distribution with mean 0 and variance \( \sigma_e^2 \). The random-effects are independent of each other.

Table 2 shows the considered parameter settings. The study design that was described in Section 4 is used (see Table 1). In total, 240 settings are simulated. The critical values are calculated for all settings using the true variances (\( \sigma_k^2 \) and \( \sigma_e^2 \)) and the estimated variances (\( \hat{\sigma}_e^2 \) and \( \hat{\sigma}_e^2 \)). The acceptable difference in mean (equivalence margin) is set to \( \Delta = \log(1.25) \), and \( \alpha = 0.05 \) is used as the significance level. No period or sequence effects are introduced. For each parameter constellation, 5000 datasets are generated and the test statistics, critical values, and the test decisions are recorded.

All calculations were performed with R 3.2.3 software. The linear mixed-effects model was fitted using the package nlme,27 the quantiles of the multivariate normal distribution were calculated with the package mvtnorm,28 and the generation of multivariate normal distributed random variables was performed with the package MASS.29
The performance of the methods is compared under the null hypothesis (to evaluate the type 1 error rate) and under the alternative (to evaluate the power). Also, the impact of using the estimated variances instead of the true variances is investigated. Table 3 shows the mean value and the 5% and 95% quantiles of the type 1 error rates over all scenarios for the estimated and true variances. The nominal significance level was $\alpha = 0.05$. If the true variances are used, the mean rejection rates are very close to the nominal significance level. If the estimated variances are used, the quadratic prediction method seems to be conservative, the rejection rates for the 2 other methods are still close to the nominal significance level. None of the methods seem to be too liberal if the estimated variances are used.

Figure 2 shows the empirical rejection rates of the methods using several scenarios. Since the difference between the rejection rates using the true variances and the estimated variances is small (mean absolute differences of rejection rates: estimation method, 0.002; distribution prediction method, 0.0086; and quadratic prediction method, 0.0180), only the rejection rates using the estimated variances are presented, which is the approach that can be used in practice. The deviation from the setting in which no switching effect is assumed ($I_{RT} = I_{TR} = 0$) is displayed on the $x$ axis. The value $\Delta = \log(1.25)$ is the acceptable difference such that the difference of the response due to switching is not clinically relevant and therefore represents the null hypothesis under which a rejection rate of 5% is anticipated. As expected, higher variances (Figure 2B,C) decrease the power and thus make it more difficult to claim switchability. A higher sample size (right-hand panels) increases the power. Changing the difference between $T$ and $R$ from 0 to $\log(1.1)$ did not influence the rejection rates (not shown). This is not surprising because the effect of $T$ in comparison with $R$ is included in the linear model for all methods. In the comparison of the rejection curves for the 3 methods, it is clear that the estimation method has the highest power under all scenarios. This was expected because the method directly estimates the increase of the switching effect, whereas the 2 methods that are based on prediction measure this increase only indirectly. The power curve for the estimation method is also steeper than for the other methods, making it easy to distinguish between the null and alternative hypotheses.

The distribution prediction method has also high power values. Especially if the switching effect is smaller than 0.05, the power is nearly as high as for the estimation method. It can be assumed that this is the region that would be a realistic assumption for power and sample size calculation in practice because it seems unlikely that a sponsor would try to claim that the products can be switched if the switching effect is not close to 0 in reality. The quadratic prediction method is less powerful. In particular, for small sample sizes ($n = 25$) and higher variances, the power is much lower than for the other 2 methods, for example, even for the setting without a switching effect ($I_{RT} = I_{TR} = 0$), the empirical power is less than 40% in the setting with a residual variance of 0.08 and a variance of the subject effect of 0.05. If the sample size per group is increased to $n = 100$ and the switching effect is smaller than 0.05, the performance is, for the considered variances, comparable with the other methods. Therefore, one needs to enroll more subjects if the quadratic prediction method is to be used.

### Table 3

Mean value and 5% and 95% quantiles (in brackets) of the type 1 error rates over all settings using the estimated and true variances

| Estimation method | Quadratic prediction method | Distribution prediction method |
|-------------------|-----------------------------|--------------------------------|
| Estimated variance| 0.0517 (0.0464, 0.0564)      | 0.0377 (0.0266, 0.0467)         | 0.0510 (0.0446, 0.0567)         |
| True variance     | 0.0504 (0.0464, 0.0542)      | 0.0497 (0.0444, 0.0554)         | 0.0469 (0.0417, 0.0528)         |
FIGURE 2 Empirical rejection rates for several data settings (Estimation: estimation method, Distribution: distribution prediction method, Quadratic: quadratic prediction method). A, $\sigma^2_e = \sigma^2_s = 0.02$. B, $\sigma^2_e = \sigma^2_s = 0.05$. C, $\sigma^2_e = 0.08, \sigma^2_s = 0.05$. The left panels use $n = 25$ subjects per sequence, and the right panels use $n = 100$ subjects per sequence. It is assumed that the difference between T and R is 0. $\Delta = \log(1.25)$ (the equivalence margin).

Considering this simulation study only, one would clearly choose the estimation method. However, it should be noted that as this method directly estimates the increase of the switching effect and the true underlying model was the same as the fitted model, the robustness against other deviations from switchability (eg, increasing variability) might be very limited. As it is unclear which data settings can be expected for products that cannot be switched (see discussion in Section 2), it might be preferable to use a method that does not reject the null and claim that switching has no impact when other deviations from switchable settings occur. This will be investigated in the next section.

6 | SENSITIVITY TO OTHER NON-SWITCHABLE SETTINGS

In this section, we consider, as an example of another type of deviation from switchability, the case of increasing variability for patients who are switching in comparison with patients with continuous treatment with T or R and compare the performance of the methods under this violation of switchability. From the individual patient's perspective, a higher variability of the response variable might make a difference: assuming that the switching effect follows a normal distribution with mean 0, a variance of the switching effect in the switching sequences of $\sigma^2_{TR} = \sigma^2_{RT} = 0.025$ corresponds to a setting in which 15% of the absolute switching effects are larger than the equivalence margin $\Delta = \log(1.25)$. For $\sigma^2_{TR} = \sigma^2_{RT} = 0.05$ this increases to 32% and for $\sigma^2_{TR} = \sigma^2_{RT} = 0.75$ is even as high as 80%.

As the main focus of the methods is to ensure that the mean value of the switching effect is small, the critical values will be based on an acceptable difference in mean of $\Delta = \log(1.25)$ as described in the previous sections. The true variances for the subject effect and the residual error are used. We assume that the switching effect is not a fixed effect, but a random effect that follows a normal distribution with mean 0 and a variance $\sigma^2_I$ (see Figure 1C). The variance depends on the type of switch and is 0 if the subjects are not switching (continuous treatment with T or R, $\sigma^2_{RR} = \sigma^2_{TT} = 0$). Furthermore, it is
Figure 3: Empirical rejection rates for the 3 methods (Estimation: estimation method, Distribution: distribution prediction method, Quadratic: quadratic prediction method). Setting: \( n = 25 \) per sequence, \( \sigma_e^2 = \sigma_s^2 = 0.02 \), difference between \( T \) and \( R \): 0, no period or sequence effect. The increasing variance of the switching effect is displayed on the \( x \) axis. The expected value of the switching effect is 0.

Assumed that the variance of the switching effect is the same for the switch from \( T \) to \( R \) and from \( R \) to \( T \) \( (\sigma_{TR}^2 = \sigma_{RT}^2) \). The aim of this analysis is to investigate if the proposed methods can detect scenarios in which the variance of the switchability effect increases even though the mean value does not change.

We chose values for the variance of the random switching effect, \( \sigma_{TR}^2 = \sigma_{RT}^2 \), between 0 and 0.75. For each value of the variance of the switching effect, 1000 datasets were generated using \( n = 25 \) subjects per sequence, \( \sigma_e^2 = \sigma_s^2 = 0.02 \) and no difference between \( T \) and \( R \). The study design given in Table 1 in Section 4 is used. The empirical rejection rates were calculated. It is desirable that the methods only reject the null hypothesis and claim switchability if the variances are low. The results are shown in Figure 3: the quadratic prediction method only decides for switchability in a high percentage of the cases if the variance of the switching effect is very close to 0. The distribution prediction method allows also for higher variances, whereas the rejection rate of the estimation method is not influenced by the increasing variances even if the additional variability is very high. If a higher variability that is due to switching is also a characteristic of interest, the estimation method might not be preferable anymore.

7 | CASE STUDY

In this section, the applicability of the proposed method will be demonstrated using data from the EGALITY study. This study was a multicentre, randomised, double-blind study in patients with moderate to severe chronic plaque-type psoriasis and was conducted between 24 June 2013 and 30 March 2015 in 74 centres across Europe and South Africa. The study design is given in Table 4. Five hundred thirty-one patients were randomised 1:1 to Erelzi® (the biosimilar) or Enbrel® (the reference product). Patients with at least 50% improvement in the Psoriasis Area and Severity Index (PASI 50) after week 12 were re-randomised to stay on the same treatment or to switch 3 times between biosimilar and reference. After week 30, all patients stay on the last treatment. Since no switch occurred in the last period, we focus our analysis on the first 4 periods (followed up until week 30), which leads to the modified study design with 4 sequences and 4 periods that is shown in Table 5. We only consider subjects with complete observations in these 4 periods. This leads to 470 subjects who were included in the assessment. In this study design, 2 of the sequences can be considered to be switching sequences (sequences 3 and 4) and 2 are non-switching sequences (sequences 1 and 2). Of the 470 subjects, 142 were in sequence 1 (TTTT), 141 in sequence 2 (RRRR), 91 in sequence 3 (RTRT), and 96 in sequence 4 (TRTR).

### Table 4: Study design of the EGALITY study

| Sequence | Follow-up | Week 12 | Week 18 | Week 24 | Week 30 | Week 52 |
|----------|-----------|---------|---------|---------|---------|---------|
| 1        | T         | T       | T       | T       | T       |         |
| 2        | R         | R       | R       | R       | R       |         |
| 3        | R         | T       | R       | T       | T       |         |
| 4        | T         | R       | T       | R       | R       |         |

T is the test treatment (the biosimilar), and R is the reference treatment.
TABLE 5 Modified study design of the EGALITY study

| Sequence | Follow-up | Week 12 | Week 18 | Week 24 | Week 30 |
|----------|-----------|---------|---------|---------|---------|
| 1        | T         | T       | T       | T       |         |
| 2        | R         |         | R       | R       |         |
| 3        | R         | T       |         | R       | T       |
| 4        | T         | R       | T       | R       |         |

T is the test treatment (the biosimilar), and R is the reference treatment.

The considered endpoint is the absolute change in the PASI score. This score is the most widely used measurement scale for the severity of psoriasis and lies between 0 (no disease) and 72 (maximal disease). The absolute change from baseline (assessment at week 0) in the PASI score cannot be assumed to be normally distributed. Therefore, a log transformation is applied to the absolute change from baseline to achieve a normally distributed variable. We consider a difference in the absolute change in PASI of approximately $\pm 3$ score points to be not clinically relevant. Since we apply our methodology to log-transformed data, we also have to transform the margins to the new scale. We approximate the margin on the log scale using the following approach: for the log-transformed response value $y_{i,j,k}$ of subject $i$ in sequence $j$ and period $k$, we assume

$$y_{i,j,k} = p_k + t_{a(i,j,k)} + t_{a(i,j,k-1),a(i,j,k)} + \kappa_{i,j} + \epsilon_{i,j,k}$$

as defined in Section 3 and, under the null hypothesis,

$$I_{RT} = I_{TR} = \Delta \text{ or } I_{RT} = I_{TR} = -\Delta$$

for the switching effects. Then, the difference in mean between an observation with a switching effect and without a switching effect on the original scale is given by

$$\Delta_1 = \exp(p_k + t_{a(i,j,k)} + I_{RT}) - \exp(p_k + t_{a(i,j,k)})$$

$$\Delta_2 = \exp(\mu) \cdot \exp(-\Delta) - \exp(\mu)$$

This formula contains the variable $\mu := p_k + t_{a(i,j,k)}$ that represents the mean value of the log-transformed response for a specific treatment and period. Since we need to derive one margin independently of a specific period or treatment, we have to choose a value for $\mu$ that best represents a typical response. For that, we use the mean value of the log-transformed responses of subjects in sequence 2 (continuous treatment with the reference product) and therefore set $\mu = 2.8696$. It should be emphasised that the log-transformed responses are fairly constant (see Figure 4). That is why the choice of the reference population is not crucial for this study. It is also important to note that either the margins on the original scale or the margins on the log scale are asymmetric. Our methodology was developed for symmetric margins. To stay conservative, we use the largest value for $\Delta$ such that the absolute values of both $\Delta_1$ and $\Delta_2$ are smaller than 3. This leads to $\Delta = 0.1571$.

The log-transformed individual patient trajectories are displayed in Figure 4. The figure already confirms that, at least from a descriptive point of view, there does not seem to be a systematic difference between the switching (upper panels) and non-switching arms (lower panels) since the pattern of the trajectories is very similar in all 4 sequences.

Before the proposed methods can be applied to the dataset, it is first necessary to estimate the variance of the subject effect $\sigma_s^2$ and the variance of the residual error $\sigma_e^2$ using the subjects in the non-switching datasets only. This leads to

$$\sigma_s^2 = 0.1555 \text{ and } \sigma_e^2 = 0.0089.$$ 

Checks of the residuals confirm that the model assumptions are met.

The results for all tests are shown in Table 6: the estimation method and the distribution prediction method reject the null hypothesis (show switchability), and the quadratic prediction method does not allow a claim of switchability. However, it should be noted that the power for this test is low: even if no switching effect is assumed ($I_{RT} = I_{TR} = 0$),
the power for the given sample sizes and the estimated variances is only 23.6%. The reason for that is the high variance of the subject effect. Table 6 also gives the smallest value $\Delta$ (tightest margin) such that the tests will still reject the null hypothesis. It is shown that the margins can be extremely small for the estimation method (absolute change in PASI score of 0.35), whereas a large margin would have been necessary to reject with the quadratic prediction method.

### TABLE 6 Critical values, observed test statistics, and test decision for the 3 proposed methods using $\Delta = 0.1571$

| Method      | Observed Value | Critical Value | Reject? | Tightest | Tightest O |
|-------------|----------------|----------------|---------|----------|------------|
| Estimation  | 0.0096         | 0.1460         | Yes     | 0.0199   | 0.3543     |
| Distribution| 0.0592         | 0.0950         | Yes     | 0.1101   | 2.0519     |
| Quadratic   | 0.0488         | −0.0140        | No      | 0.3125   | 6.4673     |

All methods reject if the observed test statistic is smaller than the critical value. Tightest: smallest margin such that the test still rejects the null hypothesis. Tightest O: smallest margins on the original scale (absolute change from baseline in Psoriasis Area and Severity Index score; back-transformed as described above).

### 8 CONCLUSION

Although the first biosimilar in Europe was already approved in 2006, there is still some debate if patients who are already on the reference product should be switched to the biosimilar or if multiple switches between the biosimilar and the reference product lead to reduced efficacy of the treatment. In this paper, we addressed this uncertainty by proposing a formal testing procedure for testing for switchability of a biosimilar (the test product) and its reference product. For that, we introduced a longitudinal statistical model that is used for defining settings in which switching has an impact on efficacy (the alternative hypothesis) and settings in which switching has no impact on the efficacy endpoint (the null hypothesis) in terms of changes in mean due to switching the product. Using this model as the foundation, we introduced 3 approaches for assessing switchability that are applicable to study designs with sequences in which the subjects are alternating between the test (T) and its reference (R) product (eg, TRTR) and sequences in which the patients are treated continuously with the same product (eg, TTTT). This class of study designs was already considered for biosimilar studies, for example, for the EGALITY study that was described by Griffiths et al. The first proposed method is based on a linear mixed-effects model and directly estimates the effect of switching in contrast to continuous treatment with the test or reference product (estimation method). The 2 other approaches use the idea of comparing the difference between the predictions.
that were obtained with a linear model assuming switchability with the actual observed values. These differences were compared between the switching sequences and sequences with continuous treatment with T or R. The comparison is based on the mean quadratic prediction error (quadratic prediction method) or on a comparison of the distributions with the Kolmogorov-Smirnov distance (distribution prediction method).

In an extensive simulation study, it was shown that all 3 methods preserve the desired α level. For the considered settings, the estimation method had the highest power. That was expected because this method uses the same statistical model that was also used for generating the data and directly targets the change in mean. The power of the distribution prediction method is slightly lower, whereas the quadratic prediction method only had high power if the sample size is large or the variability is low. Additional simulation studies that were not presented in this paper confirmed that all 3 methods are fairly robust against model misspecifications (omitted covariates and non-normal error terms).

Although the null and alternative hypotheses were defined in terms of differences in mean between the switching sequences and the continuous sequences, it might be desirable that the methods can also detect situations in which the response pattern differs between the switching and continuous sequences, but the mean values are comparable. As a sensitivity analysis, we considered the case of increasing variability in the switching sequences. It was shown that in this case, the quadratic prediction method performs best, whereas the sensitivity of the distribution prediction method is lower and the estimation method claims switchability very often if the variability in the switching sequences is high. This example shows that the high power of the estimation method comes with a price: if the mean value of the response is not influenced by switching between test and reference, but the variability is increasing, the method will, in most cases, decide that switching has no impact, although for the individual subjects, the response might change dramatically and the efficacy of the drug might be severely reduced. This is, however, a general issue with equivalence testing: as the null hypothesis is not unique and can be based on several characteristics (see Section 2), the claim concerning one characteristic (eg, expected value) does not give any confidence for the equivalence of other characteristics (eg, variance). For example, for claiming bioequivalence in the PK studies in biosimilar development, the concept of average bioequivalence is used, which compares the mean values of the test and reference products, and it is known that this method is not sensitive for detecting differences in variances as discussed by Anderson and Hauck. Therefore, analogously to the properties of the estimation method, products with similar mean, but different variances, can still meet the criterion, and this was not reported to be a concern in the past (see, for example, the discussion by Endrenyi et al). Because of limited experience with switching between biosimilars and the reference products, it is not clear if the same will be true for switchability assessment.

Which approach is preferable for assessing switchability highly depends on the expectations and experiences with the impact of switching. If switchability is defined solely in terms of differences in mean, a statistical method that directly targets this change in mean is desirable like the estimation method. However, if there is uncertainty on whether the mean values are the only measure that should be considered or how the statistical model for the null hypothesis should be specified, the quadratic prediction method that is more flexible also for other types of deviations from switchability might be more reasonable. As the sample size needs to be higher for the quadratic prediction method in comparison with the other approaches, this approach might not be feasible in all settings (see for example the extremely low power in the case study in Section 7). Then, the distribution prediction method might be preferred because it is more sensitive to general changes due to switching than the estimation method but requires a much lower sample size for showing equality in mean than the quadratic prediction method.

The applicability of the proposed methods to data was shown using the EGALITY study as an example. We were able to reject the null hypothesis and claim switchability for the endpoint “absolute change of the Psoriasis Area and Severity Index (PASI)” with the distribution prediction method and the estimation method. The null hypothesis was not rejected for the quadratic prediction method; however, the power for claiming switchability with the given sample sizes is extremely low because of the large subject effect even if we assume that the products are completely interchangeable.

The presented approaches were developed for normally distributed endpoints. It should, however, be noted that the idea of these approaches can, with small modifications, also be applied to other types of endpoints by extending the methods to the family of generalised linear models. Throughout this paper, we have assumed a compound symmetric correlation structure for the repeated measurements of each subject. That is, the correlation between 2 repeated measurements is constant irrespective of the time between the measurements. Our approach is not limited to this type of dependence structure and can be easily extended to other patterns, eg, an autoregressive correlation structure. In addition, it should be emphasised that the methods can be adjusted to other study designs than the ones discussed as long as the study arms can be divided into “switching” and “non-switching” arms.
ACKNOWLEDGEMENTS

This work was supported by funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 633567 and by the Swiss State Secretariat for Education, Research, and Innovation (SERI) under contract number 999754557. The opinions expressed and arguments used herein do not necessarily reflect the official views of the Swiss Government.

ORCID

Johanna Mielke  http://orcid.org/0000-0002-4239-5499

REFERENCES

1. Wang J, Chow SC. On the regulatory approval pathway of biosimilar products. Pharmaceuticals. 2012;5(4):353-368.
2. Tóthfalusi L, Endrényi L, Chow SC. Statistical and regulatory considerations in assessments of interchangeability of biological drug products. Eur J Health Econ. 2014;15(1):5-11.
3. Questions and answers on biosimilar medicines (similar biological medicinal products). 2012. Available at http://www.medicinesfor europ.e.com/wp-content/uploads/2016/03/WC500020062.pdf. Accessed 1 September 2018.
4. Interchangeability of biosimilars—position of Finnish Medicines Agency Fimea. 2015 Available at https://www.fimea.fi/documents/542809/838272/29197_Biosimilarien_vaihtokelpoisuus_EN.pdf (accessed 24 March 2017).
5. Guide to biosimilars for healthcare professionals and patients. Available at https://www.hpra.ie/docs/default-source/publications-forms/guidance-documents/guide-to-biosimilars-for-healthcare-professionals-and-patients-v2.pdf?sfvrsn=18 (accessed 24 March 2017); 2015.
6. Considerations for physicians on switching decisions regarding biosimilars. 2017. Available at https://www.ifpma.org/wp-content/uploads/2017/03/Considerations-for-switching-decisions_IFPMA-vF.pdf. Accessed 1 September 2018.
7. Biologics Price Competition and Innovation Act. 2009 Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/ucm216146.pdf. Accessed 29 September 2016.
8. Flandmark CE, Lilja K, Woehling H, Järvholm K. Switching from originator to biosimilar human growth hormone using dialogue teamwork: single-center experience from sweden. Biol Ther. 2013;3(1):35-43.
9. Kang YS, Moon HH, Lee SE, Lim YJ, Kang HW. Clinical experience of the use of CT-P13, a biosimilar to infliximab in patients with inflammatory bowel disease: a case series. Dig Dis Sci. 2015;60(4):951-956.
10. Yoo DH, Prodanovic N, Jaworski J, et al. Efficacy and safety of CT-P13 (biosimilar infliximab) in patients with rheumatoid arthritis: comparison between switching from reference infliximab to CT-P13 and continuing CT-P13 in the planeta extension study. Ann Rheum Dis. 2017;76(2):355-363.
11. Jørgensen KK, Olsen IC, Goll GL, et al. Switching from originator infliximab to biosimilar CT-P13 compared with maintained treatment with originator infliximab (NOR-SWITCH): a 52-week, randomised, double-blind, non-inferiority trial. Lancet. 2017;389(10086):2304-2316.
12. Faccin F, Tebbey P, Alexander E, Wang X, Cui L, Albuquerque T. The design of clinical trials to support the switching and alternation of biosimilars. Expert Opin Biol Ther. 2016;16(12):1445-1453.
13. Chow SC, Song F, Cui C. On hybrid parallel-crossover designs for assessing drug interchangeability of biosimilar products. J Biopharm Stat. 2017;27(2):265-271.
14. Chow SC, Yang LY, Starr A, Chiu ST. Statistical methods for assessing interchangeability of biosimilars. Stat Med. 2013;32(3):442-448.
15. Chow SC, Xu HL, Endrenyi L, Song FY. A new scaled criterion for drug interchangeability. Chin J Pharm Anal. 2015;35(5):844-848.
16. Li J, Chow SC. Statistical evaluation of the scaled criterion for drug interchangeability. J Biopharm Stat. 2017;27(2):282-292.
17. Zheng J, Chow SC, Song F. On safety margin for drug interchangeability. J Biopharm Stat. 2017;27(2):293-307.
18. Belleli R, Fisch R, Renard D, Woehling H, Gsteiger S. Assessing switchability for biosimilar products: modelling approaches applied to children’s growth. Pharm Stat. 2015;14(4):341-349.
19. Griffiths CEM, Thaçi D, Gerdes S, et al. The EGALITY study: a confirmatory, randomized, double-blind study comparing the efficacy, safety and immunogenicity of GP2015, a proposed etanercept biosimilar, vs. the originator product in patients with moderate-to-severe chronic plaque-type psoriasis. Br J Dermatol. 2017;176(4):928-938.
20. Anderson S, Hauck WW. Consideration of individual bioequivalence. J Pharmacokinet Biopharm. 1990;18(3):259-273.
21. Guideline on the choice of the non-inferiority margin. 2005 Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf. Accessed 30 March 2017.
22. Pinheiro J, Bates D. Mixed-Effects Models in S and S-PLUS. New York: Springer; 2006.
23. Jiang J. REML estimation: asymptotic behavior and related topics. Ann Stat. 1996;24(1):255-286.
24. Imhof JP. Computing the distribution of quadratic forms in normal variables. Biometrika. 1961;48(3/4):419-426.
25. Massey Jr. FJ. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951;46(253):68-78.
26. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna; 2015.
APPENDIX A: DERIVATION OF THE DISTRIBUTION OF THE TEST STATISTIC FOR THE ESTIMATION METHOD

As stated in Section 4.1, it is necessary to derive the variance-covariance matrix $\Sigma_1$ of the vector of the estimated switching effects, $\hat{c} = (\hat{c}_1, \hat{c}_2, -\hat{c}_1, -\hat{c}_2)^T$. The estimates for the switching effects $c_1$ and $c_2$ are asymptotically normally distributed\(^{23}\) with

$$(\hat{c}_1, \hat{c}_2)^T \sim N(\mu, \Sigma),$$

where $\mu$ is given by $(c_1, c_2)^T$. Then, the distribution of

$$(\hat{c}_1, \hat{c}_2, -\hat{c}_1, -\hat{c}_2)^T$$

is a multivariate normal distribution with mean

$$\mu_1 = (c_1, c_2, -c_1, -c_2)'$$

and covariance matrix

$$\Sigma_1 = \begin{pmatrix} \Sigma & -\Sigma \\ -\Sigma & \Sigma \end{pmatrix}.$$

The variance-covariance matrix $\Sigma$ can be calculated using known formulas for the variances of the estimates in a linear mixed-effects model. These are, for example, stated by Jacqmin-Gadda et al.\(^{33}\) There, the model is split into a part with fixed-effects and a part with random-effects, ie,

$$Y = X\beta + Z\alpha + \epsilon,$$

where $X$ denotes the fixed-effects design matrix and $Z$ is the random-effects design matrix. The vector $\beta$ is fixed, and $\alpha$ follows a multivariate normal distribution with mean 0 and covariance matrix $G$. Then, $Y$ follows a multivariate normal distribution with mean $X\beta$ and variance

$$V = ZGZ^T + \Sigma_e,$$

where $\Sigma_e$ is the covariance matrix of the residual vector $\epsilon$ (a diagonal matrix with $\sigma^2_e$ as entries). The variance of the estimates for the fixed-effects $\beta$ is then given by
\[
\text{Var}(\hat{\beta}) = (X^TV^{-1}X)^{-1}.
\]

Therefore, it is only necessary to state the design matrix \(X\) of the fixed-effects, the design matrix \(Z\) of the random-effects, the covariance matrix of the residual errors \(\Sigma_e\), and the covariance matrix of the random-effects \(G\). The residual error variance and the covariance matrix of the random-effects are in practice not known beforehand, and therefore, the true value will be replaced by the estimates in practice (see Section 4). As the switching effects are only a subset of the vector of the fixed-effects \(\beta\), the vector of the switching effects, \((\hat{c}_1, \hat{c}_2)^T\), can be determined by applying a linear transformation \(L\) to \(\hat{\beta}\) such that \((\hat{c}_1, \hat{c}_2)^T = L\hat{\beta}\). Then, the covariance matrix \(\Sigma\) of \((\hat{c}_1, \hat{c}_2)^T\) is given by

\[
\Sigma = L(X^TV^{-1}X)^{-1}L^T.
\]

With this knowledge, it is easy to set up the covariance matrix for \((\hat{c}_1, \hat{c}_2, -\hat{c}_1, -\hat{c}_2)^T\), which is given by

\[
\Sigma_1 = \begin{pmatrix}
\Sigma & -\Sigma \\
-\Sigma & \Sigma
\end{pmatrix}.
\]

**APPENDIX B: DERIVATION OF THE DISTRIBUTION OF THE DIFFERENCE BETWEEN OBSERVATION AND PREDICTION**

The vector of the differences between the observations and the predictions, \(\hat{y}\), follows a multivariate normal distribution with mean \(\mu_2\) and variance-covariance matrix \(\Sigma_2\) because both the observed values and the predicted values follow a multivariate normal distribution. For facilitating the derivation, we define a new vector \(y_M\), which consists of all observations in the modelling dataset \(M\). Let \(y_E\) be the vector with all observations in the evaluation dataset \(E\). For the variance-covariance matrix, the following equation holds true:

\[
\Sigma_2 = \text{Cov}(\hat{y}) = \text{Cov}(y_E - \hat{y}_E) = \text{Cov}(\hat{y}_E) + \text{Cov}(y_E) - \text{Cov}(y_E, y_E) - \text{Cov}(\hat{y}_E, y_E) = \text{Cov}(\hat{y}_E) + \text{Cov}(y_E) - \text{Cov}(y_E, \hat{y}_E) - \text{Cov}(y_E, y_E)^T.
\]

To derive the distribution of the differences between the observed and predicted values, it is therefore necessary to derive the mean value and the variance-covariance matrix of \(y_E\) and \(\hat{y}_E\) and the term \(\text{Cov}(\hat{y}_E, y_E)\). First, we focus on the distribution of the predictions. It is known that the distribution of the vector of the predictions, \(\hat{y}_E\), follows a multivariate normal distribution with expected value \(p_k + t_{a(i,j,k)}\) for prediction \(\hat{y}_{E,i,j,k}\) and variance-covariance matrix

\[
Z_e(Z_{m}^T Z_{m})^{-1} Z_{m}^T \text{Var}(y_M)(Z_e(Z_{m}^T Z_{m})^{-1} Z_{m}^T)^T,
\]

where \(Z_m\) is the design matrix of the modelling dataset and \(Z_e\) is the design matrix of the evaluation dataset.\(^{34}\) It is important to note that the entries in \(y_M\) are independent because only 1 observation per person was used. Therefore, the variance-covariance matrix \(\text{Var}(y_M)\) is a diagonal matrix with entries \(\sigma_e^2 + \sigma_s^2\), where \(\sigma_e^2\) is the variance of the residuals and \(\sigma_s^2\) is the variance of the subject effect. In total, the variance-covariance matrix of \(\hat{y}_E\) is given by

\[
V_{\hat{y}_E} = Z_e(Z_{m}^T Z_{m})^{-1} Z_{m}^T \text{Var}(y_M)(Z_e(Z_{m}^T Z_{m})^{-1} Z_{m}^T)^T,
\]

where \(I_{(a,b)}\) denotes the identity matrix with \(a\) rows and \(b\) columns. The observations \(y_E\) follow, according to the assumed underlying model, a multivariate normal distribution with an expectation for subject \(i\) in sequence \(j\) and period \(k\) of

\[
p_k + t_{a(i,j,k)} + t_{a(i,j,k-1),a(i,j,k)}
\]

and variance-covariance matrix \(V_{y_E}\) with the entries \(v_{l,k}\):

\[
v_{l,k} = \begin{cases}
\sigma_e^2 + \sigma_s^2 & \text{if } l = k, \\
\sigma_s^2 & \text{if } |l - k| \leq 5, l \neq k, \\
0 & \text{else}.
\end{cases}
\]

This means that the covariance between observations from the same subject (5 observations per subject) is \(\sigma_s^2\) and 0 between different subjects.
Last, it is necessary to derive $\text{Cov}(y_E, \hat{y}_E)$. The predictions and observations are not independent because, within one subject, they share the same subject effect. Without loss of generality, we assume that the first entry in $y_M$ and the first 5 observations in $y_E$ belong to the first subject. The second subject has the second observation in $y_M$ and observations 6 to 10 in $y_E$ and so on. We define the matrix

$$W := (w_{1T}, \ldots, w_{4nT}, \ldots) := \begin{pmatrix} W_{1,1} & \cdots & W_{4n,1} \\ w_{1,20n} & \cdots & w_{4n,20n} \end{pmatrix} := Z_s(Z_m^T Z_m)^{-1} Z_m \in \mathbb{R}^{20n \times 4n},$$

where $w_i$ is a vector of length $20n$ with $w_i = (w_{i,1}, \ldots, w_{i,20n})$. The prediction are given by

$$\hat{y}_E = Z_s(Z_m^T Z_m)^{-1} Z_m y_M = \begin{pmatrix} \sum_{i=1}^{4n} w_{1,1} y_{M,i} \\ \vdots \\ \sum_{i=1}^{4n} w_{1,20n} y_{M,i} \end{pmatrix}.$$ 

Before stating the general result, we would like to illustrate the derivation of the variance-covariance matrix using an example. We calculate the covariance between $y_{E,1}$ and $\hat{y}_{E,1}$, which is given by

$$\text{Cov}(y_{E,1}, \hat{y}_{E,1}) = \text{Cov} \left( y_{E,1}, \sum_{i=1}^{4n} w_{1,1} y_{M,i} \right) = \text{Cov}(y_{E,1}, w_{1,1} y_{M,1}) = w_{1,1} \sigma_s^2,$$

where the second equality holds true because $y_{E,1}$ is only correlated to $y_{M,1}$ (same subject effect), but independent of all other observations in the modelling dataset. Therefore, the only non-zero covariance is the one between $y_{E,1}$ and $y_{E,1}$. In total, this leads to the following variance-covariance between $y_E$ and $\hat{y}_E$:

$$\text{Cov}(y_E, \hat{y}_E) = C = \begin{pmatrix} \text{Cov}(y_{E,1}, \hat{y}_{E,1}) & \cdots & \text{Cov}(y_{E,1}, \hat{y}_{E,4n}) \\ \vdots & \ddots & \vdots \\ \text{Cov}(y_{E,4n}, \hat{y}_{E,1}) & \cdots & \text{Cov}(y_{E,4n}, \hat{y}_{E,4n}) \end{pmatrix} = \sigma_s^2 \begin{pmatrix} w_{1,1} \\ \vdots \\ w_{1,1} \\ w_{2,1} \\ \vdots \\ w_{4n,1} \end{pmatrix} \in \mathbb{R}^{20n \times 20n},$$

where $w_{i,1}$ is repeated 5 times because each subject appears 5 times in the evaluation dataset. The variance-covariance matrix of the vector of differences between observations and predictions is given by

$$\Sigma_2 := \text{Cov}(y_E - \hat{y}_E) = \text{Cov}(\hat{y}_E) + \text{Cov}(y_E) - \text{Cov}(y_E, \hat{y}_E) - \text{Cov}(\hat{y}_E, y_E)$$

$$= V_{\hat{y}_E} + V_{y_E} - C - C^T.$$