Received: 25 August 2020 | Revised: 26 August 2021 | Accepted: 10 December 2021

DOI: 10.1111/stan.12260

ORIGINAL ARTICLE

WILEY

Prediction intervals for all of M future observations based on linear random effects models

Max Menssen | **Frank Schaarschmidt**

Department for Biostatistics, Leibniz Universität, Hannover, Lower Saxony, Germany

Correspondence
Max Menssen, Department for Biostatistics, Leibniz Universität Hannover, Herrenhäuser Strasse 2, 30419 Hannover, Lower Saxony, Germany.
Email: menssen@cell.uni-hannover.de

In many pharmaceutical and biomedical applications such as assay validation, assessment of historical control data, or the detection of anti-drug antibodies, the calculation and interpretation of prediction intervals (PI) is of interest. The present study provides two novel methods for the calculation of prediction intervals based on linear random effects models and restricted maximum likelihood (REML) estimation. Unlike other REML-based PI found in the literature, both intervals reflect the uncertainty related with the estimation of the prediction variance. The first PI is based on Satterthwaite approximation. For the other PI, a bootstrap calibration approach that we will call *quantile-calibration* was used. Due to the calibration process this PI can be easily computed for more than one future observation and based on balanced and unbalanced data as well. In order to compare the coverage probabilities of the proposed PI with those of four intervals found in the literature, Monte Carlo simulations were run for two relatively complex random effects models and a broad range of parameter settings. The quantile-calibrated PI was implemented in the statistical software R and is available in the predint package.

KEYWORDS

anti-drug antibody, assay qualification, bootstrap calibration, historical control data, Satterthwaite approximation

1 | INTRODUCTION

Prediction intervals (PI) are statistical intervals that are computed based on an observed sample in order to contain one ore more future observations with a given degree of confidence. Usually, it is assumed that the observed sample as well as the future observation(s) descent from the same data-generating process. Hahn and Meeker (1991) and Hahn, Meeker and Escobar (2017) give a detailed review about methods for the computation of different PI based on one sample in which the observations vary around the mean. These different PI should contain either one future observation, the future mean, all of $M \geq 1$ future observations or K out of M future observations.

PI can be applied to several statistical problems and are of use in many scientific fields. In the context of pharmaceutical applications, Francq, Lin and Hoyer (2020) used PI for assay qualification. More examples for the usage of PI for process validation are given by Hahn and Meeker (1991) or in the context of gauge repeatability and reproducability experiments (Lin & Liao, 2008). Also in preclinical statistics and toxicology, PI can be useful. In this field of research, the verification of an actual control group by the use of historical control data (HCD) is heavily discussed (Elmore & Peddada, 2009; Greim, Gelbke, Reuter, Thielmann, & Edler, 2003). Nevertheless, the methods proposed for that purpose (e.g. historical range or historical mean plus minus *SD*) are rather naive and many authors are not aware, that PI for one or more future observations (depending on the purpose) can be applied to that problem. Therefore, Menssen and Schaarschmidt (2019) proposed the use of PI on HCD that is assumed to be overdispersed binomial. Nevertheless, the literature lacks methods for the computation and application of PI to other models in that research area. Another field of application occurs in early phases of drug development such as the detection of anti-drug antibodies (ADA) (Hoffman & Berger, 2011; Jaki, Allacher, & Horling, 2016). In such a bioassay, the antibody reaction is evaluated for a set of nonresponders as well as for patients with unclear status. Following Schaarschmidt, Hofmann, Jaki, Gruen, and Hothorn (2015) upper prediction limits can be computed for a sample of putative nonresponders in order to compare this limit with the outcome of the patients with unclear status. If the ADA-reaction for such a patient falls above the limit, the patient might have developed anti-drug antibodies (Hoffman & Berger, 2011).

For all the applications mentioned above, the sampling is usually done based on several factors that may influence the outcome of the study (e.g. many patients are analyzed by different experimenters in different hospitals). Since, in such applications inference is made on the level of the observations, rather than for the factors influencing them, a natural approach is the calculation of PI based on random effects models (Francq et al., 2020; Hoffman & Berger, 2011; Schaarschmidt et al., 2015). The idea of the computation of PI based on random effects models dates back to 1941. In that year Satterthwaite (1941) gave an example how to calculate "confidence limits within which we may expect an additional item" based on a one-way random effects model.

Since then, several authors worked on PI based on random effects models, but mainly focused on special cases or balanced models that are too simple for many practical applications (Jeske & Harville, 1988; Lin & Liao, 2008; Wang, 1992). A research area in which the use of PI based on complex random effects models is proposed is plant breeding. Anyhow, in this area random effects predicted by the best linear unbiased predictions are of interest (Al-Sarraj, von Brömssen, & Forkmann, 2019; Forkmann & Piepho, 2013), rather than the prediction of one or more future observations.

In the context of random effects models, PI can be computed based on mean squares (MSQ), based on generalized pivotal quantities (GPQ) or based on parameter estimates that are estimated via restricted maximum likelihood (REML). Since the estimation of variance components based on MSQ was already utilized by Satterthwaite (1941), it is the standard method to which almost all intervals that are based on more advanced methods are compared with. GPQ-based methods for the calculation of PI for $M \geq 1$ future observations were proposed by Lin and Liao (2008) for balanced data. Up to now, REML-based PI got less attention. Al-Sarraj et al. (2019) used a PI for which the variance components were estimated via REML but treated as known, following the approach of Pawitan (2001) by using a standard normal quantile. Francq, Lin, and Hoyer (2019) proposed a REML-based prediction interval for one future observation ($M = 1$) that is applicable to balanced and unbalanced data as well. However, this interval accounts only for the uncertainty of the estimated variance of the historical data but not for the prediction variance (variance of the historical data plus variance for the mean) that is used for the calculation of the corresponding PI (details are given below).

In the following sections, a REML-based approach that takes the uncertainty of the prediction into account is proposed and used for the calculation for PI for one future observation ($M = 1$). For this purpose, the degrees of freedom were approximated using the Generalized Satterthwaite method following van den Heuvel (2010). Furthermore, a bootstrap calibrated prediction interval for all of $M \geq 1$ future observations is proposed. This interval can be applied to balanced and unbalanced data as well. The coverage probabilities for the two proposed intervals, as well as for the PI of Satterthwaite (1941), Lin and Liao (2008) and Francq et al. (2019) are simulated based on two relatively complex random effects models (two-way cross-classified with interaction and two-way hierarchical) compared to the simple one-way model other simulations are based on (Lin & Liao, 2008) . Furthermore, a detailed overview about the experimental designs that occur in the research areas mentioned above is given and the PI were applied to real-life data. A user-friendly implementation of the bootstrap calibrated PI is provided by the R-package predint (Menssen, 2021).

2 | REAL-LIFE DATA

Random effects models can be applied to a wide range of experimental designs. Hence, many different designs are reported in the literature regarding assay qualification, early phase drug development such as ADA detection or the usage of HCD. For validation, a bioassay might be carried out by several experimenters on different days using samples obtained from different individuals resulting in cross-classified or hierarchical designs (Francq et al., 2020). For ADA cut point estimation, samples of several individuals may be processed by different experimenters on different plates on several days, resulting in designs that range from a simple one-way layout to complex designs with some random factors crossed and some nested (Hoffman & Berger, 2011; Jaki et al., 2016; Shen & Dai, 2021; Zhang, Zhang, Kubiak, & Yang, 2013). Data about historical controls regarding rats and mice obtained from long-time carcinogenicity studies are provided on the homepage of the National Toxicology Program (NTP, 2021). Since the compound of interest can be applied by using several different pathways and studies are carried out by several laboratories, HCD can be either cross-classified or hierarchical. Contrary to data obtained from assay qualification or used for ADA cut point estimation, HCD data can be heavily unbalanced, since different studies in which different pathways might be used are carried out over the years by different laboratories.

2.1 | Motivating examples

2.1.1 | ADA cut point estimation

In the context of ADA cut point estimation, Hoffman and Berger (2011) published a data set resulting from an electroluminescence assay in which blood plasma of 20 drug-naive mice were analyzed in three different experimental runs. In each run each plasma sample was duplicated. Hence a natural approach for modeling would be a cross-classified random effects model with an interaction term between the runs and the mice. However, since the duplicates are averaged in the reported dataset, only a cross-classified model without an interaction term can be fit to the data. Since this dataset is balanced, it will be used in order to demonstrate the calculation of PI using all six methods described below.

2.1.2 | Historical control data abouth the maximum mean weekly body weight of female mice

A dataset containing HCD about the maximum mean weekly body weight (mmwbw) of female mice (strain B6C3F1) is given in Table 1. It contains the reported mmwbw from NTP Historical Controls Reports between 2016 and 2021 (NTP, 2021) for two laboratories (Battelle Northwest and Battelle Columbus) and six pathways. Since the pathway inhalation air was used by the Battelle Northwest laboratory only and the five remaining pathways were utilized by Battelle Columbus, PI have to be calculated based on a model with pathways nested in the laboratories. Two more studies using the same strain of mice were carried out by the IIT and the Southern Research Institute for which the mmwbw of females is given in Table 2. The outcome of these two further control groups should be validated simultaneously by the data obtained from Battele Northwest and Batelle Columbus.

TABLE 1 Historical control data for female B6C3F1 mice

Study number	Laboratory	Pathway	Maximum mean weekly body weight
52060104	Battelle Northwest	inhalation_air	54.10
52072504	Battelle Northwest	inhalation_air	51.00
52051504	Battelle Northwest	inhalation_air	59.20
52052304	Battelle Northwest	inhalation_air	57.90
51047204	Battelle Northwest	inhalation_air	55.90
56031106	Battelle Northwest	inhalation_air	54.30
52000604	Battelle Columbus	gavage_corn oil	66.60
52032004	Battelle Columbus	gavage_corn oil	66.50
51098702	Battelle Columbus	oral_feed	51.50
51026002	Battelle Columbus	oral_feed	54.70
52071204	Battelle Columbus	oral_feed	51.90
50005804	Battelle Columbus	gavage_methylcellulose	58.80
52032306	Battelle Columbus	gavage_methylcellulose	58.80
52020304	Battelle Columbus	gavage_water	62.90
50303804	Battelle Columbus	oral_water	61.60
59601406	Battelle Columbus	oral_water	63.50

TABLE 2 Actual control data for female B6C3F1 mice

Study_number	Laboratory	Pathway	Maximum mean weekly body weight
52010578	IIT Research Institute	wbe_air	62.60
52020904	Southern Research Institute	gavage_corn oil	57.70

3 | METHODS

3.1 | Random effects models and PI

A general linear random effects model is given by

$$\boldsymbol{Y} = \boldsymbol{1}\mu + \boldsymbol{Z}\boldsymbol{U} + \epsilon,$$

where $\boldsymbol{Y} = (Y_1, \dots, Y_N)^T$ is the vector of random variables that represents N individual observations. The overall mean is represented by μ. $\boldsymbol{U}$ is a stacked vector containing random effects subvectors $\boldsymbol{U}_c$. In this notation, each $\boldsymbol{U}_c$ consists of all levels of a single random factor occurring in the data. Hence, the index $c = 1, \dots, C$ indicates the random factors by which the observations should be modeled (e.g., a main effects factor, an interaction term or a nested factor). The number of elements of a given random effects vector U_c is denoted by q_c. Hence, the total length of $\boldsymbol{U}$ is $q_{\text{total}} = \sum_{c=1}^{C} q_c$. $\boldsymbol{Z}$ is a design matrix and has the dimensions $Nx\, q_{\text{total}}$. The vector $\boldsymbol{\epsilon}$ represents the random errors associated with the N observations. The individual random effects can be represented as $\boldsymbol{Z}_c\boldsymbol{U}_c$ such that

$$\boldsymbol{ZU} = \begin{pmatrix} \boldsymbol{Z}_1 & \dots & \boldsymbol{Z}_C \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_1 \\ \vdots \\ \boldsymbol{U}_C \end{pmatrix} = \sum_{c=1}^{C} \boldsymbol{Z}_c\boldsymbol{U}_c,$$

with each

$$\boldsymbol{U}_c = \begin{pmatrix} U_{c,1} \\ \vdots \\ U_{c,q_c} \end{pmatrix}.$$

Each of the $\boldsymbol{U}_c$ random effects is considered to be normal distributed with $\boldsymbol{U}_c \sim N(\boldsymbol{0}_{q_c}, \boldsymbol{I}_{q_c}\sigma_c^2)$ as well as the error term $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}_N, \boldsymbol{I}_N\sigma_{C+1}^2)$ with $\boldsymbol{I}$ as an identity matrix of order q_c or N, respectively. Furthermore it is assumed that

$$\begin{gathered} \text{cov}(\mu, \boldsymbol{U}_{c,q_c}) = 0 \quad \forall \quad c = 1, \dots, C+1 \\ \text{cov}(\boldsymbol{U}_{c,q_c}, \boldsymbol{U}_{c',q_{c'}}) = 0 \quad \forall \quad c = 1, \dots, C+1, \;\; c' = 1, \dots C+1 \,:\, c \neq c', \end{gathered} \tag{1}$$

and the variance-covariance matrix of the observations is given by

$$\text{var}(\boldsymbol{Y}) = \sum_{c=1}^{C} \boldsymbol{Z}_c\boldsymbol{Z}_c^T\sigma_c^2 + \boldsymbol{I}_N\sigma_{C+1}^2,$$

with $\boldsymbol{I}_N$ as an identity matrix of order N. Further information on the model described above can be found in McCullagh and Searle (2001, pp. 156–160) or in Searle, Casella, and McCulloch (2006, pp. 233–257).

For prediction, it is assumed that the future random variable $\boldsymbol{Y}^*$ which is comprised of $M \geq 1$ observations and its historical counterpart $\boldsymbol{Y}$ are independent from each other, but descend from the same random process. Hence, the error margin of the prediction is

$$\boldsymbol{D} = \boldsymbol{Y}^* - \boldsymbol{1}\mu \sim N(\boldsymbol{0}, \text{var}(\boldsymbol{D})), \tag{2}$$

which implies that

$$\text{var}(\boldsymbol{D}) = \text{var}(\boldsymbol{Y}^* - \boldsymbol{1}\mu) = \text{var}(\boldsymbol{Y}^*), \tag{3}$$

with

$$\text{var}(\boldsymbol{Y}^*) = \sum_{c=1}^{C} \boldsymbol{Z}_c^*\boldsymbol{Z}_c^{*T}\sigma_c^2 + \boldsymbol{I}_M\sigma_{C+1}^2. \tag{4}$$

Please note that in the univariate case of $M = 1$, Equation (4) simplifies to $\text{var}(Y^*) = \sum_{c=1}^{C+1} \sigma_c^2$.

Based on observed historical data $\boldsymbol{y}$ and the fitted model

$$\boldsymbol{y} = \boldsymbol{1}\hat{\mu} + \boldsymbol{Z}\hat{\boldsymbol{u}} + \hat{\boldsymbol{e}},$$

the estimate for the prediction variance becomes

$$\widehat{\text{var}}(\boldsymbol{D}) = \widehat{\text{var}}(\boldsymbol{Y}^* - \boldsymbol{1}\hat{\mu}) = \widehat{\text{var}}(\boldsymbol{Y}^*) + \widehat{\text{var}}(\boldsymbol{1}\hat{\mu}), \tag{5}$$

with $\widehat{\text{var}}(\boldsymbol{D})$ being a square matrix of order M. Please note that Equation (5) does not consider the covariance between the mean and the variance components since Equation (1) implies that

$$\text{cov}(\hat{\mu}, \hat{\sigma}_c^2) = 0 \quad \forall \quad c = 1, \dots, C+1. \tag{6}$$

Thus, this is the standard assumption on which all methods given below rely.

A prediction interval for $M \geq 1$ future observations $\boldsymbol{y}_M^*$ with coverage probability $\Psi = P(L \leq \boldsymbol{y}_M^* \leq U) = 1 - \alpha$ is given by

$$[L, U] = \hat{\mu} \pm t_{1-\frac{\alpha}{2},\ df,\ \widehat{\text{var}}(\boldsymbol{D})}, \tag{7}$$

where $t_{1-\frac{\alpha}{2},\ df,\ \widehat{\text{var}}(\boldsymbol{D})}$ is a quantile of the multivariate t-distribution with df degrees of freedom and $\widehat{\text{var}}(\boldsymbol{D})$ as the estimated variance-covariance matrix for the prediction error. Please note that Equation (7) represents a general form for the calculation of a prediction interval for $M \geq 1$ future observations, which in the univariate case $M = 1$ simplifies to

$$[L, U] = \hat{\mu} \pm t_{1-\frac{\alpha}{2}, df} \sqrt{\widehat{\text{var}}(\hat{\mu}) + \sum_{c=1}^{C+1} \hat{\sigma}_c^2}.$$

with $\widehat{\text{var}}(\hat{\mu}) + \sum_{c=1}^{C+1} \hat{\sigma}_c^2 = \widehat{\text{var}}(D)$. Hence $\widehat{\text{var}}(\boldsymbol{D})$ denotes the variance-covariance matrix that is associated with $M > 1$ future observations and $\widehat{\text{var}}(D)$ represents the prediction variance if a PI for $M = 1$ future observation is calculated.

3.2 | Calculation of PI

3.2.1 | PI for $M = 1$ future observation based on MSQ

The estimation of PI for $M = 1$ one future observation based on MSQ was firstly described in 1941 by Satterthwaite (1941). Assuming a balanced design, $\text{var}(D)$ is estimated by $\widehat{\text{var}}(D)^{\text{Sat}} = \sum_{c=1}^{C+1} \omega_c^{\text{Sat}} MS_c^{\text{Sat}}$ and the prediction interval is given by

$$[L, U]^{\text{Sat}} = \bar{y} \pm t_{1-\frac{\alpha}{2}, df^{\text{Sat}}} \sqrt{\widehat{\text{var}}(D)^{\text{Sat}}},$$

with $\bar{y}$ as the arithmetic mean of $\boldsymbol{y}$. In this approach $t_{1-\frac{\alpha}{2},df^{\text{Sat}}}$ is the $1-\frac{\alpha}{2}$ quantile from the t distribution with approximate degrees of freedom

$$df^{\text{Sat}} = \frac{\left(\sum_{c=1}^{C+1} \omega_c^{\text{Sat}} MS_c^{\text{Sat}}\right)^2}{\sum_{c=1}^{C+1} \frac{(\omega_c^{\text{Sat}} MS_c^{\text{Sat}})^2}{df_c}},$$

and df_c as the individual degrees of freedom according to the $c = 1, \dots, C+1$ random effects.

Formulas for the calculation of weights ω_c^{Sat}, MSQ MS_c^{Sat} and individual degrees of freedom df_c are given in Tables 3 and 4 for a hierarchical as well as for a cross-classified design. In the following sections, especially in Figures 1 and 2 this interval will be referred to as Satterthwaite 1941.

3.2.2 | PI for $M = 1$ future observation based on REML

This method is based on parameter estimates that are estimated using the REML approach. Generally, the degrees of freedom associated with variance components estimated via REML can be approximated by using the Generalized Satterthwaite method (Schuetzenmeister & Dufey, 2019; van den Heuvel, 2010) which is based on the estimated variance component $\hat{\sigma}_c^2$ as well as on its estimated standard error $\widehat{SE}(\hat{\sigma}_c^2)$. The individual degrees of freedom can be approximated as

$$df^{\hat{\sigma}_c^2} = 2\left(\frac{\hat{\sigma}_c^2}{\widehat{SE}(\hat{\sigma}_c^2)}\right)^2 = 2\frac{\hat{\sigma}_c^4}{\widehat{\text{var}}(\hat{\sigma}_c^2)}. \tag{8}$$

For linear mixed models as well as for random effects models the estimates used in Equation (8) can be obtained by using the R package VCA (Schuetzenmeister & Dufey, 2019). This package provides degrees of freedom and standard errors for the individual variance components $\hat{\sigma}_c^2$ and their sum.

Recently, Francq et al. (2019) proposed a REML based PI for $M = 1$ future observation, that was applied in an assay qualification study (Francq et al., 2020). For this interval, the degrees of freedom are approximated based on the Generalized Satterthwaite method and hence, it is applicable to balanced and unbalanced data as well their. In the following sections, especially in Figures 1 and 2 this PI will be referred to as Francq et al. 2019. However, for this interval, Francq et al. (2019) approximated the degrees of freedom based on $\widehat{\text{var}}(Y^*) = \sum_{c=1}^{C+1} \hat{\sigma}_c^2$, rather than on the prediction variance $\widehat{\text{var}}(D) = \widehat{\text{var}}(\mu) + \widehat{\text{var}}(Y^*)$. Hence, the degrees of freedom used for interval calculation are

$$df^{\widehat{\text{var}}(Y^*)} = 2\frac{\left(\sum_{c=1}^{C+1} \hat{\sigma}_c^2\right)^2}{\widehat{\text{var}}\left(\sum_{c=1}^{C+1} \hat{\sigma}_c^2\right)}.$$

Consequently, the interval of Francq et al. (2019) is given by

$$[L, U]^{Franq} = \hat{\mu} \pm t_{1-\frac{\alpha}{2},df^{\widehat{\text{var}}(Y^*)}} \sqrt{\widehat{\text{var}}(\hat{\mu}) + \widehat{\text{var}}(Y^*)}. \tag{9}$$

In order to account for the degrees of freedom associated with the whole prediction variance, the approximation given in equation 18 of van den Heuvel (2010) can be utilized and therefore, the corresponding PI will be called van den Heuvel (2010) in the results section. The approximation was originally published for the calculation of confidence intervals but can be easily applied for other purposes. For balanced designs the variance of the prediction can be calculated by

$$\widehat{\text{var}}(D) = \widehat{\text{var}}(\hat{\mu}) + \widehat{\text{var}}(Y^*) = \sum_{c=1}^{C+1} \omega_c^{\text{REML}} \hat{\sigma}_c^2$$

and the variance of the prediction variance can be estimated by

$$\widehat{\text{var}}[\widehat{\text{var}}(D)] = \sum_{c=1}^{C+1} (\omega_c^{\text{REML}})^2 \widehat{\text{var}}(\hat{\sigma}_c^2).$$

Then, the approximated degrees of freedom are

$$df^{Pred} = 2\frac{\widehat{\text{var}}(D)^2}{\widehat{\text{var}}[\widehat{\text{var}}(D)]}$$
$$df^{\widehat{\text{var}}(D)} = \max[1, \min(N-1, df^{\text{Pred}})]$$

with N as the total number of historical observations. The prediction interval is given by

$$[L, U]^{vdH} = \hat{\mu} \pm t_{1-\frac{\alpha}{2}, df^{\widehat{\text{var}}(D)}} \sqrt{\widehat{\text{var}}(D)}. \qquad (10)$$

Formulas for the weights ω_c^{REML} are given in Tables 3 and 4 for a hierarchical as well as for a cross-classified design.

The main difference between the two REML-based intervals mentioned above are the variance terms for which the df-approximation is done. Because the approximation used by Francq et al. (2019) is based on $\widehat{\text{var}}(Y^*)$ rather than on the whole prediction variance $\widehat{\text{var}}(D) = \widehat{\text{var}}(\mu) + \widehat{\text{var}}(Y^*)$, the degrees of freedom used for the calculation of $[L, U]^{\text{Franq}}$ are on average higher than the degrees of freedom on which $[L, U]^{vdH}$ is based on (see Figure A1). Consequently the PI of Francq et al. (2019) is expected to be less wide than the PI given in Equation (10) in most of the cases and hence, should yield lower coverage probabilities. This effect is strongest if a relatively large variance component has few replications for estimation and is therefore associated with small df_c, but decreases with an increase of the number of observations (and higher df) due to the convergence of the t-distribution against the standard normal distribution (see Figures 1 and 2).

3.2.3 | PI for $M \geq 1$ future observations based on GPQ

The theoretical background on which this interval is based on, is given by Lin and Liao (2008). Following Lin and Liao, the interval can be calculated for $M \geq 1$ future observations $\boldsymbol{y}^*$ based on balanced designs. Their approach grounds on the finding of a GPQ for the expected MSQ. Hence, the algorithm given below, makes use of the relationship between expected mean squares EMS_c and variance components σ_c^2 which is described in many textbooks regarding ANOVA methods such as Sahai and Ageel (2000).

Following Lin and Liao, a GPQ for the expected mean squares EMS_c is given by

$$\text{GPQ}(\text{EMS}_c) = \frac{s_c^2}{R_c}$$

with s_c^2 as the observed sum of squares and $R_c \sim \chi^2_{df_c}$. A GPQ-based prediction interval can be obtained using the following algorithm:

1. For each of the $C + 1$ random factors, sample $H = 10{,}000$ mutually independent realizations $R_{c,1}, \dots, R_{c,H}$ from the χ^2-distribution with degrees of freedom df_c.
2. Calculate $\text{GPQ}(\text{EMS}_c)_h = \frac{s_c^2}{R_{c,h}}$.
3. Calculate $\text{GPQ}(\sigma_c^2)_h$ based on $\text{GPQ}(\text{EMS}_c)_h$. The formulas used for this step depend on the experimental design. Examples are given in Sections 3.3.1 and 3.3.2.
4. Calculate GPQs for the variance-covariance matrix $GPQ(\text{var}(\boldsymbol{D}))_h$ by substituting $\text{GPQ}(\sigma_c^2)_h$ into $\widehat{\text{var}}(\boldsymbol{D})$. Please note that further formulas for the calculation of $\widehat{\text{var}}(\boldsymbol{D})$ are given below in Sections 3.3.1 and 3.3.2.
5. Based on $\text{GPQ}(\text{var}(\boldsymbol{D}))_h$, compute H qunatiles from the corresponding multivariate normal distributions, such that $q_h = z_{1-\alpha/2,\mathbf{0},\text{GPQ}(\text{var}(\boldsymbol{D}))_h}$.
6. Calculate $\text{GPQ}(\boldsymbol{D}) = \text{median}(q_h)$
7. The corresponding prediction interval is given by $[L, U]^{\text{GPQ\_}M1} = \bar{y} \pm \text{GPQ}(\boldsymbol{D})$

For both, a two-way-hierarchical and a two-way cross-classified model with interaction, $\text{GPQ}(\sigma_c^2)_h$ can be obtained if the EMS_c used in Equations (16) to (18) and (21) to (24) are substituted by $\text{GPQ}(\text{EMS}_c)_h$. $\text{GPQ}(\text{var}(\boldsymbol{D}))_h$ can be obtained if σ_c^2 is substituted by $\text{GPQ}(\sigma_c^2)_h$ in Equation (15) and Equation (20).

This approach, which Lin and Liao called Method 1, is based on the the calculation of $H = 10{,}000$ quantiles from the multivariate-normal distribution $z_{1-\alpha/2,\ \mathbf{0},\ GPQ(\text{var}(\boldsymbol{D}))_h}$. Since the calculation of a multivariate-normal quantile is computationally intensive (Genz & Bretz, 2009), this approach will take too much computing time to be useful in practical applications or Monte Carlo simulations (it took around 13 min on a MacBook Pro to calculate a PI for eight future observations based on a cross-classified model).

Hence Lin and Liao gave an alternative approach which was called Method 3 in their paper: Calculate step 1–4 as described above. Then, calculate the means for the elements of $\text{GPQ}(\text{var}(\boldsymbol{D}))_h$, such that

$$\text{GPQ}(\text{var}(\boldsymbol{D})) = \frac{\sum_{h=1}^{K} \text{GPQ}(\text{var}(\boldsymbol{D}))_h}{H}. \tag{11}$$

The corresponding prediction interval is given by

$$[L, U]^{\text{GPQ\_}M3} = \bar{y} \pm z_{1-\alpha/2,\mathbf{0},\text{GPQ}(\text{var}(\boldsymbol{D}))}),$$

treating $\text{GPQ}(\text{var}(\boldsymbol{D}))$ as known. Since the quantile of the multivariate-normal distribution has to be calculated only once, this approach reduces the computing time down to a manageable level. Anyhow, if a prediction interval for only $M = 1$ future observation is needed $\text{var}(\boldsymbol{D})$ reduces to $\text{var}(\mu) + \text{var}(Y^*)$. Hence, the corresponding quantile is drawn from a univariate normal distribution. This approach is far less computational intensive, such that in this special case both methods are applicable in Monte Carlo simulations.

3.2.4 | Quantile calibrated PI for $M \geq 1$ future observations

This method is also based on REML estimates, but the quantile used for the calculation of the PI is approximated by a bootstrap procedure. This idea is related to the idea of α-calibration (Efron & Tibshirani, 1993), but, instead of calibrating the α with which the interval is calculated, the whole quantile that is used for the calculation of the prediction interval is approximated. Hence, no assumption regarding a multivariate distribution or the variance-covariance matrix of the future observations is needed. Therefore, the quantile-calibrated PI can be easily calculated for more than one future observation and based on many different experimental layouts as well as for balanced and unbalanced data.

The first step of the quantile-calibration is to fit a random effects model to the initial dataset $\boldsymbol{y}$. Then, based on the estimated model parameters $b = 1, \ldots B$ new bootstrap datasets $\boldsymbol{y}^*_{\boldsymbol{b}}$ of same sample size and structure as the original data set are drawn. Then, $m = 1, \ldots, M$ observations per bootstrap data set are randomly sampled from $\boldsymbol{y}^*_{\boldsymbol{b}}$ without replacement, resulting in a reduced set $\boldsymbol{y}^*_{\boldsymbol{bm}}$. From this M sampled future observations the minimum and the maximum

$$\min{}^*_b = \min(\boldsymbol{y}^*_{\boldsymbol{bm}})$$

$$\max{}^*_b = \max(\boldsymbol{y}^*_{\boldsymbol{bm}}),$$

will serve for the calibration in the further steps.

Then, draw B further bootstrap samples $\boldsymbol{y}^{**}_{\boldsymbol{b}}$. Fit the initial model to $\boldsymbol{y}^{**}_{\boldsymbol{b}}$ in order to obtain estimates for the variance components $\hat{\sigma}^2_{bc}$ as well as for the variance of the estimated mean $\widehat{\text{var}}(\hat{\mu}_b)$.

The second step is the calibration conditionally on $\min^*_b$ and $\max^*_b$ in order to find the coefficient λ^{calib} that results in an interval with coverage probability as close as possible to the nominal $\Psi = 1 - \alpha$. For that purpose, a bisection algorithm is used, that minimizes the distance between the observed coverage probability $\hat{\Psi}_g$ and Ψ based on $g = 1, \ldots, G$ calibration values λ_g. The bisection is stopped if the observed coverage probability falls into a tolerable area around the nominal coverage probability $\Psi \pm s$ such that $|\Psi - \hat{\Psi}_g| \leq s$ and the corresponding λ_g is set to be λ^{calib} and hence used for the calculation of the interval.

In each of the G bisection steps, the PI is calculated for each of the B bootstrap samples such that

$$\left[l_{bg}, u_{bg}\right] = \hat{\mu}_g \pm \lambda_g \sqrt{\widehat{\text{var}}(\hat{\mu}_b) + \widehat{\text{var}}(\boldsymbol{y}^{**}_{\boldsymbol{b}})}.$$

The coverage probability of the particular λ_g based intervals is estimated to be

$$\hat{\psi}_g = \frac{\sum_{b=1}^{B} I_{bg}}{B}, \text{ with}$$
$$I_{bg} = 1 \text{ if } (l_{bg} \leq \min_b^* \text{ and } \max_b^* \leq u_{bg})$$
$$I_{bg} = 0 \text{ if } (l_{bg} > \min_b^* \text{ or } \max_b^* > u_{bg}).$$

The algorithm starts by defining the start values λ_1 and λ_2 in a way that the corresponding $\hat{\Psi}_1$ is smaller than the nominal $\Psi = (1 - \alpha)$ (due to a small λ_1) and the corresponding $\hat{\Psi}_2$ is greater than Ψ (due to a high λ_2). Then the midpoint of the search interval is

$$\lambda_3 = \frac{\lambda_1 + \lambda_2}{2}. \tag{12}$$

and the coverage probability $\hat{\Psi}_3$ is calculated based on λ_3. If $\Psi - \hat{\Psi}_3$ is positive, λ_4 is calculated by replacing λ_1 in Equation (12) by λ_3 such that

$$\lambda_4 = \frac{\lambda_2 + \lambda_3}{2}.$$

If $\Psi - \Psi_3$ is negative, λ_4 is calculated by replacing λ_2 in Equation (12) by λ_3 such that

$$\lambda_4 = \frac{\lambda_1 + \lambda_3}{2}.$$

This iteration process is run until $|\Psi - \Psi_g| \leq s$ and the corresponding λ_g is set to be λ^{calib}.

The last step is the calculation of the quantile-calibrated interval based on the estimates of the initial model together with λ^{calib}

$$\left[l, u\right] = \hat{\mu} \pm \lambda^{\text{calib}} \sqrt{\widehat{\text{var}}(\hat{\mu}) + \sum_{c=1}^{C+1} \hat{\sigma}^2}.$$

3.3 | Simulation study

The coverage probabilities of the six different PI described above, were assessed by Monte Carlo simulations based on two different random effects models: A two-way-hierarchical design (h2) and a two-way cross-classified layout with interaction (c2). This two models were chosen since they are applied in real-life situations (as mentioned above) and they reflect a certain degree of complexity. On the other hand they are not too complex and hence, the computing time for the simulations were kept to a manageable level. In the following sections these models are explained in the context of patients that are analyzed in different laboratories, but of course the models can be applied to any experimental setup that fits into the scheme.

3.3.1 | Two-way hierarchical model (h2)

The h2 random effects model is given by

$$
\begin{aligned}
y_{ijk} &= \mu + a_i + b_{j(i)} + e_{k(ij)} \\
a_i &\sim N(0, \sigma_a^2), \quad i = 1, \dots, I \\
b_{j(i)} &\sim N(0, \sigma_b^2), \quad j(i) = 1, \dots, n_{j(i)} \\
e_{k(ij)} &\sim N(0, \sigma_e^2), \quad k(ij) = 1, \dots, n_{k(ij),}
\end{aligned} \tag{13}
$$

in which a random sample of $\sum_{i=1}^{I} n_{j(i)}$ patients is analyzed in $i = 1, \dots, I$ laboratories, such that in an unbalanced design different subsets of $n_{j(i)}$ patients are analyzed per laboratory with $n_{(ij)}$ observations for each of the $j(i)$ patients, for example, due to obtaining $n_{(ij)}$ technical replicates from each patient $j(i)$. In the balanced case, the total number of patients is IJ with $J = n_{j(i)}$ $\forall j(i) = 1(i), \dots, n_{j(i)}$ and the total number of observations is $N = IJK$ with $K = n_{k(ij)}$ $\forall k(ji) = 1(ji), \dots, n_{k(ij)}$.

In the model given above μ is the overall mean, a_i are the random effects for the laboratories, $b_{j(i)}$ are the random effects for the patients within the laboratories and $e_{k(ij)}$ are the residuals. Please note that a_i, $b_{j(i)}$, and $e_{k(ij)}$ are assumed to be independent from each other.

Mean squares and weights for the calculation of the PI based on the h2 model are given in Table 3. In analogy to Lin and Liao, the variance-covariance matrix used for the calculation of the GPQ-based PI for $M = I^* J^* K^*$ future observations obtained from a balanced design is given by

$$
\widehat{\mathrm{var}}(\boldsymbol{Y}^*) = \hat{\sigma}_a^2(\boldsymbol{I}_{I^*} \otimes \boldsymbol{J}_{J^*} \otimes \boldsymbol{J}_{K^*}) + \hat{\sigma}_b^2(\boldsymbol{I}_{I^*} \otimes \boldsymbol{I}_{J^*} \otimes \boldsymbol{J}_{K^*}) + \hat{\sigma}_e^2(\boldsymbol{I}_{I^*} \otimes \boldsymbol{I}_{I^*} \otimes \boldsymbol{I}_{K^*}) \tag{14}
$$

$$
\widehat{\mathrm{var}}(\boldsymbol{D}) = \widehat{\mathrm{var}}(\boldsymbol{Y}^*) + \widehat{\mathrm{var}}(\mu)(\boldsymbol{I}_{I^*} \otimes \boldsymbol{J}_{J^*} \otimes \boldsymbol{J}_{K^*}), \tag{15}
$$

with $\widehat{\mathrm{var}}(\mu) = \frac{1}{IJK}(JK\hat{\sigma}_a^2 + K\hat{\sigma}_b^2 + \hat{\sigma}_e^2)$, $\boldsymbol{I}_{I^*}$ as the identity matrix of order I^* and $\boldsymbol{J}_{J^*}$ as a square matrix of order J^* with all entries set to one and $\otimes$ as the Kronecker product. According to Sahai and Ageel (2000) the variance components can be expressed as

$$
\sigma_a^2 = \frac{1}{JK}(\mathrm{EMS}_a - \mathrm{EMS}_{b(a)}). \tag{16}
$$

$$
\sigma_{b(a)}^2 = \frac{1}{K}(\mathrm{EMS}_{b(a)} - \mathrm{EMS}_e). \tag{17}
$$

$$
\sigma_e^2 = \mathrm{EMS}_e. \tag{18}
$$

TABLE 3 Model h2 (balanced): Formulas for the calculation of prediction intervals

Effect	c	df_c	$\mathbf{MS}_c^{\mathbf{Sat}}$	$\omega_c^{\mathbf{Sat}}$	$\omega_c^{\mathbf{REML}}$
a_i	1	$I - 1$	$\frac{\sum_i (\bar{y}_{i..} - \bar{y}_{...})^2}{I-1}$	$1 + 1/I$	$1 + 1/I$
$b_{j(i)}$	2	$IJ - I$	$\frac{\sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..})^2}{IJ-I}$	$1 - 1/J$	$1 + 1/IJ$
$e_{k(ij)}$	3	$IJK - IJ$	$\frac{\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2}{IJK-IJ}$	$1 - 1/K$	$1 + 1/IJK$

3.3.2 | Two-way cross-classified model with replication (c2)

The model for the two-way cross-classified layout with replication is given by

$$y_{ijk} = \mu + a_i + b_j + ab_{ij} + e_{k(ij)}$$

$$a_i \sim N(0, \sigma_a^2), \quad i = 1, \ldots, I$$

$$b_j \sim N(0, \sigma_b^2), \quad j = 1, \ldots, J$$

$$ab_{ij} \sim N(0, \sigma_{ab}^2), \quad ij = (11, \ldots, IJ)$$

$$e_{k(ij)} \sim N(0, \sigma_e^2), \quad k(ij) = 1, \ldots, n_{k(ij)}.$$

Usually this setup is balanced such that I patients are analyzed in J laboratories exactly $K = n_{k(ij)}\ \forall k(ij) = 1(ij), \ldots, n_{k(ij)}$ times. Unbalancedness occurs if some of the possible IJ combinations of patient and laboratory are missing in the data such that $n_{(ij)} = 0$ for that particular interaction term or some of the K repetitions per combination are missing $(K \neq n_{k(ij)}\ \exists k(ij) \neq 1(ij), \ldots, n_{k(ij)})$. The total number of observations is $N = \sum_i \sum_j \sum_k n_{k(ij)}$.

In the model given above μ is the overall mean, a_i are the random effects for the patients, b_j are the random effects for the laboratories, ab_{ij} is the interaction term and $e_{k(ij)}$ are the residuals. Please note that a_i, b_j, ab_{ij}, and $e_{k(ij)}$ are assumed to be independent from each other.

MSQ and weights for the calculation of PI based on the c2 model are given in Table 4. The variance-covariance matrix used for the calculation of the GPQ-based PI for $M = I^* J^* K^*$ future observations obtained from a balanced design is given by Lin and Liao

$$\widehat{\text{var}}(\boldsymbol{Y}^*) = \hat{\sigma}_a^2(\boldsymbol{I}_{I^*} \otimes \boldsymbol{J}_{J^*} \otimes \boldsymbol{J}_{K^*}) + \hat{\sigma}_b^2(\boldsymbol{J}_{I^*} \otimes \boldsymbol{I}_{J^*} \otimes \boldsymbol{J}_{K^*}) + \hat{\sigma}_{ab}^2(\boldsymbol{I}_{I^*} \otimes \boldsymbol{I}_{J^*} \otimes \boldsymbol{J}_{K^*}) + \hat{\sigma}_e^2(\boldsymbol{I}_{I^*} \otimes \boldsymbol{I}_{I^*} \otimes \boldsymbol{I}_{K^*}), \tag{19}$$

$$\widehat{\text{var}}(\boldsymbol{D}) = \widehat{\text{var}}(\boldsymbol{Y}^*) + \widehat{\text{var}}(\mu)\boldsymbol{J}_M, \tag{20}$$

with $\widehat{\text{var}}(\mu) = \frac{1}{IJK}(JK\hat{\sigma}_a^2 + IK\hat{\sigma}_b^2 + K\hat{\sigma}_{ab}^2 + \hat{\sigma}_e^2)$, $\boldsymbol{I}_{I^*}$ as the identity matrix of order I^* and $\boldsymbol{J}_{J^*}$ as a square matrix of order J^* with all entries set to one. Following Lin and Liao (2008), the variance components are given by

$$\sigma_a^2 = \frac{1}{JK}(\text{EMS}_a - \text{EMS}_{ab}), \tag{21}$$

$$\sigma_b^2 = \frac{1}{IK}(\text{EMS}_b - \text{EMS}_{ab}), \tag{22}$$

$$\sigma_{ab}^2 = \frac{1}{K}(\text{EMS}_{ab} - \text{EMS}_e), \tag{23}$$

$$\sigma_e^2 = \text{MS}_e, \tag{24}$$

using the weights given in Table 4.

TABLE 4 Model c2 (balanced): Formulas for the calculation of prediction intervals

Effect	c	df_c	$\mathbf{MS}_c^{\mathbf{Sat}}$	$\omega_c^{\mathbf{Sat}}$	$\omega_c^{\mathbf{REML}}$
a_i	1	$I-1$	$\frac{\sum_i (\bar{y}_{i..}-\bar{y}_{...})^2}{I-1}$	$1+1/I$	$1+1/I$
b_j	2	$J-1$	$\frac{\sum_j (\bar{y}_{.j.}-\bar{y}_{...})^2}{J-1}$	$1-1/J$	$1+1/J$
ab_{ij}	3	$(I-1)(J-1)$	$\frac{\sum_i \sum_j (\bar{y}_{ij.}-(\bar{y}_{i..}+\bar{y}_{.j.}-\bar{y}_{...}))^2}{(I-1)(J-1)}$	$1-1/I-1/J-1/IJ$	$1+1/IJ$
$e_{(ij)}$	4	$IJ(K-1)$	$\frac{\sum_i \sum_j \sum_k (y_{ijk}-\bar{y}_{ij.})^2}{IJ(K-1)}$	$1-1/K$	$1+1/IJK$

3.3.3 | Simulation settings

In order to assess the coverage probabilities of the six different PI, Monte Carlo simulations were run. For that purpose, the two models described above (h2, c2) were utilized. All simulations were run independently from each other.

For the h2 model simulations were run for the $h = 1, \dots, 162$ different combinations of $I = \{5, 10, 15\}$, $J = \{2, 5, 10\}$, $K = \{2, 10\}$, $\sigma_a^2 = \{20, 2, 0.2\}$, $\sigma_b^2 = \{20, 2, 0.2\}$ and $\sigma_e^2 = 2$ for all three methods.

The simulation setting for the c2 model was comprised of $h = 1, \dots, 486$ combinations of $I = \{5, 10, 15\}, J = \{2, 5, 10\}, K = \{2, 10\}, \sigma_a^2 = \{20, 2, 0.2\}, \sigma_b^2 = \{20, 2, 0.2\}$ $\sigma_{ab}^2 = \{20, 2, 0.2\}$ and $\sigma_e^2 = 2$ for the two simple PI. But, due to the extensive computing time, this setting was reduced for the df-calibrated PI. The parameters I, J, K, and σ_e were the same as before, but the simulations were run either with $\sigma_a^2 = \{20, 2\}, \sigma_b^2 = \{20, 2\}$ $\sigma_{ab}^2 = \{20, 2\}$ or with $\sigma_a^2 = \{20, 0.2\}, \sigma_b^2 = \{20, 0.2\}$ $\sigma_{ab}^2 = \{20, 0.2\}$.

In the simulations regarding the quantile-calibrated PI, the number of bootstraps was set to $B = 1000$, $\lambda_1 = 1$ $\lambda_2 = 20$, the maximum number of bisection-steps was $D = 30$ and the tolerance was set to $s = 0.001$. If after 30 bisection steps $|\psi - \psi_d|$ was higher than the tolerance, λ_{30} was used for the calculation of that particular PI. The relatively low number of B=1000 bootstrap samples was chosen to keep the computing time of the simulation on a manageable level.

The performance of PI for one future observation ($M = 1$) was assessed for all six methods based on balanced data as well as for $M = 8$ (with $I^* = 2$, $J^* = 2$, $K^* = 2$) using the GPQ-based (Method 3) and the bootstrap-calibrated PI. Furthermore, coverage probabilities of the bootstrap calibrated interval were also simulated for $M = 5$ future observations based on unbalanced data. In this setting, the sampling of the simulation data sets was done as described before, but single observations on the lowest hierarchical level ($e_{k(ij)}$) were dropped out following a Bernoulli distribution with proportion set to 0.3. In a next step observations were dropped out on the level of the interaction terms ($b_{j(i)}$, ab_{ij}) following a Bernoulli distribution with proportion set to 0.1. This approach was done in order to generate data that is heavily unbalanced on both of the possible hierarchical levels.

For each of the simulation settings $r = 1, \dots, 5{,}000$ historical data sets were drawn. Similarly another dataset was sampled from which M observations were randomly chosen to be the actual observations $\boldsymbol{y}_{hr}^*$. For each of the historical data sets one prediction interval $[l, u]_{hr}$ was computed and the coverage probability ψ_h was estimated to be

$$\hat{\psi}_h = \frac{\sum_{s=1}^{R} I_{hr}}{R} \quad \text{with}$$

$$I_{hr} = 1 \quad \text{if} \quad \boldsymbol{y}^*_{hr} \in [l, u]_{hr}$$

$$I_{hr} = 0 \quad \text{if} \quad \boldsymbol{y}^*_{hr} \notin [l, u]_{hr}.$$

It has to be noted that the `lmer()` function threw warning messages regarding the convergence of the model for up to almost 50% of the sampled data sets (using R 3.6.2 and lme4 1.1.23 on Windows 10). Hence, the datasets on which `lmer()` threw a warning were tracked and the coverage probability was also computed based on the simulated datasets that did not result in a warning. But, since the coverage probability did not change depending on inclusion or exclusion of cases with warnings the results given below depend on all simulated datasets rather than on the datasets that do not result in a warning only.

However, due to the sampling process of unbalanced data, it was possible that in rare cases the sampled data was such small, that the model failed to converge if $I = 5$ (less than 1% per setting). In this case the coverage probabilities were computed based on the reduced set of the simulated data.

4 | RESULTS

The simulated coverage probabilities $\hat{\psi}_h$ are given in Figures 1–4 which depend on the number of replications (I, J) for the random effects. Two additional quantities are displayed to focus on settings with extreme ratios between variance components and total variance as well as between variance components and their corresponding degrees of freedom. These quantities are denoted as

$$\Omega_h = \max\left(\frac{\sigma^2_{ch}}{\sum_c \sigma^2_{ch}}\right) \quad \text{and}$$

$$\tau_h = \max\left[\frac{\sigma^2_{a,h}}{\sigma^2_{ab,h}} \Big/ \frac{df_{a,h}}{df_{ab,h}}, \frac{\sigma^2_{b,h}}{\sigma^2_{ab,h}} \Big/ \frac{df_{b,h}}{df_{ab,h}}, \frac{\sigma^2_{ab,h}}{\sigma^2_{e,h}} \Big/ \frac{df_{ab,h}}{df_{e,h}}\right].$$

Please note, that in the h2 model τ_h contains only the ratios $\frac{\sigma^2_{a,h}}{\sigma^2_{ab,h}} / \frac{df_{a,h}}{df_{ab,h}}$ and $\frac{\sigma^2_{ab,h}}{\sigma^2_{e,h}} / \frac{df_{ab,h}}{df_{e,h}}$ due to the given hierarchical order of the observations. In the simulation, the minimum τ_h was 0.01 and the maximum τ_h was 900 for the h2 model and 0.04 and 1400 for the c2 model, respectively.

In this setting, Ω_h represents the maximum ratio of the variance components to the total variance, meaning that the higher Ω_h becomes, the more one single variance component plays a dominant role in the data and vice versa (size of the dots in Figures 1–4). As described above, τ_h indicates the maximum ratio of the variance-components of higher hierarchical order to the variance-component one hierarchical level below compared to the ratio of their corresponding degrees of freedom. Hence, $\tau_h = 1$ means that the ratio between variance components equals the ratio of their corresponding degrees of freedom. If the variance components of higher hierarchical order are estimated to be high compared to the components one level below, but are estimated with relatively small degrees of freedom, τ_h will be > 1, resulting in coverage probabilities below the nominal 95% (red dots in the figures). Contrary, τ_h will be < 1 if the variance components

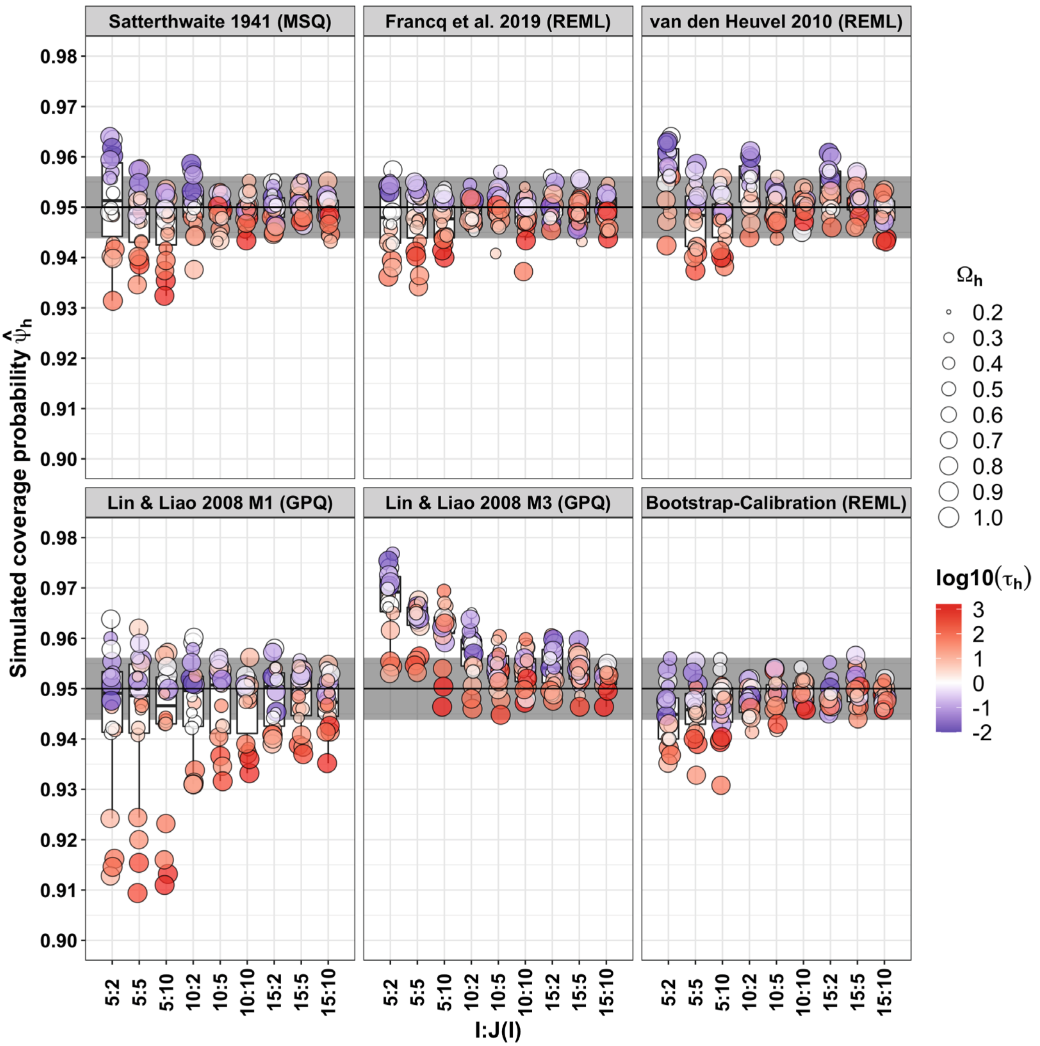

FIGURE 1 Coverage probabilities of prediction intervals for one future observation for the balanced h2 design. The nominal coverage probability $\psi = 0.95$ is indicated by the black line. The grey area indicates $\psi \pm 2se(\psi)$. The six different prediction intervals are represented by the panels

of higher hierarchical order are small compared to the components one level below, but are estimated with relatively high degrees of freedom. This results in coverage probabilities above the nominal 95% (blue dots in Figures 1–4).

The nominal coverage probability of $\psi = 0.95$ is given by the black horizontal lines. The grey area represents $\psi \pm 2se(\psi)$ with $se(\psi) = \sqrt{(0.95 \cdot 0.05)/5{,}000}$. Therefore an estimated coverage probability that falls into the grey area can not be treated to be different from the nominal 0.95.

4.1 | Coverage probabilities of PI for one future observation

The simulated coverage probabilities of PI for one future observation based on balanced h2 models are given in Figure 1. For all six methods, the coverage probabilities depend mainly

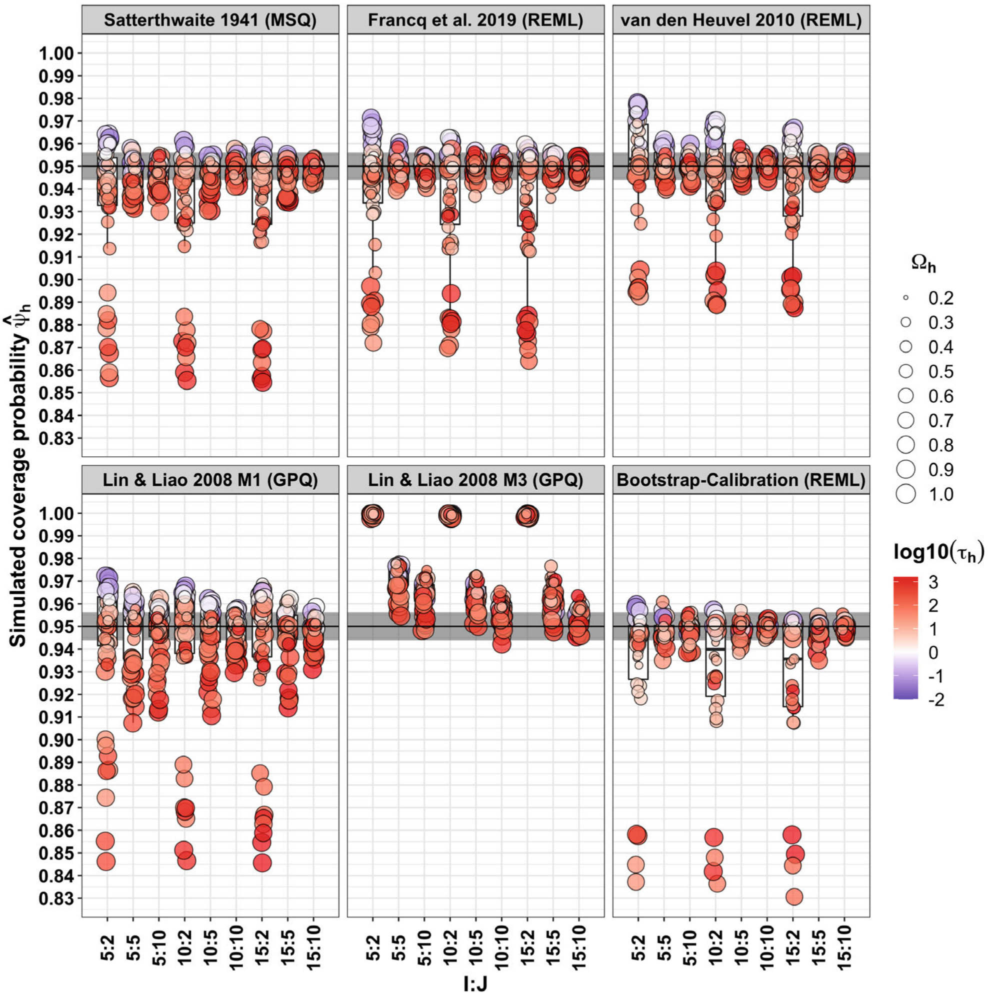

FIGURE 2 Coverage probabilities of prediction intervals for one future observation for the balanced c2 design. The nominal coverage probability $\psi = 0.95$ is indicated by the black line. The grey area indicates $\psi \pm 2se(\psi)$. The six different prediction intervals are represented by the panels

on the numbers of observations for each random effect. For the MSQ- and REML-based intervals, the simulated coverage probabilities approach the nominal 0.95 up to a satisfactory level for almost all combinations of Ω_h and τ_h, if $I > 5$ and $J(I) > 2$. Furthermore, if the number of observations for the random effect of highest hierarchical order is high (I is at least 10), the bootstrap-calibrated PI and the PI of Francq et al. (2019) approach the nominal level even for $J(I) = 2$.

The GPQ-based interval following Method 1 remains liberal if τ_h is high, even for higher I and J. The GPQ-based prediction interval following Method 3 is the only interval that approaches the coverage probabilities from above. Anyhow, especially for small τ_h, the interval remains slightly too conservative even if $I > 5$ and $J(I) > 2$.

For $I = 5$ the coverage probabilities of the MSQ based PI following Satterthwaite (1941) are too low if the variance component σ_a^2 is relatively high, but estimated based on a low number of

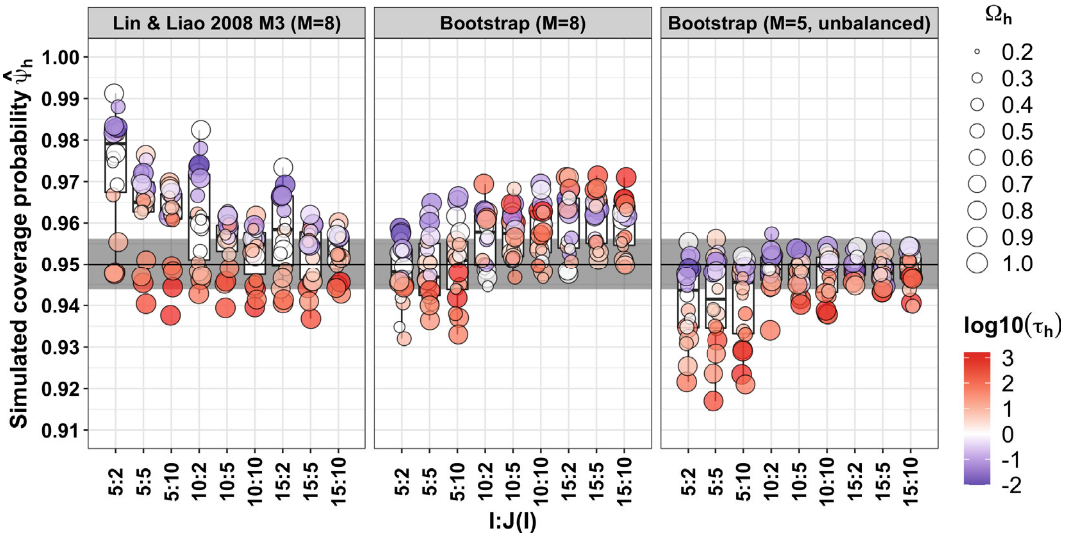

FIGURE 3 Coverage probabilities of prediction intervals for more than one future observation for the balanced h2 design. The nominal coverage probability $\psi = 0.95$ is indicated by the black line. The grey area indicates $\psi \pm 2se(\psi)$

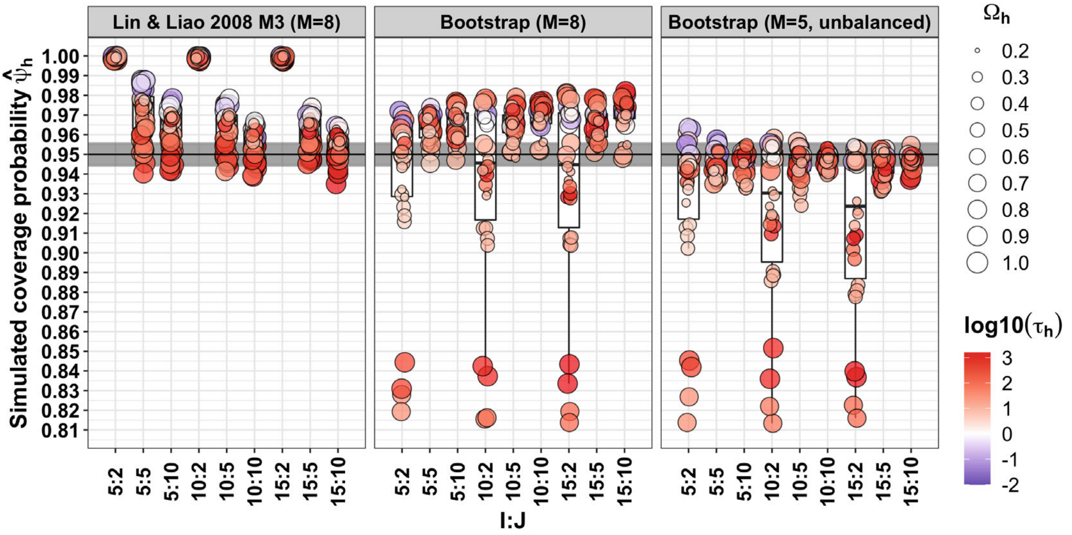

FIGURE 4 Coverage probabilities of prediction intervals for more than one future observation for the balanced h2 design. The nominal coverage probability $\psi = 0.95$ is indicated by the black line. The grey area indicates $\psi \pm 2se(\psi)$

observations (high τ_h, red dots) and too high if σ_a^2 is relatively low, but estimated based on a high number of observations (low τ_h, blue dots).

The coverage probabilities of the PI for one future observation based on balanced c2 models are given in Figure 2. Regardless of the number of observations per random effect (I and J), both GPQ-based methods do not approach the nominal coverage probability of 0.95 to a satisfactory level for most of the simulated settings. The PI based of GPQ-Method 1 remains liberal if τ_h is high, even for high I and J. Contrary, the PI calculated with GPQ-Method 3 remains conservative $J = 2$ the simulated coverage probabilities are close to one and even for high numbers of observations per random effect ($I = 15$ and $J = 10$) many observed coverage probabilities remain above the nominal level.

If both I and J are at least 5, most of the coverage probabilities all three REML-based intervals are close to 0.95 and hence approach the nominal coverage probability up to a satisfactory level. The MSQ-based PI of Satterthwaite (1941) approaches the nominal level only if I and J are at least 10, since for high τ_h the coverage probability remains liberal for smaller numbers of observations for the random effects.

4.2 | Coverage probabilities of PI for several future observations

Coverage probabilities of PI for $M = 8$ future observations were computed for the GPQ-based PI following Method 3 of Lin and Liao (2008) as well as for the bootstrap-calibrated PI. The coverage probabilities of the GPQ-based PI are slightly higher than for the PI for $M = 1$, if τ_h is small (blue dots in the left panel of Figures 3 and 4) or slightly lower if τ_h is high (red dots). In the settings were the the PI for one future observation approaches the nominal 0.95 ($I > 5$ and $J > 2$ for the h2-model or $I \geq 5$ and $J \geq 2$ in thec2-model) the coverage probabilities of the bootstrap-calibrated PI approach the nominal level or remain slightly above (middle panel of Figures 3 and 4). Contrary, the coverage probabilities reach the nominal level or remain slightly below if the bootstrap-calibrated PI is calculated based on unbalanced data (right panel of Figures 3 and 4).

5 | COMPUTATIONAL DETAILS

Except for the quantile calibrated interval, none of the methods described above are publicly available in R (R Core Team, 2019) in a user friendly form, neither as a code script that works without adaption nor as an add on package. Hence, the existing methods were implemented by hand. ANOVA-based statistics such as sum of squares or degrees of freedom used for the calculation of the intervals given by Satterthwaite (1941) and Lin and Liao (2008) were calculated using the `aov()` function from the stats package (R Core Team, 2019). The estimates $\widehat{\text{var}}(\hat{\sigma}_c^2)$ and $\widehat{\text{var}}(\widehat{\text{var}}(\boldsymbol{y}))$ used for the calculation of the uncalibrated REML-based intervals were obtained from the `remlMM()` function of the VCA package (Schuetzenmeister & Dufey, 2019). The bootstrap-calibrated PI can be applied using the `lmer_pi()` function from the predint package (Menssen, 2021).

5.1 | Quantile-calibrated PI with the predint package

As mentioned before, the `predint::lmer_pi()` function provides a user-friendly implementation of the bootstrap-calibrated prediction interval given in Section 3.2.4. Its arguments and the variables that are described by them are given in Table 5. PI as well as upper or lower prediction limits (argument `alternative`) can be computed based on a random effects model (argument `model`) fit to the historical data using `lme4::lmer()` (Bates, Maechler, Bolker, & Walker, 2015). If a dataset containing actual data is provided via `newdat`, `predint::lmer_pi()`

TABLE 5 Arguments of the `lmer_pi()` function and their description

Argument	Variable	Description
`model`		Random effects model fit with `lme4::lmer()`
`newdat`	$\boldsymbol{y}^*$	Data set with new observations
`m`	M	Number of future observations
`alternative`		Prediction interval, lower prediction limit or upper prediction limit
`alpha`	α	Defines the nominel coverage probability $1-\alpha$
`nboot`	B	Number of bootstraps
`lambda_min`	λ_1	Lower start value for bisection
`lambda_max`	λ_2	Upper start value for bisection
`traceplot`		Graphical overview about the bisection process
`n_bisec`	D	Maximum number of bisection steps

automatically marks the observations that are not covered by the interval. Alternatively, only the number of future observations for which the PI should be computed can be specified using the argument `m`.

The start values for the bisection process are given by `lambda_min` and `lambda_max`. In rare cases it might happen, that the default values (0.01, 10) for `lambda_min` and `lambda_max` result in bootstrapped coverage probabilities lower or higher than the nominal level for both start values. If the coverage is too low, the PI will be computed based on `lambda_max`. In contrary, the PI will be computed based on `lambda_min`, if the coverage is too high. Anyhow, in this cases `predint::lmer_pi()` gives a warning such that the user can define the start values by hand.

Since `predint::lmer_pi()` relies on random effects models fit with `lme4::lmer()` and `lme4::bootMer()` for bootstrapping, it can be applied to all data formats, regardless if they are balanced or unbalanced. Another feature that makes `predint::lmer_pi()` easy to apply is the fact that no variance-covariance matrix for the future observations have to be provided. The application of `predint::lmer_pi()` to real-life data is demonstrated in the following section. For a detailed description of the predint package and its other functions and fields of application, see https://cran.r-project.org/web/packages/predint/readme/README.html.

6 | APPLICATION OF PI TO REAL-LIFE DATA

As already described above, all methods were implemented by hand (except for the bootstrap calibrated PI for which the predint package was used). In order to make the application of all six PI as reproducible as possible, the R-code used for the calculation of the PI given in Tables 6 and 7 is available on GitHub under https://github.com/MaxMenssen/menssen_schaarschmidt_2021.

6.1 | ADA cut point estimation

For all six methods, predcition intervals for one future observation were calculated for the data set given by Hoffman and Berger (2011) which is comprised of data from a bioassay in which

TABLE 6 Prediction intervals based on the data set of Hoffmann and Berger (2011)

Method	*L*	*U*	Comp. time (s)
Satterthwaite (1941)	0.7556553	1.5512672	0.002
Lin and Liao (2008), M1	0.7637171	1.5348919	0.034
Lin and Liao (2008), M3	0.3139889	3.7333264	0.030
Franq et al. (2019)	0.749874	1.563227	0.049
van den Heuvel (2010)	0.7359454	1.5928128	0.049
bs-calibrated	0.7562869	1.549972	248.7

TABLE 7 Prediction interval for the historical control data

Maximum mean weekly body weight	Laboratory	Pathway	Lower	Upper	Cover
62.60	IIT Research Institute	wbe_air	43.91	75.46	TRUE
57.70	Southern Research Institute	gavage_corn oil	43.91	75.46	TRUE

electroluminescence signals (normalized mean RU) of 20 drug-naive mice were analyzed in three experimental runs. Since the normalized mean RU values are skewed, they were ln-transformed (following Hoffmann and Berger) such that

$$ln(y_{ij}) = \mu + a_i + b_j + e_{ij}$$

$$a_i \sim N(0, \sigma_a^2), \quad i = 1, \dots, I$$

$$b_j \sim N(0, \sigma_b^2), \quad j = 1, \dots, J$$

$$e_{ij} \sim N(0, \sigma_e^2),$$

with $ln(y_{ij})$ as the ln-transformed normalized mean RUs, a_i as the random effects associated with the runs, b_j as the random effects associated with the mice and e_{ij} as the residuals. The resulting PI for all six methods are given in Table 6. Please note that these intervals are already back transformed to the response scale (normalized mean RU).

Except for the GPQ-based interval calculated with Method 3 of Lin and Liao which is the widest PI by far, all PI are relatively close to each other. These findings are in line with the results obtained from the simulation studies (Figures 1 and 2) where the GPQ-based PI following Method 3 appears to be conservative.

This behavior can be explained by the fact, that the GPQ(var$(\boldsymbol{D})_k$) are averaged to yield one single estimate for the prediction variance GPQ(var($\boldsymbol{D}$)) (see Equation 11), but the distribution of GPQ(σ^2_{ck}) used for the calculation of GPQ(var($\boldsymbol{D}$)) is heavily right skewed. Hence the estimate GPQ(var($\boldsymbol{D}$)) has a positive bias. Because the estimate for the variance-covariance matrix of the error margin GPQ(var($\boldsymbol{D}$)) is treated as known, naturally one would assume that this interval shows coverage probabilities below the nominal level. Anyhow, the bias of GPQ(var($\boldsymbol{D}$)) is strong enough to contradict this effect.

6.2 | Historical control data about the maximum mean weekly bodyweight of female mice

Since the dataset containing historical controls regarding the mmwbw of mice is heavily unbalanced (Table 1), a prediction interval was calculated based on the quantile-calibrated PI only. For this purpose the `lmer_pi()` function from the predint package was used. A random effects model in which the pathways were nested in the two different laboratories (see Equation 13) was fitted to the data using the `lmer()` function from the lme4 package. Then, this model was handed over to `lmer_pi()` using the `model` argument. The two actual control groups given in Table 2 were provided by the `newdat` argument, such that a prediction interval for $M = 2$ future observations was calculated. The resulting output of `lmer_pi()` is given in Table 7. Since the prediction interval $[L, U] = [43.91, 75.46]$ covers the two actual observations, it can be assumed that they are in line with the historical mmwbws.

7 | DISCUSSION

In the sections above, two methods for the calculation of PI based on random effects models were proposed and compared to four PI that are already published. Due to the fact that MSQ-based PI occur in literature since 80 years (Satterthwaite, 1941) most of the previous research was done on that topic. Anyhow, only a few studies that use other methods than MSQ, obtained from the classical ANOVA tables, are available. Two GPQ-based methods for PI for $M \geq 1$ future observations were proposed by Lin and Liao (2008). Despite the fact, that the estimation of model parameters in random and mixed effects models via REML is available since the 1970s (Corbeil & Searle, 1976) and has become a standard method for estimation since then, only a few studies about PI that are based on REML estimates could be found in literature: Al-Sarraj et al. (2019) used a REML-based PI originally published by Pawitan (2001) and Francq et al. (2019) published a PI that is applicable to balanced or unbalanced mixed and random effects models. Anyhow, both approaches do not consider the uncertainty of the estimated prediction variance: Pawitan (2001) treats the estimated prediction variance as known and uses a standard normal quantile for the interval calculation. The interval of Francq et al. (2019) is based on a t-quantile for which the degrees of freedom are approximated based on the variance of the historical data and not on the prediction variance itself and hence neglecting a source of uncertainty (the estimated variance of the mean). Furthermore the literature lacks REML-based PI for more than one future observation.

The two proposed methods for the calculation of PI address the shortcomings mentioned above: A PI that takes the whole uncertainty that is associated with the prediction variance into account was computed by applying the df-approximation given by van den Heuvel (2010). But, with the weights presented here, this PI is only applicable to balanced data.

Furthermore, a bootstrap calibrated PI was proposed for which the whole quantile used for interval calculation was approximated. Classically, bootstrap calibration is based on the α by which the quantile used for interval calculation is defined (usually $t_{1-\alpha/2,df}$ or $\chi^2_{1-\alpha/2,df}$). In this approach, the α that is used for interval calculation is alternated by a bootstrap procedure until a value α^{calib} is found, such that the calibrated interval calculated with $t_{1-\alpha^{\text{calib}}/2,df}$ (or $\chi^2_{1-\alpha^{\text{calib}}/2,df}$) has coverage close to the nominal level $1 - \alpha$. This approach was developed in the early 1990s and was described by Efron & Tibshirani (1993) in detail. Therefore, α-calibration was applied by several authors for different purposes, such as tolerance limits (Hoffman, 2010), confidence intervals (Lee & Liao, 2012; Lee & Liao, 2014) or PI for overdispersed binomial data (Menssen

& Schaarschmidt, 2019). Anyhow, the idea of calibration can be also applied for other purposes such as the approximation of quantiles. Due to the approximation of the whole quantile (rather than a calibration of α), no assumption regarding its corresponding distribution has to be made. This circumstance makes the quantile-calibrated prediction interval easy to apply, especially if an interval for more than one future observation is needed because the formulation of the variance-covariance matrix for future observations is unnecessary. Since the bootstrap is drawn from a model fitted based on the REML approach, it does not matter if the data is balanced or unbalanced which makes the interval usable for a broad range of practical applications.

Furthermore, it has to be noted, that none of the existing methods is implemented in R (except for the quantile-calibrated PI). Hence, their application needs implementation by hand which is far beyond the scope of most applicants who are not trained in advanced programming. As far as the authors know, the quantile-calibrated PI is the only method for the calculation of PI based on random effects model, that is implemented in R and available in a general way. It could be shown, that the empirical coverage probabilities of the quantile-calibrated PI are slightly closer to the nominal level than that of the existing methods in most of the simulation settings. Anyhow, the simulated coverage probabilities do not approach the nominal level if the numbers of observations per random effect is lower than five.

Since the bootstrap calibration does not make any assumption regarding the distribution from which the quantile used for interval calculation is drawn, it can be applied to many other problems and models as long as the variance used for interval calculation can be computed. Therefore, the calibration process given above is also applicable for PI based on overdispersed binomial and count data. Further details regarding the implementation of this models can be found in the vignette of the predint package. A topic for future research that remains, is the application of the quantile calibration bootstrap to (generalized) linear mixed models.

ACKNOWLEDGMENTS

The authors have to thank Clemens Buczilowski for his technical support and Olaf Menssen for fruitful discussions. Furthermore we have to thank the reviewers for reading the manuscript and for their helpful comments. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Max Menssen https://orcid.org/0000-0003-2888-8542

REFERENCES

Al-Sarraj, R., von Brömssen, C., & Forkmann, J. (2019). Generalized prediction intervals for treatment effects in random-effects models. *Biometrical Journal*, *61*, 1242–1257.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Corbeil, R. R., & Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics.*, *18*(1), 31–38.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Elmore, A. S., & Peddada, S. D. (2009). Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. *Toxicologic Pathology*, *37*(5), 672–676.

Forkmann, J., & Piepho, H.-P. (2013). Performance of empirical BLUP and Bayesian prediction in small randomized complete block experiments. *Journal of Agricultural Science*, *151*, 381–395.

Francq, B. G., Lin, D., & Hoyer, W. (2019). Confidence, prediction, and tolerance in linear mixed models. *Statistics in Medicine*, *38*, 5603–5622.

Francq, B. G., Lin, D., & Hoyer, W. (2020). Confidence and prediction in linear mixed models: Do not concatenate the random effects. application in an assay qualification study. *Statistics in Biopharmaceutical Research.*, *12*(3), 267–272.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and T probabilities Lecture Notes in Statistics* (Vol. *195*). Heidelberg, Germany: Springer-Verlag.

Greim, H., Gelbke, H. P., Reuter, U., Thielmann, H. W., & Edler, L. (2003). Evaluation of historical control data in carcinogenicity studies. *Human & Experimental Toxicology*, *22*(10), 541–549.

Hahn, G. J., & Meeker, Q. M. (1991). *Statistical intervals* (1st ed.). Hoboken, NJ: John Wiley and Sons Inc.

Hahn, G. J., Meeker, Q. M., & Escobar, L. A. (2017). *Statistical intervals* (2nd ed.). Hoboken, NJ: John Wiley and Sons Inc.

Hoffman, D. (2010). One-sided tolerance limits for balanced and unbalanced random effects models. *Technometrics*, *52*, 303–312.

Hoffman, D., & Berger, M. (2011). Statistical considerations for calculation of immunogenicity screening assay cut points. *Journal of Immunological Methods*, *373*, 200–208.

Jaki, T., Allacher, P., & Horling, F. (2016). A false sense of security? can tiered approach be trusted to accurately classify immunogenicity samples? *Journal of Pharmaceutical and Biomedical Analysis*, *128*, 166–173.

Jeske, D. R., & Harville, D. A. (1988). Prediction interval procedures and (fixed-effects) confidence interval procedures for mixed linear models. *Communications in Statistics-Theory and Methods*, *17*(4), 1053–1087.

Lee, H. I., & Liao, C. T. (2012). Estimation for conformance proportions in a normal variance components model. *Journal of Quality Technology*, *44*, 63–79.

Lee, H. I., & Liao, C. T. (2014). Unilateral conformance proportions in balanced and unbalanced normal random effects models. *The Journal of Agricultural, Biological and Environmental Statistics*, *19*, 202–218.

Lin, T. Y., & Liao, C.-T. (2008). Prediction intervals for general balanced linear random models. *Journal of Statistical Planning and Inference*, *138*(19), 3164–3175.

McCullagh, C. E., & Searle, S. R. (2001). *Generalized, linear and mixed models*. Hoboken, NJ: John Wiley and Sons Inc.

Menssen M. (2021). predint: Prediction intervals. R package version 1.0.0. https://CRAN.R-project.org/package=predint

Menssen, M., & Schaarschmidt, F. (2019). Prediction intervals for overdispersed binomial data with application to historical controls. *Statistics in Medicine*, *38*, 2652–2663.

NTP 2021. Retrieved May 31, 2021, from https://ntp.niehs.nih.gov/data/controls/index.html

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood* (p. 433). Oxford, UK: Oxford University Press.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, Retrieved from. https://www.R-project.org/

Sahai, H., & Ageel, M. I. (2000). *The analysis of variance*. Boston: Birkhäuser.

Satterthwaite, F. E. (1941). Sythesis of variance. *Psychometrika*, *6*(5), 309–316.

Schaarschmidt, F., Hofmann, M., Jaki, T., Gruen, B., & Hothorn, L. A. (2015). Statistical approaches for the determination of cut points in anti-drug antibody bioassays. *Journal of Immunological Methods*, *418*, 84–100.

Schuetzenmeister A., Dufey F. (2019). VCA: variance component analysis. R package version 1.4.1. Retrieved from https://CRAN.R-project.org/package=VCA

Searle R.S., Casella G., McCulloch C.E. (2006): *Variance components*. 2nd, Hoboken, NJ: John Wiley and Sons Inc

Shen, M., & Dai, T. (2021). Statistical methods of screening cut point determination in immunogenicity studies. *Bioanalysis*, *13*(7), 551–563.

van den Heuvel, E. R. (2010). A comparison of estimation methods on the coverage probability of satterthwaite confidence intervals for assay precision with unbalanced data. *Communications in Statistics - Simulation and Computation*, *39*(4), 777–794. https://doi.org/10.1080/03610911003646373

Wang, C. M. (1992). Prediction intervals for balanced one-way random effects model. *Communications in Statistics - Simulation and Computation*, *21*(3), 671–687.

Zhang, L., Zhang, J. J., Kubiak, R. J., & Yang, H. (2013). Statistical methods and tool for cut point analysis in immunogenicity assays. *Journal of Immunological Methods*, *389*, 79–87.

How to cite this article: Menssen, M., & Schaarschmidt, F. (2022). Prediction intervals for all of M future observations based on linear random effects models. *Statistica Neerlandica*, *76*(3), 283–308. https://doi.org/10.1111/stan.12260

APPENDIX A. COMPARISONS BETWEEN THE DEGREES OF FREEDOM ESTIMATED WITH THE TWO VERSIONS OF THE GENERALIZED SATTERTHWAITE METHOD

In Figure A1, the average of the approximated degrees of freedom associated with the total variance of the historical data (approach of Francq et al. 2019) is compared to the degrees of freedom approximated for the prediction variance following van den Heuvel 2010. The errorbars indicate the observed minimum and maximum degrees of freedom for both methods. For each of the simulation settings the average degrees of freedom are higher for the approach of Francq et al. 2010. If the approximated degrees of freedom are low (squares in Figure A1), the difference between the two methods has an influence on the width of the corresponding PI. With rising degrees of freedom (bigger datasets) this effect becomes smaller and can be neglected for degrees of freedom higher than say 30, due to the convergence of the *t*-distribution against the standard normal.

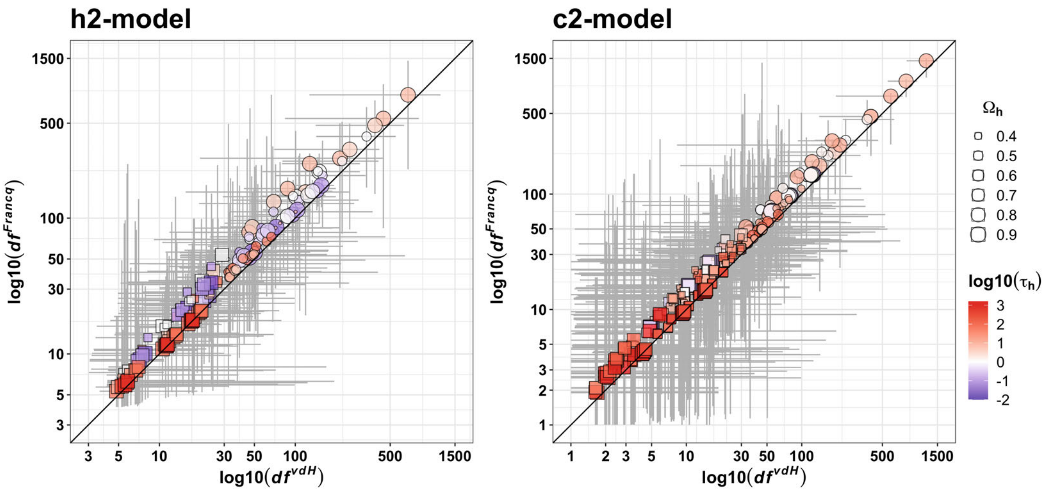

FIGURE A1 Simulated average degrees of freedom: df^{Francq} versus df^{vdH}. The black line indicates a 1:1 relationship. The grey errorbars represent the corresponding minimum and maximum obtained in the simulation. Squares indicate observations were $df^{vdH} < 30$