Prediction intervals based on historical control data obtained from bioassays

Von der Naturwissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation

von

Max Menssen, M. Sc.

2021

Referent: PD Dr. rer. hort. Frank Schaarschmidt

Korrefferent: Prof. Dr. rer. nat. habil. Dr.-Ing. Ludwig Hothorn

Tag der Promotion: 09.12.2021

For my parents Horst and Birgit Menssen

Contributing manuscripts

1. Menssen M., Schaarschmidt F. (2019):
   Prediction intervals for overdispersed binomial data with application to historical controls.
   Research article
   Statistics in Medicine. 38, 2652 - 2663.
   DOI: https://doi.org/10.1002/sim.8124

2. Menssen M., Schaarschmidt F. (2019):
   Prediction intervals for overdispersed binomial data with application to historical controls.
   Supplementary materials
   Statistics in Medicine. 38, 2652 - 2663.
   DOI: https://doi.org/10.1002/sim.8124

3. Menssen M., Schaarschmidt F. (2021):
   Prediction intervals for all of M future observations based on linear random effects models.
   Research article
   Statistica Neerlandica. Published online.
   DOI: https://doi.org/10.1111/stan.12260

4. Menssen M. (2021):
   predint: Prediction Intervals.
   Reference manual for the R-package predint.
   https://cran.r-project.org/web/packages/predint/index.html

Abstract

The calculation of prediction intervals based on historical control data obtained from bioassays is of interest in many fields of biological research. In pharmaceutical and pre-clinical applications, such as immonogenicity assays, the calculation of prediction intervals (or upper prediction limits) that distinguish between anti-drug antibody positive responders and anti-drug antibody negative non-responders are of interest. In (eco)toxicology several bioassays are run in order to study the toxicological properties of a given chemical compound on model organisms (eg. its carcinogenicity or its impact on aquatic food chains). In this field of research it is of interest to validate if the outcome of the actual untreated control (or the whole actual trial) is in line with the historical information. For that purpose, prediction intervals can be computed based on the historical control data. If the actual observations are covered by the interval, they are treated to be in line with the historical information.

The first chapter of this thesis provides a detailed overview about the use of historical control data in the context of biological trials. Furthermore, it reviews the data structure (dichotomous data, count data, continuous data) and the models on which the proposed prediction intervals ground. In the context of dichotomous or count data, special attention is given to overdispersion which is commonly present in data that has a biological background, but is usually not considered in literature regarding prediction intervals.

Hence, prediction intervals for one future observation that are based on overdispersed binomial data were proposed. The coverage probabilities of this intervals were assessed based on Monte-Carlo simulations and were substantially closer to the nominal level than prediction intervals found in literature that do not consider overdispersion (see sections 2.1 and 2.2).

In several applications the response is a continuous variable that can be assumed to be normal distributed. Anyhow, the data can be influenced by several random factors such as different laboratories that analyze probes of several patients. In this case the data can be modeled by linear random effects models and parameter estimates can be obtained based on the restricted maximum likelihood approach. For this scenario, two prediction intervals are proposed in section 2.3. One of this proposed intervals grounds on a bootstrap calibration procedure that makes it applicable even in cases where a prediction interval for more than one future observation is needed.

Section 2.4 describes the R-package 'predint' that provides the bootstrap calibrated prediction interval (as well as lower and upper prediction limits) described in section 2.3. Furthermore it provides prediction intervals for at least one future observation for overdispersed binomial or count data that make use of a similar calibration bootstrap as the prediction interval that is based on random effects models.

The key feature of this thesis is the derivation of prediction intervals for one or more future observations that are based on overdispersed binomial data, overdispersed count data or linear random effects models. To the authors knowledge, this is the first time that prediction intervals that reflect overdispersion are proposed. Furthermore, 'predint' is the first R-package available from the comprehensive R archive network that provides functions for the application of prediction intervals for the mentioned models. Hence, the methodology proposed in this thesis is publicly available and easy to apply by other researchers.

Keywords:

Bioassay, Overdispersion, Random Effects, Quasi-Likelihood, Restricted Maximum Likelihood

Zusammenfassung

Die Berechnung von Vorhersageintervallen auf der Grundlage von historischen Kontrolldaten aus Bioassays ist in vielen Bereichen der biologischen Forschung von Interesse. Bei pharmazeutischen und präklinischen Anwendungen, wie z. B. Immonogenitätstests, ist die Berechnung von Vorhersageintervallen (oder oberen Vorhersagegrenzen), die zwischen anti-drug Antikörper positiven Patienten und anti-drug Antikörper negativen Patienten unterscheiden, von Interesse. In der (Öko-)Toxikologie werden verschiedene Bioassays angewendet, um die toxikologischen Eigenschaften einer bestimmten chemischen Verbindung an Modellorganismen zu untersuchen (z. B. ihre Karzinogenität oder ihre Auswirkungen auf aquatische Nahrungsketten). In diesem Forschungsbereich ist es von Interesse zu überprüfen, ob das Ergebnis der aktuellen unbehandelten Kontrolle (oder der gesamten aktuellen Studie) mit den historischen Informationen übereinstimmt. Zu diesem Zweck können Vorhersageintervalle auf der Grundlage von historischen Kontrolldaten berechnet werden. Wenn die aktuellen Beobachtungen im Vorhersageintervall liegen, kann davon ausgegangen werden, dass sie mit den historischen Informationen übereinstimmen.
Das erste Kapitel dieser Arbeit gibt einen detaillierten Überblick über die Verwendung von historischen Kontrolldaten im Rahmen von biologischen Versuchen. Darüber hinaus wird ein Überblick über die Datenstruktur (dichotome Daten, Zähldaten, kontinuierliche Daten) und die Modelle, auf denen die vorgeschlagenen Vorhersageintervalle basieren, gegeben. Im Zusammenhang mit dichotomen Daten oder Zähldaten wird besonderes Augenmerk auf Überdispersion gelegt, die in Daten mit biologischem Hintergrund häufig vorkommt, in der Literatur zu Vorhersageintervallen jedoch meist nicht berücksichtigt wird.
Daher wurden Vorhersageintervalle für eine zukünftige Beobachtung vorgeschlagen, die auf überdispersen Binomialdaten beruhen. Die Überdeckungswahrscheinlichkeiten dieser Intervalle wurden auf der Grundlage von Monte-Carlo-Simulationen bewertet und lagen wesentlich näher am nominellen Level als die in der Literatur gefundenen Vorhersageintervalle, die keine Überdispersion berücksichtigen (siehe Abschnitte 2.1 und 2.2).
In mehreren Anwendungen ist die abhängige Variable kontinuierlich und wird als normalverteilt angenommen. Dennoch können die Daten durch verschiedene Zufallsfaktoren (zum Beispiel unterschiedliche Labore die Proben von mehreren Patienten analysieren) beeinflusst werden. In diesem Fall können die Daten durch lineare Modelle mit zufälligen Effekten modelliert werden, bei denen Parameterschätzer mittels Restricted-Maximum-Likelihood Verfahren geschätzt werden. Für dieses Szenario werden in Abschnitt 2.3 zwei Vorhersageintervalle vorgeschlagen. Eines dieser vorgeschlagenen Intervalle basiert auf einem Bootstrap-Kalibrierungsverfahren, das es auch in Fällen anwendbar macht, in denen ein Vorhersageintervall für mehr als eine zukünftige Beobachtung benötigt wird.
Abschnitt 2.4 beschreibt das R-Paket predint, in dem das in Abschnitt 2.3 beschriebene bootstrap-kalibrierte Vorhersageintervall (sowie untere und obere Vorhersagegrenzen) implementiert ist. Darüber hinaus sind Vorhersageintervalle für mindestens eine zukünftige Beobachtung für überdisperse Binomial- oder Zähldaten implementiert.
Der Kern dieser Arbeit besteht in der Berechnung von Vorhersageintervallen für eine oder mehrere zukünftige Beobachtungen, die auf überdispersen Binomialdaten, überdispersen Zähldaten oder linearen Modellen mit zufälligen Effekten basieren. Nach Kenntnis des Autors ist dies das erste Mal, dass Vorhersageintervalle, die Überdispersion berücksichtigen, vorgeschlagen werden. Darüber hinaus ist "predint" das erste über CRAN verfügbare R-Paket, das Funktionen für die Anwendung von Vorhersageintervallen für die genannten Modelle bereitstellt. Somit ist die in dieser Arbeit vorgeschlagene Methodik öffentlich zugänglich und kann von anderen Forschenden leicht angewendet werden.

Schlagworte:

Bioassay, Überdispersion, Zufällige Effekte, Quasi-Likelihood, Restricted Maximum Likelihood

Abbreviations

ADA	Anti-drug antibodies
CEBS	Chemical Effects in Biological Systems
CRAN	Comprehensive R archive network
GM	Genetically modified
HCD	Historical control data
NTP	National Toxicology Program
PI	Prediction interval
REML	Restricted maximum likelihood
RITA	Registry of Industrial Toxicology Animal-data

Contents

Chapter 1

Introduction

In many fields of biological research such as toxicology, highly standardized trials are run following predefined protocols or guidelines. Commonly, all of these trials use a certain experimental design (untreated control group vs. several groups that received a treatment), a particular model organism (e.g. a given strain of rats) and are run under standardized conditions. If the same type of trial is run several times, the knowledge about the behavior of historical control groups rises with each trial that was run. Hence, the historical control data can be used in order to verify the outcome of the actual control group or if needed the whole actual trial. This is the aim of the following sections in this thesis.

1.1 Motivating example

In 2012 Seralini et al. released their study about the long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize (GM). This paper induced a medial echo far beyond the reach of a usual scientific article. Furthermore, it induced a broad discussion among researchers that culminated in several letters to the editor, comments and other forms of communication in which the study was criticized. This huge amount of critique lead the Journal to retract the original paper [Seralini et al. 2014a]. Anyhow, the study was republished again in 2014 [Seralini et al. 2014b] in order to enable further scientific discussions.
Seralini et al. 2012 performed a long term study (24 month) in which 100 virgin albino Spraque-Dawley rats per sex were randomly assigned to one of ten cohorts. For each sex, one cohort served as an untreated control that did not receive the Roundup herbicide or the genetically modified maize. The remaining nine cohorts were treated either with the GM-maize, with the Roundup herbicide or with both. Hence, in each sex, all cohorts were comprised of ten individual rats.
After 24 month (the end of the study), three out of ten female rats in the untreated control had developed tumors, whereas fife to eight out of the ten female rats of the treatment groups developed at least one tumor (p.4224 of Seralini et al. 2012).
Although Seralini et al. 2012 did not perform any statistical comparisons between the control group and the treatment groups, the reported tumor rates and figures suggest an increase of the tumor incidence in the treatment groups compared to the untreated control. This suggestion was criticized to be only random variation [Seralini et al. 2014c].
One way to clarify such a suggestion, is the application of a statistical test in which the tumor rate of the untreated control is compared to the rates of the treatment groups. Unfortunately, Seralini et al. 2012 did not perform any statistical test nor provide the tumor rates for all of the nine treatment groups such that a test on the complete data can be applied by others. Another possibility for the evaluation of the actual tumor rates is to use historical information about the background tumor rate of female Spraque-Dawley rats. For this purpose, historical control data (HCD) obtained from the untreated control groups of similar trials can be used.
A graphical overview about tumor rates obtained from HCD together with the tumor rates reported by Seralini et al. 2012 is given in figure 1.1. The historical control data about female Spraque-Dawley rats was drawn from the Historical Controls Report 2020 of the National Toxicology Program (NTP). This report provides tumor rates of untreated control groups of long term studies (24 month) started between 2007 and 2012 [NTP 2021]. Since it is unclear if the tumor rates reported by Seralini et al. 2012 referred to the female rats that developed a tumor regardless of its kind (total tumor rate) or to mammary tumors only, HCD for both types is given.

At first glance, it seems to be obvious that the tumor rate of the untreated control of Seralini et al. 2012 is unusually low, compared to the historical control data (regardless if compared to the rates of total or mammary tumors). Furthermore, if one follows the first visual impression, one might think that the reported tumor rates of the treatment groups fit perfectly to the historical rates of mammary tumors and hence, Seralini et al. 2012 interpreted only the random variation that has to be expected for female Spraque-Dawley rats. However, this is a rather naive interpretation that neither considers the uncertainty of the observed tumor rates nor the number of rats (10, 50, 90) on which the observed rates are based on.
Anyhow, this approach tries to answer the simple question: "Which of the actual tumor rates are in line with the ones obtained in the historical data and which are not". From a statistical point of view, this question can be answered by the application of a prediction interval that defines in which range one or more future observations can be expected based on the historical data with a given error probability. Hence the application of prediction intervals aims to provide cut points in order to distinguish if future tumor rates are in line with the observations from the historical control data or not (which is the case if they are above the upper cut point or below the lower cut point).

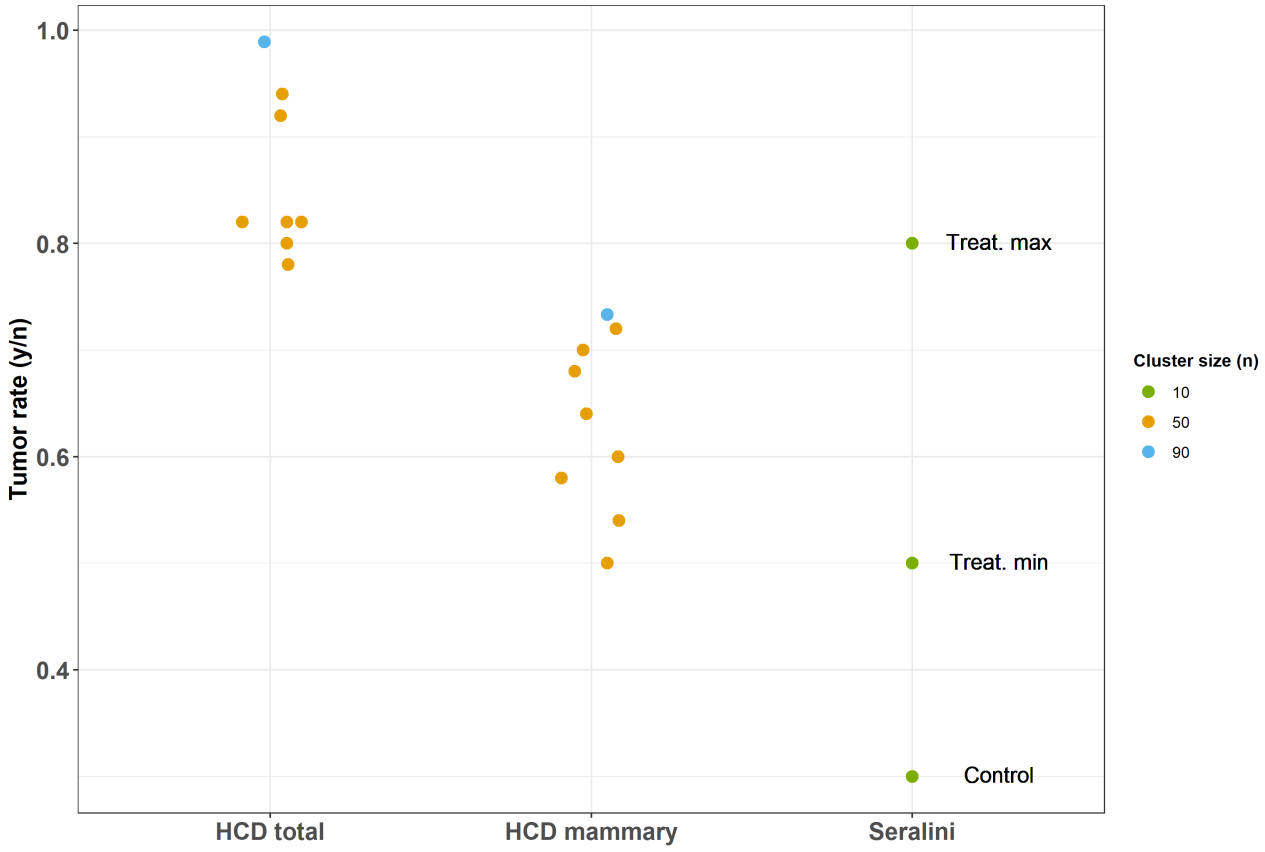

Figure 1.1: Historical control data for female Spraque Dawley rats obtained from the NTP Historical Controls Database [NTP 2021] together with the tumor rates reported by Seralini et al. 2012. **HCD total:** Total numbers of tumors; **HCD mammary:** Mammary gland (Fibroma, Fibroadenoma, Carcinoma, or Adenoma); **Seralini:** Tumor rates reported by Seralini et al. 2012; **Treat. max:** Maximum tumor rate of the treatment groups reported by Seralini et al. 2012; **Treat. min:** Minimum tumor rate of the treatment groups reported by Seralini et al. 2012; **Control:** Tumor rate of the untreated control group reported by Seralini et al. 2012; **Cluster size:** Number of rats (n) inside each treatment or control group; **Tumor rate:** Number of rats with tumor (y) divided by its corresponding cluster size (n).

1.2 Definition of HCD

According to Brooks et al. 2019, HCD can be defined as follows: "Historical control data are control data compiled from similar studies, performed either before or after the concurrent study. The basic assumption for using historical control data is that the past performance of test subjects under a particular set of conditions is a good predictor of future or previous performance. They can therefore be used, together with a concurrent control, to understand what a normal response is for a particular type of test subject under a particular set of conditions, and therefore to help determine when a treatment response may be outside the norm".
This definition implies, that theoretically, HCD can be obtained for almost all kinds of trials as long as historical control groups are comparable to the actual control group(s) with regard to experimental design, model organism, living conditions etc. Hence, HCD is used in several fields of research such as clinical trials [Viele et al. 2014], engineering [Young et al. 2016] or in mammalian toxicology [Deschl et al. 2002]. Recently, the use of HCD is discussed in ecotoxicology as well [Brooks et al. 2019, Rotolo et al. 2021]. Anyhow, the following sections will focus on the use of HCD for trials with a biological background, especially on bioassays.

1.3 Definition of bioassays

The idea that adverse effects of a chemical compound on target organisms, such as humans or species of a food chain in a given ecosystem, can be studied by its application to model organisms is more than a hundred years old: One of the first trials one can call a bioassay was reported in 1918 and was carried out in order to study if coal tar is carcinogenic for humans [Yamagiwa and Ichikawa 1918]. For this purpose coal tar was applied on the ears of domestic rabbits since no case of spontaneous tumor growth on that organ was known by the authors. Tumor rates were recorded after 70 and 150 days and it turned out, that 10 % to 100 % of the rabbit ears developed at least one tumor when treated with coal tar.
Although Yamagiwa and Ichikawa did not include an untreated control group in their study nor did any statistical comparisons, their findings led them to the conclusion that the coal tar treatment induced the observed tumors in the rabbits ears. The reason that, also nowadays, this conclusion seems to be plausible is the fact that the reported tumor rates are much higher (up to 100 %) than the ones provided by the historical knowledge available in 1918 (0 %). In this context it is noteworthy, that even one century ago, historical knowledge was used in order to verify the outcome of the actual trial.
Anyway, in the last century, many different kinds of bioassays for several purposes and scientific questions were established. Even if the scientific question that should be answered by running a certain type of bioassay can be highly different between the fields of application, all bioassays have several things in common:

1. A certain type of model organism is used for a certain type of bioassay.
2. An untreated control group is compared to at least one (usually several) treatment groups.
3. The experimental design is highly standardized.
4. The amount of information regarding the (historical) control groups rises with each run.

1.4 Fields of application

One field of research in which HCD is routinely used is toxicology where carcinogenicity studies with rats and mice are carried out routinely. Hence, the features by which HCD should be characterized to be a meaningful source of information are discussed in that field [Brooks et al. 2019]. According to Hasemann 1995, the recorded tumor rates of rats and mice can be influenced by several factors such as housing condition, body weight or inconsistencies in histopathological diagnosis. Another factor that can influence the outcome of the untreated control is genetic drift. Hence it is recommended not to use HCD regarding rats and mice that is older than fife to seven years [Hasemann 1995, Elmore and Peddada 2009]. Anyhow, wider time intervals might be appropriate if the tumor rates are stable over a longer time span [Keenan et al. 2009]. Furthermore, several sources for historical information are treated differently with regard to their origin: HCD from the same laboratory that conducted the actual study is treated to be more comparable than HCD compiled from the records of several laboratories. Furthermore, HCD that undergone a toxicological peer-review process is favored over HCD that is not. The least favorable source of HCD is data that is published elsewhere [Keenan et al. 2009].

For many laboratories HCD is available inhouse. For example, the Evans Analytical Group LLC avian toxicology laboratory maintains a data base with data for avian reproduction studies for the past 40 years [Valverde-Garcia et al. 2018]. Igl et al. 2019 used HCD regarding the rat bone marrow micronucleus test that was comprised of data obtained by four collaborating laboratories between 2001 and 2016. Nevertheless, in some cases HCD might not be available inhouse and hence has to be obtained from elsewhere (e.g. if a laboratory starts its work with a certain type of bioassay).
Tumor rates of rats and mice of the untreated control groups of long term carcinogenicity studies are publicly available from the NTP Historical Controls Database [NTP 2021]. Another source for HCD of rats and mice (and hamsters as well) is the Registry of Industrial Toxicology Animal database (RITA) which is maintained by the Fraunhofer Institute in Hannover, Germany [Deschl et al. 2002]. But, contrary to the NTP reports, this database is freely accessible for members of the RITA project as well as for members of certain research societies of toxicological pathology only [RITA 2021]. Anyhow, both data bases check their data for plausibility prior to publication by using an internal peer review [Keenan et al. 2009].
Contrary to toxicology studies on mammals, HCD is not frequently used in ecotoxicology where a verification of the impact of chemicals from antropogenic sources on wildlife is of interest (e.g. pesticides, biocides, veterinary medicines and pharmaceuticals). Anyhow, in this field of research, the use of HCD is the objective of an ongoing discussion. Several fields of application of HCD in ecotoxicology such as avian reproduction studies which are a regulatory requirement for pesticides [Valverde-Garcia et al. 2018], fish full life cycle studies or studies with non-target terrestrial plants are reviewed in Brooks et al. 2019. Furthermore, the application of HCD to study egg hatching success and larval immobilization of the calanoid copepod *Acartia tonsa* are reported by Rotolo et al. 2021.
Another field of research where HCD plays a role is pre-clinical safety assessment such as the detection of anti-drug antibody (ADA) cut points in immunogenicity assays [Hoffmann and Berger 2011]. The cut points are calculated based on a (historical) set of non-responders in order to classify future specimens into ADA positive responders or negative non-responders [Schaarschmidt et al. 2015]. Please note, that further information about the application of HCD with regard to bioassays can be found in Kluxen et al. 2021.

1.5 Statistical methods for the use of HCD reported in literature

Despite the fact that there seems to be broad agreement about the kind of historical data that has to be collected to be valid enough for verification of actual trials, there is little guidance about the methods and applications of HCD [Brooks et al. 2019]. Hence, several different approaches can be found in literature.

1.5.1 Inclusion of HCD to the statistical test procedure

The inclusion of HCD in statistical tests dates back to the early 1980ies [Tarone 1982] and was adopted by several authors such as Kitsche et al. 2012. Anyhow, comparable procedures are frequently applied in the context of clinical trials [Viele et al. 2014], but seem to play a minor role in the context of (eco)toxicological bioassays and hence, are beyond the scope of this thesis.

1.5.2 Informal graphical comparison between HCD and the actual data

The informal comparison between HCD and the observations obtained from the actual trial (figure 1.2A) as done in the motivating example (see 1.1) is in line with recommendations found in literature regarding carcinogenicity studies [Keenan et al. 2009]. Anyhow, other authors recommend the depiction of historical control data as a boxplot in order to give an overview about the properties of its empirical distribution [Elmore and Peddada 2009]. Another form of graphical comparison was proposed for ecotoxicological studies with aquatic mesocosms [Brooks et al. 2019]. This kind of studies are used to evaluate the impact of plant protection products on aquatic ecosystems. Since several model organisms are studied simultaneously over a longer period (e.g. one year) the results of the untreated control group can be highly variable. Hence, Brooks et al. 2019 used plots of model based predictions (mean curves) and their confidence intervals that represent historical control data and compared them to the observed growth curves of an actual trial. Based on this visual comparison, they concluded that the abundance of *Daphnia* was unusually low in their actual trial and hence, comparisons between the actual *Daphnia* control and their corresponding treatment groups might be misleading.

1.5.3 Calculation of intervals that define the level of "normal" background variation

Generally, three types of different statistical intervals [l,u] are reviewed in literature [Hahn et al. 2017] and are commonly used in practical applications:

1. Confidence intervals
2. Tolerance intervals
3. Prediction intervals

Confidence intervals are computed based on an observed sample y in order to contain a parameter or other property θ of the unknown population Y with a predefined coverage probability $1-\alpha$ such that $P(l \leq \theta \leq u) = 1-\alpha$. Since the estimation of the parameter estimate $\hat{\theta}$ gets more precise and its standard error decreases with increasing amount of information (sample size) also the width of the confidence interval decreases with an increase of information. Consequently, also the probability that a confidence interval covers a future observation, instead of the desired population parameter, decreases with an increase of historical observations.
Based on an observed sample, tolerance intervals are computed in order to contain a proportion of the units from the unknown populations with coverage probability $P(l \leq \gamma(Y) \leq u) = 1-\alpha$. In this notation $\gamma(\cdot)$ is the proportion of units from the unknown population Y. The use of tolerance intervals based on (historical) control data is discussed for special applications like the definition of ADA cut points [Hoffmann and Berger 2011]. Nevertheless, tolerance intervals are beyond the scope of this thesis and hence, are not considered below.
Prediction intervals (PI) are computed based on an observed sample y in order to contain one or more future observations y^* (or some function of them) with coverage probability $P(l \leq y^* \leq u) = 1-\alpha$. Prediction intervals can either contain $M = 1$ future observation, $M > 1$ future observations, K out of M future observations or the mean of M future observations. Since, in the context of HCD, prediction is usually made on the level of future observations rather than on their functions, only PI for $M \geq 1$ future observations are considered in the following sections.
Most of the literature about the application of HCD in toxicology provides simple methods to calculate intervals (or cut points) that define the level of "normal" background variation such as the historical range, the mean $\pm$ one or two times the standard deviation or simple confidence intervals for the historical mean [Elmore and Peddada 2009, Greim et al. 2003]. Furthermore, the use of boxplots as described above implicitly provides cut points such as the lower and the upper empirical quartiles of the HCD. It is noteworthy, that all these methods are proposed in order to define the "normal" background variation but what "normal" really means in terms of statistical properties is usually not defined explicitly. Hence, the following paragraphs provide a short review about these methods and their statistical properties (see figure 1.2).
Since the range reflects simply the minimum and the maximum of the historical control data, it is highly influenced by extreme values. Due to this fact, the range will broaden with a rising amount of historical data, resulting in an interval that will always cover the observations of the actual control, if the number of historical observations is high enough (figure 1.2B). Hence, the coverage probability $P(l \leq y^* \leq u) = 1-\alpha$ is not defined in this approach. Therefore, several authors dissuaded from its use [Elmore and Peddada 2009, Keenan et al. 2009, Greim et al. 2003].
If the box of a boxplot is used as an interval (figure 1.2C), only the central 50 % of the observations of the underlying distribution are considered as "normal" background variation. Hence, such a quartile based interval should treat an actual observation as "abnormal" in 50 % of the cases where it is in line with the HCD, given that the number of historical observations is high enough to properly estimate the quartiles. Therefore, the coverage probability of this method is unclear, especially if the amount of historical observations is low.
The use of confidence intervals in order to define the cut points for comparison between HCD and the actual trial was proposed in the context of long term carcinogenicity studies [Greim et al. 2003]. Anyhow, a confidence interval should encompass a true parameter of the population, rather than additional observations, as is done when HCD is compared with actual data. With rising amount of observations, the estimation of the parameter estimate gets more precise and its estimated standard error decreases. Consequently, the width of a confidence interval decreases with an increase of observations (see figure 1.2D) as well as the chance that a future observation is covered by the confidence interval.

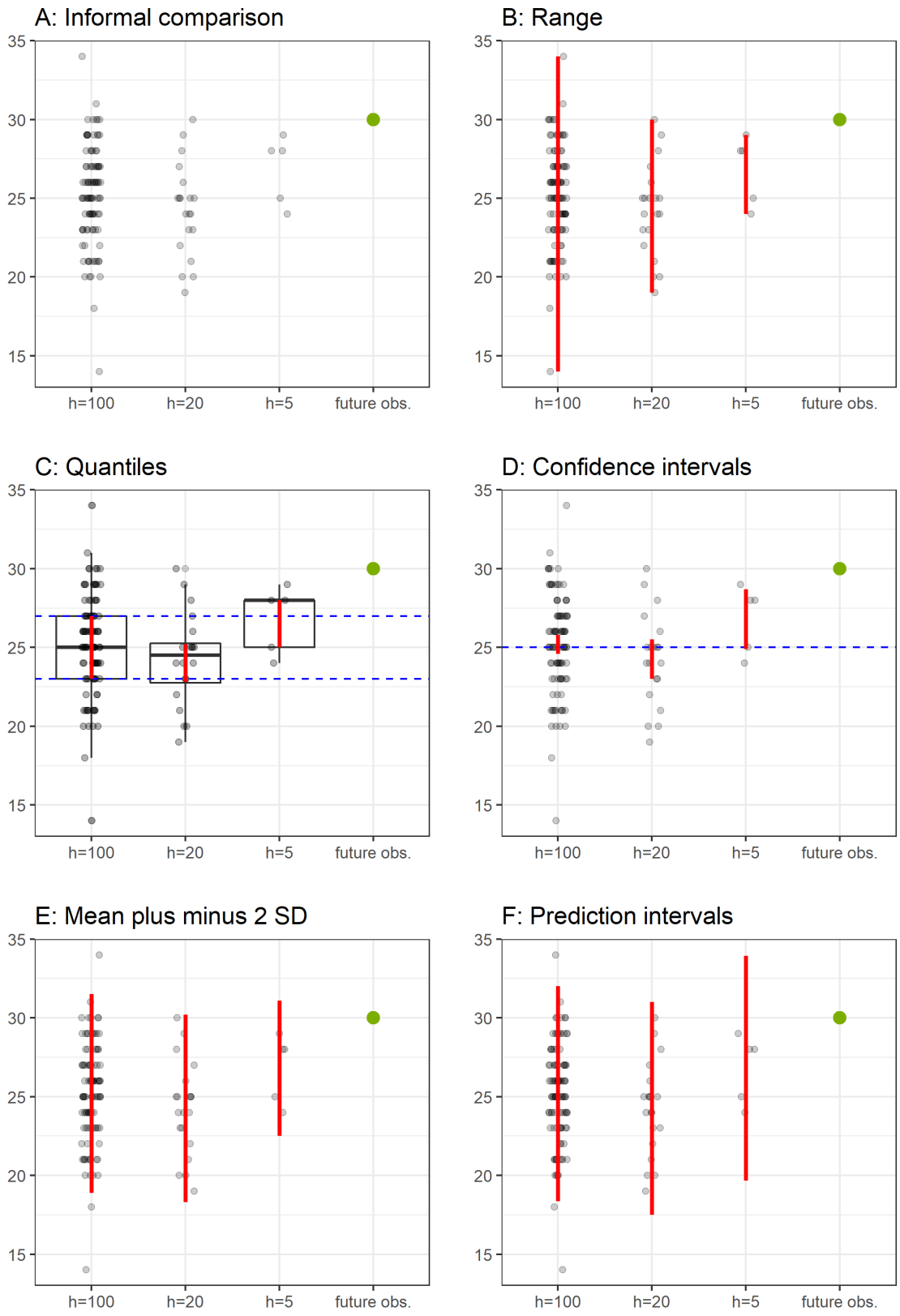

Figure 1.2: Methods for the validation of future data based on HCD. **h:** Number of historical observations; **Grey dots:** Sampled historical data; **Green dot:** Future observation (sampled from the same distribution as the HCD); **Red vertical lines:** Intervals defined by the respective method; **Blue horizontal lines:** Parameters of the underlying distribution used for sampling (mean, Q25, Q75). The data was sampled from a binomial distribution with cluster size 50 and probability of success set to 0.5.

Additionally to the assumption that both, the historical and the actual observations, descend from the same data generating process, intervals calculated as mean ± one or two times the standard deviation (fig-

ure 1.2E) assume that the data is normal distributed. If cut points are calculated as historical mean ± historical standard deviation, the actual observation should be covered in roughly 68 % of the cases. If the cut points are given as historical mean ± two times the historical standard deviation, the cut points will cover the actual control in roughly 95 % of the cases. This kind of interval was proposed in the context of long term carcinogenicity studies for the evaluation of tumor rates [Elmore and Peddada 2009, Keenan et al. 2009]. But, rates descend from a binomial process (see section 1.6.1) and are usually not normal distributed. Anyhow, according to the central limit theorem, these intervals should cover the actual tumor rate with the desired probability, if computed based on a high number of historical observations and if the rates are close to 0.5, the midpoint of their parameter space. The closer the rates get to 0 or 1, the more skewed the underlying distribution gets and hence, the coverage probability of this interval decreases. Hence its application to real life data can not be recommended.
The only statistical interval that covers future observations with a predefined probability (usually 95 %) is a prediction interval (figure 1.2F). It seems that prediction intervals are barely used in toxicology since many authors are not aware of that method [Greim et al. 2003, Elmore and Peddada 2009, Keenan et al. 2009, Rotolo et al. 2021]. Nevertheless, the use of prediction intervals for the validation of actual trials as well as reference values from guidelines was recently proposed for avian reproduction studies [Valverde-Garcia et al. 2018]. Contrary to (eco)toxicology where the application of PI seem to play a minor role, the use of prediction limits is discussed in the context of immunogenicity assays as one possible method to define anti-drug antibody cut points that distinguish between ADA responders and non-responders [Hoffmann and Berger 2011, Schaarschmidt et al. 2015]. For this purpose, the ADA reaction of non-responders is measured simultaneously together with samples of unknown status. Then an upper prediction limit is computed based on the observations from the non responders. If the observed ADA reaction of a sample with unclear status exceeds this limit it is treated to descend from a responder. Anyhow, the prediction limits found in literature are only available for several special cases but are not applicable in a general way.

1.6 Data types and model assumptions

The mayor assumption that is made for the comparison of HCD with actual data using prediction intervals is, that both kinds of observations descend from the same data generating process. Depending on the type of study, the scale of the observations of interest can differ from each other.

1.6.1 Dichotomous data

If the endpoints are dichotomous such as numbers of rats with tumors vs. number of rats without tumors, one can assume that this kind of data is binomial distributed such that

$$\begin{aligned} y_h &\sim bin(\pi, n_h) \\ var(y_h) &= n_h\pi_(1-\pi_h). \end{aligned} \tag{1.1}$$

with and $E(y_h) = n_h\pi$. In this notation π is the binomial proportion, n_h is the size of $h = 1 \dots H$ clusters (e.g. number of individuals in the hth historical study) and y_h are the number of successes obtained from the individuals of the hth cluster (e.g. rats with tumors).
Anyhow, most of the biological data that is assumed to be binomial has higher variability than possible under binomial distribution and hence exhibits extra binomial variation which is also called overdispersion [McCullagh and Nelder 1989, Demetrio et al. 2014]. Overdispersion can be caused by several reasons such as positive correlations between the individual experimental units (e.g. if the average body weight differs between treatment groups and the chance of tumor induction rises with body weight). The opposite effect that the variance of the data is smaller than binomial variance is called underdispersion. Anyhow underdispersion is thought to be implausible in biological data since it would be caused by negative correlation between experimental units (e.g. if one animal dies, the remaining animals live unusually long). Further details on that topic are given in Demetrio et al. 2014. There are two approaches to model overdispersion: The quasi-likelihood approach (which is also called quasi-binomial) and the beta-binomial assumption.

The first assumes a dispersion parameter that constantly inflates the variance for all observations such that

$$var(y_h)^{QB} = \phi^{QB} n_h\pi(1-\pi)$$

with $E(\pi_h) = \pi$ and $E(y_h) = n_h\pi$ and $\phi^{QB} > 1$. For the latter, the data is assumed to be beta-binomial

distributed such that

$$\begin{aligned} \pi_h &\sim beta(a,b) \\ y_h &\sim bin(\pi_h, n_h) \\ var(y_h)^{BB} &= n_h\pi(1-\pi)[1+(n_h-1)\rho] \end{aligned} \tag{1.2}$$

with $E(\pi_h) = \pi = a/(a+b)$, $E(y_h) = n_h\pi$ and $\rho = 1/(1+a+b)$. In this case the dispersion parameter $\phi_h^{BB} = [1+(n_h-1)\rho]$ depends on the cluster size n_h. Please note that ϕ_h^{BB} becomes constant if all of the H clusters have the same size such that $n_h = n_{h'} = n$. In this special case the quasi-likelihood approach and the beta-binomial assumption can not be distinguished from each other.

1.6.2 Count data

In several bioassays the variable of interest is comprised of count data such as eggs per hen in avian reproduction studies [Valverde-Garcia et al. 2018]. A natural approach for modeling counts is to assume them to be Poisson distributed

$$\begin{aligned} y_h &\sim Pois(\lambda) \\ E(y_h) &= var(y_h) = \lambda \end{aligned}$$

Here, y_h are the counts in several historical studies e.g. y_h eggs, counted in $h = 1 \dots H$ historical studies and λ is the Poisson mean. Similar to binomial distributed data overdispersion is usually present in such data and can be modeled as follows [Gsteiger et al. 2013, Demetrio et al. 2014]: The quasi-likelihood approach (also called quasi-Poisson) assumes that a constant dispersion parameter inflates the variance, such that

$$var(y_h)^{QP} = \phi^{QP}\lambda$$

with $\phi^{QP} > 1$ and $E(y_h) = \lambda$. Another approach for modeling overdispersed Poisson data is the negative-binomial distribution where the means of the historical studies follow a gamma distribution with parameters a and b, such that

$$\begin{aligned} \lambda_h &\sim gamma(a,b) \\ y_h &\sim Pois(\lambda_h) \\ var(y_h)^{NB} &= \lambda + \kappa\lambda^2 = \lambda(1+\kappa\lambda) \end{aligned}$$

with $E(y_h) = \lambda = a/b$ and $\kappa = 1/a$. Here, the dispersion parameter is $\phi^{NB} = (1+\kappa\lambda)$. Please note that in the case in which several counted observations y_h only vary around their expected value λ, both, the quasi-Poisson and the negative-binomial assumption are not in contradiction with each other because the dispersion parameters ϕ^{QP} and ϕ^{NB} are constant.

1.6.3 Continuous data

Several continuous measurements such as eggshell thickness or the ADA reaction can be assumed to be (log-)normal, such that

$$y_h \sim N(\mu, \sigma)$$

with y_h as the observations in $h = 1, \dots H$ historical studies, μ as the mean and σ as the standard deviation. This model is frequently considered in literature regarding statistical intervals [Hahn et al. 2017, Hothorn et al. 2009, Igl et al. 2019]. But, for applications (eg. assays regarding ADA reaction) where the data is influenced by several random factors, this model is far to simple.
If, for example, two samples from different patients were taken and the different patients were treated in different hospitals, the observations (samples) are systematically influenced by the random factors "patients" and "hospital" such that the corresponding model is given by

$$y_{ijk} = a_i + b_{j(i)} + e_{k(ij)}$$

In this case the data should be modeled by using random effects models where the total variance of the data is split into several variance components that correspond to the random factors such that $a_i \sim N(0,\sigma_a^2)$ are the random effects obtained for the hospotals, $b_{j(i)} \sim N(0,\sigma_b^2)$ are the random effects for the patients and $e_{k(ij)} \sim N(0,\sigma_e^2)$ are the residuals. A detailed description of random effects models, as well the application of PI to data that is based on such models, is given in section 2.3.

1.7 The use of prediction intervals

The idea of the application of prediction intervals dates back to 1941 where Satterthwaite proposed "confidence limits within which we may expect an additional item" as one of the motivating examples for his widely used approximation of degrees of freedom. Since then, several prediction intervals were proposed for dichotomous data, count data and for normal distributed data as well [Hahn et al. 2017]. Anyhow, these prediction intervals need further adaptions, since they do not consider several sources of variability that frequently occur in real life data.

1.7.1 Dichotomous data

Several methods for the calculation of prediction intervals for dichotomous data that ground on the binomial distribution (see eq. 1.1) were proposed in literature and are reviewed in Hahn et al. 2017 as well as in the supplementary materials of Menssen and Schaarschmidt 2019. Anyhow, none of these methods reflect the fact that HCD is usually comprised of more than one historical study. Furthermore, they do not consider overdispersion and hence, yield coverage probabilities below the nominal level, if overdispersion is present in the data (see section 2.2).
According to Hahn et al. 2017, an asymptotic prediction interval for one future binomial distributed observation y^* is based on the assumption that

$$\frac{\hat{y}-Y}{\sqrt{\widehat{var}(\hat{y}-Y)}} = \frac{n^*\hat{\pi}-Y}{\sqrt{\widehat{var}(n^*\hat{\pi}-Y)}} = \frac{n^*\hat{\pi}-Y}{\sqrt{\widehat{var}(n^*\hat{\pi})+\widehat{var}(Y)}} \tag{1.3}$$

can be approximated by a standard normal distribution. In this notation $\hat{\pi}$ is the binomial proportion estimated from the historical observations, n is the historical cluster size and n^* is the size of the future cluster. The corresponding prediction interval for one future binomial observation is given by

$$[l,u] = n^*\hat{\pi} \pm z_{1-\alpha/2}\sqrt{n^*\hat{\pi}(1-\hat{\pi})\Big(1+\frac{n^*}{n}\Big)} \tag{1.4}$$

with $\widehat{var}(Y) = n^*\hat{\pi}(1-\hat{\pi})$ and $\widehat{var}(n^*\hat{\pi}) = n^{*2}\widehat{var}(\hat{\pi}) = n^{*2}[\hat{\pi}(1-\hat{\pi})/n]$. Please note, that this interval was proposed for the application to one single historical study, rather than to HCD that is comprised of several studies. Therefore, this PI neglects the effect of overdispersion that might occur in the data (e.g. due to positive correlations between the individuals inside each historical study). Hence, the interval given in eq. 1.4 needs further adaptions.
An asymptotic prediction interval for one future observation based on the quasi-binomial assumption and observations from $h = 1,\dots H$ historical clusters is given by

$$[l,u] = n^*\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\phi^{QB}n^*\hat{\pi}(1-\hat{\pi})\Big(1+\frac{n^*}{\sum_{h=1}^{H} n_h}\Big)} \tag{1.5}$$

with $\phi^{QB} > 1$ if $\widehat{var}(Y)$ in eq. 1.3 is substituted by $\widehat{var}(Y)^{QB} = \phi^{QB}n^*\hat{\pi}(1-\hat{\pi})$.

If $\widehat{var}(Y)^{BB} = n^*\hat{\pi}(1-\hat{\pi})[1+(n^*-1)\hat{\rho}]$ is substituted into eq. 1.3, an asymptotic PI based on the beta-binomial assumption can be given by

$$[l,u] = n^*\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\Big(n^*\hat{\pi}(1-\hat{\pi})\big[1+(n^*-1)\hat{\rho}\big]\Big)\Big(1+\frac{n^*}{\sum_{h=1}^{H} n_h}\Big)} \tag{1.6}$$

in which $\widehat{var}(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{N} + \frac{N-1}{N}\hat{\pi}(1-\hat{\pi})\hat{\rho}$ with $N = \sum_{h=1}^{H} n_h$ [Moore 1987]. Another prediction interval that is based on the beta-binomial assumption can be given by an approach in which the parameters of the beta-binomial distribution are estimated from the historical data. Then the PI is given by the quantiles of the estimated beta-binomial distribution, such that

$$[l,u] = [q_{\alpha/2}(\hat{a},\hat{b},n^*), q_{1-\alpha/2}(\hat{a},\hat{b},n^*)] \tag{1.7}$$

with $q_{\alpha/2}(\cdot)$ as the $\alpha/2$-quantile of the beta-binomial distribution with parameters $\hat{a},\hat{b}$ and n^*.

1.7.2 Count data

Several methods for the calculation of prediction intervals for Poisson distributed count data are reviewed in Hahn et al. 2017. An asymptotic PI for one Poisson distributed future observation y^* is based on the assumption that

$$\frac{\hat{y} - Y}{\sqrt{\widehat{var}(\hat{y} - Y)}} = \frac{n^*\hat{\lambda} - Y}{\sqrt{\widehat{var}\left(n^*\hat{\lambda} - Y\right)}} = \frac{n^*\hat{\lambda} - Y}{\sqrt{\widehat{var}(n^*\hat{\lambda}) + \widehat{var}(Y)}} \tag{1.8}$$

is approximately standard normal [Hahn et al. 2017] and the corresponding asymptotic prediction interval is given by

$$[l, u] = n^*\hat{\lambda} \pm z_{1-\alpha/2}\sqrt{n^*\hat{\lambda}\left(1 + \frac{n^*}{n}\right)}. \tag{1.9}$$

Please note, that in this notation n is the number of units (eg. $n = 5$ hens) the historical observations y (eg. eggs) are based on. The Poisson mean is estimated using n as an offset such that $\hat{\lambda} = y/n$ and n^* is the future number of units (e.g. $n^* = 3$ hens). The adaption to historical data comprised of $h = 1, \ldots, H$ studies in which overdispersion is caused by correlations of the units between the studies is given by

$$[l, u] = n^*\hat{\lambda} \pm z_{1-\alpha/2}\sqrt{\hat{\phi}^{QB} n^*\hat{\lambda}\left(1 + \frac{n^*}{\sum_{h=1}^{H} n_h}\right)} \tag{1.10}$$

with $\hat{\phi}^{QB} > 1$.

1.7.3 Continuous data

Several prediction intervals are published in the context of normal distributed continuous data. According to Hahn et al. 2017, a PI for one future observation y^* that is based on one normal distributed historical sample is given by

$$[l, u] = \hat{\mu} \pm t_{1-\alpha/2, df=n-1}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n}\right)} \tag{1.11}$$

with $\hat{\mu}$ as the mean of the historical observations, n as the historical sample size and $\hat{\sigma}^2$ as the historical variance and $t_{1-\alpha/2, df=n-1}$ as the $1-\alpha/2$ quantile of the t-distribution with $n-1$ degrees of freedom.
Anyhow, this interval is far to simple in most of the practical applications where the observations are influenced by several random factors (e.g. different laboratories testing different patients). Hence, such data is usually modeled based on random effects models as described in section 2.3.
Prediction intervals based on random effects models can be calculated based on three different methods. The oldest method is the calculation of a PI that grounds on parameter estimates that are estimated based on mean squares and approximate degrees of freedom following Satterthwaite 1941. Other possibilities for the calculation of PI based on random effect models are generalized prediction intervals that ground on generalized pivotal quantities [Lin and Liao 2008, Al-Sarraj et al. 2019] or prediction intervals based on restricted maximum likelihood (REML) estimation as proposed by Francq et al. 2019. Anyhow, all of these methods have some drawbacks that limit their use in practical applications: The first two methods are published only for special cases and are not available in a general way that is easy to apply by a user who is not trained in programming and statistics. The third method is easy to apply in a general way, but lacks a proper approximation of the degrees of freedom that are associated with the estimation of the prediction variance.

1.8 Bootstrap calibration

Generally, a statistical interval that should encompass the variable of interest θ with nominal coverage probability

$$\Psi = P\big(l(\alpha) \leq \theta \leq u(\alpha)\big) = 1 - \alpha \tag{1.12}$$

can be expressed as

$$\big[l(\alpha), u(\alpha)\big] = \hat{\theta} \pm q_{1-\alpha/2}\sqrt{\widehat{var}(\hat{\theta})}. \tag{1.13}$$

In this notation $\hat{\theta}$ is the estimate for θ that was estimated based on the sample y and $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a certain distribution the interval grounds on (usually standard normal, t or χ^2). If θ is a parameter of the unknown population, eq. 1.13 represents a confidence interval. If θ is a proportion of the unknown population, eq. 1.13 represents a tolerance interval. A prediction interval is given if θ is a future observation.
Anyhow, in certain situations such as the application of asymptotic intervals (as described in sections 1.7.1 and 1.7.2) based on small sample sizes, the true coverage probability of the interval might not match the nominal level $1-\alpha$. A remedy for this problem can be the application of bootstrap calibration.

1.8.1 Alpha calibration

Usually bootstrap calibration focuses on the α with which the interval is computed and aims to find a coefficient δ that minimizes the difference between the the empirical coverage probability $\hat{\Psi} = P\big(l(\delta) \leq \theta \leq u(\delta)\big)$ and the nominal coverage probability $\Psi = 1-\alpha$.

The corresponding algorithm is

1. Generate $b = 1, \ldots, B$ parametric bootstrap samples based on $\hat{\theta}$ and $\widehat{var}(\hat{\theta})$, the parameter estimates of the original sample.
2. For each of the bootstrap samples estimate $\hat{\theta}_b$ and $\widehat{var}(\hat{\theta})_b$.
3. Calculate B intervals of interest $\big[l(\delta), u(\delta)\big]_b = \hat{\theta}_b \pm q_{1-\frac{\delta}{2}} \sqrt{\widehat{var}(\hat{\theta})_b}$
4. Calculate the empirical coverage probability $\hat{\Psi} = \frac{\sum_b I_b}{B}$ with
$$I = 1 \text{ if } \hat{\theta} \in \big[l(\delta), u(\delta)\big]_b$$
$$I = 0 \text{ if } \hat{\theta} \notin \big[l(\delta), u(\delta)\big]_b$$
5. Alternate δ until $\hat{\Psi}$ is satisfactory close to the nominal level of $\Psi = 1-\alpha$
6. Calculate the calibrated interval based on the initial sample estimates $\hat{\theta}$ and $\widehat{var}(\hat{\theta})$ and the value of δ for which $\hat{\Psi} - \Psi$ is minimal.

This kind of bootstrap calibration is usually referred to as alpha calibration and was proposed by Loh 1987 and reviewed in Efron and Tibshirani 1994. Please note, that a similar algorithm can be applied to intervals that are directly based on quantiles of a certain distribution (such as the PI given in eq. 1.7). Since then it was used by many authors for several fields of application such as multiple testing [Fan et al. 2007] or the calibration of confidence intervals for conformance proportions [Lee and Liao 2012]. Anyhow, the calibration algorithm given above, can be adapted to other problems.

1.8.2 Other forms of bootstrap calibration

In the context of tolerance intervals, the parameter $\gamma(\cdot)$ that defines the proportion of the unknown population that should be covered by the interval, can be calibrated in a similar fashion as α. Schuetzenmeister and Piepho 2012 provide an algorithm in which $\gamma(\cdot)$ is calibrated in order to yield simultaneous tolerance bounds for studentized residuals.
If an interval should be computed based on random effects models, it usually grounds on a quantile of the t-distribution for which the degrees of freedom are approximated following Satterthwaite 1941. Alternatively, the algorithm given above can be adapted in order to approximate the degrees of freedom associated with the standard error used for interval calculation such that

$$\big[l(\delta), u(\delta)\big] = \hat{\theta} \pm t_{1-\alpha/2, df=\delta} \sqrt{\widehat{var}(\hat{\theta})}. \tag{1.14}$$

Finally, bootstrap calibration can be applied in order to calibrate the whole quantile the interval is based on, such that

$$\big[l(\delta), u(\delta)\big] = \hat{\theta} \pm \delta \sqrt{\widehat{var}(\hat{\theta})}. \tag{1.15}$$

1.9 Software for the calculation of prediction intervals in R

R-code for the implementation of statistical methodology can be provided in several ways such as

1. Code snippets in the text of scientific publications
2. R-files or packages on GitHub
3. R-packages from the Comprehensive R Archive Network (CRAN)

Since the majority of R users only use packages that are provided via CRAN but do not download code or packages from GitHub (or elsewhere) [Wickham and Bryan 2021] and each package on CRAN has passed several quality checks, the following three sections distinguish between the source of R code that provides functions for the calculation of prediction intervals.
It has to be noted that most of the methodology described below is not mentioned in the manuscripts provided in section 2 because most methods do not match the experimental designs on which this thesis is based on. Anyhow, the following sections give a short overview about the methodology regarding prediction intervals that is implemented in R at the moment.

1.9.1 Code snippets from scientific publications

Code snippets for the calculation of several statistical intervals based on different methods or distributional assumptions are provided in the textbook of Hahn et al. 2017. R code for the calculation of the simple prediction intervals for $M = 1$ future observation based on one binomial or Poisson distributed sample (see eq. 1.4 and 1.8) can be found there. Anyhow, some of the code that is provided by Hahn et al. 2017 depends on functions of an R package that is neither available from CRAN nor from Github, but only as a zip-file from the homepage of their textbook (`wiley.com/go/meeker`). Since the package is not listed in CRAN it is de facto unavailable for many potential applicants such as toxicologists who are not trained in programming. Furthermore, the methodology provided by this package can not be easily used as a dependency for packages written by other programmers.
Francq et al. 2019 provide R code for the calculation of prediction intervals for $M = 1$ future observation based on balanced and unbalanced random effects models in the supplementary material of their paper. Anyhow, the code is not applicable anymore, since it depends on an R package (varComp) that was removed from CRAN December 2017 [CRAN 2021]. Further details about this method and its current implementation are given in section 2.3.

1.9.2 Developmental versions of R packages on Github

Prediction intervals for all of M or K out of M future observations, that are based on one historical sample were proposed by Hothorn et al 2009. Implementations of their methodology are available in the package predIntervals that can be downloaded from Github using the following code

```
devtools::install_github("daniel-gerhard/predIntervals")
```

Please note that this package also provides a generalization of the methods described by Hothorn et al 2009 to linear models fitted with `stats::lm()` that works for both, fixed effects models as well as regression models.
The use of prediction intervals that are based on linear random effects models were proposed in the context of ADA cut-point estimation [Schaarschmidt et al. 2015]. A corresponding package, called mixADA, can be downloaded from Github with

```
devtools::install_github("schaarschmidt/mixADA")
```

This package provides mean square based prediction intervals for $M = 1$ future observation, following the methodology of Satterthwaite 1941. At the moment the application is restricted to five experimental layouts that are commonly used in the context of ADA cut-point estimation (one-way, two-way hierarchical, two-way cross classified with and without interaction and a three-way layout with two factors crossed and one nested). It has to be noted, that the methodology provided by mixADA is only applicable to balanced data.

1.9.3 R packages on CRAN

Methodology for the calculation of prediction intervals that are based on one unstructured sample of random size was proposed by Barakat et al. 2014 for several continuous distributions and was recently implemented in the package PredictionR [Barakat et al. 2020]. Since the package is available from CRAN it can be downloaded with

```
install.packages("PredictionR")
```

Unfortunately, this package is barely documented, because it lacks any vigniette or readme file that provide detailed examples for its application (also the description in the reference manual is rather short). Anyhow, since the prediction intervals proposed by Barakat et al. 2014 depend on one unstructured sample, rather than on a complex experimental design in which the historical data is influenced by several random factors, it is not considered below.
The package predint (see section 2.4) provides prediction intervals as well as upper or lower prediction limits for $M \geq 1$ future observations that are based on linear random effects models or on overdispersed binomial or count data. It can be loaded from CRAN by running the following code:

```
install.packages("predint")
```

The prediction interval that is based on linear random effects models is a direct implementation of the bootstrap calibrated prediction interval that is proposed in 2.3. Therefore it is applicable to balanced and unbalanced experimental layouts as well.
The functions that provide prediction intervals for overdispersed binomial data, are based on the asymptotic PI that were derived in equations 1.5 and 1.6. In order to enable prediction for $M > 1$, also these intervals were bootstrap calibrated in the same way as the one for random effects models. Currently, bootstrap calibrated prediction intervals for $M \geq 1$ future observations based on overdispersed Poisson data are implemented. But, contrary to the interval given in eq. 1.10, the implementation does not consider offsets in its current form. An overview about the functionality of the predint package is given in table 1.1.

Table 1.1: Functions provided by the predint package

Function	**Description**
`lmer_pi()`	Prediction intervals or limits for $M \geq 1$ future observations based on random effects models
`beta_bin_pi()`	Asymptotic prediction intervals or limits for $M \geq 1$ future observations based on the beta-binomial distribution
`quasi_bin_pi()`	Asymptotic prediction intervals or limits for $M \geq 1$ future observations based on the quasi-binomial assumption
`quasi_pois_pi()`	Asymptotic prediction intervals or limits for $M \geq 1$ future observations based on the quasi-poisson assumption
`rbbinom()`	Sampling of beta-binomial data
`rqbinom()`	Sampling of quasi-binomial data
`rqpois()`	Sampling of quasi-poisson data

1.10 Prediction intervals for the motivating example

Figure 1.3A shows exactly the same data as the graphic used as a motivating example (figure 1.1). Furthermore, prediction intervals for future proportions (tumor rates) that are based on the HCD indicate the plausible background variation for $M = 10$ future control groups that are comprized of $n = 10$ female Spraque Dawley rats, each. This approach reflects the experimental design that was used by Seralini et al. 2012 (ten cohorts, each comprised of ten rats). Prediction intervals that are based on the beta-binomial distribution (solid lines) were computed using the `predint::beta_bin_pi()` function. PI using the quasi binomial assumption (dashed lines) were calculated based on `predint::quasi_bin_pi()`. Since it was unclear if the tumor rates reported by Seralini et al. 2012 correspond to the total number of rats that developed a tumor (HCD total) or to the number of rats that developed mammary tumors only (HCD mammary), PI for both data sources were computed.
If the rates reported by Seralini et al. 2012 reflect the total number of rats with tumors (HCD total), the conclusion regarding the tumor rate of the untreated control depends on the distributional assumption: Based on the beta-binomial PI it has to be treated as unusually low. Based on the quasi-binomial assumption it is in line with the rate that can be expected for a cluster size of 10 rats. It is noteworthy that the quasi-binomial PI for $M = 10$ future observations, each based on cluster size 10, comprises the whole parameter space $[l, u]^{QB} = [0, 1]$.
If the rate of mammary tumors is considered, the control falls into the prediction intervals regardles of the distributional assumption. Or, in other words: The tumor rates reported by Seralini et al. 2012 are in line with the range that can be expected for 10 clusters, each of size 10, and hence, are simply caused by random variation. Anyhow, also here the beta-binomial prediction interval is slightly shorter than the one that is based on the quasi-binomial assumption.
The main problem that occurs if one wants to interpret the tumor rates of female Spraque Dawley rats given by Seralini et al. 2012 is caused by the low number of rats per treatment group (cluster size of 10): The estimation of tumor rates based on such a small amount of information is heavily imprecise which is reflected by the extraordinary wide prediction intervals. This leads to the conclusion, that the differences between the tumor rates reported by Seralini et al. 2012. are caused by random variation only.
In order to demonstrate the effect of the cluster size, prediction intervals for 10 future clusters comprised of 50 rats each, are given in figure 1.3B. The higher cluster size has lead to prediction intervals that are considerably shorter than the ones computed based on a cluster size of 10. Anyhow, another factor that influences the width of the prediction intervals, is the number of future clusters. If a prediction interval for $M = 5$ instead of $M = 10$ future clusters (each of size 50) is computed based on the HCD for the total tumor rate (HCD total), the lower interval border increases to 0.546. If the PI is calculated for only one future observation, the lower border is 0.665 (and the upper border remains to be 1). This two effects show, that the experimental design of Seralini et al. 2012 (high number of clusters, low cluster size) was unsuitable for any profound statement regarding the tumor incidence caused by Roundup or Roundup-tolerant GM-maize.
The R-code used for this analysis is available on GitHub `https://github.com/MaxMenssen/menssen_2021_dissertation`.

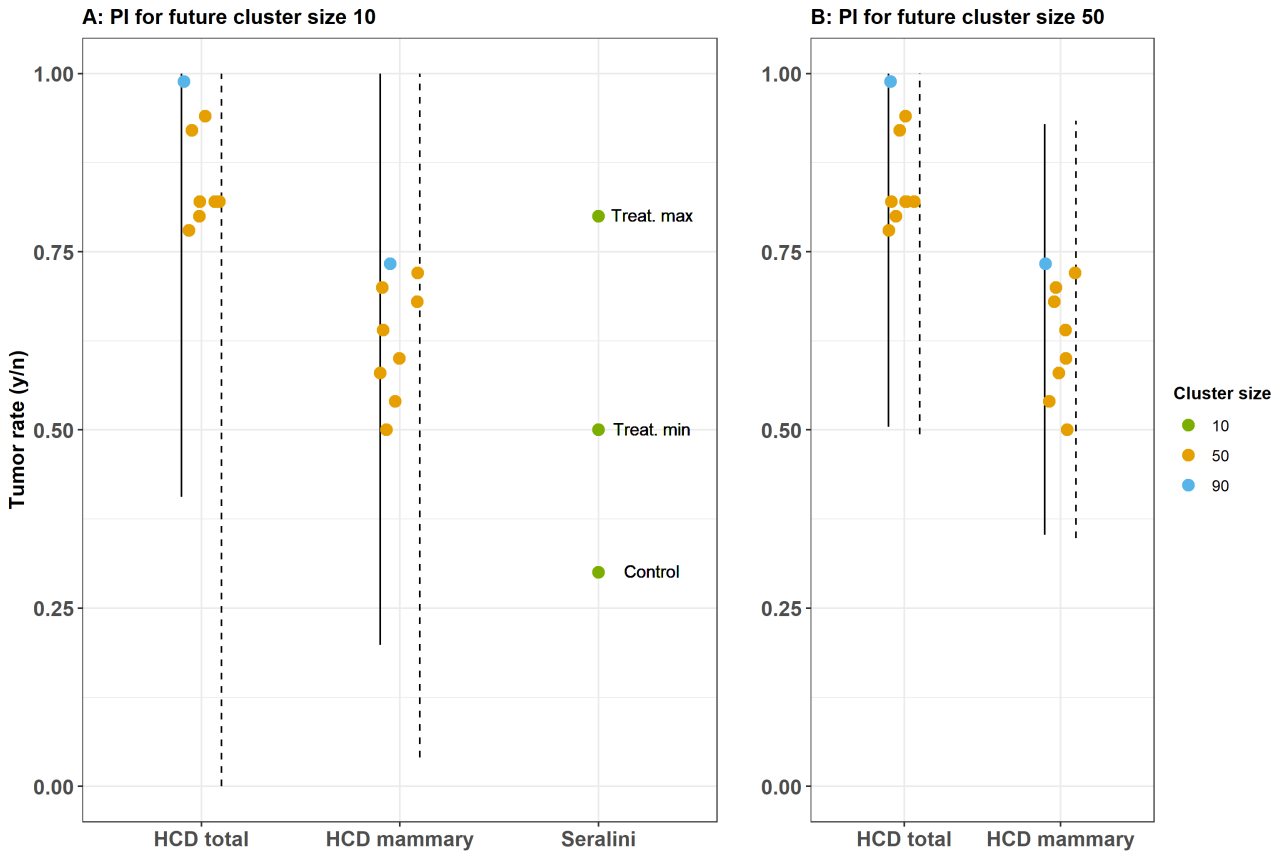

Figure 1.3: **A**: Prediction intervals for 10 future observations each of cluster size 10, based on historical control data for female Spraque Dawley rats obtained from the NTP Historical Controls Database [NTP 2021]. **B**: Prediction intervals for 10 future observations each of cluster size 50 using the same HCD as in A. **HCD total:** Total numbers of tumors; **HCD mammary:** Mammary gland (Fibroma, Fibroadenoma, Carcinoma, or Adenoma); **Seralini:** Tumor rates reported by Seralini et al. 2012; **Treat. max:** Maximum tumor rate of the treatment groups reported by Seralini et al. 2012; **Treat. min:** Minimum tumor rate of the treatment groups reported by Seralini et al. 2012; **Control:** Tumor rate of the untreated control group reported by Seralini et al. 2012; **Cluster size:** Number of rats (n) inside each treatment or control group; **Tumor rate:** Number of rats with tumor (y) divided by its corresponding cluster size (n); **Solid vertical lines:** Prediction intervals based on the beta-binomial assumption; **Dashed vertical lines:** Prediction intervals based on the quasi-binomial assumption.

1.11 Contributions to the field

Prediction intervals for one future observation based on overdispersed binomial data are proposed in section 2.1. For that purpose the asymptotic prediction interval for binomial data given in eq.1.3 was adapted to overdispersion using the quasi-likelihood assumption (see eq. 1.5). Another PI that was proposed is based on quantiles of the beta-binomial distribution (see eq. 1.7). Furthermore, both intervals were alpha-calibrated using an approach in which δ was alternated based on a bisection algorithm (see step 5 of the calibration process described in section 1.8.1). The coverage probability of the proposed PI was assessed based on Monte-Carlo simulations. The simulated coverage probabilities of the proposed PI were substantially closer to the nominal $1-\alpha = 0.95$ than the ones for prediction intervals that do not consider overdispersion (for a review of PI for binomial distributed data that neglect overdispersion see section 2.2).
Two prediction intervals based on linear random effects models and the estimation of variance components via REML are proposed in section 2.3. The first PI grounds on generalized Satterthwaite approximation of degrees of freedom and is only applicable to balanced data and usable for the prediction of one future observation. The second PI is based on a quantile calibration bootstrap and hence, is applicable to balanced and unbalanced data as well. Furthermore the calibration bootstrap enables its application also for cases where more than one future observation should be predicted ($M \geq 1$). The coverage probability of the two proposed prediction intervals as well as for four existing PI found in literature was assessed via Monte-Carlo simulation. It could be shown that the simulated coverage probabilities of the quantile calibrated PI approached the nominal $1-\alpha = 0.95$ in most of the cases or was at least comparable to the PI proposed by other researchers.
Up to 2021 there was no R-package available from CRAN, that provides user friendly implementations of prediction intervals for $M \geq 1$ future observations based on overdispersed binomial or count data as well as on linear random effects models. This gap was filled with the upload of the predint package for which the reference manual is given in section 2.4. By now, methodology for the reevaluation of bioassays on the base of prediction intervals is available and is easy to apply for other researchers (as demonstrated in the motivating example).

1.12 Conclusions and future research

The present work was focused on prediction intervals that are either based on linear random effects models or on overdispersed binomial and count data. Prediction intervals for one future observation based on overdispersed binomial data were proposed in section 2.1. Furthermore, an alpha-calibration bootstrap procedure, which was later adapted for the purpose of quantile-calibration, is described in that section.
This quantile-calibration bootstrap was proposed in order to yield prediction intervals for $M \geq 1$ future observations based on linear random effects models. Monte-Carlo simulations regarding the coverage probability of the proposed PI are given in section 2.3.
The quantile calibration bootstrap was also applied to yield the prediction intervals for overdispersed binomial and count data that are provided by the R-package predint. As far as I know, these are the only publicly available prediction intervals that are applicable to that kind of data. Anyhow, Monte-Carlo simulations regarding their coverage probabilities are not available yet and are the concern of future work.
Several sources of historical control data such as the NTP reports or the HCD regarding avian reproduction described by Valverde-Garcia 2018 are based on summary statistics such as the mean and standard deviation of the historical control groups, rather than on the original raw data. This approach might increase the amount of residual variance that can not be explained, because several random factors such as cages or pens by which a certain amount of variation could be explained are not present in the historical data anymore (due to the averaging). Anyhow, in several laboratories or research institutes, the raw data of historical trials might be available inhouse. Hence, prediction intervals that ground on generalized random effects models, in which a predictor that is comprised of binomial or count data is modeled based on several random factors, are the matter of the ongoing research.

1.13 Bibliography

[Al-Sarraj et al. 2019] Al-Sarraj R, von Brömssen C, Forkmann J (2019): Generalized prediction intervals for treatment effects in random-effects models. Biometrical Journal. 61:1242-1257, DOI:10.1002/bimj.201700255

[Barakat et al. 2014] Barakat HM, El-Adll ME, Aly AE (2014): Prediction intervals of future observations for a sample of random size from any continuous distribution. Mathematics and Computers in Simulation. 97:1-13, DOI:10.1016/j.matcom.2013.06.007

[Barakat et al. 2020] Barakat HM, Khaled OM, Ghonem HA (2020): PredictionR: Prediction for Future Data from any Continuous Distribution. R package version 1.0-12. https://CRAN.R-project.org/package=PredictionR

[Brooks et al. 2019] Brooks AC, Foudoulakis M, Schuster HS, Wheeler JR (2019): Historical control data for the interpretation of ecotoxicity data: are we missing a trick? Ecotoxicology. 28:1198-1209, DOI: 10.1007/s10646-019-02128-9

[Bucher 2006] Bucher JR. (2006): The National Toxicology Program rodent bioassay. Annals of the New York Academy of Sciences. 982(1):198-207. DOI:10.1111/j.1749-6632.2002.tb04934.x

[CRAN 2021] `https://CRAN.R-project.org/package=varComp`, visited 16.9.2021

[Demetrio et al. 2014] Demetrio CGB, Hinde J, Moral RA (2014): Models for overdispersed data in entomology. In: Ferreira CP, Godoy WAC (eds.) Ecological modelling applied to entomology. Cham, Switzerland: Springer International Publishing; 219-259, DOI:10.1007/978-3-319-06877-0_9

[Deschl et al. 2002] Deschl U, Kittel B, Rittinghausen S, Morawietz G, Kohler M, Mohr U, Keenan C (2002): The value of historical control data - scientific advantages for pathologists, industry and agencies. Toxicologic Pathology. 30(1):80-87, DOI: 10.1080/01926230252824743

[Efron and Tibshirani 1994] Efron B, Tibshirani RJ (1994): An introduction to the bootstrap. Chapman and Hall, New York. DOI:10.1201/9780429246593

[Elmore and Peddada 2009] Elmore AS, Peddada SD (2009): Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. Toxicologic Pathology. 37(5):672-676. DOI: 10.1177/0192623309339606

[Fan et al. 2007] Fan J, Hall P, Yao Q (2007):To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? Journal of the American Statistical Association. 102(480):1282-1288, DOI:10.1198/016214507000000969

[Francq et al. 2019] Francq B.G., Lin D., HoyerW. (2019): Confidence, prediction, and tolerance in linear mixed models. Statistics in Medicine. 38:5603-5622, DOI:doi.org/10.1002/sim.8386

[Greim et al. 2003] Greim H, Gelbke H-P, Reuter U, Thielmann HW, Edler L (2003): Evaluation of historical control data in carcinogenicity studies. Human and Experimental Toxicology. 22:541-549, DOI:10.1191/0960327103ht394oa

[Gsteiger et al. 2013] Gsteiger S, Neuenschwander B, Mercier F, Schmidli H (2013): Using historical control information for the design and analysis of clinical trials with overdispersed count data. Statistics in Medicine. 32:3609-3622, DOI:10.1002/sim.5851

[Hahn et al. 2017] Hahn JG, Meeker WQ, Escobar LA (2017): Statistical intervals. Second Edition. Wiley and Sons Inc. Hoboken, Jew Jersey, USA

[Hasemann 1995] Hasemann JK (1995): Data analysis: Statistical analysis and use of historical control data. Regulatory Toxicology and Pharmacology. 21:52-59, DOI: 10.1006/rtph.1995.1009

[Hoffmann and Berger 2011] Hoffman D, Berger M (2011): Statistical considerations for calculation of immunogenicity screening assay cut points. Journal of Immunological Methods. 373:200-208, DOI: 10.1016/j.jim.2011.08.019

[Hothorn et al. 2009] Hothorm LA, Gerhardt D, Hofmann M (2009): Parametric and non-parametric prediction intervals based phase II control charts for repeated bioassay data. Biologicals. 37:323-330, DOI:10.1016/j.biologicals.2009.07.001

[Igl et al. 2019] Igl B-W, Bitsch A, Bringezu F, Chang S, Dammann M, Frötschl R, Harm V, Kellner R, Krzykalla V, Lott J, Nern M, Pfuhler M, Queisser N, Schulz M, Sutter A, Vaas L, Vonk R, Zellner D, Ziemann C (2019): The rat bone marrow micronucleus test: Statistical considerations on historical negative control data. Regulatory Toxicology and Pharmacology. 102:13-22, DOI:10.1016/j.yrtph.2018.12.009

[Keenan et al. 2009] Keenan C, Elmore S, Francke-Carroll S, Kemp R, Kerlin R, Peddada S, Pletcher J, Rinke M, Schmidt SP, Taylor I, Wolf DC (2009): Best practices for use of historical control data of proliferative rodent lesions. Toxicologic Pathology. 37:679-693, DOI:10.1177/0192623309336154

[Kitsche et al. 2012] Kitsche A, Hothorn LA, Schaarschmidt F (2012): The use of historical controls in estimating simultaneous confidence intervals for comparisons against a concurrent control. Computational Statistics and Data Analysis. 56(12):3865-3875, DOI:10.1016/j.csda.2012.05.010

[Kluxen et al. 2021] Kluxen FM, Weber K, Strupp C, Jensen SM, Hothorn LA, Garcin JC, Hofmann T (2021): Using historical control data in bioassays for regulatory toxicology. Regulatory Toxicology and Pharmacology. 125:105024, DOI:10.1016/j.yrtph.2021.105024

[Lee and Liao 2012] Lee HI, Liao CT (2012): Estimation for conformance proportions in a normal variance components model. Journal of Quality Technology. 44:63-79, DOI:10.1080/00224065.2012.11917882

[Lin and Liao 2008] Lin TY, Liao CT (2008): Prediction intervals for general balanced linear random models. Journal of Statistical Planning and Inference. 138(10):3164-3175, DOI:10.1016/j.jspi.2008.01.001

[Loh 1987] Loh W-Y (1987): Calibrating confidence coefficients. Journal of the American Statistical Association. 82:155-162, DOI:10.1080/01621459.1987.10478408

[McCullagh and Nelder 1989] McCullagh P, Nelder JA (1989): Generalized Linear Models. 2nd Edition, Chapman and Hall, London. DOI:10.1007/978-1-4899-3242-6

[Menssen and Schaarschmidt 2019] Menssen M, Schaarschmidt F (2019): Prediction intervals for overdispersed binomial data with application to historical controls. Statistics in Medicine. 38:2652-2663, DOI:10.1002/sim.8124

[Moore 1987] Moore, DF (1987): Modelling the extraneous variance in the presence of extra-binomial variation. Journal of the Royal Statistical Society. 36(1)8-14, DOI:10.2307/2347840

[NTP 2021] NTP historical controls report by route and vehicle Harlan Sprague-Dawley rats. `https://ntp.niehs.nih.gov/data/controls/index.html`, visited 12.7.2021

[RITA 2021] Homepage of the Registry of Industrial Toxicology Animal-data. `https://reni.item.fraunhofer.de/reni/public/rita/#public_access`, visited 26.7.2021

[Rotolo et al. 2021] Rotolo F, Vitiello V, Pellegrini D, Carotenuto Y, Buttino I (2021): Historical control data in ecotoxicology: Eight years of tests with the copepod Acartia tonsa. Environmental Pollution. 284:117468, DOI: 10.1016/j.envpol.2021.117468

[Satterthwaite 1941] Satterthwaite FE (1941): Synthesis of variance. Psychometrika. 6(5):309-316, DOI:10.1007/BF02288586

[Schaarschmidt et al. 2015] Schaarschmidt F, Hofmann M, Jaki T, Grün B, Hothorn LA (2015): Statistical approaches for the determination of cut points in anti-drug antibody bioassays. Journal of Immunological Methods. 418:84-100, DOI:10.1016/j.jim.2015.02.004

[Schützenmeister and Piepho 2012] Schützenmeister A, Piepho H-P (2012): Residual analysis of linear mixed models using a simulation approach. Computational Statistics & Data Analysis. 56(6):1405-1416, DOI:10.1016/j.csda.2011.11.006

[Seralini et al. 2012] Seralini G-E, Clair E, Mesnage R, Gress S, Defarge N, Malatesta M, Hennequin D, Vendomois JS (2012): RETRACTED: Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. Food and Chemical Toxicology. 50(11):4221-4231, DOI: 10.1016/j.fct.2012.08.005

[Seralini et al. 2014a] Seralini G-E, Clair E, Mesnage R, Gress S, Defarge N, Malatesta M, Hennequin D, Vendomois JS (2014): Retraction notice to "Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize" [Food Chem. Toxicol. 50 (2012) 4221-4231]. Food and Chemical Toxicology 63:244. DOI: 10.1016/j.fct.2012.08.005

[Seralini et al. 2014b] Seralini G-E, Clair E, Mesnage R, Gress S, Defarge N, Malatesta M, Hennequin D, Vendomois JS (2014): Republished study: Long-term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. Environmental Sciences Europe. 26:14, DOI: 10.1186/s12302-014-0014-5

[Seralini et al. 2014c] Seralini G-E, Clair E, Mesnage R, Gress S, Defarge N, Malatesta M, Hennequin D, Vendomois JS (2014): Conclusiveness of toxicity data and double standards. Food and Chemical Toxicology. 69:357-359. DOI:10.1016/j.fct.2014.04.018

[Tarone 1982] Tarone RE (1982): The use of historical control information in testing for a trend in proportions. Biometrics. 38:215-220, DOI:10.2307/2530304

[Viele et al. 2014] Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, Hobbs B, Ibrahim JG, Kinnersley N, Lindborg S, Micallef S, Roychoudhury S, Thompson L (2014): Use of historical control data for assessing treatment effects in clinical trials. Pharmaceutical Statistics. 13(1):41-54, DOI:10.1002/pst.1589

[Valverde-Garcia et al. 2018] Valverde-Garcia P, Springer T, Kramer V, Foudoulakis M, Wheeler JR (2018): An avian reproduction study historical control database: A tool for data interpretation. Regulatory Toxicology and Pharmacology. 92:295-302, DOI:10.1016/j.yrtph.2017.12.004

[Wickham and Bryan 2021] R packages: organize, test, document and share your code. O'Reilly. Work-in-progress 2nd edition. `https://r-pkgs.org/`, visited 9.8.2021

[Yamagiwa and Ichikawa 1918] Yamagiwa K, Ichikawa K (1918): Experimental study of the pathogenesis of carcinoma. The Journal of Cancer Research. 3(1):1-29, DOI:10.1158/jcr.1918.1

[Young et al. 2016] Young DS, Gordon CM, Zhu S, Olin BD (2016): Sample size determination in strategies for normal tolerance intervals using historical data. Quality Engineering. 28(3):337-351, DOI:10.1080/08982112.2015.1124279

Chapter 2

Publications and Manuscripts

2.1 Prediction intervals for overdispersed binomial data with application to historical controls

Max Menssen[1], Frank Schaarschmidt[1]

1. Institut für Zellbiologie und Biophysik, Abteilung Biostatistik, Leinbiz Universität Hannover, Herrenhäuser Str. 2, 30419 Hannover

Type of authorship:	First author
Type of article:	Research article
Journal:	Statistics in Medicine
Impact factor:	2.373
Number of citations:	1 (web of science)
DOI:	10.1002/sim.8124

Contributions

Max Menssen

1. Derivation of the two proposed prediction intervals for overdispersed binomial data
2. Implementation of all prediction intervals
3. Design and implementation of the Monte-Carlo simulations
4. Writing the paper

Frank Schaarschmidt

1. Derivation of the two proposed prediction intervals for overdispersed binomial data
2. Design of the Monte-Carlo simulations
3. Writing the paper

Received: 20 July 2018 | Revised: 24 January 2019 | Accepted: 25 January 2019

DOI: 10.1002/sim.8124

RESEARCH ARTICLE

WILEY Statistics in Medicine

Prediction intervals for overdispersed binomial data with application to historical controls

Max Menssen | Frank Schaarschmidt

Abteilung Biostatistik, Institut für Zellbiologie und Biophysik, Leibniz Universität Hannover, Hannover, Germany

Correspondence
Max Menssen, Abteilung Biostatistik, Institut für Zellbiologie und Biophysik, Leibniz Universität Hannover, 30419 Hannover, Germany.
Email: menssen@cell.uni-hannover.de

Present Address
Max Menssen, Abteilung Biostatistik Herrenhäuser Straße 2, 30419 Hannover

Bioassays are highly standardized trials for assessing the impact of a chemical compound on a model organism. In that context, it is standard to compare several treatment groups with an untreated control. If the same type of bioassay is carried out several times, the amount of information about the historical controls rises with every new study. This information can be applied to predict the outcome of one future control using a prediction interval. Since the observations are counts of success out of a given sample size, like mortality or histopathological findings, the data can be assumed to be binomial but may exhibit overdispersion caused by the variability between historical studies. We describe two approaches that account for overdispersion: asymptotic prediction intervals using the quasi-binomial assumption and prediction intervals based on the quantiles of the beta-binomial distribution. Both interval types were α-calibrated using bootstrap methods. For an assessment of the intervals coverage probabilities, a simulation study based on various numbers of historical studies and sample sizes as well as different binomial proportions and varying levels of overdispersion was run. It could be shown that α-calibration can improve the coverage probabilities of both interval types. The coverage probability of the calibrated intervals, calculated based on at least 10 historical studies, was satisfactory close to the nominal 95%. In a last step, the intervals were computed based on a real data set from the NTP homepage, using historical controls from bioassays with the mice strain B6C3F1.

KEYWORDS

alpha-calibration bootstrap, beta-binomial, bioassay, extra binomial variation, quasi-binomial

1 | INTRODUCTION

Bioassays are standard procedures for the assessment of the toxicological properties of chemical compounds. Such trials are carried out under standardized conditions, using well-known model organisms that are exposed to increasing dosages of a certain compound. Potential hazardous effects can be assessed by comparing dosage groups vs an untreated control. In many cases, the treatment effect on the model organism is quantified like "rats with and without tumors" or "presence and absence" of disease symptoms or histopathological findings. Hence, we assume the outcome of such studies to be counts of two categories and therefore a dichotomous variable.

Due to restricted sample size, the inference drawn from these comparisons may have high rates of type-2 error. Because the observed outcome of the study is a random variable, it may occur that the mean of the control group is extraordinarily low, resulting in a test statistic that supports significant differences to the treatment groups although the substance under control is not hazardous (type-1 error). In other words, this would mean that a safe substance is treated falsely to be harmful. On the other hand, it could be possible that the outcome of the control is such high, and that the test statistic could not state any significant differences to the treatment group, although in reality there are differences and the compound is hazardous (type-2 error), meaning that a harmful compound is considered to be safe. Because the aim of most bioassays is to show that a chemical substance is not hazardous, high rates of type-2 error have to be avoided. However, the conclusion drawn from such studies depends strongly on the outcome of the control group, leading to a considerable lack of power, especially if the baseline proportion and the effect size is small (see figure (1) of the supplementary material).

If the same type of bioassay is carried out several times using the same model organism, the amount of historical information regarding the untreated control groups rises. This pool of information can be used to be compared with the outcome of the actual control.

Simple strategies of summarizing historical control data are the computation of the range, the mean plus-minus the standard deviation or a 95% confidence interval for the historical mean.[1] However, these approaches have several disadvantages: Since the range is sensitive for extreme values, one single historical control with an extremely high or low outcome can change the inference drawn from this method dramatically. An example for this is given by Keenan et al.[2] Due to the robustness against extreme values, Elmore and Peddada[3] proposed an informal comparison of the actual control group against the quantiles and the interquartile range (IQR) of the historical control groups. Neither the inference drawn from the comparison of the actual control to the historical range (or IQR) nor the conclusion using the historical mean plus-minus the standard deviation or quantiles considers any statistical error. Moreover, if the number of historical studies is rather limited, the sampling error of quantiles leading to the IQR will be high. In addition, confidence intervals for the expected mean of the historical data are inappropriate for comparisons with actual or future observations. Since a confidence interval reflects only the variability of the mean and not the variability of the data itself, it becomes more narrow if the number of historical studies is rising. Therefore, it will be more likely that a confidence interval excludes the future observation than defined by the nominal $1 - \alpha$.

One appropriate way for taking historical information into account are Bayesian methods. Tarone[4] provided a Bayesian procedure to test for a trend in proportions that incorporates historical control data. This approach was adapted by Kitsche et al[5] to yield simultaneous confidence intervals for multiple contrasts between the actual proportions of the treatment groups and the control. For this approach, the proportions were modeled with a beta-prior of which the parameters were estimated from historical control data. Another approach that incorporates historical control data in a trend test is given by Peddada et al.[6] Furthermore, Leon-Novelo et al[7] proposed a Bayesian procedure, incorporating historical control data as prior information, to determine whether an association between the exposure to the chemical compound and tumor incidence exists or not.

An alternative method to account for historical information is the computation of prediction intervals that provide lower and upper limits to encompass a single future observation with a prespecified probability. Several approaches for prediction of one future binomial random variable based on one historical sample are given in the literature. An exact procedure for interval calculation was proposed by Faulkenberry[8] and was applied by Bain and Patel[9] to binomial random variables. This approach yields a hypergeometric distribution, for which they gave a normal approximation. A widely used large sample approximation interval was given by Hahn and Nelson.[10] This interval was reviewed by Nelson[11] and Hahn and Meeker.[12] Wang[13] developed a method to calculate the coverage probability of the Nelson interval for a given set of parameters. Based on this calculation, he proposed an α-calibration method for the Nelson interval. Furthermore, Wang[14] published a prediction interval that was constructed similar to the score confidence interval for the binomial proportion given by Wilson.[15] Another prediction interval that is based on the joint sampling distribution of both, the actual and the historical observation, was proposed by Krishnamoorthy and Peng.[16]

The prediction intervals mentioned above may be disadvantageous for application in bioassays because they are based on the assumption that only one historical sample is available, which is assumed to follow the binomial distribution. However, experimental conditions may differ between historical studies, such that variability of historical control counts may exceed the variability assumed in the binomial distribution (overdispersion or extrabinomial variability[17]). Several different causes for overdispersion, like positive correlation between individual responses or variability between experimental units, are reviewed in the literature.[18,19] A simplistic attempt to deal with historical controls from several studies might be to pool several historical studies treating them as one sample and construct the prediction interval based on this pooled sample using the binomial methods above. If the data is indeed overdispersed, the resulting intervals will tend to

be too narrow, resulting in coverage probabilities lower than the nominal level $(1 - \alpha)$. In other words, for the intervals that are supposed to exclude a future value in only 5% of the cases, such an exclusion may occur considerably more often if substantial overdispersion is ignored. Section 3 of the supplementary material illustrates this problem. Thus, prediction interval methods are needed that account for possibly overdispersed binomial data.

More general approaches for the calculation of prediction intervals are based on predictive likelihood methods or on quantiles derived from the predictive distribution for which Beran[20] applied the idea of α-calibration, described, eg, in the works of Efron and Tibshirani[21] or Loh.[22] This idea was used by Hall et al[23] to yield improved prediction intervals resulting in comparable coverage probabilities for both approaches. A bootstrap-based approach for α-calibration of quantile-based prediction intervals for discrete observations was described by Fonseca et al.[24] Although several general approaches for the calculation of prediction intervals are given, none of them has been explicitly described and validated for the application to overdispersed binomial data and sample sizes that are typical for toxicological applications.

In the following, we describe and compare methods for the computation of prediction limits that are based on historical control data which is composed of several studies possibly exhibiting overdispersion. We start with a computationally simple extension of the interval given by Hahn and Nelson[10] to overdispersed binomial data. We will use the approach of α-calibration via bootstrap to improve the coverage probability of this interval. Furthermore, we will show a bootstrap-based approach that is based on α-calibrated quantiles of a beta-binomial distribution. The coverage probability of these prediction intervals is assessed by a simulation study for a broad range of parameter settings, reflecting realistic scenarios in toxicological applications. Finally, the methods are applied to real data sets containing historical controls from bioassays using the mice strain B6C3F1. R-code to compute the proposed intervals as well as the data set regarding our simulation results is available on github: https://github.com/MaxMenssen/Prediction-intervals.

2 | MATERIAL AND METHODS

2.1 | Notation

Let x_k be the number of successes out of the historical samples of size n_k with $k = 1, \dots K$ as the index of the particular study and K as the total number of studies available. Furthermore, let y denote the number of successes out of the future sample of size m. We assume that x_k and y are derived from the same distribution with the unknown proportion π. Our objective is to determine prediction interval limits $[l, u]$ such that $P(l \leq y \leq u) = 1 - \alpha$. Computing the limits $[l, u]$ on the scale of a future count y is in line with the literature concerning prediction intervals for binomials. However, because the future sample size m is assumed to be a known quantity, one may divide the limits $[l, u]$ by m to yield limits $[l/m, u/m]$ of a prediction interval for a future observed proportion, y/m, for all methods described in the following.

2.2 | Overdispersed binomial data

Two different approaches are known to model overdispersed binomial data: The quasi-binomial approach assumes the variance to be $var^{QB}(x_k) = \phi^{QB} n_k \pi(1 - \pi)$ with ϕ^{QB} as a constant dispersion parameter that inflates the binomial variance independently of the sample size. In this parametrization, the binomial assumption is fulfilled, if $\phi^{QB} = 1$.

On the other hand, it is possible to model overdispersed count data by sampling from the beta-binomial distribution. In this approach, the binomial proportions π_k descend from the $Beta(a, b)$ distribution resulting in different proportions for each of the K historical studies such that

$$x_k \sim Bin(n_k, \pi_k) \quad \text{with} \quad \pi_k \sim Beta(a, b),$$

where $E(\pi_k) = \pi = a/(a + b)$ and $E(x_k) = n_k a/(a + b)$. Then, the beta-binomial variance of the counts is expressed by

$$\begin{aligned} var^{BB}(x_k) &= \phi_k^{BB} n_k \pi(1 - \pi) \quad \text{with} \\ \phi_k^{BB} &= 1 + (n_k - 1)\rho \quad \text{and} \\ \rho &= \frac{1}{1 + a + b}, \end{aligned}$$

with ρ being the intra class correlation.[25]

If all n_k are equal, ϕ_k^{BB} is a common factor for all observations x_k, meaning that the quasi-binomial assumption is not in contradiction with the mean-variance relation of the beta-binomial distribution.[26] Or, in other words, in this special case,

$\phi_k^{BB} = \phi^{BB} = \phi^{QB} = \phi$. To achieve data with a predefined amount of dispersion, given all n_k are equal, we set $n_k = n$ resulting in

$$a + b = \frac{(\phi - n)}{(1 - \phi)}, \tag{1}$$

$$a = \pi(a + b). \tag{2}$$

2.3 | The Nelson interval for binomial data (Nelson)

Assuming the absence of overdispersion, the prediction interval given by Nelson[11] is based on the assumption that the distribution of the prediction error is approximately normal

$$\frac{\hat{y} - y}{\sqrt{var(\hat{y} - y)}} \sim N(0, 1),$$

with $\hat{y} = m\hat{\pi}$ being the predicted number of successes out of m future observations. Assuming binomial distribution ($\phi = 1$), one might pool

$$\hat{\pi} = \frac{\sum_{k=1}^{K} x_k}{\sum_{k=1}^{K} n_k}. \tag{3}$$

The prediction interval is given by $\hat{y} \pm z_{1-\alpha}\sqrt{var(\hat{y} - y)}$ with $var(\hat{y} - y) = var(y) + var(\hat{y})$. In the binomial assumption that does not incorporate overdispersion, the variance of y is $var(y) = m\pi(1 - \pi)$ and $var(\hat{y}) = \frac{m^2\pi(1-\pi)}{\sum_{k=1}^{K} n_k}$. Hence, the Nelson prediction interval, applicable to data sets with more than one historical study, is given by

$$\hat{y} \pm z_{1-\alpha/2}\sqrt{m\hat{\pi}(1 - \hat{\pi})\left(1 + \frac{m}{\sum_{k=1}^{K} n_k}\right)}.$$

2.3.1 | Adaption of the Nelson interval to overdispersed data (Nelsonphi, Nelsonphi1)

If overdispersion is taken into account to calculate an improved Nelson interval, based on the quasi-binomial assumption, the variance terms change to $var(y) = \phi^{QB} m\pi(1 - \pi)$ and $var(\hat{y}) = \frac{\phi^{QB} m^2\pi(1-\pi)}{\sum_{k=1}^{K} n_k}$. Therefore, the variance for $\hat{y} - y$ becomes

$$var(\hat{y} - y) = \phi^{QB}\left[m\pi(1 - \pi) + \frac{m^2\pi(1 - \pi)}{\sum_{k=1}^{K} n_k}\right].$$

For the calculation of the interval, the unknown parameters are substituted by their estimates $\hat{\pi}$ and $\hat{\phi}^{QB}$. Following McCullagh and Nelder,[27] $\hat{\phi}^{QB}$ was estimated from the data set as follows:

$$\hat{\phi}^{QB} = \frac{1}{K - 1}\sum_{k=1}^{K}\frac{(x_k - n_k\hat{\pi})^2}{n_k\hat{\pi}(1 - \hat{\pi})}. \tag{4}$$

Therefore, the interval is given by

$$\hat{y} \pm z_{1-\alpha/2}\sqrt{\hat{\phi}^{QB}\left[m\hat{\pi}(1 - \hat{\pi}) + m^2\hat{\pi}(1 - \hat{\pi})\frac{1}{\sum_{k=1}^{K} n_k}\right]}.$$

In the following sections of this paper, this interval is mentioned as the Nelsonphi-interval.

If all $x_k = n_k$ or all $x_k = 0$, $\hat{\pi}(1 - \hat{\pi})$ becomes 0, consequently, $\hat{\phi}^{QB} = 0$. Thus, the Nelson intervals are not defined in such cases. To overcome this problem heuristically, we used the following correction: If all $x_k = n_k$, we replaced x_1 by $n_1 - 0.5$ and $(n_1 - x_1)$ by $(n_1 - x_1) + 0.5$. If all $x_k = 0$, we replaced x_1 by 0.5 and n_1 by $(n_1 - 0.5)$.

Underdispersion ($\phi < 1$) is thought to be implausible in the toxicological setting we consider because underdispersion would be evoked by negatively correlated events within studies. That is, the death or presence of histopathological findings in one animal would decrease the risk of death or histopathological findings for the remaining animals in the group. Furthermore, the lower limit of variance in the beta-binomial distribution is the binomial variance, meaning that the beta-binomial distribution does not allow to assume or simulate underdispersed data. Additionally, both estimates $\hat{\phi}^{QB}$

and $\hat{\phi}^{BB}$ are known to be biased.[25,27] A simulation for the bias of both estimates is given in section 2 of the supplementary material.

For those reasons, we investigated a modified version of the Nelsonphi-interval denoted as Nelsonphi1, setting $\hat{\phi} = 1$ if underdispersion was observed in the historical data set. The Nelsonphi1-interval is given by

$$[l, u] = \hat{y} \pm z_{1-\alpha/2} \sqrt{\max(1, \hat{\phi}^{QB}) \left[m\hat{\pi}(1-\hat{\pi}) + m^2\hat{\pi}(1-\hat{\pi}) \frac{1}{\sum_{k=1}^{K} n_k} \right]}. \tag{5}$$

2.3.2 | The α-calibrated Nelsonphi1-interval (Nelsonphi1_bisec)

We applied the idea of α-calibration to develop a bootstrap calibrated version of the Nelsonphi1-interval. The idea of this approach is the bootstrap estimation of the coverage probabilities of intervals, which are calculated based on α-calibration values λ instead of using the nominal α. Therefore, we replaced $z_{1-\alpha/2}$ by $z_{1-\lambda/2}$ in Equation (5) and searched for the value of λ that brings the estimate for $P(l < y < u)$ as close as possible to the nominal $1 - \alpha$ coverage probability. In the following section, the estimated λ will be referred to as α^{calib}.

Firstly, we draw $b = 1, \ldots, B$ parametric bootstrap samples. For that purpose, $\hat{\pi}$ and $\hat{\phi}^{QB}$ were estimated from the historical data set $\tau = \{x_k, n_k\}$ according to Equation (3) and Equation (4) and the bootstrap samples $\tau_b^* = \{x_{kb}^*, n_k\}$ were drawn from the beta-binomial distribution as follows: The beta parameters $\hat{a}$ and $\hat{b}$ were estimated by plugging $\hat{\phi}^{QB}$ into Equation (1) and $\hat{\pi}$ into Equation (2). Subsequently, $\hat{a}$ and $\hat{b}$ were applied to the beta-distribution to sample π_{kb}^*. Then, the numbers of success x_{kb}^* were drawn from the binomial distribution using n_k and π_{kb}^*. Simultaneously, a validation sample y_b^* for each of τ_b^* that contained only one beta-binomial observation was drawn with the same mechanism, such that

$$\begin{aligned} x_{kb}^* &\sim Bin\left(n_k, \pi_{kb}^*\right) \quad \text{with} \quad \pi_{kb}^* \sim Beta(\hat{a}, \hat{b}) \\ y_b^* &\sim Bin\left(m, \pi_b^*\right) \quad \text{with} \quad \pi_b^* \sim Beta(\hat{a}, \hat{b}). \end{aligned}$$

Then, for each τ_b^*, $\hat{\pi}_b^*$ and $\hat{\phi}_b^{QB*}$ were estimated according to Equation (3) and Equation (4) and $\hat{y}_b^*$ was calculated as $\hat{y}_b^* = m\hat{\pi}_b^*$.

The second step is the calibration of the interval, conditional on the bootstrap samples. For that purpose, we used a bisection algorithm that evaluates $c = 1, \ldots, C$ calibration values (λ_c) in order to minimize the positive distance between the observed coverage probability $\hat{\Psi}_c$ of the corresponding λ_c-interval and the nominal coverage probability Ψ, until this distance is smaller than or equal to a given tolerance t such that $0 \leq (\hat{\Psi}_c - \Psi) \leq t$.

In each of the c bisection steps, the Nelsonphi1-interval was calculated for each of the B bootstrap samples based on the respective λ_c, such that

$$[l_{bc}, u_{bc}] = \hat{y}_b^* \pm z_{1-\lambda_c/2} \sqrt{\max\left(1, \hat{\phi}_b^{QB*}\right) \left[m\hat{\pi}_b^* \left(1-\hat{\pi}_b^*\right) + m^2\hat{\pi}_b^* \left(1-\hat{\pi}_b^*\right) \frac{1}{\sum_{k=1}^{K} n_{kb}} \right]}. \tag{6}$$

The coverage probability of the λ_c α-calibrated interval $\hat{\Psi}_c$ for covering y_b^* across the B bootstrap samples was calculated as

$$\begin{aligned} &\hat{\Psi}_c = \frac{\sum_{b=1}^{B} I_{bc}}{B}, \quad \text{with} \\ &I_{bc} = 1 \quad if \quad \left(l_{bc} \leq y_b^* \leq u_{bc}\right) \\ &I_{bc} = 0 \quad if \quad \left(l_{bc} > y_b^* \cup u_{bc} < y_b^*\right). \end{aligned}$$

The bisection started by defining an initial search interval for λ_c, such that $\lambda^l < \alpha < \lambda^u$. In each step of the algorithm, the midpoint of the search interval was calculated as follows:

$$\lambda_c = \frac{\lambda^l + \lambda^u}{2}.$$

By updating either λ^l or λ^u with the λ_c calculated in the previous iteration, the search interval was bisected until $(\hat{\Psi}_c - \Psi)$ was minimized to a satisfactory level.

Firstly, start values $\lambda_1 = \lambda^l$ and $\lambda_2 = \lambda^u$ were chosen for which the coverage probabilities $\hat{\Psi}_1$ and $\hat{\Psi}_2$ were estimated as mentioned before. In the next step, λ_3 was calculated as the midpoint of the interval, like

$$\lambda_3 = \frac{\lambda_1 + \lambda_2}{2}, \tag{7}$$

and $\hat{\Psi}_3$ was estimated. If $\hat{\Psi}_3 - \Psi$ was positive, λ_4 was calculated by replacing λ_1 in Equation (7) by λ_3, such that

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

If $\hat{\Psi}_3 - \Psi$ was negative, λ_4 was calculated by replacing λ_2 in Equation (7) by λ_3, such that

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

This iteration process was repeated until $\hat{\Psi}_c - \Psi$ was minimized to a sufficient level and the corresponding λ_c was set to be α^{calib}. Due to the discreteness of the function of $\hat{\Psi}_c - \Psi$, we had to distinguish between six different cases for getting the correct λ_c for setting $\lambda_c = \alpha^{\text{calib}}$:

1. It is possible that both $\hat{\Psi}_1$ and $\hat{\Psi}_2$ are smaller than α. In this case, $\hat{\Psi}_1 - \Psi$ is the smallest value that can be detected with the bisection approach and λ_1 is the calibration value with the coverage probability closest to the nominal α. Therefore, we set
$$\lambda_1 = \alpha^{\text{calib}} \quad \text{if} \quad (\hat{\Psi}_1 - \Psi) < 0, (\hat{\Psi}_2 - \Psi) < 0.$$
2. This is the opposite of case 1 with $\hat{\Psi}_1$ and $\hat{\Psi}_2$ bigger than α. Since λ_2 is the calibration value with the coverage probability closest to the nominal α, we set
$$\lambda_2 = \alpha^{\text{calib}} \quad \text{if} \quad (\hat{\Psi}_1 - \Psi) > 0, (\hat{\Psi}_2 - \Psi) > 0.$$
3. If the difference $\hat{\Psi}_c - \Psi$ is minimized to an adequate level, such that $(\hat{\Psi}_c - \Psi) \in [0, t]$, with t as the given tolerance, stop the iteration process and take the corresponding λ_c to become
$$\lambda_c = \alpha^{\text{calib}} \quad \text{if} \quad (\hat{\Psi}_c - \Psi) \in [0, t].$$
4. The given maximum of iteration steps C is reached and none of the three options mentioned above came true, leading to three different cases.
    - 4.1 If there are one or more λ_c for which the corresponding $(\hat{\Psi}_c - \Psi) \in [0, -t]$, take the smallest λ_c for which this condition is true.
    - 4.2 If $(\hat{\Psi}_{c>1} - \Psi) < -t$ but $(\hat{\Psi}_1 - \Psi) > t$ set $\lambda_1 = \alpha^{\text{calib}}$.
    - 4.3 If some $(\hat{\Psi}_c - \Psi) < -t$ and some $(\hat{\Psi}_c - \Psi) > t$ but none of the $(\hat{\Psi}_c - \Psi) \in [-t, t]$, take the biggest λ_c with $[(\hat{\Psi}_c - \Psi) - t] > 0$.

The last step is the calculation of the calibrated Nelsonphi1-interval by using $\hat{\pi}$ and $\hat{\phi}^{QB}$ estimated from the original sample τ and α^{calib} in stead of the nominal α.

$$[l^{\text{calib}}, u^{\text{calib}}] = \hat{y} \pm z_{1-\alpha^{\text{calib}}/2} \sqrt{\hat{\phi}^{QB} \left[m\hat{\pi}(1-\hat{\pi}) + m^2\hat{\pi}(1-\hat{\pi}) \frac{1}{\sum_{k=1}^{K} n_k} \right]}$$

This interval will be referred to as the Nelsonphi1_bisec-interval in the following sections.

2.4 | Quantile-based prediction intervals (qBB, qBB_bisec)

Since the assumption was made that the future observation y descend from the same distribution as the historical observations x_k, simple prediction intervals $[l, u]$ can be given by the quantiles of the corresponding beta-binomial distribution, such that $l = q(\alpha/2)$ and $u = q(1 - \alpha/2)$. In this case, the overall proportion is given by $\pi = E(\pi_k) = \frac{a}{a+b}$ and $\phi_k^{BB} = 1 + (n_k - 1)\rho$, which depends on n_k and the intra class correlation ρ. Since all n_k are equal, such that $n_k = n$, the dispersion parameter ϕ_k^{BB} is a common factor for all K observations, resulting in $\phi_k^{BB} = \phi^{BB}$. Therefore, we can calculate the interval depending on the estimates for the parameters π and ϕ^{BB} such that

$$[l, u] = \left[q(\alpha/2, m, \hat{\pi}, \hat{\phi}^{BB}), q(1 - \alpha/2, m, \hat{\pi}, \hat{\phi}^{BB}) \right]. \tag{8}$$

The proportion $\hat{\pi}$ can be computed according to Equation (3) and $\hat{\phi}^{BB}$ can be estimated as follows:

$$\hat{\phi}^{BB} = \max[1.001, 1 + (n-1)\hat{\rho}],$$

with $\hat{\rho}$ as an estimator for the intraclass correlation that is calculated according to Lui et al.[28] The calculation of $\hat{\rho}$ is given in the supplementary material. In the following sections, this simple quantile-based interval is referred to as the qBB-interval.

We used the same α-calibration approach as described in Section 2.3.2 for increasing the coverage accuracy of the qBB-interval, except that $[l_{bc}, u_{bc}]$ were computed based on Equation (8) instead of Equation (6). The calibrated interval will be called qBB_bisec-interval in the further sections of this paper.

We used the `qBB()` function from the `gamlss` R-package[29] for the calculation of quantiles from the beta-binomial distribution. Since Rigby and Stasinopoulos[29] worked with a slightly different parametrization, depending on the parameter σ and not on ϕ^{BB}, we used the following equation to convert $\hat{\phi}^{BB}$ into $\hat{\sigma}$:

$$\hat{\sigma} = \frac{\hat{\phi}^{BB} - 1}{n - \hat{\phi}^{BB}}.$$

This reparametrization based on the estimator of Lui[28] was necessary because the estimator for σ implemented in the `gamlss` package showed convergence problems in a nonnegligible proportion of cases in the simulation study.

2.5 | Simulation study

To assess the coverage probabilities of the different prediction intervals under various conditions, a simulation study was carried out for R different parameter settings. For that purpose, historical data sets $\tau_{rs} = \{x_{rsk}, n_{rsk}\}$ were drawn from the beta-binomial distribution with $r = 1 \dots R$ as the index for the parameter setting, $s = 1 \dots S$ as the index for the number of replications, and $k = 1 \dots K$ as the index for the historical studies in each τ_{rs}.

The beta parameters a_r and b_r were calculated according to Equation (1) and Equation (2) using predefined values for ϕ_r, π_r, and $n_r = m_r$. Subsequently, the proportions π_{rsk} were drawn from the beta distribution. Finally, the numbers of success x_{rsk} were taken from the $Bin(n_{rsk}, \pi_{rsk})$ distribution. A validation sample y_{rs} containing only one beta-binomial number of success was drawn from the same sampling process simultaneously with τ_{rs}, such that

$$\begin{aligned} x_{rsk} &\sim Bin(n_{rsk}, \pi_{rsk}) \quad \text{with} \quad \pi_{rsk} \sim Beta(a_r, b_r) \\ y_{rsk} &\sim Bin(m_{rs}, \pi_{rsk}) \quad \text{with} \quad \pi_{rsk} \sim Beta(a_r, b_r) \quad \text{and} \\ & n_{rsk} = n_{rs} = m_{rs}. \end{aligned}$$

Then, for each of the historical data sets τ_{rs}, one prediction interval $[l_{rs}, u_{rs}]$ was computed and the coverage probability $\hat{\Psi}_r$ was calculated as

$$\hat{\Psi}_r = \frac{\sum_{s=1}^{S} I_{rs}}{S} \quad \text{with} \tag{9}$$

$$\begin{aligned} I_{rs} &= 1 \quad \text{if} \quad y_{rs} \in [l_{rs}, u_{rs}] \\ I_{rs} &= 0 \quad \text{if} \quad y_{rs} \notin [l_{rs}, u_{rs}]. \end{aligned}$$

The calculation of the coverage probabilities $\hat{\Psi}_r$ of the different intervals was based on $S = 5000$ independent observations for each of the $R = 256$ combinations of $\phi = (1.01, 1.5, 2.0, 3)$, $\pi = (0.01, 0.05, 0.1, 0.2)$, $n = (30, 50, 100, 150)$, and $k = (5, 10, 20, 100)$.

The start values of the bisection were set to $\lambda_1 = 0.00001$ and $\lambda_2 = 0.3$ and the tolerance was given with $t = 0.003$. The maximum number of iteration steps was $C = 15$. The α-calibration was done based on $B = 1000$ bootstraps. We are aware that $B = 1000$ is a relatively small number of bootstraps, but we chose it with respect for computing time. For an improvement of bootstrap accuracy in practical applications, where computing time is of less importance, we recommend an increased number of bootstraps.

3 | RESULTS

If overdispersion is absent and, hence, the historical data can be assumed to be binomial, the simple Nelson interval works reasonably well, except for $\pi \leq 0.05$, as depicted in figure 3 in the supplementary material. However, as far as ϕ increases, the observed coverage probabilities decrease rapidly.

The simulated coverage probabilities of the intervals that consider for overdispersion are depicted in Figure 1. The coverage probability itself can be assumed to be a binomial proportion because it is drawn from a dichotomous process (see Equation 9). Hence, the standard error of the nominal coverage probability $\Psi = 0.95$ can be computed as $se(\Psi) = \sqrt{\frac{0.95 \cdot 0.05}{5000}} = 0.00308$, which is depicted by the dashed lines in the graphic.

If the dispersion parameter was taken into account, the coverage probabilities came closer to the nominal $1 - \alpha = 0.95$. The coverage probabilities of the Nelsonphi-interval raised with the increase of K but tend to be too liberal for almost all settings we draw simulations from, except for $K = 100$ (overall median coverage probability of 0.950). The lowest coverage probabilities were observed for $K = 5$ and ranged between 0.867 and 0.917. For $K = 10$, the range was found to be 0.904 to 0.937. The coverage probabilities for $K = 20$ lay between 0.932 and 0.960 with a median of 0.937.

Restricting the estimated dispersion parameter to $\hat{\phi} \geq 1$ (Nelsonphi1-interval) improved the coverage probabilities for almost all given proportions, compared to the Nelsonphi-interval. The Nelsonphi1-interval tends to be slightly too conservative, if the historical data was de facto binomial distributed ($\phi = 1.01$) since the coverage probabilities varied around 0.96 in this setting. As far as overdispersion plays a role in the historical data, the interval became far too liberal, but this effect was reduced with an increasing number of historical studies. If the number of historical studies was high ($K = 100$), the observed overall median coverage probability (0.951) approached the nominal 0.95 coverage probability.

The α-calibration of the Nelsonphi1-interval (Nelsonphi1_bisec) results in coverage probabilities, closest to the nominal 0.95. However, if the number of historical studies is low ($K = 5$), the interval remains to be too liberal for higher values of ϕ. If the historical data can be assumed to be practically binomial ($\phi = 1.01$), the interval is slightly too conservative

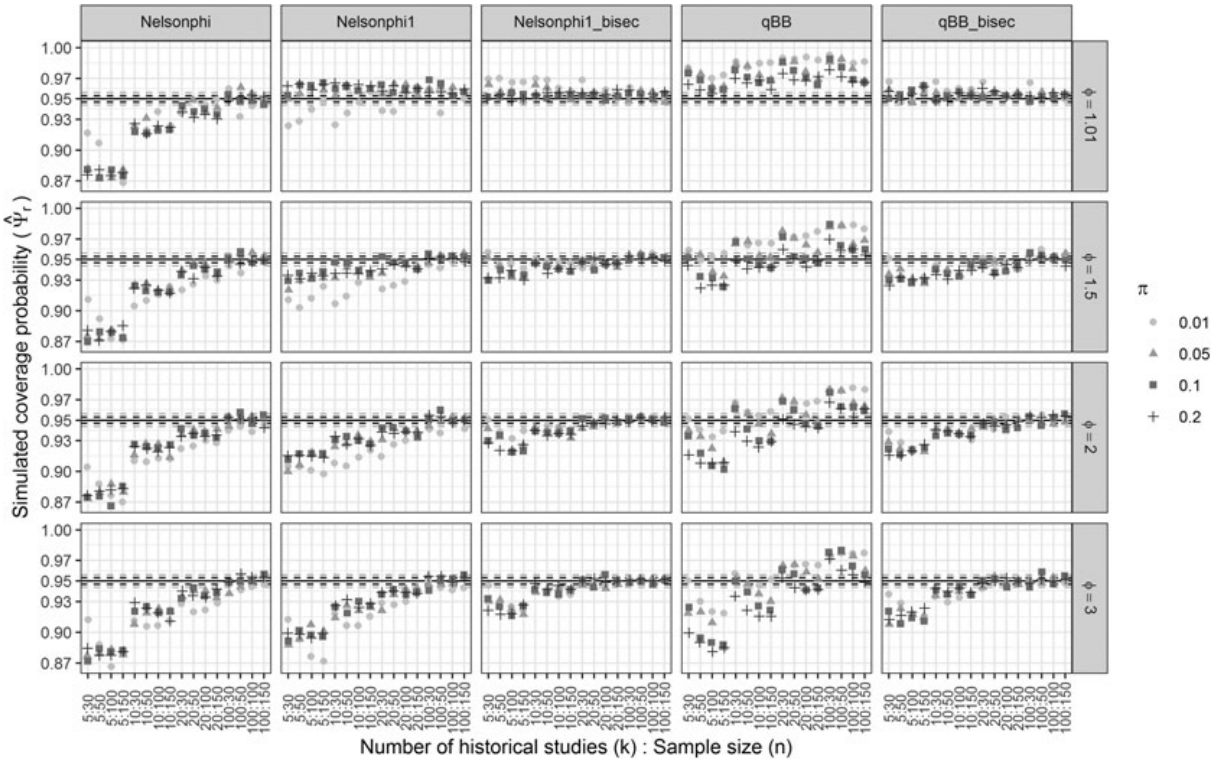

FIGURE 1 Simulated coverage probabilities for the different prediction intervals. Panels: representation of the calculation method on the x-axis and the given amount of overdispersion (ϕ) on the y-axis; colors: different binomial proportions (π); solid line: nominal 0.95 coverage probability (Ψ); black dashed lines: $0.95 \pm se(\Psi)$; gray dashed lines: $0.95 \pm 2se(\Psi)$

for lower numbers of historical studies. With an increasing number of historical studies, this effect reduces, such that for $k = 20$, the nominal coverage probability is almost achieved (overall median coverage of 0.949) and most of the observed coverage probabilities lay in the interval $0.95 \pm 2se$ (gray lines in Figure 1).

If the number of historical studies k is small, the simple quantile-based approach (qBB-interval) results in coverage probabilities that are far smaller or larger than the nominal 95%. For increasing k, the interval becomes too conservative in the settings considered here: For more than $k = 20$ historical studies, the coverage probability is mainly affected by the sample size (n) and the binomial proportion (π) in a way that the coverage probability approaches 0.95 with rising sample size and less extreme proportions. This effect was also present in a simulation using $k = (500, 1000, 5000)$ historical studies (data not shown) and can be explained as follows: As long as π and n_k are small, the lower prediction limit will be 0 with very high probability, such that only the upper prediction limit can actually exclude future counts y. Consequently, the coverage probability will become close to 0.975 for nominal 0.95 intervals. This effect only decreases if either n_k is substantially increased or π is not close to zero but is unaffected by an increasing number of historical controls, k.

The effect of α-calibration on the qBB-interval (qBB_bisec) is comparable to the effect on the Nelsonphi1-interval, except that the coverage probabilities for historical data, in which overdispersion is practically absent ($\phi = 1.01$), are closer to 0.95. On the other hand, for higher dispersion parameters, the qBB_bisec-interval is a little bit more liberal than the calibrated Nelsonphi1-interval.

It has to be noted that the qBB_bisec-interval could not be computed in all 5000 simulation steps of eight parameter settings. Six of these settings were combinations of $N = 30$, $K = 5$, $\phi = \{2, 3\}$, and $\pi = \{0.1, 0.5, 0.1\}$. The remaining two settings were $N = 30$, $\pi = 0.1$, and $\phi = 3$ combined with $K = \{10, 20\}$. In these settings, the computation failed between 2 and 22 times out of the 5000 simulation steps. If the computation of an interval failed, the interval was treated as noncovering and therefore as $I_{rs} = 0$ in Equation (9).

3.1 | Real data example

For the application of the different intervals to real data, the NTP Historical Control Reports from 2013 to 2016 about the mice strain B6C3F1 were downloaded from the NTP homepage.[30] The reports contained long-term studies, started between 2003 and 2011. If a study was present in more than one report, it was considered only once. To get information about rarely, moderately, and frequently occurring events, we analyzed the data about hemangioma, malignant tumors and mortality.

The data were given for different laboratories, sex, and pathways of exposure (gavage-corn oil, gavage-methyl, inoculation-air, oral water, skin-acetone, and skin-ethanol). Each of the studies was conducted using 50 mice per sex in the control group. The animals in the control group were treated similarly to those in the dose groups (same application pathway), except that they were not exposed to the substance under study. To summarize the data, subsets for each combination of pathway and sex, containing studies from different laboratories, were built (Figure 2). For each of the subsets, the binomial proportion $\hat{\pi}$ and the dispersion parameter $\hat{\phi}^{QB}$ were estimated according to Equation (3) and Equation (4). Because only one study was available for the pathways skin-acetone and gavage-methyl, it was impossible to calculate the dispersion parameter for both of the pathways. Therefore, both studies were excluded from further analysis.

Prediction intervals were calculated for the mortality of male mice that were exposed to the inoculation-air pathway. The data set is given in table 1 of the supplemental material. For this data set, the binomial proportion was estimated to be $\hat{\pi} = 0.276$ and the estimate for the dispersion parameter was $\hat{\phi} = 1.308$.

The calculation of the α-calibrated intervals was done using the same start values, maximum number of iteration steps, and tolerance that were used in the simulation study. The only change was an increased number of bootstraps from $B = 1000$ to $B = 10000$ to increase the bootstrap accuracy.

The intervals are given in Table 1. Because the simple Nelson method does not consider overdispersion, it results in the most narrow interval [7.30, 20.29]. Due to the fact that the estimated dispersion parameter is higher than one, the Nelsonphi- and Nelsonphi1-intervals are exactly the same [6.37, 21.23]. The alpha calibration of the Nelsonphi1-interval (Nelsonpi1_bisec) yields the widest interval [6.09, 21.50] using $\alpha^{\text{calib}} = 0.03985$ instead of $\alpha = 0.05$.

4 | DISCUSSION

In literature regarding practical applications, it is common sense that the best approach to assess the toxicity of a chemical compound is the comparison of exposed animals with a concurrent control group.[3] Nevertheless, historical data is seen

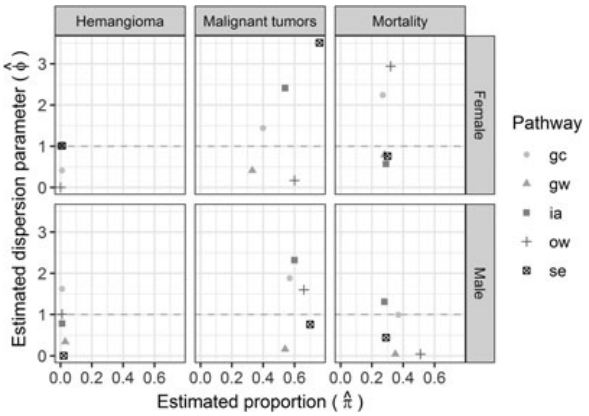

FIGURE 2 Estimated binomial proportion $\hat{\pi}$ and dispersion parameter $\hat{\phi}$ for the mice control data. gc, gavage cornoil; gw; gavage water; ia, inhalation air; ow, oral water; se, skin ethanol

TABLE 1 Prediction intervals for the mortality of male B6C3F1-mice exposed to the inoculation air pathway ($m = 50$)

Names	Lower	Upper	α
Nelson	7.30	20.29	0.05
Nelsonphi	6.37	21.23	0.05
Nelsonphi1	6.37	21.23	0.05
Nelson_bisec	6.09	21.50	0.0399
qBB	7	21	0.05
qBB_bisec	7	21	0.0516

Note that the solution of the two calibrated intervals depend on a bootstrap procedure. Hence, the results of a second run can be slightly different for both intervals.

as an additional source of information, but no consistent way of usage is formulated in guidelines yet.[1] A procedure for using historical control data in a standardized way is the calculation of a prediction interval since it is a method with well-defined statistical properties that reflects both the variability of the mean and the variability of the complete data. In literature, different prediction intervals for dichotomous data can be found. If several historical studies are available, one might pool these studies and apply methods assuming binomial distribution. However, this approach neglects the extravariability that can be noticed between the studies (overdispersion), resulting in prediction intervals with coverage probabilities far less than the nominal $1 - \alpha = 0.95$ (see figure 3 in the supplementary materials). It was possible to show that overdispersion is an important factor that has to be considered because the intervals coverage probabilities came closer to the nominal $1 - \alpha = 0.95$ (Figure 1) or reached that level, if the amount of historical studies was high.

For modeling overdispersion, we used two different approaches: The prediction interval of Hahn and Nelson[10] was extended, assuming an overdispersion parameter that is equal for all observations using the quasi-binomial assumption. In a second approach, prediction intervals are based on the quantiles of the beta-binomial distribution for which the dispersion is assumed to depend also on the sample sizes. The α-calibration of both interval types results in coverage probabilities notably closer to the nominal level of 95% than for the uncalibrated intervals. The coverage probabilities of the proposed intervals depend on the binomial proportion of the events, on the extent of extravariability (overdispersion) and on the number and sample size of available historical control groups. If overdispersion was practically absent, the α-calibrated quantile-based interval yields prediction bounds that reach the nominal coverage probability or that are slightly too conservative. In all other settings, the coverage probabilities of the α-calibrated Nelsonphi1-interval are closer to the nominal level than for the interval that was based on calibrated quantiles. With an increasing number of historical studies, the coverage probabilities of the intervals depicted in Figure 1 converge to the nominal level (except for the simple qBB-interval). The coverage probabilities of the α-calibrated intervals approached the nominal $1 - \alpha = 0.95$

to a satisfactory level. If the intervals were based on 10 historical studies, the coverage probability was at least 93% and higher than 94% if 20 historical controls were considered. If the intervals are computed based on five studies, the nominal level is not reached by any of the methods. Based on the simulation results, we recommend to always use one of the two α-calibrated methods. If less than 10 studies are available, the α-calibrated methods still are the best. However, the coverage probability may be considerably lower than the nominal 95%.

It may seem useful to calculate prediction intervals based on the maximum number of historical studies available, following the simple principle "the more the better." However, improvidently increasing the number of historical studies may lead to the inclusion of studies with genetically different strains, handling or feeding. In consequence, the actual dispersion in the data will increase and cause unnecessarily wide and less precise intervals. Although there is a demand for a high number of historical studies, this amount of historical information should be chosen with regard to biological reasons in a way that the impact of biological variation, such as changing population genetics of the model organism over time,[1] is not increased.

Following the research of Keenan et al,[2] the maximum age of historical studies, which should be used for analysis, varies between two and five years, depending on the guideline or the field of research. Hence, the maximum age of the studies used for the calculation of prediction intervals should be a compromise that, on the one hand, reflects the statistical demand for a large number of historical control groups and that, on the other hand, carefully consults the given biological restrictions.

In our simulation, we sampled data with equal sample size across historical studies. In highly standardized bioassays, equal sample sizes are the most relevant case. Furthermore, this setting allows to jointly consider prediction intervals that either rely on the quasi-binomial or the beta-binomial assumption. In consequence, our recommendations and our estimation of the coverage probability are restricted to the case of equal sample sizes. However, it should be noted that the uncalibrated overdispersed Nelson-type intervals as well as methods based on raw and calibrated beta-binomial quantiles can be applied in situations with varying sample size n_k between historical studies. Only the calibrated Nelson interval would need some adjustment for variable sample sizes because, in the current definition, the sampling step in the bootstrap calibration relies on different assumptions that are used for the computation of the uncalibrated prediction limits and $\hat{\phi}_{QB}$.

Due to the fact that most model organisms are assigned to subgroups (for example, cages or origin from the same litter), the randomization structure can be more complex than assumed in the present study. Therefore, an extension of the given prediction intervals or intervals drawn from generalized linear mixed models, reflecting more complicated randomization structure of the experiments will be an issue for future development.

ACKNOWLEDGEMENTS

We want to thank Prof Dr Ludwig Hothorn for giving helpful suggestions and Clemens Buczilowski for his technical support. Furthermore, we want to thank the reviewers for reading the manuscript and for their helpful comments.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

FINANCIAL DISCLOSURE

No specific grant from funding agencies in the public, commercial, or not-for-profit sectors was received for this research.

ORCID

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

REFERENCES

1. Greim H, Gelbke HP, Reuter U, Thielmann HW, Edler L. Evaluation of historical control data in carcinogenicity studies. *Hum Exp Toxicol*. 2003;22(10):541-549.
2. Keenan C, Elmore S, Francke-Carroll S, et al. Best practices for use of historical control data of proliferative rodent lesions. *Toxicol Pathol*. 2009;37(5):679-693.

3. Elmore AS, Peddada SD. Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. *Toxicol Pathol*. 2009;37(5):672-676.
4. Tarone RE. The use of historical control information in testing for a trend in proportions. *Biometrics*. 1982;38(1):215-220.
5. Kitsche A, Hothorn LA, Schaarschmidt F. The use of historical controls in estimating simultaneous confidence intervals for comparisons against a concurrent control. *Comput Stat Data Anal*. 2012;56(12):3865-3875.
6. Peddada SD, Dinse GE, Kissling GE. Incorporating historical control data when comparing tumor incidence rates. *J Am Stat Assoc*. 2007;102(480):1212-1220.
7. Leon-Novelo LG, Womack A, Zhu H, Wu X. A Bayesian analysis of quantal bioassay experiments incorporating historical controls via Bayes factors. *Statist Med*. 2017;36(12):1907-1923.
8. Faulkenberry GD. A method of obtaining prediction intervals. *J Am Stat Assoc*. 1973;68(342):433-435.
9. Bain LJ, Patel JK. Prediction intervals based on partial observations for some discrete distributions. *IEEE Trans Reliab*. 1993;42(3):459-463.
10. Hahn GJ, Nelson W. A survey of prediction intervals and their applications. *J Qual Techno*. 1973;5(4):178-188.
11. Nelson W. *Applied Life Data Analysis*. New York, NY: John Wiley & Sons Inc.; 1982. ISBN 0-471-09458-7.
12. Hahn GJ, Meeker WQ. *Statistical Intervals: A Guide for Practitioners*. New York, NY: John Wiley & Sons Inc.; 1991. ISBN 0-471-88769-2.
13. Wang H. Coverage probability of prediction intervals for discrete random variables. *Comput Stat Data Anal*. 2008;53(1):17-26.
14. Wang H. Closed form prediction intervals applied for disease counts. *Am Stat*. 2010;64(3):250-256.
15. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc*. 1927;22(158):209-212.
16. Krishnamoorthy K, Peng J. Improved closed-form prediction intervals for binomial and poisson distributions. *J Stat Plan Inference*. 2011;141(5):1709-1718.
17. Hinde J, Demétrio CGB. Overdispersion: models and estimation. *Comput Stat Data Anal*. 1998;27(2):151-170.
18. Luo R, Sudhir P. Estimation for zero-inflated beta-binomial regression model with missing response data. *Statist Med*. 2018;37(26):3789-3813.
19. Demétrio CGB, Hinde J, Moral RA. Models for overdispersed data in entomology. In: Ferreira CP, Godoy WAC, eds. *Ecological Modelling Applied to Entomology*. Cham, Switzerland: Springer International Publishing; 2014:219-259.
20. Beran R. Calibrating prediction regions. *J Am Stat Assoc*. 1990;85(411):715-723.
21. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall; 1993.
22. Loh WY. Calibrating confidence coefficients. *J Am Stat Assoc*. 1987;82(397):155-162.
23. Hall P, Peng L, Tajvidi N. On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika*. 1999;86(4):871-880.
24. Fonseca G, Giummole F, Vidoni P. Calibrating predictive distributions. *J Stat Comput Simul*. 2014;84(2):373-383.
25. Saha KK, Paul SR. Bias-corrected maximum likelihood estimator of the intraclass correlation parameter for binary data. *Statist Med*. 2005;24(22):3497-3512.
26. Venables WN, Ripley BD. *Modern Applied Statistics With S*. 4th ed. New York, NY: Springer; 2002. ISBN 0-387-95457-0.
27. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. New York, NY: Chapman & Hall; 1989. ISBN 0-412-31760-5.
28. Lui KJ, Mayer JA, Eckhardt L. Confidence intervals for the risk ratio under cluster sampling based on the beta-binomial model. *Statist Med*. 2000;19(21):2933-2942.
29. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc Ser C Appl Stat*. 2005;54(3):507-554.
30. NTP. Tables of historical controls: pathology tables by route/vehicle. https://ntp.niehs.nih.gov/results/dbsearch/historical/index.html. Accessed May 17, 2017.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Menssen M, Schaarschmidt F. Prediction intervals for overdispersed binomial data with application to historical controls. *Statistics in Medicine*. 2019;38:2652–2663. https://doi.org/10.1002/sim.8124

2.2 Prediction intervals for overdispersed binomial data with application to historical controls (supplementary materials)

Max Menssen[1], Frank Schaarschmidt[1]

1. Institut für Zellbiologie und Biophysik, Abteilung Biostatistik, Leinbiz Universität Hannover, Herrenhäuser Str. 2, 30419 Hannover

Type of authorship:	First author
Type of article:	Supplementary material
Journal:	Statistics in Medicine
Impact factor:	2.373
Number of citations:	1 (web of science)
DOI:	10.1002/sim.8124

Contributions

Max Menssen

1. Implementation of all prediction intervals
2. Design and implementation of the Monte-Carlo simulations
3. Writing the manuscript

Frank Schaarschmidt

1. Design and implementation of the Monte-Carlo simulations
2. Writing the manuscript

Max Menssen, Frank Schaarschmidt

Leibniz Universität Hannover: Institut für Biostatistik

Prediction intervals for overdispersed binomial data with application to historical controls

---

Supplementary materials

1 Power to detect difference to concurrent control

Bretz and Hothorn (2002) provide methods to compute power of a Dunnett test for binomial proportion. Figure 1 shows the any-pair-power of a one-sided Dunnett test for increasing proportion in three treatments compared to a control, with proportions $\pi_i, i = 1, ..., 4$, and sample sizes $n_i, i = 1, ..., 4$. The proportion in the control π_1 is set to small values: $\pi_1 = 0.01, 0.02, 0.05$ or 0.1 while in one treatment (e.g. highest dose, π_4) or two treatments (e.g. mid and high doses, π_3, π_4) the proportion is assumed to be increeased by factor 2, 3, or 4 compared to the control proportion.

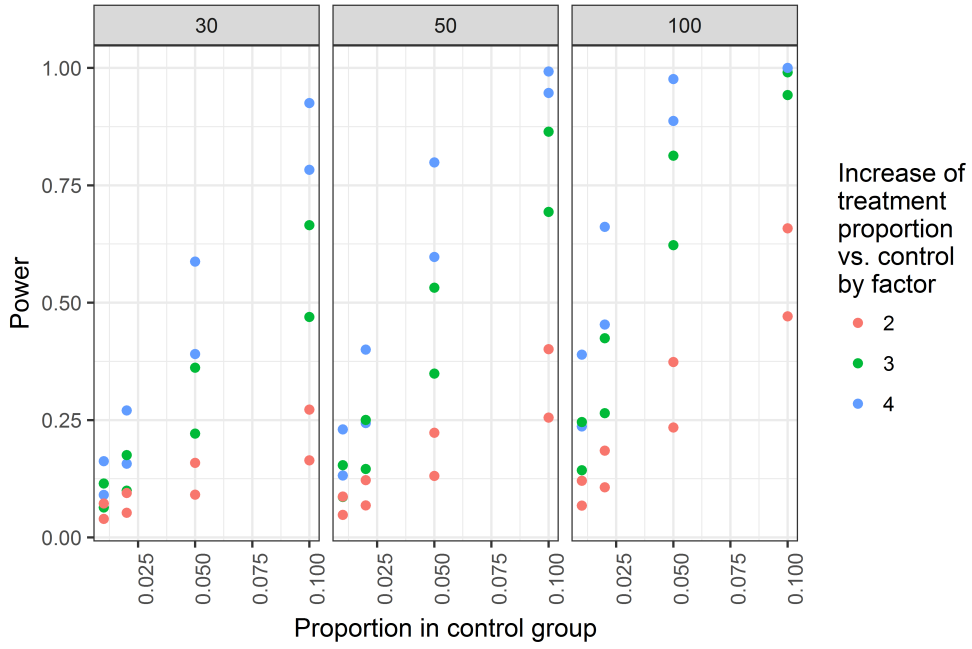

Figure 1: Anypair power of one-sided Dunnett-type test for comparing three treatment proportions to a control proportion.

2 Bias of $\hat{\phi}$

To assess the bias of $\hat{\phi}$, data sets were drawn from the beta binomial distribution as described in section 2.5 of the paper. The data $\boldsymbol{\tau} = \tau_{11}...\tau_{RS}$ was generated with $r = 1, ..., R$ as the index of all possible combinations of $\phi = (1.01, 1.5, 2.0, 2.5, 3)$, $\pi = (0.01, 0.05, 0.1, 0, 2, 0.3, 0.4, 0.6, 0.8)$, $n = (30, 50, 100, 150)$ and $k = (5, 10, 20, 100)$ and $s = 1, ..., S$ as the number of simulations with $S = 5000$. For each data set the dispersion parameter $\hat{\phi}_{rs}$ was computed. Then, the estimated mean dispersion was computed over all simulations as $\bar{\hat{\phi}}_r = \frac{\sum_{s=1}^{S} \hat{\phi}_{rs}}{S}$. Subsequently, the relative bias was calculated as $\bar{\hat{\phi}}_r / \phi_r$.

The estimate, based on the quasi-binomial assumption $\hat{\phi}^{QB}$ was calculated according to equation (11) of the paper and was restricted to $\hat{\phi}^{QBAdj} = max(1, \hat{\phi}^{QB})$. Based on the assumption of beta-binomial distributed observations the estimate $\hat{\phi}^{BB}$ was calculated as $\hat{\phi}^{BB} = 1 + (n-1)\hat{\rho}$ and was set to $\hat{\phi}^{BBAdj} = max(1.001, \hat{\phi}^{BB})$. The relative bias of all four estimates is depicted in figure (2).

Except for the case, that the amount of historical information is considerably high (K=100) all four estimates are negatively biased for small π. If overdispersion plays no role in the historical data, both estimates that do not account for underdispersion ($\hat{\phi} < 1$) are positively biased. This effect decrease if the amount of historical studies is increasing.

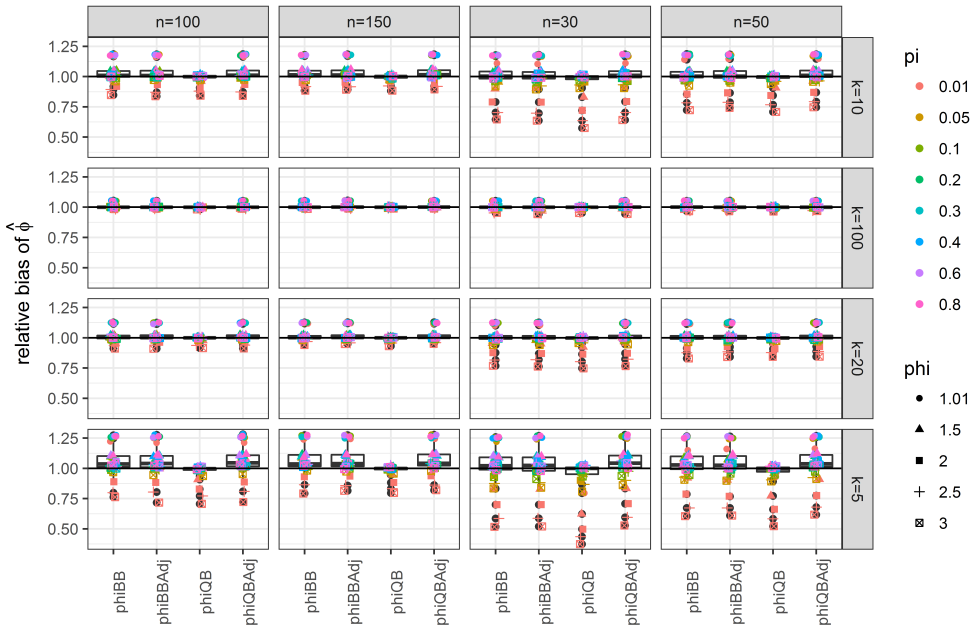

Figure 2: Relative bias of $\hat{\phi}$.

Columns of the subfigures represent different sample sizes (n). The number of historical studies (k) is given by the rows of the sub-figures. Different binomial proportions are indicated by colors. The different amount of overdispersion is given by the shapes.

3 Coverage probabilities of simple prediction intervals

The simulation for getting the coverage probabilities of the simple intervals was carried out as described in section 2.5 of the paper. The interval mentioned as the baipat-interval was calculated according to eq. (11, 12) in Bain & Patel (1993). The Wang-interval was computed according to Wang (2010) eq. (6, 8). The krishpeng-interval was given by Krishnamoorthy & Peng (2011) eq. (12) and the Nelson-interval was calculated according to eq. (9) in section 2.3 of this paper. Since the intervals were developed for one binomial sample, the historical information was pooled such that $n = \sum_{k=1}^{K} n_k$ and $\pi = \frac{\sum_{k=1}^{K} y_k}{\sum_{k=1}^{K} n_k}$. Since the approach of pooling the historical information to one sample does not account for overdispersion, the coverage probabilities decrease with an increasing dispersion parameter. The estimated coverage probabilities are depicted in fig (3).

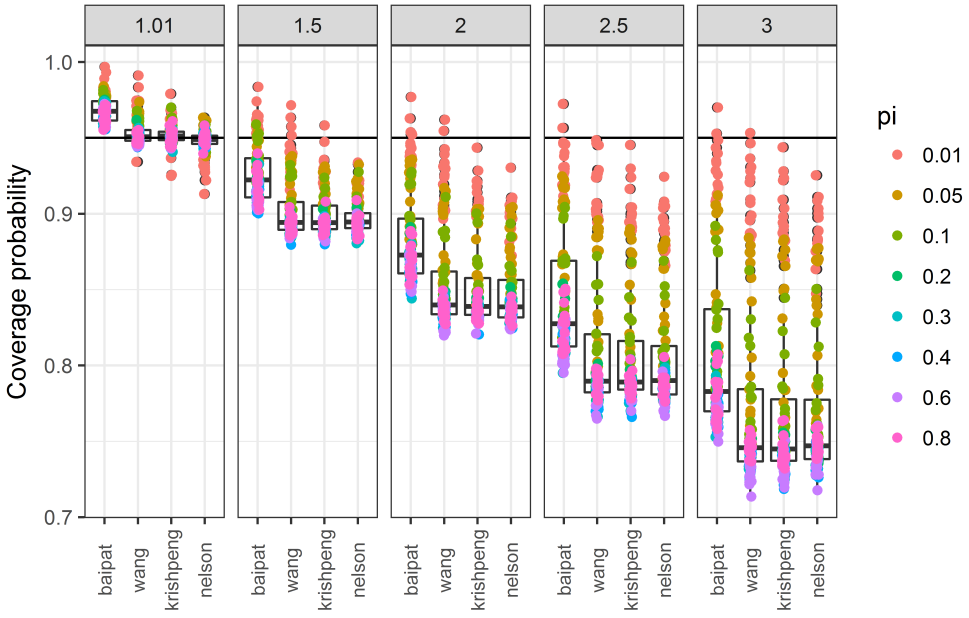

Figure 3: Coverage probabilities of the simple prediction intervals.
For each amount of overdispersion one subfigure is given. Different binomial proportions are indicated by colors.

4 Historical data

The data set that was used for the example (section 3.1) is given in table 1. The data was downloaded from the NTP-homepage and was processed as described in section 3.1.

Table 1: Mortality of male B6C3F1-mice exposed to the inoculation air pathway

Report	Start	Lab	Dead	Total
2016	2009	bn	15	50
2016	2011	bn	10	50
2015	2008	bn	12	50
2015	2009	bn	12	50
2013	2003	bn	13	50
2013	2006	bn	11	50
2013	2003	bn	19	50
2013	2008	bn	11	50
2013	2003	bn	14	50
2013	2005	bn	21	50

5 The calculation of $\hat{\rho}$

The estimated intra class correlation Lui et al. (2000) can be calculated as

$$\hat{\rho} = \frac{BMS - WMS}{BMS + WMS(\omega - 1)} \quad with \tag{1}$$

$$BMS = \frac{\sum_{k=1}^{K}(x_k^2/n_k) - (\sum_{k=1}^{K} x_k)^2/(\sum_{k=1}^{K} n_k)}{k-1}, \tag{2}$$

$$WMS = \frac{\sum_{k=1}^{K} x_k - \sum_{k=1}^{K}(x_k^2/n_k)}{\sum_{k=1}^{K} n_k - 1} \quad and \tag{3}$$

$$\omega = \frac{(\sum_{k=1}^{K} n_k)^2 - \sum_{k=1}^{K} n_k^2}{(k-1)\sum_{k=1}^{K} n_k} \tag{4}$$

6 Reference

Bretz F., Hothorn L.A. (2002): Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. Statistics in Medicine, 21(22):3325-3335. doi:10.1002/sim.1324

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

Max Menssen[1], Frank Schaarschmidt[1]

1. Institut für Zellbiologie und Biophysik, Abteilung Biostatistik, Leinbiz Universität Hannover, Herrenhäuser Str. 2, 30419 Hannover

Type of authorship:	First author
Type of article:	Research article
Journal:	Statistica Neerlandica
Impact factor:	1.190
Number of citations:	0
DOI	10.1111/stan.12260

Contributions

Max Menssen

1. Derivation of the two proposed REML based prediction intervals
2. Implementation of the six different prediction intervals
3. Design and implementation of the Monte-Carlo simulations
4. Writing the paper

Frank Schaarschmidt

1. Derivation of mean square based prediction intervals for the models used for simulation
2. Design of the Monte-Carlo simulations

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

Max Menssen* | Frank Schaarschmidt

[1]Department for Biostatistics, Leibniz Universität Hannover, Lower Saxony, Germany

Correspondence
*Max Menssen, Herrenhäuser Strasse 2 30419 Hannover Email: menssen@cell.uni-hannover.de

Summary

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 REML estimation. Unlike other REML based PI found in 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 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:
Satterthwaite approximation, bootstrap calibration, historical control data, anti-drug antibody, assay qualification

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 & Meeker[1] and Hahn et al.[2] 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 prediction intervals should contain either one future observation, the future mean, all of $M \geq 1$ future observations or K out of M future observations.

Prediction intervals can be applied to several statistical problems and are of use in many scientific fields. In the context of pharmaceutical applications, Francq et al.[3] used prediction intervals for assay qualification. More examples for the usage of prediction intervals for process validation are given by Hahn & Meeker[1] or in the context of gauge repeatability and reproducability experiments[4]. Also in preclenical 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[5,6]. Nevertheless, the methods proposed for that purpose (e.g. historical range or historical mean plus minus standard deviation) are rather naive and many authors are not aware, that prediction intervals for one or more future observations (depending on the purpose) can be applied to that problem.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/stan.12260

Another field of application occurs in early phases of drug development such as the detection of anti-drug antibodies (ADA)[8,9]. In such a bioassay, the antibody reaction is evaluated for a set of non-responders as well as for patients with unclear status. Following Schaarschmidt et al.[10] upper prediction limits can be computed for a sample of putative non-responders 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[8].

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 prediction intervals based on random effects models[3,8,10]. The idea of the computation of prediction intervals based on random effects models dates back to 1941. In that year Satterthwaithe[11] 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[4,12,13]. 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 (BLUP) are of interest[14,15], rather than the prediction of one or more future observations.

In the context of random effects models, prediction intervals 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 mean squares was already utilized by Satterthwaite[11], 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 prediction intervals for $M \geq 1$ future observations were proposed by Lin & Liao[4] for balanced data. Up to now, REML based PI got less attention. Al-Sarraj et al.[15] used a PI for which the variance components were estimated via REML but treated as known, following the approach of Pawitan[16] by using a standard normal quantile. Francq et al.[17] 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 prediction intervals for one future observation ($M = 1$). For this purpose, the degrees of freedom were approximated using the Generalized Satterthwaite method following van den Heuvel[18]. 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[11], Lin and Liao[4] and Francq et al.[17] 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[4] . 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[19].

2 | REAL LIFE DATA

Random effects models can be applied to a wide range of experimental designs. Hence, many different designs are reported in literature regarding assay qualification, early phase drug development such as ADA detection or the usage of historical control data. 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[3]. 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[8,9,20,21]. Data about historical controls regarding rats and mice obtained from long time carcinogenicity studies are provided on the homepage of the National Toxicology Program[22]. Since the compound of interest can be applied by using several different pathways and studies are carried out by several laboratories, historical control data 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[8] published a data set resulting from an electroluminescence assay in which blood plasma of twenty 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 data set, only a cross-classified model without an interaction term can be fit to the data. Since this data set is balanced, it will be used in order to demonstrate the calculation of prediction intervals using all six methods described below.

2.1.2 | Historical control data Maximum mean weekly body weight of female mice

A data set containing historical control data 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[22] 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, prediction intervals 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 maximum mean weekly body weight 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	mmwbw
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	mmwbw
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

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

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 sub-vectors $\boldsymbol{U}_c$. 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 $q_{total} = \sum_{c=1}^{C} q_c$. $\boldsymbol{Z}$ is a design matrix and has the dimensions $N\,x\,q_{total}$. 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

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

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

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

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

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

with $\boldsymbol{I}_N$ as an identity matrix of order N. Further information on the model described above can be found in McCullogh and Searle[23] (pp. 156-160) or in Searle et al.[24] (pp. 233-257).

For prediction, it is assumed that the future random variable $\boldsymbol{Y}^*$ which is comprised of $\boldsymbol{M} \geq 1$ 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

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

which implies that

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

with

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

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

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

$$\boldsymbol{y} = \mathbf{1}\hat{\mu} + \boldsymbol{Z}\hat{\boldsymbol{u}} + \hat{\boldsymbol{\epsilon}}$$

the estimate for the prediction variance becomes

$$\widehat{v\hat{a}r}(\boldsymbol{D}) = \widehat{v\hat{a}r}(\boldsymbol{Y}^* - \mathbf{1}\hat{\mu}) = \widehat{v\hat{a}r}(\boldsymbol{Y}^*) + \widehat{v\hat{a}r}(\mathbf{1}\hat{\mu}) \tag{5}$$

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

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

Anyhow, this is the standard assumption on which all methods given below rely.
A prediction interval for $\boldsymbol{M} \geq 1$ future observations $\boldsymbol{y}_M^*$ with coverage probability $\Psi = P(L \leq \boldsymbol{y}_M^* \leq U) = 1 - \alpha$ is given by

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

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

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

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

3.2 | Calculation of prediction intervals

3.2.1 | PI for $\boldsymbol{M} = 1$ future observation based on mean squares

The estimation of prediction intervals for $\boldsymbol{M} = 1$ one future observation based on mean squares was firstly described 1941 by Satterthwaite [11]. Assuming a balanced design, $var(D)$ is estimated by $\widehat{v\hat{a}r}(D)^{Sat} = \sum_{c=1}^{C+1} \omega_c^{Sat} \boldsymbol{M}\boldsymbol{S}_c^{Sat}$ and the prediction interval is given by

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

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

$$df^{Sat} = \frac{(\sum_{c=1}^{C+1} \omega_c^{Sat} \boldsymbol{M}\boldsymbol{S}_c^{Sat})^2}{\sum_{c=1}^{C+1} \frac{(\omega_c^{Sat} \boldsymbol{M}\boldsymbol{S}_c^{Sat})^2}{df_c}}$$

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}, mean squares $\boldsymbol{M}\boldsymbol{S}_c^{Sat}$ 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 $\boldsymbol{M} = 1$ future observation based on REML

This method is based on parameter estimates that are estimated using the restricted maximum likelihood (REML) approach. Generally, the degrees of freedom associated with variance components estimated via REML can be approximated by using

the Generalized Satterthwaite method[18,28] 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

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

For linear mixed models as well as for random effects models the estimates used in eq. 8 can be obtained by using the R package VCA[28]. This package provides degrees of freedom and standard errors for the individual variance components $\hat{\sigma}_c^2$ and its sum.

Recently, Francq et al.[17] proposed a REML based PI for $M = 1$ future observation, that was applied in an assay qualification study[3]. 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 as . 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.[17] approximated the degrees of freedom based on $\widehat{var}(Y^*) = \sum_{c=1}^{C+1} \hat{\sigma}_c^2$, rather than on the prediction variance $\widehat{var}(D) = \widehat{var}(\mu) + \widehat{var}(Y^*)$. Hence, the degrees of freedom used for interval calculation are

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

Consequently, the interval of Francq et al.[17] is given by

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

In order to account for the degrees of freedom associated with the whole prediction variance, the approximation given in eq. 18 of van den Heuvel[18] 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

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

and the variance of the prediction variance can be estimated by

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

Then, the approximated degrees of freedom are

$$df^{Pred} = 2\frac{\widehat{var}(D)^2}{\widehat{var}[\widehat{var}(D)]}$$

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

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

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

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.[17] is based on $\widehat{var}(Y^*)$ rather than on the whole prediction variance $\widehat{var}(D) = \widehat{var}(\mu) + \widehat{var}(Y^*)$, the degrees of freedom used for the calculation of $[L, U]^{Franq}$ 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.[17] is expected to be less wide than the PI given in eq. 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 generalized pivotal quantities

The theoretical background on which this interval is based on, is given by Lin and Liao[4]. 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 generalized pivotal quantity (GPQ) for the expected mean squares. 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[27].
Following Lin and Liao, a GPQ for the expected mean squares EMS_c is given by

$$GPQ(EMS_c) = \frac{s_c^2}{R_c}$$

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 $H = 10000$ mutually independent realizations $R_{c,1}, \dots, R_{c,H}$ from the χ^2-distribution with degrees of freedom df_c.
2. Calculate $GPQ(EMS_c)_h = \frac{s_c^2}{R_{c,h}}$.
3. Calculate $GPQ(\sigma_c^2)_h$ based on $GPQ(EMS_c)_h$. 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 $GPQ(var(\boldsymbol{D}))_h$ by substituting $GPQ(\sigma_c^2)_h$ into $\widehat{var}(\boldsymbol{D})$. Please note that further formulas for the calculation of $\widehat{var}(\boldsymbol{D})$ are given below in sections 3.3.1 and 3.3.2.
5. Based on $GPQ(var(\boldsymbol{D}))_h$, compute H qunatiles from the corresponding multivariate normal distributions, such that $q_h = z_{1-\alpha/2,\, \boldsymbol{0},\, GPQ(var(\boldsymbol{D}))_h}$.
6. Calculate $GPQ(\boldsymbol{D}) = median(q_h)$
7. The corresponding prediction interval is given by $[L, U]^{GPQ\_M1} = \bar{y} \pm GPQ(\boldsymbol{D})$

For both, a two-way-hierarchical and a two-way cross-classified model with interaction, $GPQ(\sigma_c^2)_h$ can be obtained if the EMS_c used in equations 16 to 18 and 21 to 24 are substituted by $GPQ(EMS_c)_h$. $GPQ(var(\boldsymbol{D}))_h$ can be obtained if σ_c^2 is substituted by $GPQ(\sigma_c^2)_h$ in eq. 15 and eq. 20.
This approach, which Lin and Liao called **Method 1**, is based on the the calculation of $H = 10000$ quantiles from the multivariate-normal distribution $z_{1-\alpha/2,\, \boldsymbol{0},\, GPQ(var(\boldsymbol{D}))_h}$. Since the calculation of a multivariate-normal quantile is computationally intensive[25], this approach will take too much computing time to be useful in practical applications or Monte Carlo simulations (it took around 13 minutes 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 $GPQ(var(\boldsymbol{D}))_h$, such that

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

The corresponding prediction interval is given by

$$[L, U]^{GPQ\_M3} = \bar{y} \pm z_{1-\alpha/2,\, \boldsymbol{0},\, GPQ(var(\boldsymbol{D}))}$$

treating $GPQ(var(\boldsymbol{D}))$ 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 $var(\boldsymbol{D})$ reduces to $var(\mu) + var(Y^*)$. 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[33], 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 data set $\boldsymbol{y}$. Then, based on the estimated model parameters $b = 1, \dots B$ new bootstrap data sets $\boldsymbol{y}_{\boldsymbol{b}}^*$ of same sample size and structure as the original data set are drawn. Then, $m = 1, \dots, M$ 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 $\hat{\sigma}_{bc}^2$ as well as for the variance of the estimated mean $\widehat{var}(\hat{\mu}_b)$.

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 $g = 1, \dots, G$ 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

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

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

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

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 eq. 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 eq. 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}

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

3.3 | Simulation study

The coverage probabilities of the six different prediction intervals 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, e.g. 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 prediction intervals 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

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

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

with $\widehat{var}(\mu) = \frac{1}{IJK}(JK\hat{\sigma}_a^2 + K\hat{\sigma}_b^2 + \hat{\sigma}_e^2)$, $\boldsymbol{I}_{I^*}$ 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[27] the variance components can be expressed as

$$\sigma_a^2 = \frac{1}{JK}(EMS_a - EMS_{b(a)}) \tag{16}$$

$$\sigma_{b(a)}^2 = \frac{1}{K}(EMS_{b(a)} - EMS_e) \tag{17}$$

$$\sigma_e^2 = EMS_e. \tag{18}$$

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

Effect	c	df_c	MS_c^{Sat}	ω_c^{Sat}	ω_c^{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

$$\begin{aligned} y_{ijk} &= \mu + a_i + b_j + ab_{ij} + e_{k(ij)} \\ a_i &\sim N(0,\sigma_a^2), \quad i = 1,\dots,I \\ b_j &\sim N(0,\sigma_b^2), \quad j = 1,\dots,J \\ ab_{ij} &\sim N(0,\sigma_{ab}^2), \quad ij = (11,\dots,IJ) \\ e_{k(ij)} &\sim N(0,\sigma_e^2), \quad k(ij) = 1,\dots,n_{k(ij)} \end{aligned}$$

Usually this setup is balanced such that I patients are analyzed in J laboratories exactly $K = n_{k(ij)}\ \forall\ k(ij) = 1(ij),\dots,n_{k(ij)}$ 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 ($K \neq n_{k(ij)}\ \exists\ k(ij) \neq 1(ij),\dots,n_{k(ij)}$). 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.

Mean squares and weights for the calculation of prediction intervals 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

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

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

with $\widehat{var}(\mu) = \frac{1}{IJK}(JK\hat{\sigma}_a^2 + IK\hat{\sigma}_b^2 + K\hat{\sigma}_{ab}^2 + \hat{\sigma}_e^2)$, $\boldsymbol{I}_{I^*}$ 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[4], the variance components are given by

$$\sigma_a^2 = \frac{1}{JK}(EMS_a - EMS_{ab}) \tag{21}$$

$$\sigma_b^2 = \frac{1}{IK}(EMS_b - EMS_{ab}) \tag{22}$$

$$\sigma_{ab}^2 = \frac{1}{K}(EMS_{ab} - EMS_e) \tag{23}$$

$$\sigma_e^2 = MS_e \tag{24}$$

using the weights given in Table (4)

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

Effect	c	df_c	MS_c^{Sat}	ω_c^{Sat}	ω_c^{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 prediction intervals, 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 prediction intervals. 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 prediction intervals 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 $r = 1, \dots, 5000$ historical data sets were drawn. Similarly another data set 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 data sets on which `lmer()` threw a warning were tracked and the coverage probability was also computed based on the simulated data sets 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 data sets rather than on the data sets 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 to 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

$$\Omega_h = max(\frac{\sigma^2_{ch}}{\sum_c \sigma^2_{ch}}) \text{ and}$$
$$\tau_h = max[\frac{\sigma^2_{a,h}}{\sigma^2_{ab,h}} / \frac{df_{a,h}}{df_{ab,h}}, \frac{\sigma^2_{b,h}}{\sigma^2_{ab,h}} / \frac{df_{b,h}}{df_{ab,h}}, \frac{\sigma^2_{ab,h}}{\sigma^2_{e,h}} / \frac{df_{ab,h}}{df_{e,h}}]$$

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 to 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 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 to 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 $se(\psi) = \sqrt{(0.95 \cdot 0.05)/5000}$. 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 prediction intervals for one future observation

The simulated coverage probabilities of prediction intervals for one future observation based on balanced h2 models are given in Figure 1. For all six methods, the coverage probabilities depend mainly 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.[17] 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[11] are too low if the variance component σ^2_a is relatively high, but estimated based on a low number of observations (high τ_h, red dots) and too high if σ^2_a is relatively low, but estimated based on a high number of observations (low τ_h, blue dots).

The coverage probabilities of the prediction intervals 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[11] 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 prediction intervals 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[4] 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[30] 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[11] and Lin and Liao[4] were calculated using the `aov()` function from the stats package[30]. The estimates $\widehat{var}(\hat{\sigma}_c^2)$ and $\widehat{var}(\widehat{var}(\boldsymbol{y}))$ used for the calculation of the uncalibrated REML-based intervals were obtained from the `remlMM()` function of the VCA package[28]. The bootstrap-calibrated PI can be applied using the `lmer_pi()` function from the predint package[19].

5.1 | Quantile-calibrated prediction intervals 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. Prediction intervals 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()`[29]. If a data set containing actual data is provided via `newdat`, `predint::lmer_pi()` 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`. 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.

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`	$\boldsymbol{M}$	Number of future observations
`alternative`		Prediction interval, lower prediction limit or upper prediction limit
`alpha`	α	Defines the nominel coverage probability $1-\alpha$
`nboot`	$\boldsymbol{B}$	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`	$\boldsymbol{D}$	Maximum number of bisection steps

6 | APPLICATION OF PREDICTION INTERVALS 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 prediction intervals given in Tables 6 and 7 is available on GitHub under https://github.com/MaxMenssen/menssen_schaarschmidt_2021.

6.0.1 | 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[8] which is comprised of data from a bioassay in which 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

$$
\begin{aligned}
ln(y_{ij}) &= \mu + a_i + b_j + e_{ij} \\
a_i &\sim N(0, \sigma_a^2), \quad i = 1, \dots, I \\
b_j &\sim N(0, \sigma_b^2), \quad j = 1, \dots, J \\
e_{ij} &\sim N(0, \sigma_e^2)
\end{aligned}
$$

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 prediction intervals 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).

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

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

Except for the GPQ-based interval calculated with Method 3 of Lin and Liao which is the widest PI by far, all prediction intervals 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 eq. 11), but the distribution of $GPQ(\sigma^2_{ck})$ 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.

6.0.2 | Historical control data Maximum mean weekly body weight of female mice

Since the data set containing historical controls regarding the maximum mean weekly body weight 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 eq. 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 maximum mean weekly body weights.

TABLE 7 Prediction interval for the historical control data

mmwbw	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

7 | DISCUSSION

In the sections above, two methods for the calculation of prediction intervals based on random effects models were proposed and compared to four prediction intervals that are already published. Due to the fact that mean square based prediction intervals occur in literature since almost 80 years [11] most of the previous research was done on that topic. Anyhow, only a few studies that use other methods than Mean Squares, obtained from the classical ANOVA tables, are available. Two GPQ-based methods for prediction intervals for $M \geq 1$ future observations were proposed by Lin and Liao [4]. Despite the fact, that the estimation of model parameters in random and mixed effects models via REML is available since the 1970ies [32] 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. [15] used a REML based PI originally published by Pawitan [16] and Francq et al. [17] 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 [16] treats the estimated prediction variance as known and uses a standard normal quantile for the interval calculation. The interval of Francq et al. [17] 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 prediction intervals 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 [18]. 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 $t_{1-\alpha^{calib}/2,df}$ (or $\chi^2_{1-\alpha^{calib}/2,df}$) has coverage close to the nominal level $1-\alpha$. This approach was developed in the early 1990ies and was described by Efron and Tibshirani[33] in detail. Therefore, α-calibration was applied by several authors for different purposes, such as tolerance limits[34], confidence intervals[35,36] or prediction intervals for overdispersed binomial data[7]. 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, 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.

Data availability statement

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

Conflict of interest

The authors declare no potential conflict of interests.

References

1. Hahn G.J., Meeker Q.M. (1991): Statistical intervals. First edition. John Wiley and Sons Inc.
2. Hahn G.J., Meeker Q.M., Escobar L.A. (2017): Statistical intervals. Second edition. John Wiley and Sons Inc.
3. 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 Stat Biopharm Res. 12(3):267-272
4. Lin T.Y., Liao C-T. (2008): Prediction intervals for general balanced linear random models. J. Stat. Plan. Inference. 138(19):3164-3175
5. Greim H, Gelbke HP, Reuter U, Thielmann HW, Edler L. (2003): Evaluation of historical control data in carcinogenicity studies. Hum Exp Toxicol. 22(10):541-549.
6. Elmore AS, Peddada SD (2009). Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. Toxicol Pathol. 37(5):672-676.

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

8. Hoffman D., Berger M. (2011): Statistical considerations for calculation of immunogenicity screening assay cut points. J Immonol. Methods. 373:200-208

9. Jaki T., Allacher P., Horling F. (2016): A false sense of security? Can tiered approach be trusted to accurately classify immunogenicity samples? J. Pharm. Biomed. Anal. 128:166-173

10. 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. J. Immunol. Methods. 418:84-100

11. Satterthwaite F.E. (1941): Sythesis of variance. Psychometrika. 6(5):309-316.

12. Jeske D.R., Harville D.A. (1988): Prediction interval procedures and (fixed-effects) confidence interval procedures for mixed linear models. Commun. Statist. Theory Meth. 17(4):1053-1087

13. Wang C.M. (1992): Prediction intervals for balanced one-way random effects model. Commun. Statist. Simula. 21(3):671-687

14. 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

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

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

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

18. 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, DOI: 10.1080/03610911003646373

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

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

21. Zhang L., Zhang JJ., Kubiak RJ., Yang H. (2013) Statistical methods and tool for cut point analysis in immunogenicity assays. J. Immunol. Methods 389:79-87

22. NTP 2021: https://ntp.niehs.nih.gov/data/controls/index.html visited 31.05.2021

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

24. Searle R.S., Casella G., McCulloch C.E. (2006): Variance components. Second edition, John Wiley and Sons Inc. New Jersey

25. Genz A., Bretz F. (2009): Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195., Springer-Verlag, Heidelberg.

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

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

28. Schuetzenmeister A., Dufey F. (2019). VCA: Variance Component Analysis. R package version 1.4.1. https://CRAN.R-project.org/package=VCA

29. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1):1-48.

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

31. Giesbrecht, F.G. and Burns, J.C. (1985), Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results, Biometrics 41:477-486

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

33. Efron B., Tibshirani R.J. (1993): An introduction to the bootstrap. Chapman & Hall/CRC: Boca Raton Florida

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

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

36. Lee, H. I, and Liao, C. T. (2014): Unilateral conformance proportions in balanced and unbalanced normal random effects models. J. Agric. Biol. Environ. Stat. 19:202-218.

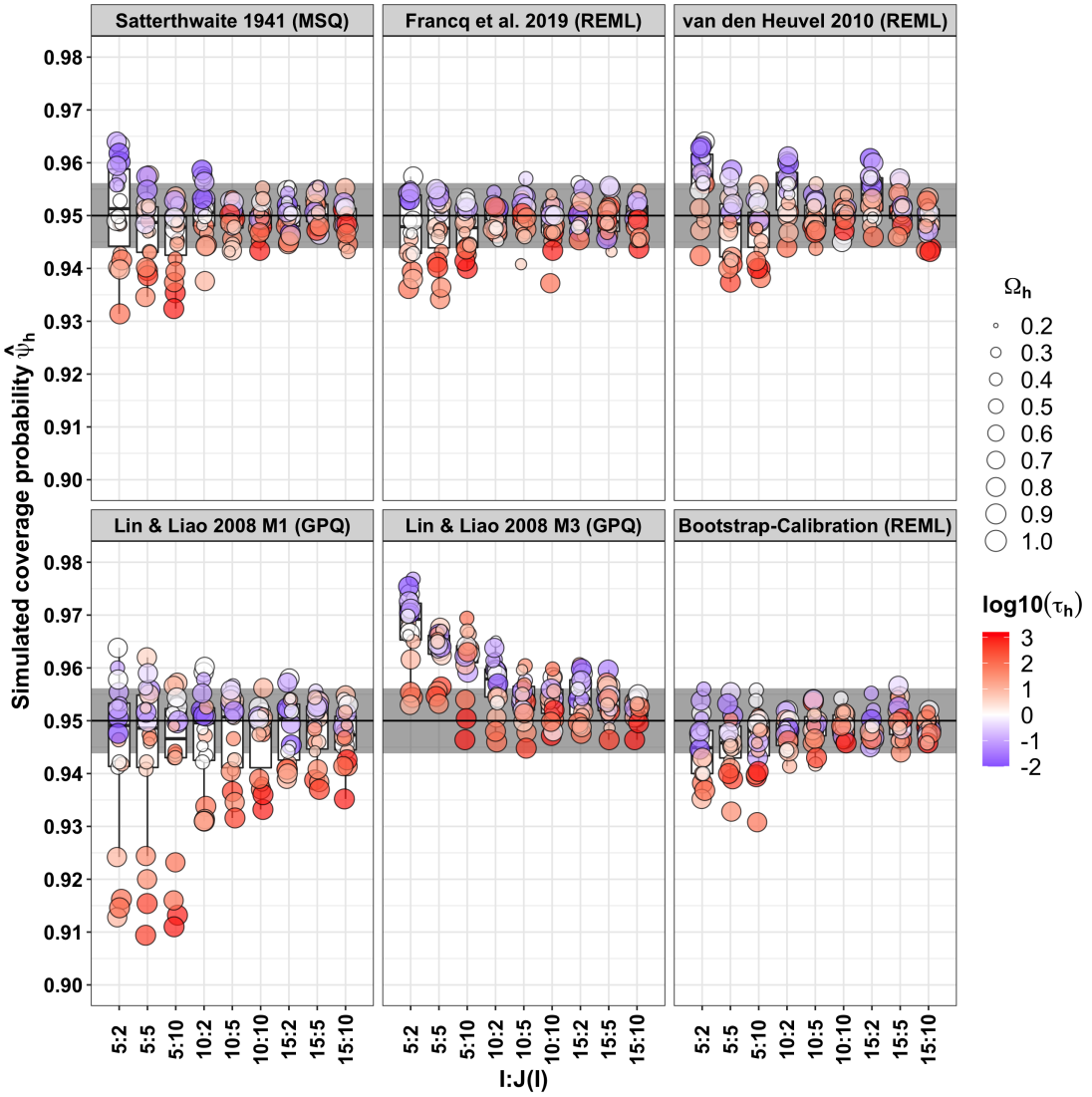

FIGURE 1 Coverage probabilities of PI for one future observation for the balanced h2 design. The nominal coverage probability $\psi = 0.95$ is indicated by the black line. The grey area indicates $\psi \pm 2se(\psi)$. The six different prediction intervals are represented by the panels.

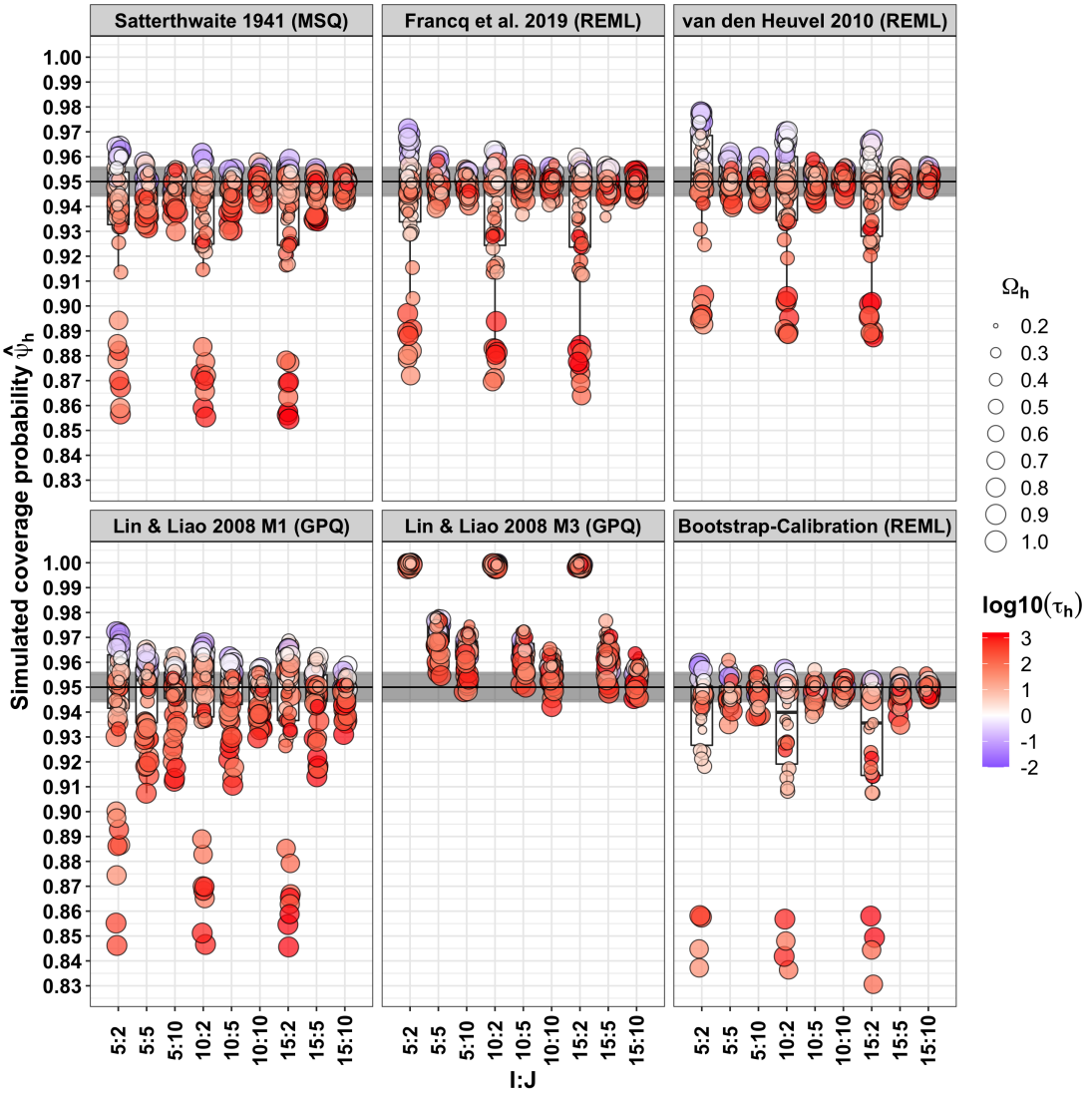

FIGURE 2 Coverage probabilities of PI for one future observation for the balanced c2 design. The nominal coverage probability $\psi = 0.95$ is indicated by the black line. The grey area indicates $\psi \pm 2se(\psi)$. The six different prediction intervals are represented by the panels.

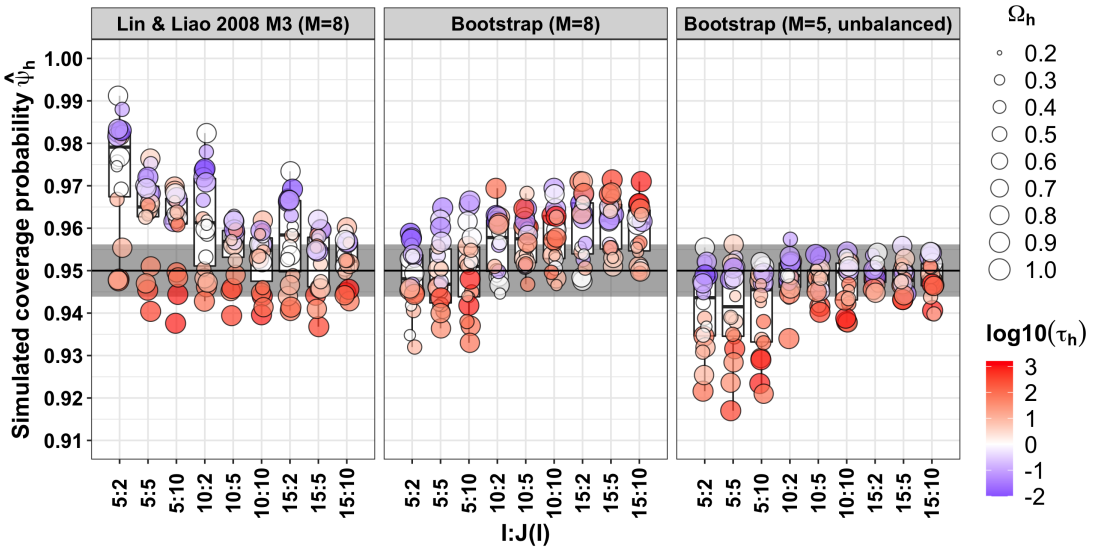

FIGURE 3 Coverage probabilities of PI for more than one future observation for the balanced h2 design. The nominal coverage probability $\psi = 0.95$ is indicated by the black line. The grey area indicates $\psi \pm 2se(\psi)$. The six different prediction intervals are represented by the panels.

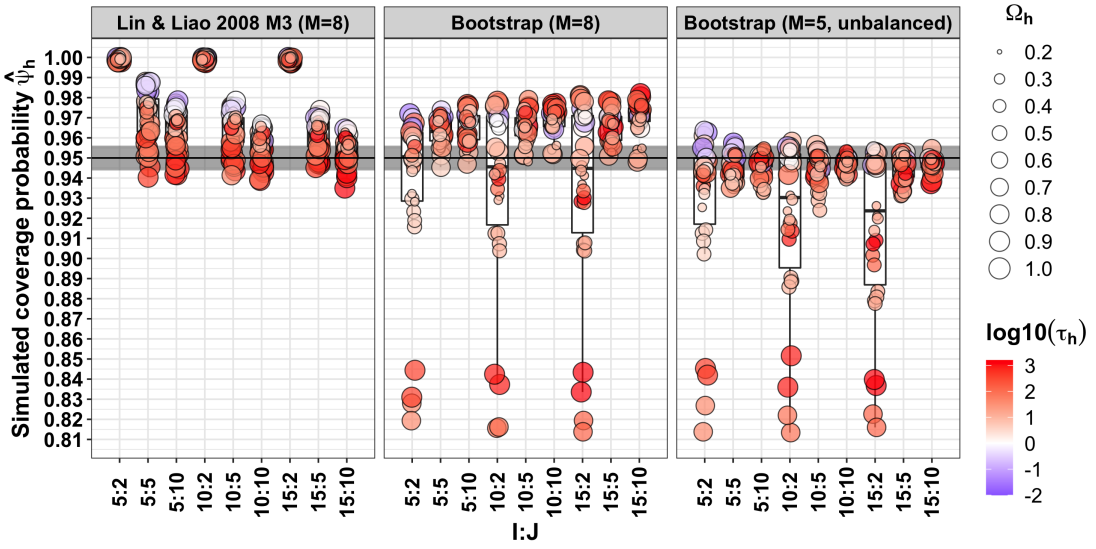

FIGURE 4 Coverage probabilities of PI for more than one future observation for the balanced h2 design. The nominal coverage probability $\psi = 0.95$ is indicated by the black line. The grey area indicates $\psi \pm 2se(\psi)$. The six different prediction intervals are represented by the panels.

How to cite this article: Menssen M., Schaarschmidt F. (2021), Prediction intervals for all of M future observations based on linear random effects models, , .

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 prediction intervals. With rising degrees of freedom (bigger data sets) this effect becomes smaller and can be neglected for degrees of freedom higher than lets say 30, due to the convergence of the t-distribution against the standard normal.

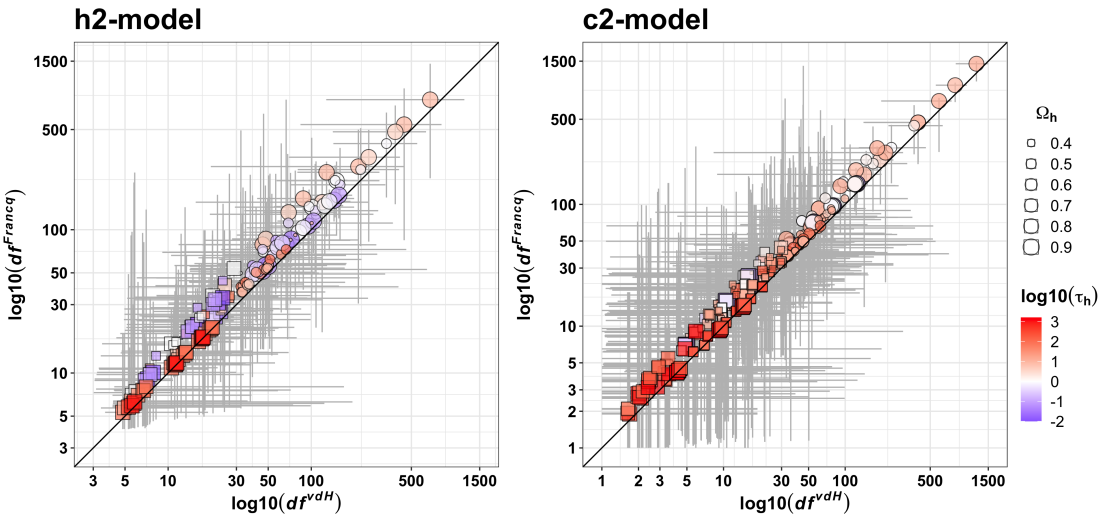

FIGURE A1 Simulated average degrees of freedom: df^{Francq} vs. df^{vdH}. The black line indicates a 1:1 relationship. The grey errorbars represent the corresponding minimum and maximum obtained in the simulation. Squares indicate observations were $df^{vdH} < 30$.

2.4 predint: Prediction intervals

Max Menssen[1]

1. Institut für Zellbiologie und Biophysik, Abteilung Biostatistik, Leinbiz Universität Hannover, Herrenhäuser Str. 2, 30419 Hannover

Type of authorship:	First author
Type of article:	Reference manual for the predint R package
Available on CRAN:	`https://cran.r-project.org/web/packages/predint/index.html`
Number of citations:	0
Contribution:	100 %

Contributions

Max Menssen

1. Derivation and implementation of prediction intervals for beta-binomial and quasi-binomial
2. Derivation and implementation of a prediction interval for quasi-poisson data
3. Implementation of the bootstrap-calibrated prediction interval of Menssen and Schaarschmidt 2021 (as described in section 2.3)

Package 'predint'

May 12, 2021

Type Package

Title Prediction Intervals

Version 1.0.0

Description An implementation of prediction intervals for overdispersed count data, for overdispersed binomial data and for linear random effects models.

License GPL (>= 2)

Encoding UTF-8

LazyData true

Imports lme4, dplyr, stats, graphics

RoxygenNote 7.1.1

Suggests rmarkdown, knitr, testthat (>= 3.0.0)

Config/testthat/edition 3

Depends R (>= 3.5.0)

NeedsCompilation no

Author Max Menssen [aut, cre]

Maintainer Max Menssen <menssen@cell.uni-hannover.de>

Repository CRAN

Date/Publication 2021-05-12 12:00:02 UTC

R topics documented:

`bb_dat1` *Beta-binomial data (example 1)*

Description

This data set contains sampled beta-binomial data from 10 clusters each of size 50. The data set was sampled with `rbbinom(n=10,size=50,prob=0.1,rho=0.06)`.

Usage

```
bb_dat1
```

Format

A `data.frame` with 10 rows and 2 columns:

succ number of successes

fail number of failures

Examples

```
# Upper prediction limit for m=3 future number of successes
# that are based on cluster sizes 40, 50, 60 respectively
beta_bin_pi(histdat=bb_dat1, newsize=c(40, 50, 60), alternative="upper", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

bb_dat2 *Beta-binomial data (example 2)*

Description

This data set contains sampled beta-binomial data from 3 clusters with different size, each. The data set was sampled with `rbbinom(n=3,size=c(40,50,60),prob=0.1,rho=0.06)`.

Usage

```
bb_dat2
```

Format

A `data.frame` with 3 rows and 2 columns:

succ number of successes

fail number of failures

Examples

```
# Prediction interval using bb_dat2 as future data
beta_bin_pi(histdat=bb_dat1, newdat=bb_dat2, nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

beta_bin_pi *Prediction intervals for beta-binomial data*

Description

beta_bin_pi calculates bootstrap calibrated prediction intervals for beta-binomial data

Usage

```
beta_bin_pi(
  histdat,
  newdat = NULL,
  newsize = NULL,
  alternative = "both",
  alpha = 0.05,
  nboot = 10000,
  lambda_min = 0.01,
  lambda_max = 10,
```

```
  traceplot = TRUE,
  n_bisec = 30
)
```

Arguments

histdat	a data.frame with two columns (number of successes and number of failures) containing the historical data
newdat	a data.frame with two columns (number of successes and number of failures) containing the future data
newsize	a vector containing the future cluster sizes
alternative	either "both", "upper" or "lower". alternative specifies if a prediction interval or an upper or a lower prediction limit should be computed
alpha	defines the level of confidence (1-alpha)
nboot	number of bootstraps
lambda_min	lower start value for bisection
lambda_max	upper start value for bisection
traceplot	plot for visualization of the bisection process
n_bisec	maximal number of bisection steps

Details

This function returns bootstrap calibrated prediction intervals

$$[l, u]_m = \hat{y}_m \pm q\sqrt{\hat{var}(\hat{y}_m - y_m)}$$

with $\hat{y}_m$ as the predicted future number of successes for $m = 1, ..., M$ future clusters, y_m as the observed future number of successes, $\sqrt{\hat{var}(\hat{y}_m - y_m)}$ as the prediction standard error and q as the bootstrap calibrated coefficient that approximates a quantile from a multivariate normal distribution. Please note that the predicted future number of successes is based on the future cluster size n_m and the success probability estimated from the historical data π^{hist} such that $\hat{y}_m = \pi^{hist} n_m$. Hence, the prediction intervals $[l, u]_m$ are different for each of the m future clusters, if their size is not the same.
If traceplot=TRUE, a graphical overview about the bisection process is given.

Value

If newdat is specified: A data.frame that contains the future data, the historical proportion (hist_prob), the calibrated coefficient (quant_calib), the prediction standard error (pred_se), the prediction interval (lower and upper) and a statement if the prediction interval covers the future observation (cover).

If newsize is specified: A data.frame that contains the future cluster sizes (total) the historical proportion (hist_prob), the calibrated coefficient (quant_calib), the prediction standard error (pred_se) and the prediction interval (lower and upper).

If alternative is set to "lower": Lower prediction bounds are computed instead of a prediction interval.

If alternative is set to "upper": Upper prediction bounds are computed instead of a prediction interval.

Examples

```
# Historical data
bb_dat1

# Future data
bb_dat2

# Prediction interval using bb_dat2 as future data
beta_bin_pi(histdat=bb_dat1, newdat=bb_dat2, nboot=100)

# Upper prediction bound for m=3 future number of successes
# that are based on cluster sizes 40, 50, 60 respectively
beta_bin_pi(histdat=bb_dat1, newsize=c(40, 50, 60), alternative="upper", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

---

c2_dat1 *Cross-classified data (example 1)*

Description

c2_dat1 contains data that is sampled from a balanced cross-classified design.

Usage

```
c2_dat1
```

Format

A `data.frame` with 27 rows and 3 columns:

y_ijk observations

a treatment a

b treatment b

Examples

```
# loading lme4
library(lme4)

# Fitting a random effects model based on c2_dat_1
fit <- lmer(y_ijk~(1|a)+(1|b)+(1|a:b), c2_dat1)
summary(fit)

# Prediction interval using c2_dat2 as future data
```

```
lmer_pi(model=fit, newdat=c2_dat2, alternative="both", nboot=100)

# Upper prediction limit for m=3 future observations
lmer_pi(model=fit, m=3, alternative="upper", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

`c2_dat2` *Cross-classified data (example 2)*

Description

c2_dat2 contains data that was sampled from an unbalanced cross-classified design.

Usage

```
c2_dat2
```

Format

A `data.frame` with 21 rows and 3 columns:

y_ijk observations

a treatment a

b treatment b

Examples

```
# loading lme4
library(lme4)

# Fitting a random effects model based on c2_dat_1
fit <- lmer(y_ijk~(1|a)+(1|b)+(1|a:b), c2_dat1)
summary(fit)

# Prediction interval using c2_dat2 as future data
lmer_pi(model=fit, newdat=c2_dat2, alternative="both", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

`lmer_pi` *Prediction intervals for future observations based on linear random effects models*

Description

lmer_pi calculates a bootstrap calibrated prediction interval for one or more future observation(s) based on linear random effects models

Usage

```
lmer_pi(
  model,
  newdat = NULL,
  m = NULL,
  alternative = "both",
  alpha = 0.05,
  nboot = 10000,
  lambda_min = 0.01,
  lambda_max = 10,
  traceplot = TRUE,
  n_bisec = 30
)
```

Arguments

`model`	a random effects model of class lmerMod
`newdat`	a `data.frame` with the same column names as the historical data on which the model depends
`m`	number of future observations
`alternative`	either "both", "upper" or "lower". `alternative` specifies if a prediction interval or an upper or a lower prediction limit should be computed
`alpha`	defines the level of confidence (1-`alpha`)
`nboot`	number of bootstraps
`lambda_min`	lower start value for bisection
`lambda_max`	upper start value for bisection
`traceplot`	plot for visualization of the bisection process
`n_bisec`	maximal number of bisection steps

Details

This function returns a bootstrap calibrated prediction interval

$$[l, u] = \hat{y} \pm q\sqrt{\hat{var}(\hat{y} - y)}$$

with $\hat{y}$ as the predicted future observation, y as the observed future observations, $\sqrt{\hat{var}(\hat{y}-y)}$ as the prediction standard error and q as the bootstrap calibrated coefficient that approximates a multivariate t-distribution.
Please note that this function relies on linear random effects models that are fitted with lmer() from the lme4 package.Random effects have to be specified as `(1|random_effect)`.
If traceplot=TRUE, a graphical overview about the bisection process is given.

Value

If `newdat` is specified: A `data.frame` that contains the future data, the historical mean (hist_mean), the calibrated coefficient (quant_calib), the prediction standard error (pred_se), the prediction interval (lower and upper) and a statement if the prediction interval covers the future observation (cover).

If `m` is specified: A `data.frame` that contains the number of future observations (m) the historical mean (hist_mean), the calibrated coefficient (quant_calib), the prediction standard error (pred_se) and the prediction interval (lower and upper).

If `alternative` is set to "lower": Lower prediction limits are computed instead of a prediction interval.

If `alternative` is set to "upper": Upper prediction limits are computed instead of a prediction interval.

Examples

```
# loading lme4
library(lme4)

# Fitting a random effects model based on c2_dat_1
fit <- lmer(y_ijk~(1|a)+(1|b)+(1|a:b), c2_dat1)
summary(fit)

# Prediction interval using c2_dat2 as future data
lmer_pi(model=fit, newdat=c2_dat2, alternative="both", nboot=100)

# Upper prediction limit for m=3 future observations
lmer_pi(model=fit, m=3, alternative="upper", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

`pi_rho_est` *Estimation of the binomial proportion and the intra class correlation.*

Description

pi_rho_est estimates the overall binomial proportion $\hat{\pi}$ and the intra class correlation $\hat{\rho}$ of data that is assumed to follow the beta-binomial distribution. The estimation of $\hat{\pi}$ and $\hat{\rho}$ is done following the approach of Lui et al. 2000.

Usage

```
pi_rho_est(dat)
```

Arguments

dat	a data.frame with two columns (successes and failures)

Value

a vector containing estimates for π and ρ

References

Lui, K.-J., Mayer, J.A. and Eckhardt, L: Confidence intervals for the risk ratio under cluster sampling based on the beta-binomial model. Statistics in Medicine.2000;19:2933-2942. https://doi.org/10.1002/1097-0258(20001115)19:21<2933::AID-SIM591>3.0.CO;2-Q

Examples

```
# Estimates for bb_dat1
pi_rho_est(bb_dat1)
```

---

qb_dat1 *Quasi-binomial data (example 1)*

Description

This data set contains sampled quasi-binomial data from from 10 clusters each of size 50. The data set was sampled with `rqbinom(n=10,size=50,prob=0.1,phi=3)`.

Usage

```
qb_dat1
```

Format

A data.frame with 3 rows and 2 columns:

succ number of successes

fail number of failures

Examples

```
# Upper prediction limit for m=3 future observations
# that are based on cluster sizes 40, 50, 60 respectively
quasi_bin_pi(histdat=qb_dat1, newsize=c(40, 50, 60), alternative="upper", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

---

qb_dat2 *Quasi-binomial data (example 2)*

Description

This data set contains sampled quasi binomial data from 3 clusters with different size.The data set was sampled with `rqbinom(n=3,size=c(40,50,60),prob=0.1,phi=3)`.

Usage

```
qb_dat2
```

Format

A `data.frame` with 3 rows and 2 columns:

succ number of successes

fail number of failures

Examples

```
# Prediction interval using qb_dat2 as future data
quasi_bin_pi(histdat=qb_dat1, newdat=qb_dat2, nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

qp_dat1 *Quasi-poisson data (example 1)*

Description

This data set contains sampled quasi-poisson data for 10 clusters. The data set was sampled with `rqpois(n=10,lambda=50,phi=3)`.

Usage

```
qp_dat1
```

Format

An integer vector with ten entries containing quasi poisson data

Examples

```
# Upper prediction limit for m=3 future observations
quasi_pois_pi(histdat=data.frame(qp_dat1), m=3, alternative="upper", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

qp_dat2 *Quasi-poisson data (example 2)*

Description

This data set contains sampled quasi-poisson data for 3 clusters. The data set was sampled with `rqpois(n=3,lambda=50,phi=3)`.

Usage

```
qp_dat2
```

Format

An integer vector with three entries containing quasi poisson data

Examples

```
# Prediction interval using qp_dat2 as future data
quasi_pois_pi(histdat=data.frame(qp_dat1), newdat=data.frame(qp_dat2), nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

quasi_bin_pi *Prediction intervals for quasi-binomial data*

Description

quasi_bin_pi calculates bootstrap calibrated prediction intervals for binomial data with constant overdispersion (quasi-binomial assumption).

Usage

```
quasi_bin_pi(
  histdat,
  newdat = NULL,
  newsize = NULL,
  alternative = "both",
  alpha = 0.05,
  nboot = 10000,
  lambda_min = 0.01,
  lambda_max = 10,
  traceplot = TRUE,
  n_bisec = 30
)
```

Arguments

histdat	a data.frame with two columns (success and failures) containing the historical data
newdat	a data.frame with two columns (success and failures) containing the future data
newsize	a vector containing the future cluster sizes
alternative	either "both", "upper" or "lower". alternative specifies if a prediction interval or an upper or a lower prediction limit should be computed
alpha	defines the level of confidence (1-alpha)
nboot	number of bootstraps
lambda_min	lower start value for bisection
lambda_max	upper start value for bisection
traceplot	plot for visualization of the bisection process
n_bisec	maximal number of bisection steps

Details

This function returns bootstrap calibrated prediction intervals

$$[l, u]_m = \hat{y}_m \pm q\sqrt{\hat{var}(\hat{y}_m - y_m)}$$

with $\hat{y}_m$ as the predicted future number of successes for $m = 1, ..., M$ future clusters, y_m as the observed future number of successes, $\sqrt{v\hat{a}r(\hat{y}_m - y_m)}$ as the prediction standard error and q as the bootstrap calibrated coefficient that approximates a quantile from a multivariate normal.
Please note that the predicted future number of successes is based on the future cluster size n_m and the success probability estimated from the historical data π^{hist} such that $\hat{y}_m = \pi^{hist} n_m$. Hence, the prediction intervals are different for each of the m future clusters, if their size is not the same.
If traceplot=TRUE, a graphical overview about the bisection process is given.

Value

If `newdat` is specified: A `data.frame` that contains the future data, the the historical proportion (hist_prob), the calibrated coefficient (quant_calib), the prediction standard error (pred_se), the prediction interval (lower and upper) and a statement if the prediction interval covers the future observation (cover).

If `newsize` is specified: A `data.frame` that contains the future cluster sizes (total) the the historical proportion (hist_prob), the calibrated coefficient (quant_calib), the prediction standard error (pred_se) and the prediction interval (lower and upper).

If `alternative` is set to "lower": Lower prediction bounds are computed instead of a prediction interval.

If `alternative` is set to "upper": Upper prediction bounds are computed instead of a prediction interval.

Examples

```
#' # Historical data
qb_dat1

# Future data
qb_dat2

# Prediction interval using qb_dat2 as future data
quasi_bin_pi(histdat=qb_dat1, newdat=qb_dat1, nboot=100)

# Upper prediction bound for m=3 future observations
# that are based on cluster sizes 40, 50, 60 respectively
quasi_bin_pi(histdat=qb_dat1, newsize=c(40, 50, 60), alternative="upper", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

`quasi_pois_pi` *Prediction intervals for quasi-poisson data*

Description

quasi_pois_pi calculates bootstrap calibrated prediction intervals for poisson data with constant overdispersion (quasi-poisson).

Usage

```
quasi_pois_pi(
  histdat,
  newdat = NULL,
  m = NULL,
  alternative = "both",
  alpha = 0.05,
  nboot = 10000,
  lambda_min = 0.01,
  lambda_max = 10,
  traceplot = TRUE,
  n_bisec = 30
)
```

Arguments

histdat	a data.frame with one column containing the historical data
newdat	a data.frame with one column containing the future data
m	number of future clusters
alternative	either "both", "upper" or "lower". alternative specifies if a prediction interval or an upper or a lower prediction limit should be computed
alpha	defines the level of confidence (1-alpha)
nboot	number of bootstraps
lambda_min	lower start value for bisection
lambda_max	upper start value for bisection
traceplot	plot for visualization of the bisection process
n_bisec	maximal number of bisection steps

Details

This function returns a bootstrap calibrated prediction interval

$$[l, u] = \hat{y} \pm q\sqrt{\hat{var}(\hat{y} - y)}$$

with $\hat{y}$ as the predicted future observation, y as the observed future observations, $\sqrt{\hat{var}(\hat{y} - y)}$ as the prediction error and q as the bootstrap calibrated coefficient that approximates a quantile of a multivariate normal distribution.
If traceplot=TRUE, a graphical overview about the bisection process is given.

Value

If newdat is specified: A data.frame that contains the future data, the historical mean (hist_mean), the calibrated coefficient (quant_calib), the prediction error (pred_se), the prediction interval (lower and upper) and a statement if the prediction interval covers the future observation (cover).

If m is specified: A data.frame that contains the number of future observations (m) the historical mean (hist_mean), the calibrated coefficient (quant_calib), the prediction error (pred_se) and the prediction interval (lower and upper).

If `alternative` is set to "lower": Lower prediction bounds are computed instead of a prediction interval.

If `alternative` is set to "upper": Upper prediction bounds are computed instead of a prediction interval.

Examples

```
#' # Historical data
qp_dat1

# Future data
qp_dat2

# Prediction interval using bb_dat2 as future data
quasi_pois_pi(histdat=data.frame(qp_dat1), newdat=data.frame(qp_dat2), nboot=100)

# Upper prediction bound for m=3 future observations
quasi_pois_pi(histdat=data.frame(qp_dat1), m=3, alternative="upper", nboot=100)

# Please note that nboot was set to 100 in order to increase computing time
# of the example. For a valid analysis set nboot=10000.
```

---

rbbinom *Sampling of beta-binomial data*

Description

rbbinom samples beta-binomial data according to Menssen and Schaarschmidt (2019).

Usage

```
rbbinom(n, size, prob, rho)
```

Arguments

`n`	defines the number of clusters (i)
`size`	integer vector defining the number of trials per cluster (n_i)
`prob`	probability of success on each trial (π)
`rho`	intra class correlation (ρ)

Details

For beta binomial data with $i = 1, ...I$ clusters, the variance is

$$var(y_i) = n_i\pi(1-\pi)(1+(n_i-1)\rho)$$

with ρ as the intra class correlation coefficient

$$\rho = 1/(1 + a + b).$$

For the sampling $(a + b)$ is defined as

$$(a + b) = (1 - \rho)/(\rho)$$

where $a = \pi(a + b)$ and $b = (a + b) - a$. Then, the binomial proportions for each cluster are sampled from the beta distribution

$$\pi_i \sim Beta(a, b)$$

and the number of successes for each cluster are sampled to be

$$y_i \sim Bin(n_i, \pi_i).$$

In this parametrization $E(\pi_i) = \pi = a/(a + b)$ and $E(y_i) = n_i\pi$. Please note, that $1 + (n_i - 1)\rho$ is a constant if all cluster sizes are the same and hence, in this special case, also the quasi binomial assumption is fulfilled.

Value

a `data.frame` with two columns (succ, fail)

References

Menssen M, Schaarschmidt F.: Prediction intervals for overdispersed binomial data with application to historical controls. Statistics in Medicine. 2019;38:2652-2663. https://doi.org/10.1002/sim.8124

Examples

```
# Sampling of example data
set.seed(234)
bb_dat1 <- rbbinom(n=10, size=50, prob=0.1, rho=0.06)
bb_dat1


set.seed(234)
bb_dat2 <- rbbinom(n=3, size=c(40, 50, 60), prob=0.1, rho=0.06)
bb_dat2
```

`rqbinom` *Sampling of overdispersed binomial data with constant overdispersion*

Description

rqbinom samples overdispersed binomial data with constant overdispersion from the beta-binomial distribution such that the quasi-binomial assumption is fulfilled.

Usage

```
rqbinom(n, size, prob, phi)
```

Arguments

n	defines the number of clusters (i)
size	integer vector defining the number of trials per cluster (n_i)
prob	probability of success on each trial (π)
phi	dispersion parameter (Φ)

Details

It is assumed that the dispersion parameter (Φ) is constant for all $i = 1, ...I$ clusters, such that the variance becomes

$$var(y_i) = \Phi n_i \pi (1 - \pi).$$

For the sampling $(a + b)_i$ is defined as

$$(a + b)_i = (\Phi - n_i)/(1 - \Phi)$$

where $a_i = \pi(a + b)_i$ and $b_i = (a + b)_i - a_i$. Then, the binomial proportions for each cluster are sampled from the beta distribution

$$\pi_i \sim Beta(a_i, b_i)$$

and the numbers of succes for each cluster are sampled to be

$$y_i \sim Bin(n_i, \pi_i).$$

In this parametrization $E(\pi_i) = \pi$ and $E(y_i) = n_i \pi$. Please note, the quasi-binomial assumption is not in contradiction with the beta-binomial distribution if all cluster sizes are the same.

Value

a `data.frame` with two columns (succ, fail)

Examples

```
# Sampling of example data
set.seed(456)
qb_dat1 <- rqbinom(n=10, size=50, prob=0.1, phi=3)
qb_dat1

set.seed(456)
qb_dat2 <- rqbinom(n=3, size=c(40, 50, 60), prob=0.1, phi=3)
qb_dat2
```

`rqpois` *Sampling of overdispersed poisson data with constant overdispersion*

Description

rqpois samples overdispersed poisson data with constant overdispersion from the negative-binomial distribution such that the quasi-poisson assumption is fulfilled. The following description of the sampling process is based on the parametrization used by Gsteiger et al. 2013.

Usage

```
rqpois(n, lambda, phi)
```

Arguments

`n`	defines the number of clusters (i)
`lambda`	defines the overall poisson mean (λ)
`phi`	dispersion parameter (Φ)

Details

It is assumed that the dispersion parameter (Φ) is constant for all $i = 1, ...I$ clusters, such that the variance becomes

$$var(y_i) = \lambda(1 + \lambda\kappa) = \Phi\lambda.$$

For the sampling κ is defined as

$$\kappa = (\Phi - 1)/(\lambda)$$

where $a = 1/\kappa$ and $b = 1/(\kappa\lambda)$. Then, the poisson means for each cluster are sampled from the gamma distribution

$$\lambda_i \sim Gamma(a, b)$$

and the observations per cluster are sampled to be

$$y_i \sim Pois(\lambda_i).$$

Please note, that the quasi-poisson assumption is not in contradiction with the negative-binomial distribution if the data structure is defined by the number of clusters only (which is the case here), rather than by a complex randomization structure.

Value

a vector containing the sampled observations (one per cluster)

References

Gsteiger, S., Neuenschwander, B., Mercier, F. and Schmidli, H. (2013): Using historical control information for the design and analysis of clinical trials with overdispersed count data. Statist. Med., 32: 3609-3622. https://doi.org/10.1002/sim.5851

Examples

```
set.seed(123)
qp_dat1 <- rqpois(n=10, lambda=50, phi=3)
qp_dat1

set.seed(123)
qp_dat2 <- rqpois(n=3, lambda=50, phi=3)
qp_dat2
```

Index

Chapter 3

Appendix

3.1 Curriculum Vitae

Career

since 12/2016	PhD candidate Department of Biostatistics, Institute of Cell Biology and Biophysics Leibniz University Hanover
10/2014 - 10/2016	Master of Science "Horticultural Science" Leibniz University Hanover Thesis: Diversity studies on 15 East African accessions of cowpea (*Vigna unguiculata*)
09/2011 - 08/2014	Bachelor of Science "Horticultural Production" University of Applied Sciences Osnabrück Thesis: Wirkung von *Mlo*-Knockout-Mutationen bei Tabak
03/2011 - 07/2011	Internship at the tree nursery Magni Piante in Pistoia, Italy Horticultural production for the European market
07/2010 - 12/2010	Community service Christopherus-Werk Lingen e.V.
08/2009 - 06/2010	University of applied sciences entrance qualification (Fachhochschulreife) Berufsbildende Schulen Meppen Landwirtschaftliche und Hauswirtschaftliche Fachrichtungen
08/2006 - 06/2009	Apprenticeship: Horticulturist (Gärtner der Fachrichtung Baumschule) Diderk Heinje Baumschulen Edewecht
08/2001 - 07/2006	Secondary school certificate (Erweiterter Sekundarabschluss I) Friedensschule Lingen

International experience

02/2019 - 04/2019	Research stay in Christchurch, New Zealand University of Canterbury College of Engineering, School of Mathematics and Statistics In collaboration with Dr. Daniel Gerhardt Topic: Confidence intervals for the effective dose in nonlinear hierarchical models Funding: IP@Leibniz Scholarship
10/2015 - 12/2015	Research stay in Juja, Kenya Jomo Kenyatta University of Agriculture and Technology Department of Horticulture In collaboration with Prof. Dr. Mary Abukutsa-Onyango Topic: Field trial with cowpea (*Vigna unguiculata*) for my Master-Thesis Funding: Promos Scholarship
09/2013	Summerschool in Hefei, China Anhui Agriculture University Topic: Sustainable Land Use and Resource Protection

3.2 List of publications

1. Menssen M, Schaarschmidt F (2021):
   Prediction intervals for all of M future observations based on linear random effects models.
   Statistica Neerlandica. Published online,
   DOI: https://doi.org/10.1111/stan.12260

2. Niemann J-U, Menssen M, Poehling H-M (2021):
   Manipulation of landing behaviour of two whitefly species by reflective foils. Journal of Plant Diseases and Protection. 128:97-108,
   DOI: https://doi.org/10.1007/s41348-020-00394-y

3. Grunewaldt-Stöcker G, Popp C, Baumann A, Fricke S, Menssen M, Winkelmann T, Maiss E (2020):
   Observations on early fungal infections with relevance for replant disease in fine roots of the rose rootstock Rosa corymbifera 'Laxa'. Scientific Reports. 10(1):22410,
   DOI: https://doi.org/10.1038/s41598-020-79878-8

4. Dilger N, Neehus A-L, Grieger K, Menssen M, Ngezahayo A (2020):
   Gap Junction Dependent Cell Communication Is Modulated During Transdifferentiation of Mesenchymal Stem/Stromal Cells Towards Neuron-Like Cells. Frontiers in Cell and Developmental Biology. 8:869,
   DOI: https://doi.org/10.3389/fcell.2020.00869

5. Menssen M, Schaarschmidt F (2019):
   Prediction intervals for overdispersed binomial data with application to historical controls. Statistics in Medicine. 38:2652-2663,
   DOI: https://doi.org/10.1002/sim.8124

6. Menssen M, Linde M, Omondi EO, Abukutsa-Onyango M, Dinssa FF, Winkelmann T (2017):
   Genetic and morphological diversity of cowpea (*Vigna unguiculata* (L.) Walp.) entries from East Africa. Scientia Horticulturae. 226:268-276,
   DOI: https://doi.org/10.1016/j.scienta.2017.08.003

3.3 Oral presentations

1. Schätzung der Effektiven Dosis und deren Konfidenzintervalle basierend auf hierarchischen nichtlinearen Modellen.
   10.10.2019, WeGa-Doktorandentag at the University of Applied Science, Osnabrück, Germany

2. Prediction intervals for overdispersed binomial data with application to historical controls.
   04.04.2019, research seminar of the Department of Mathematics and Statistics, University of Canterbury in Christchurch, New Zealand

3. Vorhersage-Intervalle für die Risikobewertung von Pestiziden.
   12.10.2018, WeGa-Doktorandentag at the Leibniz University Hannover, Germany

3.4 Poster presentation

1. Prediction intervals for overdispersed binomial data.
   12.10.2017, WeGa-Doktorandentag at the Leibniz University Hannover, Germany

3.5 Danksagung

Da eine wissenschaftliche Arbeit - egal welcher Form oder Disziplin - nie einfach nur die Leistung einer einzelnen Person ist, sondern vor allem von fachlichem und ideellem Austausch lebt und die Unterstützung von Freunden, Kollegen und Familie manchmal unendlich viel mehr wert ist als alle Paper der Welt, bin ich diversen Leuten zu Dank verpflichtet:

Frank: Als erstes muss ich dir dafür danken, dass du mir die Chance gegeben hast bei dir zu promovieren. Du hast super viel Zeit, Arbeit und Mühe in meine Ausbildung gesteckt und hattest für jede noch so schräge Frage ein offenes Ohr. Das war (und ist) echt Spitze! Danke dafür! Die vielen Kaffees und die damit verbundenen lustigen Gespräche über Gott und die Welt dürfen an dieser Stelle natürlich nicht unerwähnt bleiben. Dementsprechend habe ich von dir auch abseits von Statistik so einiges gelernt auf das ich von selbst nicht gekommen wäre. Auch dafür hab Dank!

Herr Hothorn: Ihnen gebührt mein Dank dafür, dass Sie meine Einstellung als Doktorand ermöglicht haben und mir damit die Chance zur Promotion gaben. Bedanken muss ich mich natürlich auch dafür, dass Sie meine Arbeit begutachtet haben.

Papa: Auch wenn du nicht mehr bei uns bist, muss ich dir an dieser Stelle trotzdem für alles danken. Du warst der beste Vater den man sich nur wünschen kann und du bist und bleibst mein Vorbild in den meisten Dingen. Durch dein Tun und Handeln hast du mir beigebracht neugierig zu sein und damit den Grundstein für diese Arbeit gelegt.

Christin: Aus Platzgründen mache ich es kurz. Vielen Dank für alles!!! Du bist die Beste! Dass du in allen Lebenslagen für mich da bist, ist eins der größten Geschenke die mir das Leben bisher gemacht hat. Ohne dich wäre das alles hier nicht möglich gewesen.

Olaf: Danke für das Gegenlesen vom zweiten Paper und all die Tipps und Tricks zum Programmieren! Damit hast du mir echt geholfen.

Clemens: Besten Dank für deinen IT-Support und die vielen netten Gespräche. Da du dich ja demnächst in den wohlverdienten Ruhestand verabschiedest, wünsche ich dir jetzt schon mal viel Spaß im neuen Lebensabschnitt.

Daniel: Besten Dank dafür, dass ich dich in Neuseeland besuchen konnte. Der Aufenthalt am Ende der Welt hat mich in vielerlei Hinsicht echt weiter gebracht!

David: Besten Dank für das Gegengelesen der Arbeit und die vielen guten Anregungen und Kommentare!

Darüber hinaus danke ich allen Mitarbeitenden vom Institut, den Leuten von WeGa-PhD und allen anderen die hier an der Uni ihre Zeit mit mir verbracht haben, für die vielen guten Stunden, die vielen netten Gespräche und die vielen gemeinsam getrunkenen Kaffees. Ohne euch wäre meine Doktorandenzeit lange nicht so gut gewesen wie sie es letztendlich war.