UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Bayesian Prediction Intervals for Clinical Assay Quality Monitoring

Stefan Gehrig
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Christel FAES

SUPERVISOR :
Mikaël LE BOUTER

2024
2025

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Bayesian Prediction Intervals for Clinical Assay Quality Monitoring

Stefan Gehrig

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Christel FAES

SUPERVISOR :
Mikaël LE BOUTER

Bayesian Prediction Intervals for Clinical Assay Quality Monitoring

Master Thesis Report
2024-2025

Master of Statistics and Data Science
Hasselt University

Student:

Stefan Gehrig (2158857)

Internal supervisor:

Prof. Dr. Christel Faes

External supervisors:

An Tran Ly Binh (GSK)
Mikaël Le Bouter (GSK)

Submission Date: 17 June, 2025

Abstract

In the field of chemistry, manufacturing and controls, biostatistical methods contribute to characterization, quality control and monitoring of processes. One such application concerns the construction of a quality control (QC) range for clinical assays. The QC range indicates within which limits future measurements of a QC sample are expected to lie, based on an estimate of assay precision. The QC range provides a way of identifying unreliable measurement runs or an "out of control" analytical assay. Here, we study how Bayesian methods, in particular Bayesian hierarchical models, can be used to construct QC ranges for a vaccine assay via posterior predictive distributions. Various weakly informative and one non-informative prior specifications for the variance components are proposed, and the method is implemented via R interfaces to the probabilistic programming language Stan. In a simulation study, we evaluate the performance of the Bayesian QC range in terms of expected prediction coverage, expected interval width, and root mean squared prediction coverage error. For comparison, frequentist prediction intervals based on more classical modeling approaches are also evaluated as candidates for the QC range. This includes a simple, single-level model that ignores clustering of assay measurements within runs, as well as a hierarchical model estimated with restricted maximum likelihood. The methods are applied in 30 realistic data-generation scenarios that vary according to number of total observations, number of clusters, balancedness, assay variability, and strength of intra-run correlation. Results show generally good QC range performance of the hierarchical models, both frequentist and Bayesian, whereas the single-level model leads to too narrow intervals in most scenarios. In settings with only a few runs from a highly heterogeneous population, which are challenging for any estimation method, only the Bayesian methods prevent predidction undercoverage. Yet, the Bayesian QC ranges have a tendency to become too wide in other scenarios. Based on our results, weakly informative priors, inducing soft constraints on plausible ranges based on scientific expertise, should be preferred over non-informative priors. We discuss nuances with respect to prior choice and suggest several extensions and routes for further investigation. Due to their performance and flexibility, Bayesian hierarchical models represent a promising approach for vaccine assay quality control.

Contents

1 Introduction

Biostatistical methods contribute significantly to the science of chemistry, manufacturing and controls (CMC), a cornerstone of biopharmaceutical product development (Faya & Pourmohamad, 2022b). CMC focuses on activities like the characterization or control of products or processes. In CMC, statistical theory and methods play a big part in decisions-making about, for example, shelf-life durations, the release of product batches, the experimental designs used for quality assurance, or the implementation of biological measurement methods (bioassays).

It has been argued that Bayesian statistics, with its flexibility, its opportunity to incorporate external information in sparse-data settings, and its natural way of dynamically updating inferences, is particularly suited for CMC applications (Faya & Pourmohamad, 2022b; Peterson, 2020). Nevertheless, Bayesian methods are probably still underused in CMC relative to the advantages they bring for some applications. In fact, prominent regulatory documents for CMC "are silent on the use of Bayesian methods" (Faya & Pourmohamad, 2022a, p. 2). This is in contrast to the long-standing adoption of Bayesian statistics for clinical trial design and analysis (e.g., Spiegelhalter et al., 2004), as well as the growing body of research applying Bayesian statistics for CMC. In particular, the *Bayesian hierarchical model* (BHM) has been picked up increasingly by researchers for addressing various challenges in biopharmaceutical product development. The reason is that scientists commonly deal with multiple levels and sources of variation in processes and products, or with multiple related populations of processes and products (e.g., Lewis & Hudson-Curtis, 2022; Schach et al., 2025). Such settings call for methods for clustered data, information borrowing, and variance decomposition – which is precisely what (Bayesian) hierarchical models can provide (Gelman & Hill, 2006).

One area of CMC in which the BHM promises to be useful is the validation and monitoring of analytical assays (Lebrun & Rozet, 2020; Novick et al., 2021). These are methods (e.g., bioassays) to quantify specific chemical substances in a sample. It could be a human biological sample that is measured to inform clinical care or clinical development (i.e., clinical assays). Sources of random variation between assay measurements include different labs, analysts, or equipment (see Section 2.1). The BHM provides a way to analyze and account for them during inference and decision-making.

In this study, we will specify and implement a BHM for quality control and monitoring of a clinical assay in the field of vaccine development (see Section 2.1) and evaluate it in a simulation study, along with commonly used alternative methods. The statistical requirements and data-generating scenarios are developed in close collaboration with the external partner GSK. The GSK Vx-Clinical Assays Statistics team currently mainly relies on frequentist statistical methods for assay monitoring. The computation of control ranges for assay monitoring might benefit from extending the toolbox to Bayesian methods. To provide first steps in this direction, this study provides examples of possible model definition, choices of prior distributions, and techniques for posterior inference about predictive distributions used as quality control ranges.

In Section 2, we provide the necessary background on analytical assay precision and quality control and summarize our research objectives. Section 3 first explains the overall design of the simulation study (Section 3.1), then the different statistical models and approaches for constructing quality control ranges (Section 3.2), along with the software implementation (Section 3.3), and the analysis of simulation results for performance evaluation (Section 3.4). After presenting the results (Section 4), we discuss them in light of the application and the broader literature (Section 5), concluding with recommendations for method usage (Section 6).

2 Background on application

2.1 Assay precision metrics and measurement scales

The methods examined in this study apply to the quality control of vaccine immunoassays. These clinical bioassays are intended to quantify immune response by measuring the concentration of antibodies from the human body in a biological sample. They play an important role during clinical development of a vaccine candidate. Mechanistically, these could be ligand-binding assays like ELISAs (enzyme-linked immunosorbent assays) or functional assays (Dessy et al., 2024). The measurements of the immunoassay exhibit variability due to observed, but also unobserved factors and should hence be modelled as arising from a stochastic process. The variability of any assay must be understood and controlled to allow sound decision-making according to the intended purpose of the assay (Lebrun & Rozet, 2020). In principle, the exact value of the next result produced by the assay is always uncertain and statistical methods are necessary to cope with this uncertainty.

The *total error* is a core performance characteristic and a critical quality attribute of analytical assays (DeSilva et al., 2003; Dessy et al., 2024; Junker et al., 2015): How closely does the measurement procued by the assay agree with the underlying theoretical value of the analyte's concentration? It can further be decomposed into *systematic error* (often also called accuracy, bias or trueness) and *random error* (precision). The definition of the systematic error is somewhat delicate for vaccine immunoassays: "a 'true' accuracy assessment is often impossible, because highly purified and/or fully characterized reference material does not usually exist" (Dessy et al., 2024, p. 1070). Therefore, in practice, one typically accepts certain reference samples of unknown true composition as a standard of comparison.

This study focuses on statistical methods for inference on the random component of the total error, i.e., assay precision. There is not only one precision. Clearly, measurements produced in the same lab, by the same person, on the same day will vary less than measurements taken by, say, different labs or different persons or on different days, even if all assessments measure the same biological sample. This notion is captured in the definitions of different assay precision components (Dessy et al., 2024, sec. 3.2). Repeatability, or within-run precision, refers to variability of measurements under constant conditions, usually taken within a short time frame. On the other extreme, reproducibility refers to variability of measurements under vastly different conditions like different labs. In-between these extremes, the so-called *intermediate precision* of the assay is an important characteristic. It refers to the variability of measurements from the same lab, but taken during different *runs* of the assay. Runs may differ in terms of, for example, time, analyst, equipment, or reagent lots. Intermediate precision is important because it reflects the variability introduced by the way assays are operated during routine testing. Intermediate precision arises from random variation at two different levels:

$$\text{Intermediate precision variance} = \text{Within-run variance} + \text{Between-run variance}.$$

As it results from summing multiple variances, intermediate precision is also sometimes referred to as "total random error" (DeSilva et al., 2003) or "total variance/variability" (Francq et al., 2019) of an assay. We adopt the latter term in some instances where its meaning is unambiguous.

Intermediate precision is often expressed in terms of the intermediate precision percent coefficient of variation (CV), which we denote by $\%\text{CV}_{\text{IP}}$. The CV is defined as the standard deviation of a distribution divided by the mean. The ratio is often multiplied by 100, and we denote the resulting CV on the percentage scale by %CV. The metric quantifies relative variability (Lewontin, 1966). For analytical assays, the metric generalizes over a larger range of concentration levels than does the standard

deviation. The reason is that concentration measurements usually vary more at greater magnitudes, leading to the standard deviation increasing (approximately) in proportion with the mean (Dessy et al., 2024, sec. 3.2.7), at least for some concetration range.[1]

Conveniently, the assumption of constant CV, i.e., constant relative variability, on the original measurement scale implies that after taking logarithms, the transformed measurements will exhibit constant variance, irrespective of the mean (e.g., Canchola et al., 2017; Lewontin, 1966). Hence, log transformation allows the subsequent application of statistical models with independent location and scale parameters, most prominently the normal distribution. Due to better fit of the normal model, log-transformed values are commonly used for analyzing laboratory results (e.g., West, 2022), including results from analytical assays. As Dessy et al. (2024, p. 1074) note, the "use of a pre-specified data transformation (e.g., logarithm) may [...] improve the performance of the statistical analyses. For example, ELISAs are typically assumed to follow a log-normal distribution".

A $\%\mathrm{CV}_{\mathrm{IP}}$ on the original measurement scale can be mathematically transformed into an intermediate precision variance on the log-transformed scale, which we denote by $\mathrm{Var}_{\mathrm{IP}}$. This transformation and back-transformation will be useful when specifying and interpreting our statistical models for quality control range determination (Section 3). We use logarithms with base 10.[2] The formula for transformation, derived under a log-normal assumption for the distribution of measurements is (Canchola et al., 2017; Dessy et al., 2024, p. 1078)

$$\%\mathrm{CV}_{\mathrm{IP}} = \sqrt{\exp\left(\mathrm{Var}_{\mathrm{IP}} \cdot \log(10)^2\right) - 1} \times 100, \tag{1}$$

with inverse transformation

$$\mathrm{Var}_{\mathrm{IP}} = \frac{\log\left(\left(\frac{\%\mathrm{CV}_{\mathrm{IP}}}{100}\right)^2 + 1\right)}{\log(10)^2}.$$

Equation 1 shows that knowing the variance of the $\log_{10}$-transformed measurements is sufficient to compute the CV of the measurements on their original scale. We emphasize that $\%\mathrm{CV}_{\mathrm{IP}}$ always refers to variation among *untransformed* measurements coming *from different runs*.

2.2 Quality control range for vaccine assays

A vaccine assay that has been qualified for its purpose needs quality control and monitoring to ensure that performance consistently remains within acceptance limits (Dessy et al., 2024, sec. 2.2.3). Even if certain assay characteristics, like limit of detection, analytical range, or precision (e.g., in terms of $\%\mathrm{CV}_{\mathrm{IP}}$), have been described and validated during assay development, they might not be guaranteed over longer time frames of "real-life" usage. Imagine a vaccine clinical development program that takes place over multiple months or years and requires many repeated measurements of immune response across multiple samples. Quality control is important to ascertain that measurements remain reliable. Otherwise, this will ultimately pose risks to patients.

Therefore, different regulatory guidances and quality management frameworks for CMC apply directly or indirectly to the development and control of analytical methods. For example, Quality-by-Design

[1] A more general variance function, which usually applies in practice, poses that the standard deviation scales as a power function of the mean concentration (Lebrun & Rozet, 2020, p. 383). Only for power 1 this implies strictly constant CV across all concentration levels. For power 0, we have a constant standard deviation across all concentration levels, which is not realistic for bioassays.

[2] Throughout, we will write $\log_{10}$ for the logarithm with base 10 and log for the natural logarithm.

is an important regulatory and scientific initiative in the pharmaceutical industry that "begins with predefined objectives and emphasizes product and process understanding and process control, based on sound science and quality risk management" (ICH, 2009, p. 16). Despite its focus on *manufacturing processes*, the objectives and components of the approach can one-to-one be adapted to *measurement processes* like analytical assays (Junker et al., 2015; Lebrun & Rozet, 2020). For example, process validation guidance for industry by FDA (2011, p. 4) emphasizes the need for "continued process verification" during commercial production to ensure "that the process remains in a state of control". Similarly, guideline ICH Q9 calls for "monitoring systems that are capable of detecting departures from a state of control and deficiencies in manufacturing processes" (ICH, 2023, p. 11). This can be translated into the realm of analytical assays as the need for "monitoring and assessing the method's state of control" (Junker et al., 2015, p. 494).

As one means of assay quality control, it is standard practice to include quality control (QC) samples in every run of analytical assays (DeSilva et al., 2003; Dessy et al., 2024; Junker et al., 2015; Lebrun & Rozet, 2020). These samples contain a previously established concentration of the analyte and are expected to deliver consistent results over time, subject only to random error. Their inclusion serves two primary purposes. First, they provide a run-to-run check for the reliability of results from test samples, guiding the decision to accept or reject a run. Second, accumulated data over time allows for the detection of long-term trends in assay results, which may indicate that the assay has systematically gone "out of control". Both is only possible if an acceptable range of results for the QC samples has been clearly defined, as it provides the benchmark against which deviations and trends can be identified. Such a range should respect that results are inherently random and governed by the precision of the assay (see Section 2.1). We call this the *QC range*, and the statistical methods to construct this range are the main topic of this study.

A QC range is best used in conjunction with control charts, a popular tool for statistical process control, including the real-time monitoring of analytical methods (Junker et al., 2015): Measurements taken over time are entered in a chart, in which the limits of the QC range constitute horizontal control boundaries. Control charts provide a quick and easy way to monitor individual results, but also long-term trends in results. For a stylized example of a control chart, containing multiple QC ranges and simulated data, see our results Figure 2.[3]

Since the QC range expresses the range within which future measurements of QC samples are expected to lie, it is natural to draw on statistical theory for prediction intervals and predictive distributions for building it. This is also the route taken in this study. From a risk perspective for both decision-makers and patients, it is crucial that the next measurement is likely of acceptable quality – or, in general terms, that the *process* (i.e., assay) yields *products* (i.e., measurements) within specifications. For precisely this logic, predictive distributions are an important target of inference in statistical process control (Boulanger & Mutsvari, 2020; Schach et al., 2025) – sometimes more so than the underlying model parameters of the process themselves (Boulanger & Mutsvari, 2020, sec. 18.2). Following the classical definition (e.g., Casella & Berger, 2002; Tian et al., 2022), a prediction interval is a function of the random sample $\boldsymbol{Y}$ that satisfies

$$P_{\boldsymbol{\theta}}\left[L(\boldsymbol{Y}) \leq \tilde{Y} \leq U(\boldsymbol{Y})\right] = \beta, \tag{2}$$

[3] Of course, control charts can be applied either on the log-transformed or on the original measurements, if QC range limits are also transformed accordingly – acceptance and rejection of results is unaffected by the monotone transformation. This is in contrast to the statistical methods for establishing QC ranges in the first place, which usually draw on modeling assumptions that are more likely to hold under the log transformation (Section 2.1).

where $\tilde{Y}$ is the future random variable, $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ is the prediction interval with lower and upper limit, respectively, $\boldsymbol{\theta}$ are model parameters of the joint distribution of $(\boldsymbol{Y}, \tilde{Y})$, and $\beta \in [0, 1]$ is a nominal confidence level. Put simply, we require that the interval constructed from observing the sample of assay measurements $\boldsymbol{Y}$ will contain a proportion of $\beta \cdot 100\%$ of future measurements governed by parameters $\boldsymbol{\theta}$ of the process (e.g., assay precision, QC sample concentration). In practice, $\boldsymbol{\theta}$ is unknown and will be estimated. Equation 2 implies a fixed parameter $\boldsymbol{\theta}$, which contrasts with Bayesian prediction intervals, for which we marginalize over the distribution of the random parameter $\boldsymbol{\theta}$. The distinction will be made explicit in Section 3.2, when we describe our statistical models and approaches. For the sake of motivating the use of prediction intervals as QC ranges, it is not yet of prime importance.

The use of prediction intervals as QC ranges has been recommended in the literature on assay quality control, though with some diversity in the statistical approaches considered. For example, Lebrun & Rozet (2020, p. 387) state that data from "performed experiments can define Bayesian prediction intervals that can be used as initial control limits when building analytical procedure control charts". In contrast, Francq et al. (2019, sec. 5.2) draw on frequentist mixed model theory to build intervals for predicted concentration values. Irrespective of the precise approach, it is important that statistical methods for constructing the QC range are well able to capture the precision of the assay governed by the multiple variance components outlined in Section 2.1.

For process control, prediction intervals are sometimes extended to also satisfy a requirement on the confidence/probability with which a given proportion of future values are captured by the interval. This leads to so-called (type II, or "β-content") tolerance intervals (Casella & Berger, 2002, sec. 9.5.4; Francq et al., 2019; Lewis & Hudson-Curtis, 2022; Patel, 1986), which are not further considered in this study.[4]

2.3 Research objectives

The first objective of this study is to develop theoretically and practically how Bayesian methods can be used for constructing a QC range for assay quality control. This includes aspects like model definition, prior choice, interval estimation, and software. So far, external collaborators at GSK have relied on more classical, frequentist methods for constructing the QC range. Therefore, proposing and implementing a Bayesian approach for the QC range problem constitutes the first important research objective.

As second objective, we evaluate the performance of different methods to construct QC ranges. In a statistical sense, the QC ranges are prediction intervals, and they can hence be evaluated in terms of their prediction coverage and interval width. Coverage and width are conventional metrics for evaluating interval estimators (Casella & Berger, 2002, sec. 9.3).

Throughout the study, we examine the two-sided $\beta \cdot 100\%$ prediction intervals for $\beta = 0.99$. We will ask the questions: How well do the prediction intervals obtained with the different methods respect the probability of 99% to include future concentration measurements, both on average and in terms of mean deviation, and how wide are the QC ranges on average? The questions will be answered with Monte Carlo estimates of the relevant quantities from a simulation study. The evaluation metrics used in the current study are explained in more detail in Section 3.4.

[4] See Lewis & Hudson-Curtis (2022) for a Bayesian model and simulation-based implementation similar in spirit to the one used in our study, though for estimating tolerance intervals.

In particular, we compare

- a "naive" approach, which proposes a non-hierarchical (single-level) iid normal model. The model and QC range can be estimated easily, but the approach could be too simplistic under realistic settings,
- a frequentist approach to Gaussian hierarchical modeling,
- and the proposed Bayesian approach to Gaussian hierarchical modeling, using different prior distributions.

Based on the comparisons, recommendations about the future use of QC range methods for the investigated settings shall be derived.

3 Methods

3.1 Study design

3.1.1 Overview

In this simulation study, both the data-generating process (30 scenarios) and the statistical method for determining the QC range (7 methods) are varied systematically. One simulation repetition produces a random data set for each scenario. On each data set, all QC range methods are applied.

Each run of the simulation broadly resembles a stylized real-life workflow for determining and using an assay QC range:

(1) Collect baseline data to estimate mean for the QC sample and assay variability $\%\text{CV}_{\text{IP}}$.
(2) Construct a QC range based on the estimates.
(3) Apply the QC range for testing future measurement of the same QC sample.

For step (1), we manipulate the number of runs (i.e., clusters), total number of observations, and balancedness of the baseline data collection in a systematic way, leading to 5 different designs. Within each design, the true degree of total variability and strength of clustering is varied, leading to the in total 30 different scenarios (Section 3.1.2).

For step (2), we vary the assumed statistical model (2 models: single-level and hierarchical model), the estimation approach for the hierarchical model (frequentist and Bayesian approach), and, for the Bayesian approach, the prior distribution (5 prior distributions), leading to in total 7 different methods (Section 3.2). Consequently, in one simulation repetition, $30 \cdot 7 = 210$ QC ranges are computed.

For step (3), we can make use of the fact that the marginal distribution of the population of future observations – i.e., of random draws from random runs – is known in the simulation study. By using the quantiles of the known distribution, it can be exactly stated which proportion of future measurements is covered by the QC range constructed from the baseline data collected in step (1).

A stylized visualization of the results from these steps with a simulated data example is also shown in Figure 2. We simulate 750 data sets (repetitions) per scenario and method to obtain Monte Carlo estimates of QC range evaluation metrics for each scenario and method (Section 3.4).

3.1.2 Data generation scenarios

Table 1 lists all 30 investigated scenarios for generating the baseline data for the QC sample.

The scenarios fall into five broader designs with specific run configurations:

- The "Standard" case describes the design that is considered desirable in practice according to standard operating procedures: there are exactly two measurements for each run and $m = 20$ runs in total, adding up to $n = 40$ measurements in total.
- In the "Unbalanced 1" design, there are also $n = 40$ measurements, but they spread unequally across only $m = 10$ runs.
- In the "Sparse 1" design, there are less measurements ($n = 20$), and fewer runs ($m = 5$). All runs have the same size.

- In the "Unbalanced 2" design, measurements are again spread unequally across runs as in the "Unbalanced 1" design, but m and n are held constant relative to the "Standard" design.
- In the "Sparse 2" design, there are fewer runs ($m = 5$) as in the "Standard" design, but n is held constant relative to the "Standard" design.

The first three designs represent typical designs encountered in practice. The last two designs are included to allow comparisons that are better interpretable by not varying multiple parameters at the same time. We emphasize that all examined designs have relatively few and small clusters compared to many other applications of hierarchical models in, for example, clinical or social science research (e.g., Browne & Draper, 2006; Gelman & Hill, 2006).

Table 1: Settings of the data-generating process for the simulated collection of baseline data and all scenarios. Column name n refers to the total number of assay measurements (i.e., total observations); m to the number of runs (i.e., clusters); ρ to the intra-run correlation (i.e., strength of clustering); and $\%\text{CV}_{\text{IP}}$ to the intermediate precision percent coefficient of variation (i.e., total variability; see Section 2.1).

	Design configuration					
Design	n	m	Balance	Scenario	$\%\text{CV}_{\text{IP}}$	ρ
Standard	40	20	Yes	1	10	0.2
Standard	40	20	Yes	2	10	0.5
Standard	40	20	Yes	3	10	0.8
Standard	40	20	Yes	4	40	0.2
Standard	40	20	Yes	5	40	0.5
Standard	40	20	Yes	6	40	0.8
Unbalanced 1	40	10	No	7	10	0.2
Unbalanced 1	40	10	No	8	10	0.5
Unbalanced 1	40	10	No	9	10	0.8
Unbalanced 1	40	10	No	10	40	0.2
Unbalanced 1	40	10	No	11	40	0.5
Unbalanced 1	40	10	No	12	40	0.8
Sparse 1	20	5	Yes	13	10	0.2
Sparse 1	20	5	Yes	14	10	0.5
Sparse 1	20	5	Yes	15	10	0.8
Sparse 1	20	5	Yes	16	40	0.2
Sparse 1	20	5	Yes	17	40	0.5
Sparse 1	20	5	Yes	18	40	0.8
Unbalanced 2	40	20	No	19	10	0.2
Unbalanced 2	40	20	No	20	10	0.5
Unbalanced 2	40	20	No	21	10	0.8
Unbalanced 2	40	20	No	22	40	0.2
Unbalanced 2	40	20	No	23	40	0.5
Unbalanced 2	40	20	No	24	40	0.8
Sparse 2	40	5	Yes	25	10	0.2
Sparse 2	40	5	Yes	26	10	0.5
Sparse 2	40	5	Yes	27	10	0.8
Sparse 2	40	5	Yes	28	40	0.2
Sparse 2	40	5	Yes	29	40	0.5
Sparse 2	40	5	Yes	30	40	0.8

For unbalanced designs ("Balance: No" in Table 1), we decide not to fix the distribution of run sizes (i.e., cluster sizes). Rather, run sizes are randomly drawn from a distribution. The reason is that in practice the designs result by compiling data from different sources, like runs conducted during assay

development. There is no experimental protocol that governs run sizes, but run sizes vary due to which records are available to the scientists to be used as baseline data for the assay at hand. Therefore, the degree of the unbalancedness in our study varies randomly across simulated data sets, but according to a known process (see Appendix A.1). The generated run sizes are recorded and can be analyzed.

Within each of these five designs, we vary the true intermediate precision $\%\text{CV}_{\text{IP}}$ (small or large) and the intra-run correlation ρ (small, medium, or large; for a formal definition of ρ in the context of our statistical models, see Section 3.2.2) in a fully factorial fashion. Informally speaking, these two parameters control how large the total variability is in the population, and how it distributes across within-run and between-run variability.

Whereas $\%\text{CV}_{\text{IP}}$ reflects the precision of the assay, the population mean of measurements on the original measurement scale reflects the concentration level of the QC sample. In our study, it is the same in all scenarios and fixed at 100.

Simulated measurements are generated as described by Algorithm 1. Step 2 in Algorithm 1 is based on the assumption of a log-normal distribution of the observations on the original measurement scale. The step transforms the expected value of measurements on the original scale to the expected value of measurements on the $\log_{10}$ scale (e.g., Casella & Berger, 2002, p. 109). By directly sampling from the $\log_{10}$-transformed measurements, the data are already on the suitable scale for Gaussian modeling and QC range estimation (see Section 3.2).

Algorithm 1: Simulation of measurements on the $\log_{10}$ scale according to scenario parameters.

1 **Input:** Parameters of the scenario (n, m, $\%\text{CV}_{\text{IP}}$, ρ); QC sample concentration (100);
2 Transform $\%\text{CV}_{\text{IP}}$ into Var_{IP} (see section 2.1);
3 Compute mean on $\log_{10}$ scale as $\mu = \log_{10}(100) - \frac{\text{Var}_{\text{IP}}}{2} \cdot \log(10)$;
4 Sample m run-specific intercepts from $\mathcal{N}(\mu, \text{Var}_{\text{IP}} \cdot \rho)$;
5 Sample n within-run residuals from $\mathcal{N}(0, \text{Var}_{\text{IP}} \cdot (1 - \rho))$;
6 Sum run-specific intercept and within-run residual for each measurement;

3.2 Statistical models

In what follows, we describe the specification and estimation of the different statistical models to analyze the baseline data for the QC sample. From the estimated models, we compute prediction intervals for a random future observation, drawn from a random future run. The limits of the prediction intervals can be used as an analytical assay's QC range (see Section 2.2).

3.2.1 Single-level model

Denote by y_{ij} the $\log_{10}$-transformed observation (QC sample measurement) $j = 1, \dots, k_i$ from run $i = 1, \dots, m$. By n, we denote the overall number of observations $n = \sum_{i=1}^{m} k_i$. All observations from cluster i are contained in k_i-dimensional vector $\boldsymbol{y_i}$, and all observations from all clusters are contained in n-dimensional vector $\boldsymbol{y}$. A simple way to model the data-generating process would be to assume identically and independently distributed observations in a single-level model. Observations following a normal distribution with a common mean parameter μ_{s} and a common variance parameter σ_{s}^2,

$$y_{ij} \sim \mathcal{N}(\mu_{\text{s}}, \sigma_{\text{s}}^2). \tag{3}$$

Variation among data points y_{ij} is modelled with a single variance component σ_s^2. The single parameter captures both random within-run and random between-run variation, which jointly make up the intermediate precision variance of interest (Var_{IP}; see Section 2.1).

By $\bar{y}$ we denote the overall sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k_i} y_{ij}$. Parameter estimators for the assumed single-level model are easy to obtain:

$$\begin{aligned} \hat{\mu}_s &= \bar{y}, \\ \hat{\sigma}_s^2 &= \frac{1}{n-1}\sum_{i,j}(y_{ij} - \bar{y})^2. \end{aligned} \tag{4}$$

The normal variance σ_s^2 is estimated by the sample variance, also called ANOVA estimator, which, other than the maximum likelihood estimator, is unbiased in case Model 3 holds (Casella & Berger, 2002, p. 331; Searle et al., 1992, p. 249).

We can call this *pooled* estimation, since observations from all runs are pooled during parameter estimation rather than contributing to the estimation of run-specific parameters. This leads to $\hat{\sigma}_s^2$ capturing in one parameter the variability from all levels in the true data-generating process. Clustering is ignored and all observations are assumed independent. Thus, since observations from the same run are correlated under the true model (Section 3.1.2), Model 3 is misspecified. Accordingly, $\hat{\sigma}_s^2$ might not a good estimator of Var_{IP}.

The model can also be read as a simple linear regression model with only an intercept. A two-sided $\beta \cdot 100\%$ prediction interval for the random future observation $\tilde{y}$ can be obtained with standard formulae (e.g., Casella & Berger, 2002, sec. 11.3.5) and simplifies to

$$\hat{\mu}_s \pm t_{n-1,(1-\beta)/2} \sqrt{\underbrace{(1 + 1/n)\,\hat{\sigma}_s^2}_{\substack{\text{intermediate precision} \\ \text{variance estimate}}}}, \tag{5}$$

where $t_{n-1,(1-\beta)/2}$ is the $(1-\beta)/2$ quantile of the t distribution with $n-1$ degrees of freedom.

3.2.2 Hierarchical model

The following two-level Gaussian hierarchical model is proposed:

$$\begin{aligned} y_{ij} \mid \alpha_i &\sim \mathcal{N}(\alpha_i, \sigma^2), \\ \alpha_i &\sim \mathcal{N}(\mu, \tau^2). \end{aligned} \tag{6}$$

This model reflects a hierarchical structure, where observations (level 1) are nested within runs (level 2). Not only observations within runs, but also run-specific intercepts are assumed to vary independently according to a normal distribution. In the literature, this is also referred to as multilevel model, random-effects model (due to run-specific intercepts α_i being "random effects"), or mixed model – although strictly speaking, the only "fixed effect" in Equation (6) is the global intercept term μ. The model explicitly takes clustering of observations into account (e.g., Gelman & Hill, 2006). It is common to assume hierarchical models when analyzing results from analytical assays (e.g., Francq et al., 2019; Junker et al., 2015; Lebrun & Rozet, 2020; Novick et al., 2021).

Variation among data points y_{ij} is modelled as coming from two sources: random within-run (σ^2) and random between-run variation (τ^2). In this model, the intermediate precision variance (Var_{IP};

see Section 2.1) is the sum $\sigma^2 + \tau^2$ (Lebrun & Rozet, 2020, p. 381). Since generally any conditional random effects model implies a marginal model (Verbeke & Molenberghs, 2000, sec. 3.3), a marginal formulation is also possible for Model (6), given by the multivariate normal distribution

$$\boldsymbol{y}_i \sim \mathcal{N}\left(\boldsymbol{\mu}, \begin{bmatrix} \sigma^2+\tau^2 & \tau^2 & \dots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \dots & \tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 & \tau^2 & \dots & \sigma^2+\tau^2 \end{bmatrix}\right), \tag{7}$$

where vectors and matrices have dimension k_i and $k_i \times k_i$, respectively. Marginally, the population of y_{ij} follows the intermediate precision variance, shown on the diagonal of the variance-covariance matrix in Equation 7. The marginal formulation will be useful for estimating prediction intervals for randomly drawn observations from randomly drawn runs.

Further, the intra-cluster correlation defined by ratio

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2}$$

is an informative quantity. It expresses the proportion of the intermediate precision variance that is due to variation between runs, or, alternatively, the correlation between any two observations from the same run (Verbeke & Molenberghs, 2000, sec. 3.3). Higher values indicate a stronger degree of clustering.

Model (6) can be estimated in a frequentist (Section 3.2.2.1) or Bayesian framework (Section 3.2.2.2), and both will be part of the present study. Both estimation approaches can be viewed as *partially pooled*, or shrinkage estimation (Gelman & Hill, 2006, sec. 12.2): In contrast to Model (3) described in Section 3.2.1, separate, run-specific means α_i are assumed. Yet, during estimation, there is still borrowing of information between runs, as their means are modeled as coming from one common distribution. The degree of pooling will generally be larger for lower values of ρ (Lesaffre & Lawson, 2012, sec. 9.4.3). In this study, the run-specific means α_i are not of scientific interest and their estimates will not be analyzed. Specific runs represent unique applications of the assay in a specific lab at a specific time, and it is impossible to observe future data from the same run ever again.

Note that Model (6) correctly specifies the data-generating process outlined in Section 3.1.2. Outside of a simulation study, it is unlikely that analysts will fully correctly specify the model, but the proposed model might approximately hold.

3.2.2.1 Frequentist estimation and prediction

In the frequentist tradition, parameters in the model in Equation (6) are typically estimated via Restricted Maximum Likelihood (REML) estimation. In brief, REML estimation maximizes the likelihood for a linear transformation of the original data that eliminates the "fixed effects" from the likelihood (Searle et al., 1992, sec. 6.6). The reduced set of transformed data contains only linearly independent observations. In that way, when estimating variance components, REML respects that some degrees of freedom are involved in estimating the "fixed effects" – in the present case of Model (6), the global intercept μ. REML prevents possible downward bias of ML for variance estimation in finite samples.[5]

[5] In fact, the single-level model parameter estimator $\hat{\sigma}_s^2$ in Equation 4 is also a REML estimator.

Having obtained REML estimates for parameters in the hierarchical model, a two-sided $\beta \cdot 100\%$ prediction interval for the random future observation $\tilde{y}$ from a random future run can be estimated as

$$\hat{\mu} \pm t_{r,(1-\beta)/2} \sqrt{\widehat{\text{Var}}(\hat{\mu}) + \underbrace{\hat{\sigma}^2 + \hat{\tau}^2}_{\substack{\text{intermediate precision} \\ \text{variance estimate}}}}, \tag{8}$$

where r refers to the appropriate degrees of freedom of the t distribution, approximated according to Francq et al. (2019, p. 5608) as

$$r = 2\frac{(\hat{\sigma}^2 + \hat{\tau}^2)^2}{\widehat{\text{Var}}\left(\hat{\sigma}^2 + \hat{\tau}^2\right)}. \tag{9}$$

The estimate $\widehat{\text{Var}}\left(\hat{\sigma}^2 + \hat{\tau}^2\right)$ can be obtained via summing over all four cells of the estimated variance-covariance matrix of the REML estimators $\hat{\sigma}^2$ and $\hat{\tau}^2$, i.e., $\widehat{\text{Var}}\left(\hat{\sigma}^2 + \hat{\tau}^2\right) = \widehat{\text{Var}}\left(\hat{\sigma}^2\right) + \widehat{\text{Var}}\left(\hat{\tau}^2\right) + 2\,\widehat{\text{Cov}}\left(\hat{\sigma}^2, \hat{\tau}^2\right)$. The variance-covariance matrix is obtained by inverting the observed Fisher information matrix, i.e., the negative Hessian of the REML log-likelihood evaluated at the estimates. The prediction interval hence takes into account non-independent variance component estimators and adjusts the degrees of freedom accordingly.

We recognize the marginal model formulation from Equation (7) in the prediction interval in Equation (8). In fact, asymptotically, the bounds defined by the estimator in Equation (8) converge to the true $(1 \pm \beta)/2$ quantiles of the population of y_{ij} described by the distribution in Equation (7) (Francq et al., 2019, p. 5608).

3.2.2.2 Bayesian estimation and prediction

Bayesian hierarchical models are a suitable approach to analyze clustered data produced by multiple variance components (Box & Tiao, 1973, Chapter 5; Gelman et al., 2013, Chapter 5; Lesaffre & Lawson, 2012, Chapter 9). Based on the Bayesian hierarchical model, it is also straightforward to derive a full and exact distribution for a future observation, the posterior predictive distribution (PPD), rather than having to rely on an approximation of prediction interval limits (Section 3.2.2.1).

The likelihood part of the BHM is described by the two-level model in Equation (6), equivalent to the frequentist approach. The three parameters are assigned a prior distribution $p(\mu, \sigma^2, \tau^2)$. For this study, we will mostly, but not always, choose independent priors that allow the factorization $p(\mu)\ p(\sigma^2)\ p(\tau^2)$. The overall mean parameter is assigned a very dispersed prior with

$$\mu \sim \mathcal{N}(0, 100^2).$$

It is essentially uninformative, considering that μ is the mean concentration level on the $\log_{10}$-transformed scale. This could reflect a sensible default if no prior information about the concentration level should be incorporated, but an improper prior, like $p(\mu) \propto 1$, should be avoided. We will not vary it as part of the current study. In conventional measurement units and before $\log_{10}$ transformation, concentrations of QC samples are typically in the tens, hundreds (as in our simulated data, see Section 3.1.2), or thousands, which are orders of magnitude comfortably covered by the prior distribution above. There is typically enough information in the data to estimate the overall mean with good precision. For variance components, instead, different choices of prior distributions are evaluated as part of the current research design (Section 3.1). Precise estimation of the multiple variance components is important for estimating assay precision (and, hence, the QC range), but it

is usually more difficult than estimation of the mean, given the limited sample size of clusters and observations per cluster that would inform about within- and between-cluster variability. Therefore, deciding between different prior distributions for variance components can play a crucial role for inference in the BHM (see also Browne & Draper, 2006). We introduce them in Section 3.2.2.3.

The Markov Chain Monte Carlo (MCMC) technique will be used for inference about the joint posterior distribution (Section 3.3), which is given by (see Lesaffre & Lawson, 2012, sec. 9.4.2)

$$p\left(\boldsymbol{\alpha}, \sigma^2, \mu, \tau^2 \mid \boldsymbol{y}\right) \propto \prod_{i=1}^{m} \prod_{j=1}^{k_i} \mathcal{N}\left(y_{ij} \mid \alpha_i, \sigma^2\right) \prod_{i=1}^{m} \mathcal{N}\left(\alpha_i \mid \mu, \tau^2\right) p\left(\sigma^2, \mu, \tau^2\right).$$

The PPD for the random future observation $\tilde{y}$ from a random future run with intercept $\tilde{\alpha}$ is derived from the marginal posterior distribution $p\left(\sigma^2, \mu, \tau^2 \mid \boldsymbol{y}\right)$ as

$$p(\tilde{y} \mid \boldsymbol{y}) = \int\int\int\int p(\tilde{y} \mid \tilde{\alpha}, \sigma^2)\, p(\tilde{\alpha} \mid \mu, \tau^2)\, p\left(\sigma^2, \mu, \tau^2 \mid \boldsymbol{y}\right)\, d\tilde{\alpha}\, d\sigma^2\, d\mu\, d\tau^2. \tag{10}$$

Simple Monte Carlo simulation based on the posterior draws from MCMC allows us to approximate the PPD (see Section 3.3), as well as its $(1 \pm \beta)/2$ quantiles. These PPD quantiles constitute the limits of a $\beta \cdot 100\%$ equal-tailed prediction interval.

3.2.2.3 Prior distributions for the variance components

Prior distributions for variance parameters in a Gaussian BHM have been discussed extensively in the applied statistical literature (e.g., Gelman, 2006; Lesaffre & Lawson, 2012, sec. 9.5.7; Spiegelhalter et al., 2004, Chapter 5), including in the literature on Bayesian random effects meta-analysis (Hamaguchi et al., 2021; e.g., Röver et al., 2021).[6]

We will investigate both *weakly informative priors* and a *non-informative prior* (or, *reference prior*) for comparison. Weakly informative priors can be defined as prior distributions that "attempt to let the data speak while being strong enough to exclude various 'unphysical' possibilities which, if not blocked, can take over a posterior distribution in settings with sparse data" (Gelman, 2009, p. 176). For example, in this study we will use weakly informative priors to exclude unrealistically large values for the sum of between-run and within-run variance. In the same vein, McElreath (2020, p. 407) argues that priors which "express only a rough notion of an average standard deviation and regularize towards zero" are often suitable default priors for variance components in the BHM. This can stabilize estimation in particular in settings like the current: Only few runs or repeats per run might have been measured, some of which might exhibit unusual values. Weakly informative priors can then have a regularizing effect by relying on historical records or theoretical knowledge about the possible range for assay precision. Still, we would like to mainly let the data speak, and hence put only little weight on the prior information. In brief, "weakly informative priors can strike a balance between fidelity to a strong signal, and shrinkage of a weak signal" (Simpson et al., 2017, p. 5). This philosophy of prior specification within a Bayesian analysis has been called the "regularization point-of-view" by Röver et al. (2021, p. 450), though we note that, ultimately, the purpose of prior choice for us is pragmatic: The QC range should exhibit good performance characteristics for QC. This will be the yardstick to compare different priors.

[6] In random-effects meta-analysis, it is commonly assumed that study-specific standard errors are known, which implies that the (common or study-specific) population variance is known. This is not a reasonable assumption in our study, where also within-run variance needs to be estimated from the observed data.

Gelman (2006, p. 517) distinguishes weakly informative from non-informative priors for the random effects variance in hierarchical models, which are often improper: "We characterize a prior distribution as weakly informative if it is proper but is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available." He argues that "any problem has some natural constraints that would allow a weakly-informative model." This is certainly true for the variability of an analytical assay. Accordingly, previous research has used weakly informative priors for variance components of assay precision (Lebrun & Rozet, 2020; Novick et al., 2021; Wang & Cheng, 2022).

The current study will investigate four different (proper) weakly informative prior distributions (P1 to P4) and one (improper) non-informative prior distribution (P5). The weakly informative priors will draw on the following prior knowledge elicited from scientists at GSK:

- For a typical, properly applied vaccine immunoassay, which has passed assay qualification and validation, an intermediate precision of $\%\text{CV}_{\text{IP}} > 200$ can essentially be rejected already *a priori*. The usual thresholds for assays to be accepted for a particular concentration range are lower (e.g., 50 or 70, depending on assay type). Generally, it can be expected that most $\%\text{CV}_{\text{IP}}$ fall in the range from 10 to 50.
- Assays differ considerably in how total variance splits into within- and between-run variance, so ρ should not be restricted to or pulled strongly towards a particular range on the $[0, 1]$ interval *a priori*.

Despite the focus of this study on non- and weakly informative priors, their extension to informative priors will be discussed and is straightforward within the same modeling framework and distributional families (Section 5.2). Informative priors can be powerful for example when assay-specific prior knowledge is available.

The investigated priors P1 to P5 are defined as follows.[7]

Prior 1 (P1): Restricted uniform priors on σ, τ

A uniform prior on a restricted range has been recommended for the level-2 standard deviation (Lesaffre & Lawson, 2012, sec. 9.5.7). Putting the same, independent prior distribution on both σ and τ will always maintain a symmetric prior on ρ with prior expectation of $\mathbb{E}(\rho) = 0.5$ for the intra-run correlation, broadly in line with the idea that neither small values of ρ should be favored over large values *a priori*, nor vice versa.[8] Therefore, we choose

$$\sigma,\ \tau \sim U(0, c).$$

The upper limit c is chosen such that $\%\text{CV}_{\text{IP}}$ is capped at 200. No prior probability mass is assigned to the event that $\%\text{CV}_{\text{IP}}$ exceeds 200. Thus, based on Equation 1, we need to pick c to satisfy $\sqrt{\exp\left((c^2 + c^2) \cdot \log(10)^2\right) - 1} = 2$. The solution for c under the constraint is $c = 0.390$, after rounding to three significant digits.

Prior 2 (P2): Half-normal priors on σ, τ

Another common choice for the prior distribution for the level-2 standard deviation is the half-normal distribution (McElreath, 2020, Chapter 13; Röver et al., 2021; Spiegelhalter et al., 2004, Chapter 5).

[7] Though it should go without saying, we emphasize that no information about the values of the parameters of true data-generating process, which in this study is known (Section 3.1.2), was "leaked" into the procedure for specifying priors.

[8] If σ and τ follow the same distribution, the random variables $\sigma^2/(\sigma^2 + \tau^2)$ and $\tau^2/(\sigma^2 + \tau^2)$ have the same expectation, and these expectations must some to 1.

Different from the uniform prior, and similar to the half-Cauchy prior (discussed below), it comes with advantageous technical characteristics like "roughly uniform behavior near zero (implying indifference among small heterogeneity values on the τ scale and ensuring integrability in the lower tail), and a monotonically decaying tail with increasing heterogeneity values (implying decreasing probability for increasing τ values and ensuring integrability in the upper tail)" (Röver et al., 2021, p. 452).

For the same reasons as above, σ and τ get the same, independent half-normal prior:

$$\sigma,\ \tau \sim \text{Half-}\mathcal{N}(0, \omega).$$

The scale parameter ω is chosen such that 99% of probability mass is assigned to values of $\%\text{CV}_{\text{IP}}$ below 200. The solution found via simulations is $\omega = 0.181$. Thus, our approach for choosing the hyperparameter of the prior is to declare the plausible upper bound based on substantive knowledge followed by "matching [...] the upper bound to an upper percentile such as the 99th" (Gelman et al., 2013, p. 117).

Prior 3 (P3): Half-Cauchy priors on σ, τ

Another alternative is the half-Cauchy prior, a half-t distribution with 1 degree of freedom and, again, one free scale parameter. The half-Cauchy distribution, and related half-t distributions with degrees of freedom > 1, have been recommended and used as priors for random effect variances (Bürkner, 2017; Gelman, 2006; Novick et al., 2021; Röver et al., 2021). For example, Polson & Scott (2012, p. 896) investigate its technical properties and summarize that "the half-Cauchy is a sensible default prior for scale parameters in hierarchical models".

For the same reason as above (ensuring $\mathbb{E}(\rho) = 0.5$), σ and τ get the same, independent half-Cauchy prior:

$$\sigma,\ \tau \sim \text{Half-Cauchy}(0, \psi).$$

The scale parameter ψ is chosen by matching the median of the half-Cauchy prior distribution to that of the half-normal prior distribution specified above, a procedure used for comparisons between priors also by Röver et al. (2021). This requires to multiply the scale parameter ω from the half-normal distribution by the 75th percentile of the standard normal distribution (≈ 0.674). We obtain $\psi = \omega \cdot 0.674 = 0.181 \cdot 0.674 = 0.122$.

The main ways in which the resulting prior differs from the half-normal prior specified above are (i) that slightly more prior probability is assigned to values close to zero, and (ii) the heavier upper tail. Otherwise, the shape is quite similar (Figure 1A,B). The gentler slope in the upper tail, however, might be an important advantage, since it does not by default preclude "occasionally large" values of the standard deviation parameter (Gelman, 2006).

Prior 4 (P4): Gamma prior on $\%\mathbf{CV_{IP}}/100$, uniform prior on ρ

Priors P1 to P3 included probability distributions directly assigned to σ and τ. Indirectly, they imply priors also on quantities derived from σ and τ, like the intra-run correlation ρ. For example, even though P1 to P3 guarantee the prior expectation $\mathbb{E}(\rho) = 0.5$, they put relatively more prior weight on extreme values of ρ (Figure 1D). This might or might not be appropriate for the problem at hand. Also, the prior for the assay precision parameter that is arguably most easily interpreted and most widely cited by the practicing scientist, $\%\text{CV}_{\text{IP}}$, can only rather indirectly be controlled via $p(\sigma)$ and $p(\tau)$.

Another parametrization of the variance components priors, in contrast, allows to flexibly assign prior probabilities *directly* on values of $\%\text{CV}_{\text{IP}}$ and ρ. In turn, this induces joint prior distributions that cannot be factorized (different from the independently assigned prior distributions, as for priors P1 to P3) on the variance components σ^2 and τ^2. The procedure is closely related to the hierarchical decomposition priors proposed by Fuglstad et al. (2020), which also avoid explicit and independent priors on the variance components in favor of priors on total variance and variance splits. They show some favorable properties of such priors for the BHM, among which is also the opportunity to "include expert knowledge through interpretable statements on the total variance and the distribution of variance" (p. 1112) – a point we return to in Section 5.2.2.

An example of suitable distributional families in the present case are the Gamma distribution (after dividing $\%\text{CV}_{\text{IP}}$ by 100 to return from the percentage to the ratio scale) and Beta distribution, respectively. These will be used here, but other families that respect the domain of the parameters are similarly suitable (e.g., a log-normal distribution for $\%\text{CV}_{\text{IP}}$). The two independent priors are

$$\begin{aligned} \%\text{CV}_{\text{IP}}/100 &\sim \text{Gamma}(\alpha_G, \beta_G), \\ \rho &\sim \text{Beta}(\alpha_B, \beta_B). \end{aligned}$$

The two hyperparameters per distribution give good flexibility in specifying scale and shape (with the Gamma distribution parametrized by the shape α_G and the inverse scale parameter β_G). How could they be chosen for reasonable weakly informative priors? Again, we apply constraints from substantive knowledge, which is straightforward in this case, because $\%\text{CV}_{\text{IP}}$ expresses prevision of bioassays on the original measurement scale. For $\%\text{CV}_{\text{IP}}$, we decide to fix the mode of the prior distribution at 20 – a typical and representative value. In addition, as previously for prior P2, we match the elicited upper bound of the plausible range, $\%\text{CV}_{\text{IP}} = 200$, to the 99th percentile of the distribution. The solution for hyperparameters of the Gamma distribution to satify these constraints, found via numerical optimization, is $\alpha_G = 1.58$ and $\beta_G = 2.92$. For ρ, we encode exact indifference over all values in the $[0, 1]$ interval with a uniform prior distribution ($\alpha_B = \beta_B = 1$). This maintains the prior expectation that no values of intra-run correlation should be favored over others – not even those close to 0 or 1, as was induced by priors P1 to P3 (Figure 1D).

Prior 5 (P5): Reference prior

For the level-1 variance, we pick the standard Jeffrey's prior for the variance of a non-hierarchical normal model (e.g., Lesaffre & Lawson, 2012, p. 116), i.e.,

$$p(\sigma) \propto 1/\sigma, \tag{11}$$

which implies a uniform prior on the log-standard deviation scale $p\left(\log(\sigma)\right) \propto 1$.

For the level-2 standard deviation, we pick a uniform distribution over the positive real line (e.g., Gelman, 2006, p. 521), i.e.,

$$p(\tau) \propto 1,$$

equivalent with $p(\tau^2) \propto 1/\tau$. These independent prior distributions for the two variance components are both improper, but, in combination with sufficient data (e.g., ≥ 3 runs, see Gelman, 2006), yield a proper joint posterior. Importantly, using Equation (11) as independent prior also for the level-2 variance would not lead to a proper posterior (Browne & Draper, 2006, p. 483; Gelman, 2006, sec. 4.2; Lesaffre & Lawson, 2012, sec. 9.5.7).

The reference prior is constructed to be uninformative, acting as an objective standard for comparison. It is not recommended for practice, as it fully ignores any knowledge about plausible parameter values, risks leading to improper or too broad posterior distributions (e.g., when there are few clusters), and can also lead to MCMC convergence problems when the data is little informative about the joint posterior. It must also be acknowledged that a truly and universally uninformative prior is essentially unattainable (and undesirable). Spiegelhalter et al. (2004, p. 171) note about the uniform distribution of the level-2 standard deviation that "it would be inappropriate to term this 'non-informative', as it is a fairly strong statement to declare that small values […] are as likely as large values."

Figure 1 plots marginal densities of all weakly informative, proper prior distributions in terms of σ, τ, ρ, and $\%CV_{IP}$. All black densities were specified directly for the respective prior, while the colored densities are induced by them indirectly.[9]

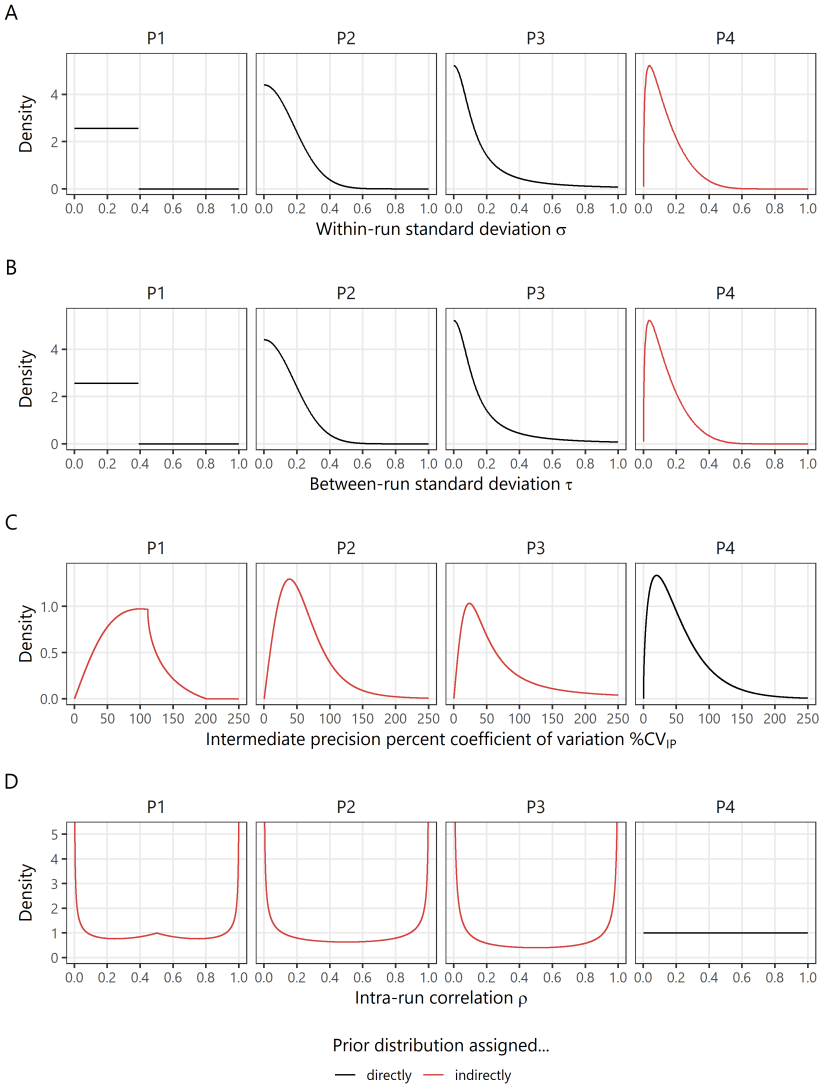

Figure 1: Marginal densities of (A, B) the variance components and (C, D) their derived quantities for the weakly informative prior distributions P1 to P4. Black lines indicate densities that are specified directly on the respective scale. Red lines indicate densities that follow from directly specified priors via transformations.

[9]For Figure 1, the indirectly specified densities were either calculated via the change of variable method (requiring also numerical integration for the convolution of random variables when adding up the independent variance component priors) or Monte Carlo simulations.

Table 2 shows the prior medians for all parameters. The prior QC range, i.e., the QC range implied by the prior before seeing the data and given by the $(1 \pm \beta)/2$ quantiles of the prior predictive distribution, could in principle be also be investigated and compared – for example, in terms of their width or plausibility. However, the prior QC range is strongly dominated by the very dispersed prior for μ. Differences in the priors for the variance components play only a small part. This is easily seen when comparing the small prior medians of the standard deviations σ and τ with the vast prior QC ranges, which integrate over the prior for μ. Hence, the comparison of prior QC ranges is not very informative for choosing between priors P1 to P4. Only the half-Cauchy prior has a visible impact on the prior QC range due to its heavier tails compared to the other distributions (Figure 1).[10]

Table 2: Prior summaries for parameters related to assay precision and prior QC ranges, based on the weakly formative prior distributions P1 to P4. The prior QC range is defined by the $(1 \pm \beta)/2$ quantiles of the prior predictive distribution with $\beta = 0.99$. The shown QC range is on the same scale as the standard deviation parameters (measurements after $\log_{10}$ transformation). Prior P3 is the half-Cauchy distribution that has undefined mean and variance.

					Prior 90% equal-tailed credible interval				
	Prior median		Prior mean	Prior variance	Lower limit	Upper limit	Prior correlation	Prior QC range	
Prior	σ, τ	$\%CV_{IP}$	σ, τ	σ, τ	σ, τ	σ, τ	$\text{Cor}(\sigma^2, \tau^2)$	Lower limit	Upper limit
P1	0.195	82	0.195	0.013	0.020	0.370	0.00	-257.6	257.6
P2	0.122	52	0.144	0.012	0.011	0.355	0.00	-257.6	257.6
P3	0.122	68	-	-	0.010	1.550	0.00	-259.7	259.7
P4	0.109	43	0.135	0.011	0.014	0.342	0.27	-257.6	257.6

3.3 Software and computation

All simulations and data analyses were conducted in `R` v4.5.0 (R Core Team, 2025). The REML estimators for the frequentist hierarchical models were obtained with the `mgcv` package (Wood, 2017). It provides the necessary robustness to small values of the variance components, because the optimization is conducted on the scale of the log precision. In addition, it is easy to extract the necessary optimization results, like the Hessian matrix including the variance parameters, to calculate Equation 9.

For posterior inference in the Bayesian models, we used the `brms` package for priors P1 to P3 (Bürkner, 2017), and, to gain flexibility in prior specification, the `cmdstanr` interface for priors P4 and P5 (Gabry et al., 2025). Both employ v2.36 of the Stan software (Carpenter et al., 2017) in the backend. Stan is a probabilistic programming language that allows fast and efficient MCMC sampling via the NUTS algorithm, an adaptive variant of Hamiltonian Monte Carlo (Stan Development Team, 2024). A sampling-based technique is necessary because Gaussian Bayesian hierarchical models require approximations for most except the simplest prior distributions and posterior properties (discussed by Box & Tiao, 1973, Chapter 5). We run two Markov chains per model with a length of $15,000$ samples per chain, obtained after $1,000$ warm-up iterations.[11]. MCMC diagnostics are reported as part of the results (Section 4.2.1) and were computed with functions from the `posterior` (Bürkner et al., 2025) and `coda` (Plummer et al., 2006) packages.

Monte Carlo simulation based on the $30,000$ posterior draws from MCMC allowed us to approximate the PPD $p(\tilde{y} \mid \boldsymbol{y})$. That is, we generated a Monte Carlo sample of the PPD by repeatedly sampling

[10] Prior QC ranges in Table 2 are computed by sampling form the prior predictive distributions via Monte Carlo simulations. The prior predictive distribution $p(\tilde{y})$ is obtained by replacing the joint posterior $p(\sigma^2, \mu, \tau^2 \mid \boldsymbol{y})$ with $p(\sigma^2, \mu, \tau^2)$ in Equation (10).

[11] Further MCMC algorithm settings were held fixed at `adapt_delta = 0.99`, and `max_treedepth = 10`.

from $p\left(\sigma^2, \mu, \tau^2 \mid \boldsymbol{y}\right)$ via MCMC, and subsequently from the conditional distribution $p(\tilde{y} \mid \sigma^2, \mu, \tau^2)$ via Monte Carlo (one sample per sample from the joint posterior distribution), creating a mixture distribution over the range of likely parameter values. The $(1 \pm \beta)/2$ quantiles of this distribution provide the prediction interval used as QC range for all methods based on the BHM.

We simulated 750 data sets per scenario (Table 1), and analyzed each with 7 different methods (Section 3.2). That is, in total, we calculated $750 \cdot 30 \cdot 7 = 157,500$ QC ranges, $750 \cdot 30 \cdot 5 = 112,500$ of which were based on the BHM. On a Dell XPS 15 9510 machine (64 GB RAM, 11th Gen Intel Core i7, 2304 Mhz, 8 pyhsical cores) using 15 logical cores, the runtime of the simulations approximated 48 hours.

From the $112,500$ BHM fits, we removed 2 fits that caused an error during program execution, and further 20 fits that we deemed unreliable (they were, however, included in the diagnostics in Section 4.2.1). This was decided if either only one of the two chains finished successfully, or an excessive number of divergent sampler transitions ($> 10,000$) was observed. In general, convergence was good, but, unsurprisingly, depended on the prior distributions (Section 4.2.1). Most of the 20 fits we had to discard used one of the uniform priors for τ^2 and belonged to one of the sparse data scenarios. The final data set analyzed hence consisted of $157,500 - 22 = 157,478$ QC ranges.

3.4 Analysis

The objective of the simulation study was to obtain estimates of the expected prediction coverage and expected width of the QC ranges for the 7 methods and nominal level $\beta = 0.99$. In Section 3.2, we described in detail how the different prediction intervals are obtained given the data $\boldsymbol{y}$ and the assumed models. Here, we describe the estimation of their expected prediction coverage and expected interval width based on the simulation study with simulated data sets $r = 1, 2, \ldots, R$ per scenario.

Denote by $L(\boldsymbol{y}_r)$ and $U(\boldsymbol{y}_r)$ the lower and upper prediction interval limits, respectively, of simulation repetition r, obtained with one of the methods after having observed data $\boldsymbol{y}_r$. Since the data $\boldsymbol{y}_r$ are random, also the intervals are random. We observe a single interval per data set and method – leading to a single observed value of both prediction coverage probability C_r and interval width W_r for each repetition r – both of which are also random.[12] They are calculated as

$$C_r = \Phi\left(\frac{U(\boldsymbol{y}_r) - \mu}{\sqrt{\text{Var}_{\text{IP}}}}\right) - \Phi\left(\frac{L(\boldsymbol{y}_r) - \mu}{\sqrt{\text{Var}_{\text{IP}}}}\right),$$

where μ and Var_{IP} are the true parameter values used for data generation in the respective simulation scenario, and

$$W_r = U(\boldsymbol{y}_r) - L(\boldsymbol{y}_r),$$

respectively. The Monte Carlo estimate of expected prediction coverage is then calculated as $R^{-1}\sum_{r=1}^{R} C_r$, and the Monte Carlo estimate of expected interval width is calculated as $R^{-1}\sum_{r=1}^{R} W_r$.

Accordingly, we compute for each simulated data set the probability that a normally distributed random variable, following the population marginal distribution of measurements, falls within the estimated lower and upper limits of the prediction interval. The estimand of interest is an expectation, and we estimate it with averaging over the observed prediction coverages from all simulation repetitions,

[12] The definition of nominal prediction interval coverage by Patel (1986, p. 2723) as an expectation over repeated samples makes this randomness explicit.

but we also report the distributions of observed prediction coverages themselves (Appendix A.5).[13] The evaluation metrics can be reported with Monte Carlo standard errors, which, due to independently drawn Monte Carlo samples, scale with $\sqrt{R}^{-1}$. It is noteworthy that, due to the monotonocity of the transformation, the obtained estimate of expected prediction coverage holds also after back-transforming the interval limits to the original measurement scale – a usual practice when applying assay control charts (stylized example in Figure 2).

Besides expected prediction coverage probability, it is also relevant how strongly prediction intervals deviate on average from the nominal level. For example, one of our methods might produce slight overcoverage on average compared to the nominal level, but observed prediction coverages might deviate less from the nominal level across repeated samples than for other methods. Therefore, we introduce the *root mean squared coverage error* (RMSCE), similar to the notion of "mean squared conditional error" from Kiyani et al. (2024). It is estimated as

$$\sqrt{\frac{1}{R}\sum_{r=1}^{R}(C_r - \beta)^2}.$$

We believe this criterion is especially important when comparing frequentist and Bayesian estimators in hierarchical models. Regarding point estimators of variance components, Bayesian methods might reduce overall error for the cost of some bias (e.g., Chaloner, 1987).

In addition to the evaluation of performance in terms prediction coverage probability and interval width, we analyze MCMC diagnostics in the aggregate for all Bayesian models fit as part of the simulation study. This is important, since the proposed Bayesian methods are only a viable alternative in the current setting if posterior distributions can be obtained reliably and efficiently. Where diagnostics are calculated by parameter (e.g., $\hat{R}$), we focus results on the two assay precision parameters σ and τ. For some diagnostics (e.g., Geweke statistic), we show results for the first chain only for brevity.

Finally, we analyze the point estimates for various model parameters (or their derived quantities) produced by the different methods with simple descriptive statistics. This supplements the interpretation of results from the performance evaluation of QC ranges.

[13] As a side note, the approach implies that we evaluate also the *Bayesian* prediction intervals in terms of a fundamentally *frequentist* property (see Tian et al., 2022): If we sampled repeatedly from the same data-generating process and constructed an equal-tailed $\beta \cdot 100\%$ posterior predictive interval, which proportion of the population of future values is on average covered by it? It is not uncommon to evaluate Bayesian methods in terms of frequentist operating characteristics, for example to guide prior choice under pragmatic performance considerations (Röver et al., 2021, p. 450). See also other simulation studies on Bayesian estimators like Browne & Draper (2006) or Hamaguchi et al. (2021).

4 Results

4.1 Illustration of results for one data set

Before turning to the main results of the study, which concern the performance evaluation by aggregating over simulated data sets, it is instructive to analyze a single randomly simulated data set in more detail. This will give some more insight into the methods used. The simulated baseline data, using the "Standard" design and parameters $\%\mathrm{CV}_{\mathrm{IP}} = 40$ and $\rho = 0.8$ (i.e., scenario 6 in Table 1), is visualized on the left in Figure 2: We have 20 runs, each consisting of two observed concentration values. Observations are transformed back to the original measurement units by taking $10^{y_{ij}}$, but the plot axis is itself $\log_{10}$ scaled. This reflects how QC ranges and control charts are routinely displayed in practice. The high intra-run correlation is clearly visible in terms of relatively narrow clusters of observations from the same run.

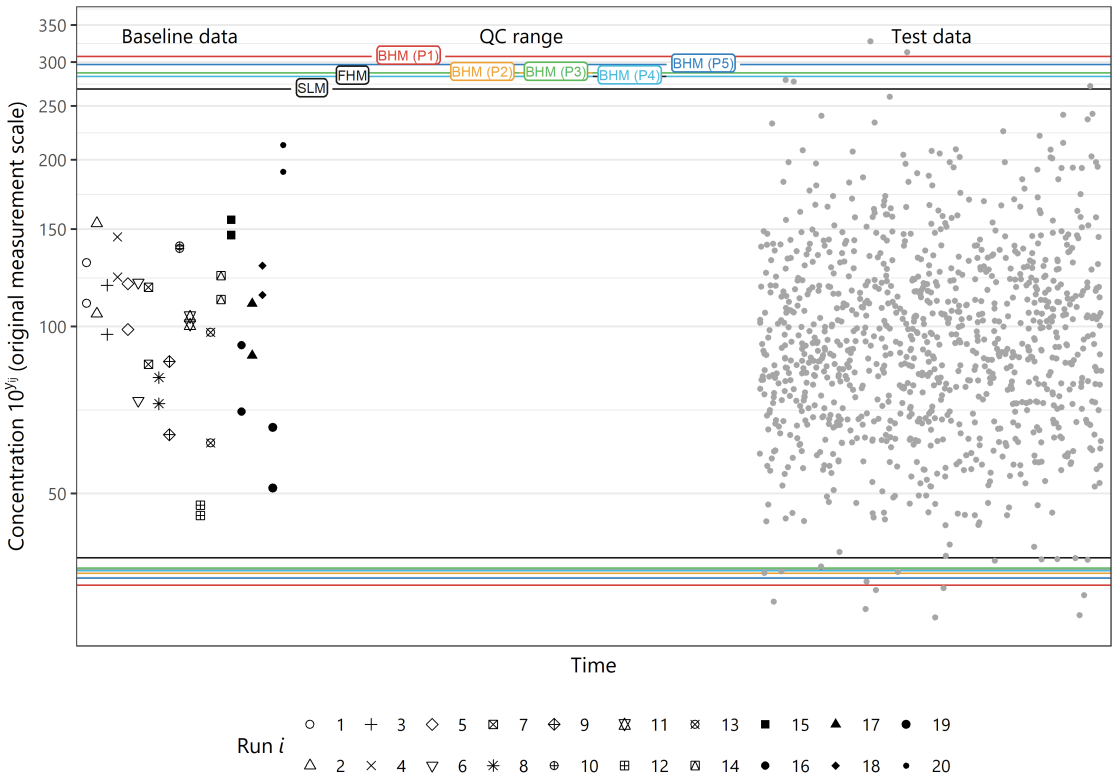

Figure 2: Control chart example with simulated baseline data for the QC sample, nested in runs (black points with different shapes on the left), and the resulting QC ranges (horizontal lines), based on the 7 methods compared in this study. Simulated future observations drawn from random future runs serving as hypothetical test data of the QC sample are also displayed (grey points). Observations are on the original measurement unit scale. The vertical axis is $\log_{10}$ scaled. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

Now, we apply all 7 methods of QC range estimation by using the baseline data and our various statistical models from Section 3.2. The resulting lower and upper limits are shown as colored horizontal lines in Figure 2. For example, the BHM with prior distribution P1 gives the widest QC range for the present data set (red lines labelled "BHM (P1)"), and the single-level model gives the smallest (black lines labelled "SLM"). We randomly sample $1,000$ independent future QC sample measurements that will be tested against the estimated QC range ("Test data" shown in grey in Figure 2). This illustrates

the use case of the QC range. As expected, most of the future measurements are covered by the QC ranges. The observed QC ranges, their prediction coverage and interval width for the example data set are shown in Table 3. Limits and widths are on the $\log_{10}$ scale, i.e., the scale on which model and interval estimation take place. In this example, the narrowest interval that respects the nominal coverage requirement is based on the frequentist hierarchical model. Even for the same example data set, coverage and width show no monotone relationship across methods.

Table 3: Observed QC ranges with their prediction coverage probabilities and interval widths for the simulated example of baseline data, based on the 7 QC range methods compared in this study. Interval limits and widths refer to the $\log_{10}$ scale.

	QC range			
Method	Lower limit	Upper limit	Prediction coverage	Interval width
Single-level model	1.5836	2.4286	0.9862	0.8450
Frequentist hierarchical model	1.5608	2.4514	0.9906	0.8906
Bayesian hierarchical model (P1)	1.5337	2.4874	0.9943	0.9537
Bayesian hierarchical model (P2)	1.5555	2.4579	0.9914	0.9024
Bayesian hierarchical model (P3)	1.5646	2.4580	0.9903	0.8934
Bayesian hierarchical model (P4)	1.5601	2.4517	0.9907	0.8916
Bayesian hierarchical model (P5)	1.5465	2.4730	0.9928	0.9265

Ultimately, the prediction intervals from hierarchical models, both frequentist and Bayesian, make use of estimates of mean μ, and within-run and between-run variances σ^2 and τ^2, respectively, or their posterior distributions. Figure 3 presents inference about these parameters based on the different hierarchical modeling approaches.

Figure 3A compares REML estimates and 90% confidence intervals with Bayesian marginal posterior distributions under the five different prior distributions. True parameter values are printed as red vertical lines in the upper panels, together with the frequentist confidence intervals. Over the posterior densities, 90% equal-tailed credible intervals (marked by the 5% and 95% quantiles) and posterior mode (obtained from a kernel density estimate), median, and mean are plotted. There is generally agreement between inferences from the frequentist and from the Bayesian hierarchical models. For the example data set, all intervals cover the true parameter values. All models correctly portioned most of the total variance into the between-run component. Marginal posterior densities for the standard deviations are right-skewed, consistently leading to the mode, median, and mean to be the smallest, medium, and largest posterior summary measure of central tendency, respectively.[14] The largest effect of the different priors can be observed for Bayesian inference about the between-run standard deviation τ. Also the credible intervals are widest for τ. For the mean parameter μ, neither is there much variation between different prior specifications, nor across different posterior summary measures. In any case, we emphasize that the PPD integrates over the full posterior distribution (Equation 10). Therefore, for the current study, characteristics of different Bayesian point estimators of marginal densities are not of primary importance.

[14] For prior P1, which assigns restricted uniform distributions to both standard deviation parameters, we see quite close agreement between REML estimates and the posterior modes of σ and τ. This does not come as a surprise, given that the posterior density should in these cases more closely follow the shape of the (restricted) likelihood as for other priors. When averaging over all simulation repetitions, this pattern largely holds: Under prior P1, in all 30 scenarios, the posterior mode is the posterior summary measure of central tendency closest to the REML estimate of σ. In 16 out of 30 scenarios, the posterior mode is the posterior summary measure of central tendency closest to the REML estimate of τ (in the other 14 scenarios, it is the posterior median). Pick et al. (2023, p. 2558) write that, for variances, "use of the posterior mode is often justified as being the closest to the maximum likelihood estimate (MLE) when uninformative priors are used", but also caution against simplifications for mixed-effect models, where multiple priors and multi-dimensional joint posteriors are involved. Lesaffre & Lawson (2012, sec. 9.8.2) discuss the equivalency in more detail.

Figure 3B shows that both the frequentist and Bayesian hierarchical models are able to capture the dependency of inferential uncertainty for the two variance components with similar results. The left panels show draws from the joint posterior $p\left(\sigma^2, \tau^2 \mid \boldsymbol{y}\right)$ for the BHM based on the five different prior distributions and the example data set. The overlaid black line highlights the quantile-based 90% probability contours. In the right panel of Figure 3B, the bivariate normal ellipse summarizes the estimated variance-covariance matrix of the variance components' estimators from the frequentist hierarchical model. All methods give a similar, negative correlation between the variance components in terms of posterior densities (for the BHM), or REML point estimators (for the frequentist hierarchical model), with correlation values shown in the upper right corner.

After this illustrative analysis of a single simulated data set, we now turn to the main results of this study.

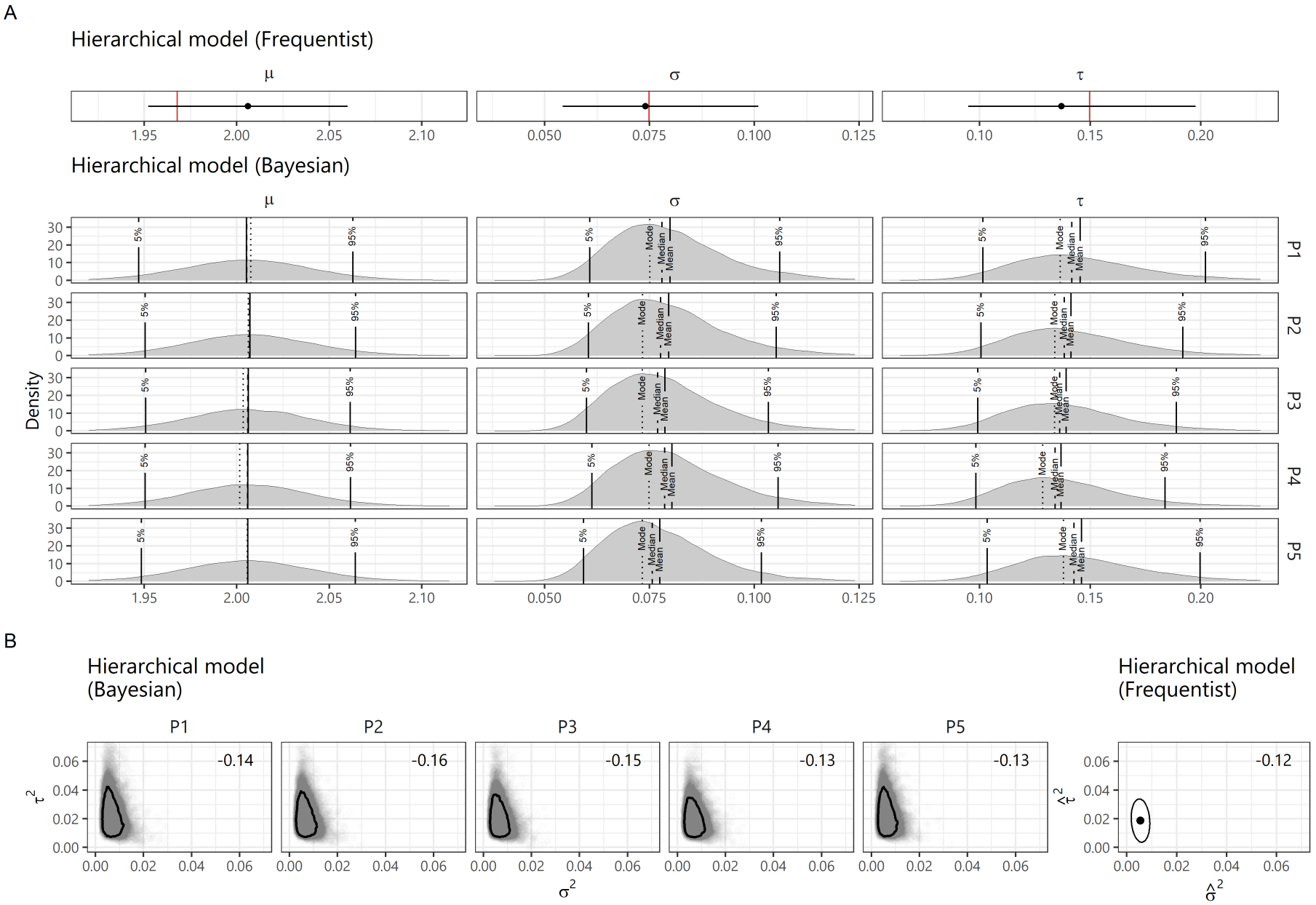

Figure 3: Parameter estimation in the simulated example of baseline data for all methods based on hierarchical models. (A) Point estimates and 90% confidence intervals for the frequentiest hierarchical model, along with the true values used for data generation (red vertical lines), as well as marginal posterior densities with posterior summaries for the Bayesian hierarchical model, with its five different prior distributions. (B) Scatter and contour plot of joint posterior densities of the variance components (left panels), and bivariate normal ellipse of the covariance of REML esimators of the variance components (right panel). Correlations are shown shown in the upper right corner of each panel.

4.2 Performance evaluation

4.2.1 Diagnostics for the BHM

MCMC diagnostics of the Bayesian hierarchical models with different prior distributions aggregated over scenarios show no serious issues (Figure 4). For by far the most simulated data sets and priors, the number of divergent transitions of the NUTS sampler was zero after warm-up (Figure 4A). Priors P4 and P5 led to most divergences, but with problematic frequencies essentially only occurring for the reference prior P5 (more than $1,000/30,000 \approx 3.3\%$), and only rarely so (also note that diagnostics were calculated before removing the most problematic model fits; see Section 3.3). For each chain, an estimated Bayesian fraction of missing information (E-BFMI) can be calculated, which is, broadly speaking, a measure for the efficacy of the exploration of the posterior under Hamiltonian Monte Carlo (Betancourt, 2016). Values < 0.2 are usually taken to indicate problems. They occurred rarely and, if they did, were concentrated among priors P4 and P5. By running two chains per model, it is possible to calculate the $\hat{R}$ statistic (revised version by Vehtari et al., 2021), which should be close to 1 if chains reached stationarity. The majority of $\hat{R}$ values for the parameters governing assay precision (σ and τ) was indeed practically 1, with the highest proportion of larger values for prior P5 (Figure 4C). The autocorrelation function (ACF) estimated at lag 3 from the first chain showed some variation in mixing according to prior and parameter (Figure 4D). For σ, autocorrelation was generally smaller than for τ, so sampling efficiency tended to be lower for τ. Relatedly, τ had lower tail effective sample size (ESS Tail) estimates than σ (Figure 4E). ESS Tail refers specifically to the sampling efficiency in the distributions' tails, which is most relevant for getting interval estimates from the posterior (Vehtari et al., 2021), as required for QC range construction. Over most of the aforementioned diagnostics, priors P1 to P3 exhibited similar results, often less spread out than results for priors P4 and P5. Arguably, priors P1 and P3 are quite similar to each other (proper, weakly informative, with *a priori* independent variance components; see Section 3.2.2.3). Further explorations could shed more light on why and how this translated into the observed "clustering" of diagnostic results.[15]

In Figure 4F, we provide the frequencies of Geweke statistics that fell into the two-sided rejection region for a 0.05 significance level (see Lesaffre & Lawson, 2012, sec. 7.2.4). They tested equality of the means of the first 10% and last 50% of samples from the first chain. Frequencies of rejection were close to where they should be under the null hypothesis of convergence ($\approx$ 5-6$\%$). There was slight concentration of more rejections when sampling the posterior under reference prior P5 ($\approx$ 7-8$\%$).

Finally, the Monte Carlo standard error (MCSE) for the $(1 \pm 0.99)/2$ quantiles of the PPD, i.e., of the limits of the QC range, was computed for each model fit following Vehtari et al. (2021), taking the dependency of samples into account. MCSE was < 0.01 most of time, and tended to be largest under reference prior P5 (Figure 4G). Since it needs to be interpreted on the scale of the predictive distribution ($\log_{10}$-transformed concentration measurements), it should be evaluated relative to the width of the estimated prediction interval. Roughly, average widths were between 0.2 and 2, depending on scenario (Figure 7), i.e., much larger than MCSE estimates. This indicates acceptable reliability of the MCMC-based approximation of the exact PPD's quantiles used as QC range limits.

[15] Initial informal checks suggest that priors P4 and P5 led to more variable MCMC efficiency because they support more efficient sampling of τ in some scenarios but less efficient sampling in others. Specifically, they work less well than the other priors in scenarios with *low* intra-run correlation ($\rho = 0.2$) and *many* available runs ($m \geq 10$); that is, when there is a moderate (but not excessive) empirical basis for estimating τ, while at the same time, τ is much smaller than σ. The pattern could hence emerge due to a mild form of prior-data conflict: Compared to priors P1 to P3, the priors P4 and P5 put much less probability on small values of ρ, yet the data speak moderately in favor of a small value of ρ in the aforementioned scenarios. In all other scenarios, no such conflict arises: Either the true ρ is larger, or there is less information on between-run variability in the data to contradict the prior.

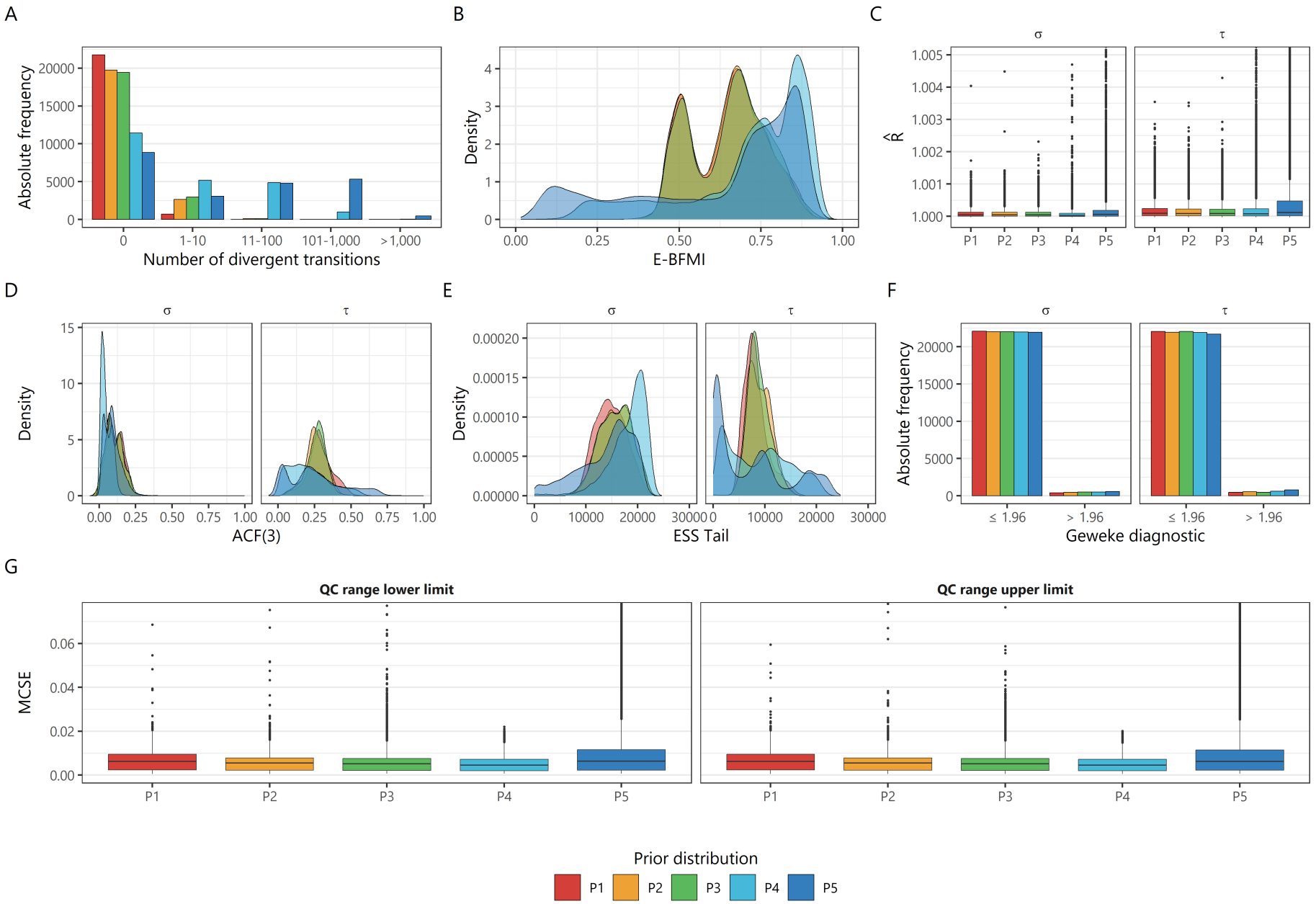

Figure 4: Distribution of MCMC diagnostic results for the hierarchical Bayesian models with five different prior distributions. (A) Numer of divergent transitions of MCMC algorithm. (B) Estimated Bayesian fraction of missing information (E-BFMI). (C) $\hat{R}$ convergence diagnostic for the two standard deviation parameters. (D) Autocorrelation function (ACF) at lag 3 for the two standard deviation parameters (first chain). (E) Estimated teil effective sample size (ESS Tail) for the two standard deviation parameters. (F) Geweke diagnostic test statistic for the two standard deviation parameters (first chain). (G) Monte Carlo standard error (MCSE) for the lower and upper limit of the QC range (i.e., the $(1 \pm 0.99)/2$ quantiles of the posterior predictive distribution).

4.2.2 Interval coverage and width

Figure 5 presents the Monte Carlo estimates of expected prediction coverage of QC ranges for all scenarios and methods. Appendix A.2 presents them along with standard errors. Figure 6 presents the Monte Carlo estimate of prediction coverage error in terms of RMSCE, with results in more detail in Appendix A.4. The distributions of observed prediction coverages over repeated samples are visualized in Appendix A.5. Turning to estimates of expected interval width, Figure 7 summarizes the results. Appendix A.3 presents them along with standard errors. A number of main results on QC range performance in terms of prediction coverage and width can be distilled from the outputs.

The single-level model (SLM) performs worst and provides severe undercoverage for multiple scenarios (e.g., $< 96\%$ for highly correlated measurements at $\rho = 0.8$ when runs are sparse). It yields satisfactory expected coverage only under low intra-run correlation ρ, and, even then, is compromised if baseline data comes from few or unbalanced runs. The deficiency of the SLM can also be read from the expected width estimates: When ρ increases, and hence the amount of information in the data decreases, they tend to become smaller (Figure 7). The pattern is reversed for the hierarchical models, mostly so when there are only few runs (designs "Unbalanced 1", "Sparse 1", "Sparse 2").

All hierarchical models deliver QC ranges with expected prediction coverage that respect the nominal level better than the SLM (e.g., no estimates of expected coverage $< 98\%$). The frequentist hierarchical model (FHM), however, shows greater dataset-to-dataset variability in prediction coverage than the Bayesian methods. Extremely low prediction coverages in some data sets (e.g., $< 75\%$) only occurred for the FHM, not the BHM (Appendix A.5). Accordingly, RMSCE is also higher for the FHM than the BHM across scenarios, irrespective of the prior distribution.

Almost all methods have lower expected coverage when within-run correlation is stronger, but the decrease is substantially smaller for the QC ranges derived from hierarchical models than those derived from the SLM. For the hierarchical models, the decrease in coverage with increasing correlation is minor as long as the number of runs does not become too small (e.g., $m = 20$ in designs "Standard" and "Unbalanced 2" appears to suffice). Average deviation from nominal prediction coverage in terms of RMSCE is also higher at higher correlations (Figure 6). Caution is needed when interpreting and comparing the four estimates of expected coverage for the FHM in sparse data settings at $\rho = 0.8$, as they come with relatively high standard errors (Appendix A.2) due to their high variability (Appendix A.5). For example, at $\rho = 0.8$ and $\%\text{CV}_{\text{IP}} = 40$ in the "Sparse 1" design, the range of coverage estimate ± 2 SE for the FHM is roughly from 98.2% to 98.7%.

In almost all scenarios, QC ranges based on the BHM exhibit average overcoverage to some extent, and most strongly so for the priors that are flat for at least one standard deviation parameter (restricted uniform prior P1 and reference prior P5). But the prior distribution makes a difference mainly when the total variability in terms of $\%\text{CV}_{\text{IP}}$ is high. Priors P2 and P3, based on independent Half-normal and half-Cauchy distributions for the standard deviations, respectively, often lead to highly similar results. Most markedly, at high values of correlation $\rho = 0.8$, the BHM with prior P4 behaves differently from the others, providing lower expected coverage, in some scenarios even slightly $< 99\%$. This could result from its stronger shrinkage of ρ towards 0.5: Compared to the other priors, we are effectively "imputing" more information about between-run variability, on which we have few data, from the relatively abundant information on within-run variability.[16] Since the latter is low at $\rho = 0.8$,

[16] Whereas all other priors assume *a priori* independence between variance components, Prior P4 induces positive correlation between variance components (Table 2).

estimates of total variability (Figure 9) and prediction coverage are smaller than for all other priors; and, in fact, mostly closer to the true values in the explored scenarios.

There is no large effect on the methods' performances when varying between assays with higher ($\%\text{CV}_{\text{IP}} = 10$) or lower precision ($\%\text{CV}_{\text{IP}} = 40$). Under very few runs ($m = 5$ in designs "Sparse 1" and "Sparse 2"), a lower assay precision reduces some of the expected overcoverage of the BHM for priors P2 to P4 – bringing their prediction coverage closer to the nominal level. There is less conflict between the weakly informative priors and the data at lower precision. Indeed, for P2 to P4, prior modes and prior medians of intermediate precision are relatively close to the true intermediate precision in the low-precision scenario (Figure 1, Table 2).

Unbalancedness of the baseline data for the QC sample does not seem to be of concern *per se* if hierarchical models are used, at least for the explored parameter settings. As long as the number and size of runs remain the same (designs "Standard" and "Unbalanced 2"), expected prediction coverage and RMSCE of QC ranges essentially do not change. However, if unbalancedness is accompanied by a reduced number of runs – even if these runs are larger (design "Unbalanced 1") – performance deteriorates to some degree when these few, unbalanced runs are also highly heterogeneous ($\rho = 0.8$).

The built-in ceiling of prediction coverage at 100%, which is close to the 99% level used in this study, affects our results in ways that need to be examined more closely. For example, in designs "Sparse 1" and "Sparse 2", when information on between-run heterogeneity is scarce, QC ranges based on the BHM often hit that ceiling, especially when true variability is low (Appendix A.5). Although intervals become too wide, their deviation from the nominal level is capped at 100%. This leads to relatively low RMSCE for the Bayesian QC ranges (Figure 6), while at the same time these QC ranges tend to be relatively wide compared to the FHM (Figure 7), in line with their expected overcoverage for many scenarios (Figure 5). For the same reason, the BHM with prior P5 generally leads to the widest intervals and highest overcoverage of future measurements while simultaneously never having highest RMSCE in any of the data-generating scenarios (Figure 6).

A related phenomenon creates the following seeming paradox among QC ranges from hierarchical models: When moving from lower to higher values of ρ in the "challenging" designs with fewer runs ("Unbalanced 1", "Sparse 1", "Sparse 2"), estimates of expected interval width *increase*, some even substantially so, while, at the same time, estimates of expected prediction coverage *decrease*. This is despite the fact that the estimates of μ (roughly the center point of the intervals, and exactly so for the frequentist methods) do not change meaningfully across these scenarios (Table A5). Figure 8 helps to understand this. As a showcase, it depicts QC ranges from the first 75 simulated data sets for each value of ρ for the "Sparse 1" design under low total variability (scenarios 13, 14, 15, respectively, in Table 1). Single QC ranges that do not reach the nominal level are colored in red.[17] We see that for all methods using hierarchical models, *variation* in interval width between random data sets is increasing with ρ: Some QC ranges become very small (with very low coverage), but most become wide (with high coverage). Since coverage cannot exceed 100%, the negative effect that few very small intervals have on expected coverage outweighs the positive effect that many wide intervals have. Ultimately, such effects, and more generally the balancing of different types (under- vs. overcoverage), frequency, and sizes of coverage errors need to be discussed with a perspective on the costs that false rejections or acceptances of measurements have for vaccine assay quality control.

[17] Note that if one wished to guarantee that a certain fraction of QC ranges covered at least 99% of future measurements (intervals in black in Figure 8), rather than that QC ranges covered 99% of future measurements on average, as we do here, this would lead to the "β-content" tolerance intervals mentioned in Section 2.2.

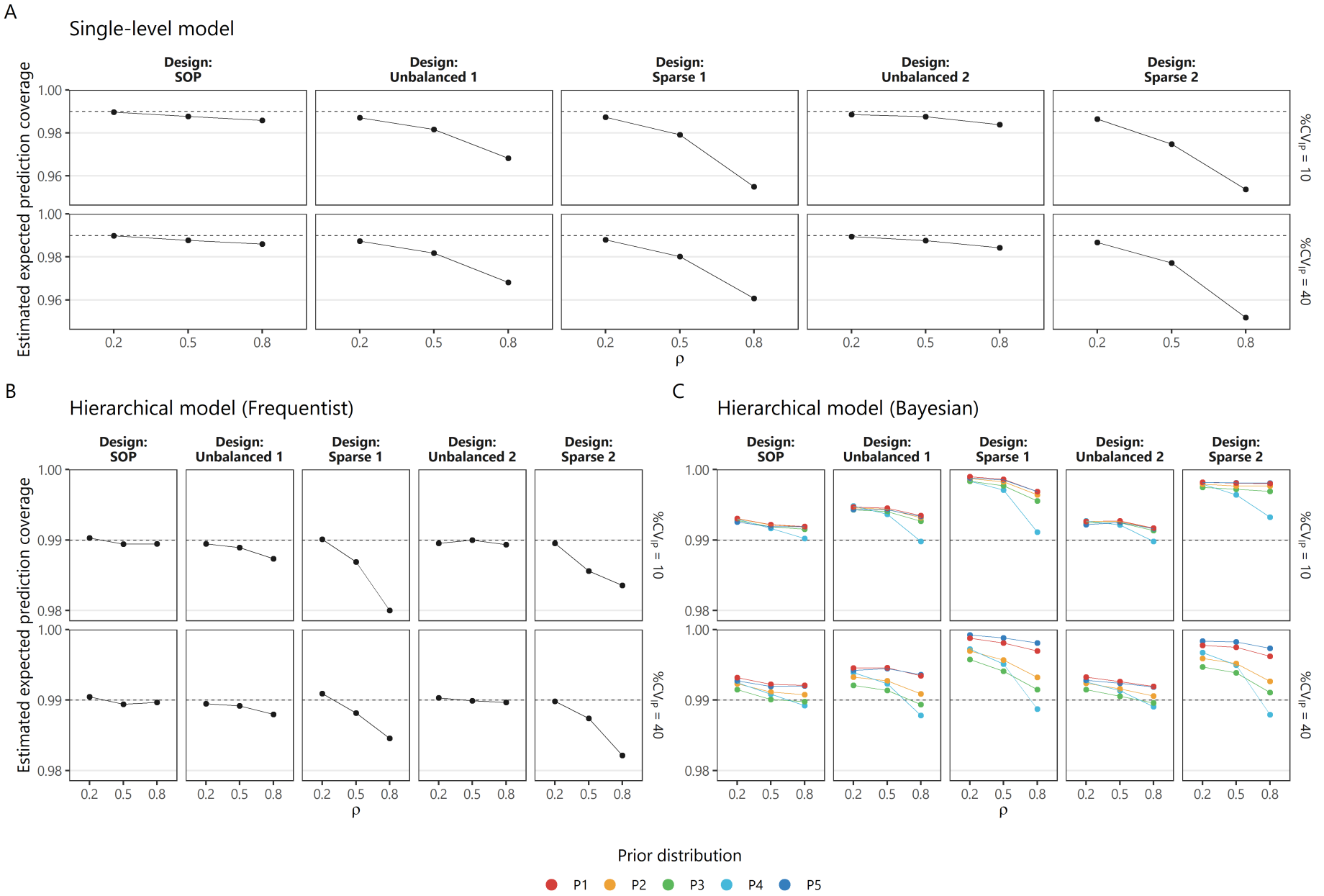

Figure 5: Monte Carlo estimates of expected prediction coverage of the QC ranges for all scenarios based on (A) the single-level model, (B) the frequentist hierarchical model, and (C) the Bayesian hierarchical model, with its five different prior distributions.

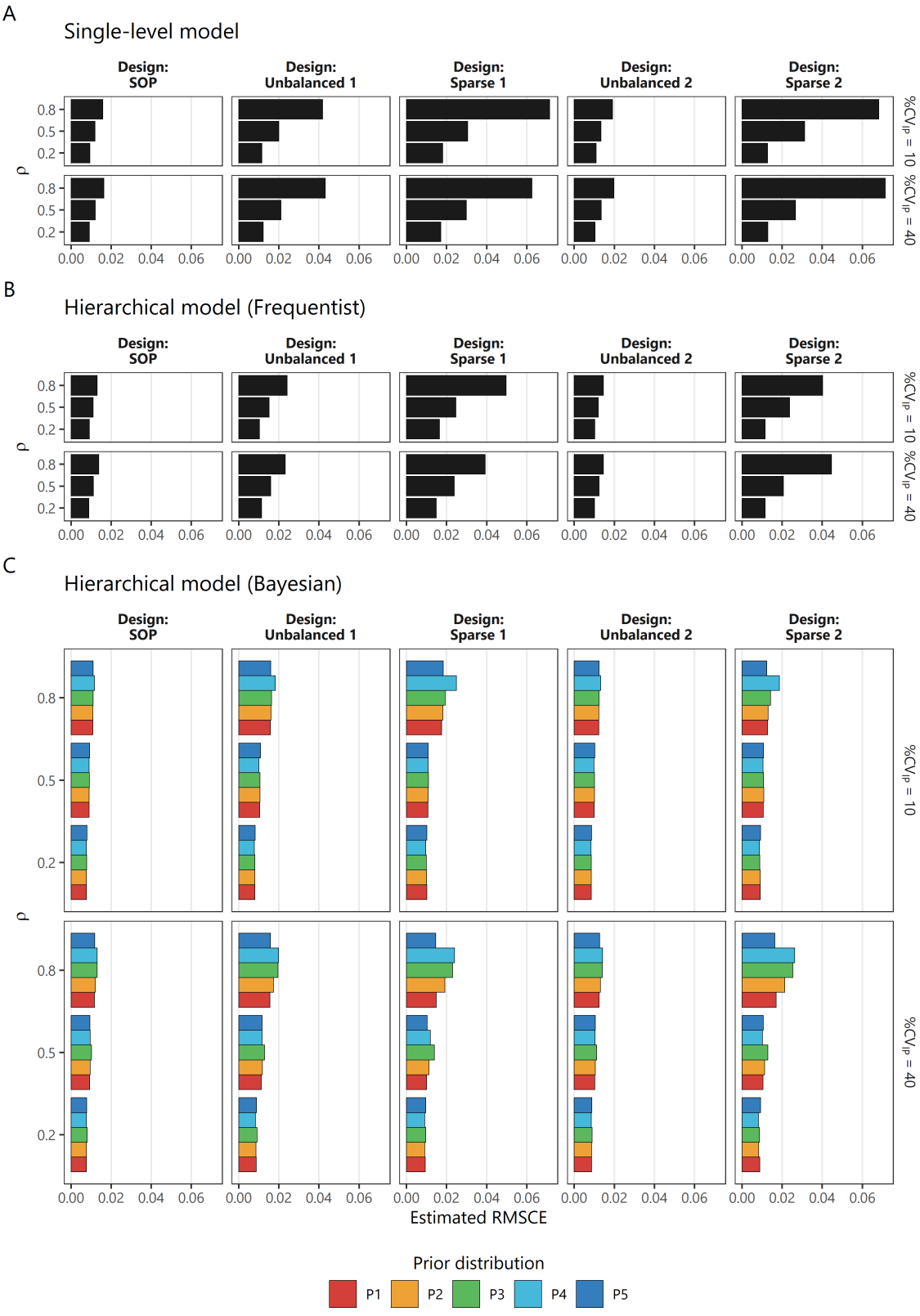

Figure 6: Monte Carlo estimate of root mean squared coverage error (RMSCE) of the QC ranges for all scenarios based on (A) the single-level model, (B) the frequentist hierarchical model, and (C) the Bayesian hierarchical model, with its five different prior distributions.

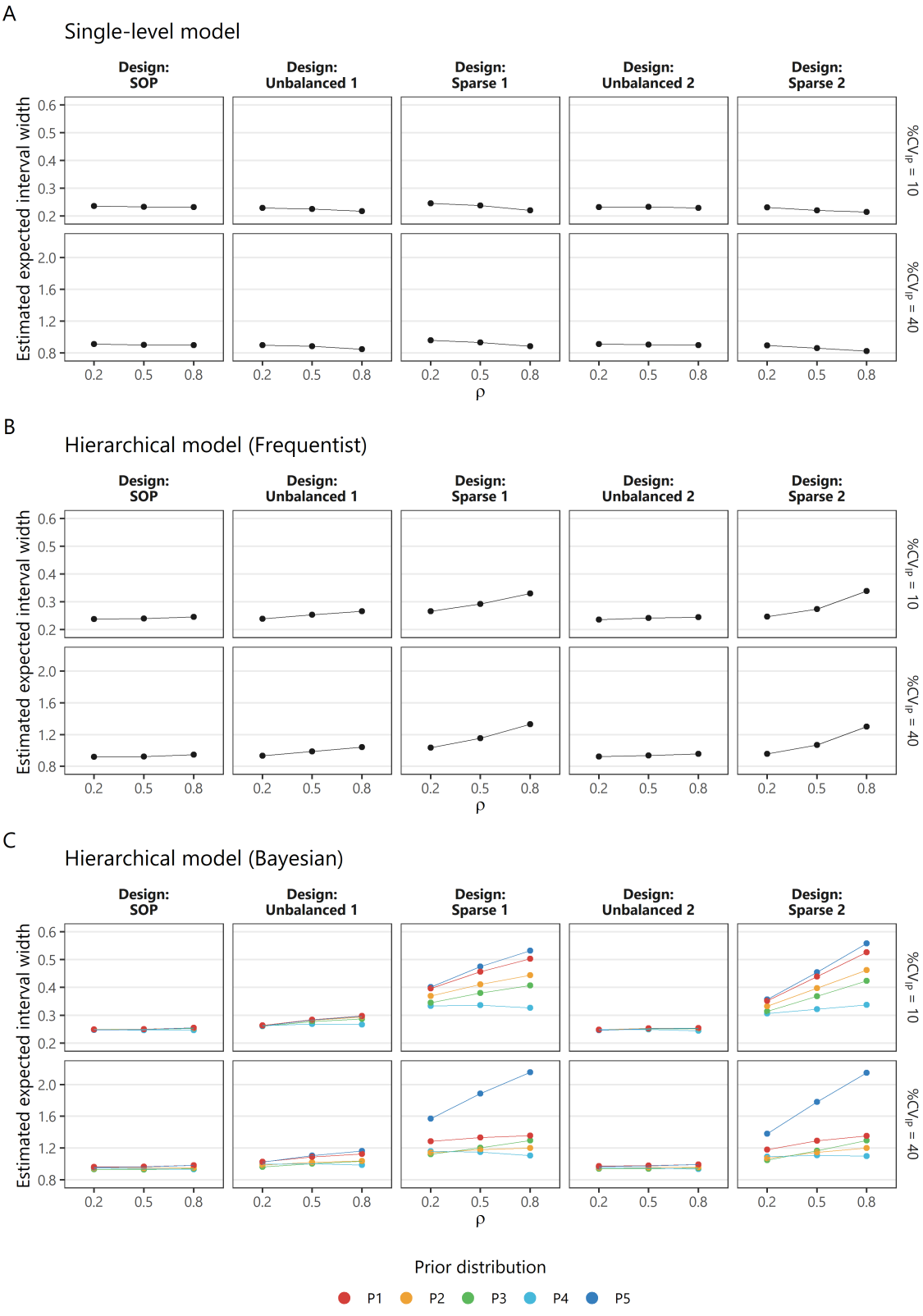

Figure 7: Monte Carlo estimate of expected interval width of the QC ranges for all scenarios based on (A) the single-level model, (B) the frequentist hierarchical model, and (C) the Bayesian hierarchical model, with its five different prior distributions.

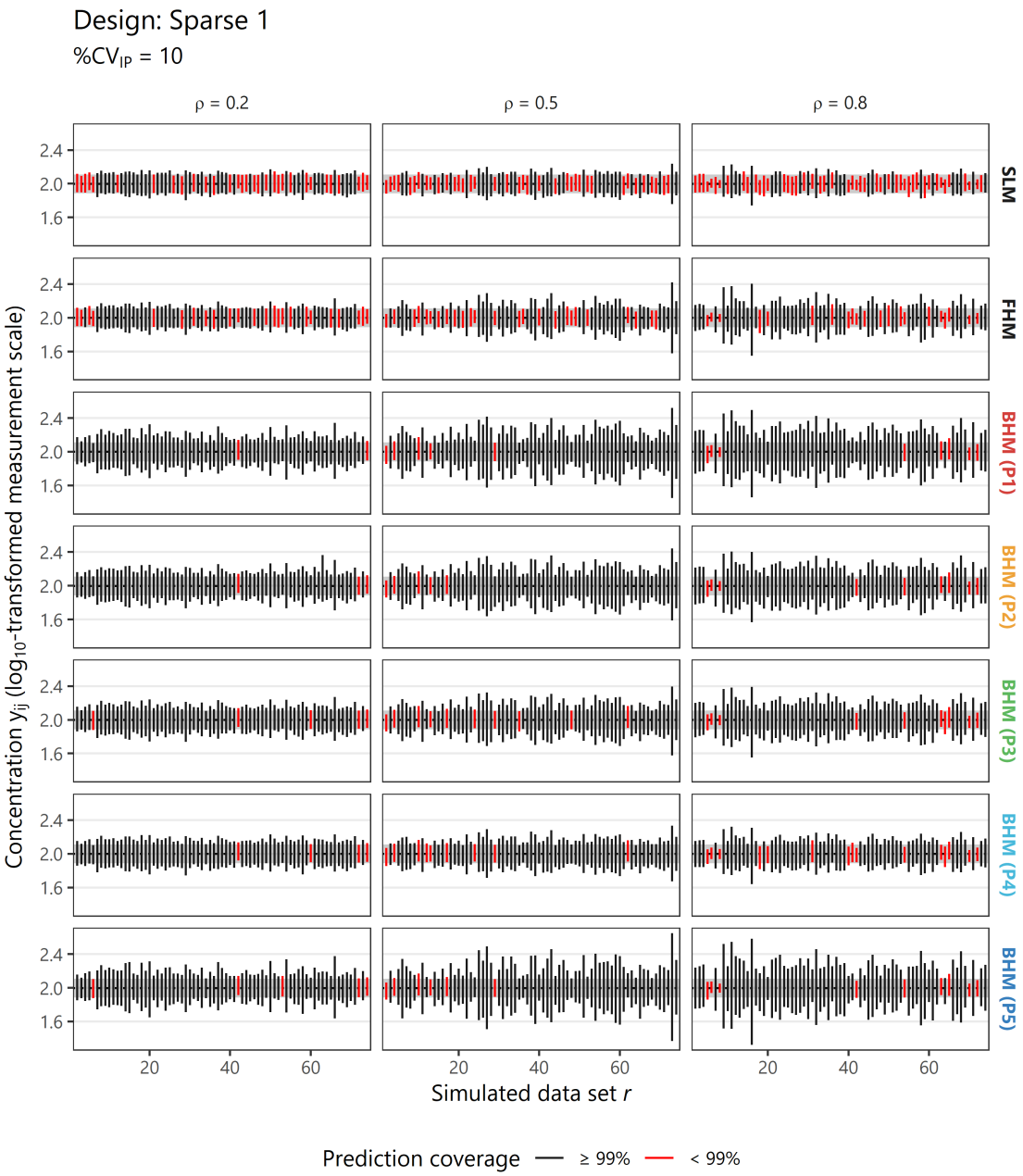

Figure 8: QC ranges for 75 simulated data sets for each of three scenarios, all based on the "Sparse 1" design. The $(1\pm0.99)/2$ quantiles of the true population of future observations drawn from random future runs are shown as grey band in the background. Red intervals indicate QC ranges which fail to reach the desired prediction coverage of 99%, i.e., which cover less than 99% of the vertical range of the grey band. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

4.2.3 Parameter estimates

The QC range performance in terms of prediction coverage and interval width is the most important outcome in this study. Yet, we also report point estimates for model parameters (or their derived quantities) produced by the different methods. First, some of them are relevant in their own right for the practicing scientists, because they encode assay quality characteristics (e.g., the intermediate precision). Therefore, obtaining "good" estimates is relevant. Second, they can help to interpret performance of the QC range, which is, of course, dependent on inferences about model parameters.

Appendix A.6 reports the median point estimate (median over all simulated data sets) for all parameters, methods, and scenarios. The median point estimate is compared to the true parameter value used during data generation. Here, truth refers to the data-generating process described in Section 3.1.2. The estimators are those described in Section 3.2. For example, for the true mean μ (Section 3.1.2), the single-level model estimator is $\hat{\mu}_s$ (Section 3.2.1). Some cells are empty for the single-level model. For example, for the true within-run standard deviation $\sigma = \sqrt{\text{Var}_{\text{IP}} \cdot (1 - \rho)}$ (as applied during data generation in Section 3.1.2), there is no single-level model estimator, because there is only one standard deviation estimate $\hat{\sigma}_s$, which combines both within- and between-run variability (see Section 3.2.1). For the frequentist hierarchical model, the point estimates are the REML estimates (Section 3.2.2.1). For the BHM (Section 3.2.2.2), the posterior median is used as point estimate.[18] Figures 9 and 10 visualize distributions of point estimates for $\%\text{CV}_{\text{IP}}$ and ρ, respectively, where available.

We observe a couple of results. First, the mean of the QC sample is estimated well across all methods and scenarios (Table A5). It is rather the assay precision where scenarios and methods exhibit variation (Tables A6, A7). Generally speaking, when the number of runs becomes limited (designs "Sparse 1", "Sparse 2", "Unbalanced 1"), the FHM will tend to underestimate total variability, while the BHM tends to overestimate total variability – in line with results on QC range prediction coverage (Section 4.2.2). The most accurate Bayesian estimates of total variability are achieved under prior P4 and $\rho = 0.8$. As explained earlier (Section 4.2.2), overestimation is prevented under prior P4, probably due to its stronger "partial pooling" of variance components.

Second, point estimates of $\%\text{CV}_{\text{IP}}$ from the BHM (posterior median) tend to show a larger variation across simulated data sets than those from the FHM, especially so under priors P1 and P5. These impose uniform prior densities for τ. Intuitively, then, they also tend to produce wider posterior densities for τ (see also Figure 3A), with less stable posterior medians across random data sets. The FHM, instead, only maximizes a (restricted) likelihood, circumventing the problem (but also skipping the advantages) of estimating a whole distribution or its quantiles.

Finally, all hierarchical models are in principle able to capture the general pattern of partitioning of variances (Figure 10). The variability of point estimates is consistently higher for lower values of the true intra-run correlation, irrespective of Bayesian or frequentist inference. In these scenarios, we estimate small variances from little information on the group level, which is difficult. Again, prior P4 behaves a bit differently than the other priors: Small values of ρ are slightly overestimated, high values of ρ are slightly underestimated, medium values of ρ are estimated well – an effect of the mild shrinkage towards $\rho = 0.5$ prior P4 induces. Not surprisingly, estimates of ρ are most variable when there are few runs and also few observations (design "Sparse 1").

To close this section, we emphasize that our comparisons between Bayesian and frequentist point estimators should be interpreted with caution, as different results could be obtained if we used other posterior summaries than the posterior median (Pick et al., 2023). It is not generally advisable to summarize a Bayesian posterior distribution in a single point, in particular when estimating variance components. How well estimators based on different posterior summaries correspond to the true variance parameters is highly sensitive to multiple factors, not least the prior distribution (Browne & Draper, 2006). Accordingly, our QC ranges constructed from Bayesian models use the *full* joint posterior distribution.

[18] In particular for variances, the posterior median can be preferable as a point estimator over the posterior mean (e.g., Browne & Draper, 2006; Pick et al., 2023).

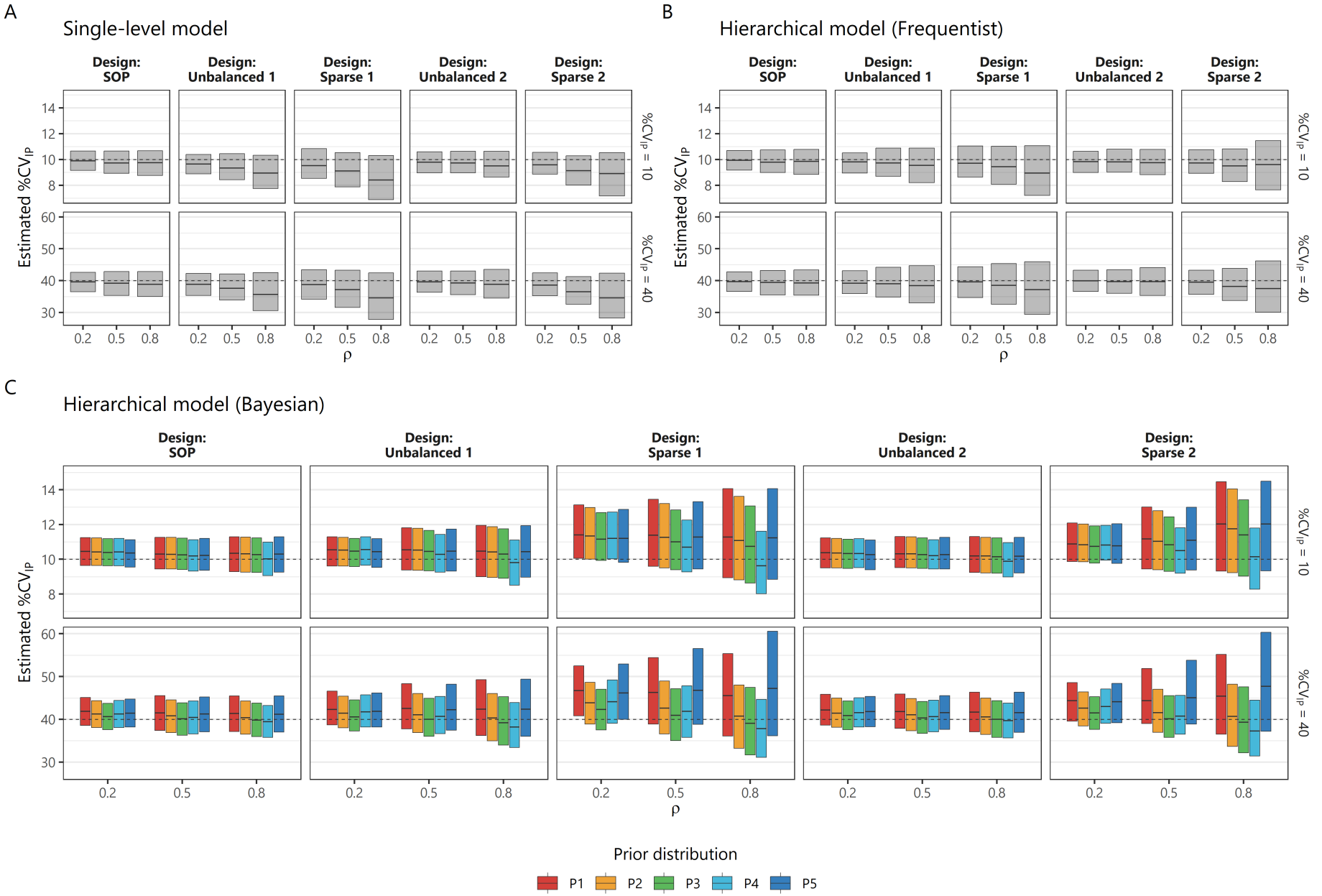

Figure 9: Distribution (median and interquartile range) of point estimates of the intermediate precision percent coefficient of variation for all scenarios based on (A) the single-level model, (B) the frequentist hierarchical model, and (C) the Bayesian hierarchical model, with its five different prior distributions. For the BHM, the posterior median is used as point estimator. Dashed lines indicate the true value. Across methods, the scaling of the plots is held constant.

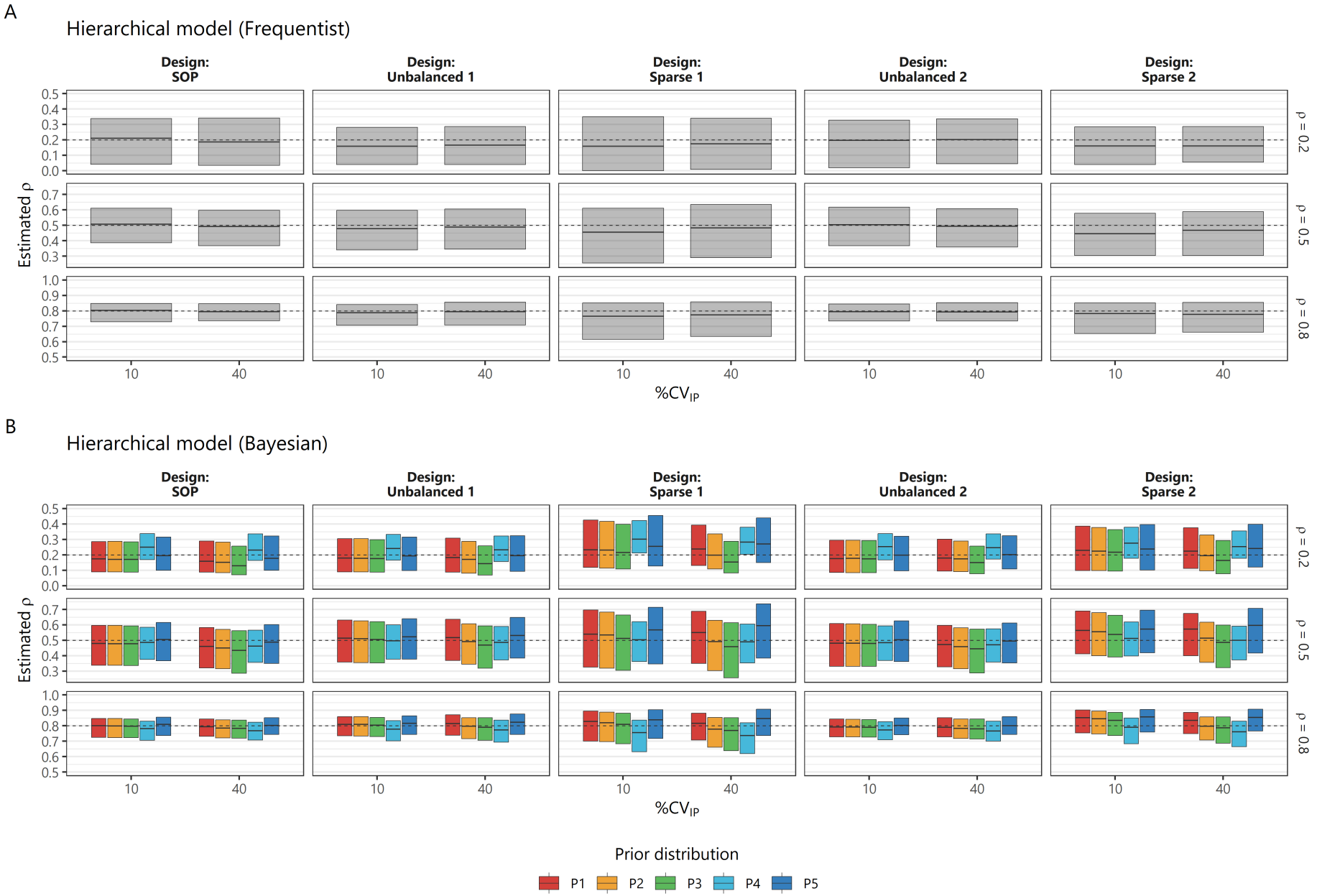

Figure 10: Distribution (median and interquartile range) of point estimates of the intra-run correlation for all scenarios based on (A) the frequentist hierarchical model, and (B) the Bayesian hierarchical model, with its five different prior distributions. For the BHM, the posterior median is used as point estimator. Dashed lines indicate the true value. Across methods, the scaling of the plots is held constant.

5 Discussion

5.1 Choosing between modeling approaches

Results show that the SLM that assumes no correlation structure fails to deliver suitable QC ranges for vaccine assay quality monitoring under realistic settings. Under moderate clustering of assay measurements in runs, the assay's intermediate precision is underestimated. QC ranges become too narrow, which, in practice, triggers too many rejections of measurements or runs that are, in fact, within specification. Indeed, the approach of pooling measurements from different assay batches or runs and ignoring the variance components structure, "is known to slightly underestimate the true interbatch imprecision" (DeSilva et al., 2003, p. 1892). Such underestimates result consistently in our simulation study (Table A8). Dessy et al. (2024, p. 1077), in their article on vaccine assay validation, also report that "total variability obtained by combining well characterized variance components [...] is a better estimate than simple variability estimates ignoring the interaction between these variables" and advocate the use of mixed models. Intuitively, correlated data reduce the amount of information in the data. This can also be viewed as a reduction of "effective sample size" due to clustering (Faes et al., 2009). A model that assumes independence is misspecified and will make too confident inferences. The misspecification becomes more severe under higher ratios of between- to within-run variance, with accordingly negative effects on prediction interval performance. In line with this, we see that observing only few independent runs in the baseline data collection further reduces performance of the SLM drastically under moderate to high correlation.

It is more advisable to estimate prediction intervals for assay quality control with hierarchical models. In our study, their prediction coverage is in all scenarios in the ballpark of nominal level ± 1 percentage points. Still, the frequentist approach to hierarchical modeling, even after degree of freedom approximations for prediction interval estimation (Francq et al., 2019), delivers some undercoverage on average when estimated based on few, highly correlated runs. Nevertheless, it can safely be used for constructing the QC range when runs are unbalanced, yet abundant (design "Unbalanced 2"), even under high correlation. The slightly worse performance in "Unbalanced 1" designs, compared to "Unbalanced 2" designs, can result from two sources: The former design provides less information on between-run variability (fewer runs), but also yields more extreme imbalance of run sizes.[19] This resonates with results from Browne & Draper (2006). The researchers compared the frequentist and Bayesian approach for Gaussian hierarchical modeling in a simulation study, albeit with somewhat larger sample sizes and more diffuse priors. They analyzed bias of different point estimators and (confidence/credible) interval coverage for parameters. Fewer clusters (their minimum was 6) led to less reliable point and interval estimates. A negative effect of unbalanced clusters could be observed in some scenarios, but was only small.

Bayesian approaches to hierarchical modeling bring a couple of advantages over the FHM. Some are more on the theoretical side. Estimation of model parameters for the FHM is based on a marginalized form of the hierarchical model, in which explicit clusters found in the data are not represented. Random effects can only be predicted after estimating the hyperparameters, which has been called Empirical Bayes. Instead, the BHM estimates hyperparameters and random effects jointly, having led Molenberghs & Verbeke (2005, p. 39) to the conclusion that "[a]rguably, a satisfactory treatment of

[19] When the constant total number of measurements is distributed across fewer runs, there is more potential for large run size heterogeneity under our data-generating mechanism for run size distributions (Appendix A.1). Indeed, mean standard deviations of generated run sizes in the "Unbalanced 1" and "Unbalanced 2" designs were 2.9 and 1.1, respectively.

the random-effects model is only possible within a Bayesian context". It is well-established that the FHM discards uncertainty about variance components (i.e., hyperparameters) for further inferences, like standard errors of fixed effects or prediction intervals, by fixing them at their point estimates (Gelman et al., 2013, Chapter 5; Lesaffre & Lawson, 2012, Chapter 9). Inference in the FHM is deeply grounded in large-sample theory (e.g., the asymptotic distribution of the likelihood-based estimator). Such shortcomings are what makes *post hoc* corrections for finite samples, like the use of complex degrees-of-freedom approximations, necessary in the first place. Instead, the Bayesian approach delivers exact posterior distributions that account for uncertainty in all parameters (though possibly in form of a sampling-based approximation). Lebrun & Rozet (2020, p. 382) emphasize the importance of this when validating analytical assays via predictive distributions Also note that, when the level-2 variance is very small or zero, only the Bayesian approach protects against negative variance estimates (Gelman et al., 2013, sec. 5.4; Lesaffre & Lawson, 2012, sec. 9.8.1).

Against this theoretical background, our results confirm good performance of the QC ranges derived from PPD quantiles of the BHM. Relevant undercoverage is, on average, not observed in any scenario, and prediction coverage varies less from sample to sample than for QC ranges derived from the FHM. The most striking result, though, is the average *over*-coverage of the QC range from the BHM in many scenarios. This needs to be considered when choosing between approaches. It can systematically result in runs and measurements being accepted during routine assay operations, although the measured QC sample concentrations are, in fact, much less likely than the nominal β level would suggest. If, for a certain assay development stage or type of sample, false *acceptances* are more costly or risky for the pharmaceutical or clinical user than false *rejections*, slight undercoverage could be preferable over slight overcoverage. The extent of overcoverage of the BHM is dependent on the chosen prior distribution, which leads us to the most important theoretical *and* practical advantage of the BHM for CMC applications (and beyond), namely the ability to include prior information. In the following Section 5.2, we discuss prior choice in more detail. In Section 6, we will briefly review some further practical extensions for assay QC range construction that can be conveniently implemented only in a Bayesian framework.

5.2 Specifying priors for the BHM

5.2.1 Non- and weakly informative priors

In the studied assay QC setting, in particular relative to other applications outside CMC, the data sets tend to be small, independent clusters are only few, and they provide limited information. In this type of setting, the importance of the level-2 variance prior is increased and needs close examination (Spiegelhalter et al., 2004, sec. 5.7.3). Our priors P1 to P4 were weakly informative. They used a weak notion of a plausible range for the intermediate precision variance, elicited from GSK's scientists, to provide some regularization. Yet, they were still sufficiently diffuse to let the likelihood dominate posterior inference about variance parameters. This can be seen when comparing the wide and long-tailed prior densities of standard deviation parameters (Figure 1) with the much more narrow posterior densities obtained from a typical data set (Figure 3). In most simulated scenarios, the expected prediction coverage, but also the interval widths and parameter point estimates, did not differ substantially across priors P1 to P4, although in some instances they did strikingly (see below). In particular prior specifications P2 (Half-normal for σ, τ) and P3 (Half-Cauchy for σ, τ), both parametrized to have the same median, showed satisfactory performance throughout and also quite similar performance – with minor advantages in terms of expected coverage for P3, but in terms of coverage error for P2. The

two prior distributions differ little in the parameter range to which they assign most of the probability mass, which likely explains their similar results.

We observe some problems with weakly informative prior P1 (restricted uniform for σ, τ) and would not recommend it for assay QC range construction, at least in the low-information settings investigated here. Just as for the non-informative reference prior P5, average coverage was too high throughout, even close to 100% in designs with very few runs (Table A.2), and the estimate (posterior median) of intermediate precision variance was both too large to a relevant degree (Table A8) and more variable than with other priors (Figure 9). The restricted uniform prior densities induce a relatively high prior median for the intermediate precision variance (Table 2), which is higher than what scientists usually expect for a vaccine assay (see Section 3.2.2.3). In sum, P1 puts too much weight on high values of the standard deviations. Hence, it also provides only little regularization. Gelman (2006, p. 521) calls this tendency of the restricted uniform prior for the level-2 variance a "slightly disagreeable miscalibration toward positive values". When there are few clusters (≤ 5), he discourages its use. Our results for the "Sparse 1" and "Sparse 2" designs confirm this. Again, a perspective on the costs of different types of errors is instructive. For example, in random effects meta-analysis, a uniform prior for the between-study variance has been described as conservative, as it leads to less shrinkage of study effects, and more uncertainty about the mean effect (Röver et al., 2021). When a QC range is built from posterior predictions of a hierarchical model, weak shrinkage could be interpreted as *anti-conservative*, instead:[20] Arguably, it would be the most conservative approach to reject most assay measurements, except the ones that deliver results for the QC samples closest to the expectation (i.e., to apply a very narrow QC range). These problems are even exacerbated with reference prior P5 (uniform over whole positive real line for τ). It can be viewed as non-informative, but, with its non-negative support, puts quite heavy weight on large values of the between-group standard deviation (Browne & Draper, 2006; Spiegelhalter et al., 2004). Hence, P5 should also not be used for assay quality monitoring. Further evidence in the same direction is provided by Hamaguchi et al. (2021), who study the coverage of Bayesian prediction intervals for random future cluster intercepts under different level-2 variance priors. Since they worked in a meta-analysis context, within-cluster variance was assumed known. When few clusters were available for model fitting (roughly ≤ 5), the improper uniform prior for the level-2 variance they used led to severe overcoverage and excessively wide prediction intervals, and more so for lower values of the true level-2 variance. We also found that MCMC sampling was least efficient for P5, and frequency of divergences was highest (Figure 4). The relatively "good" RMSCE values for priors P1 and P5 are a consequence of ceiling effects at excessive interval widths, and should not be taken to indicate truly low prediction coverage error. In fact, priors P1 and P5 often produced the most variable variance estimates.[21]

A point of discussion is how to choose parameters of the prior density functions for the variance components (i.e., hyperpriors), given that a certain distributional family is used as a weakly informative prior. In most cases, like our priors P2 and P3, this boils down to choosing a scale hyperparameter (e.g., for the half-normal distribution). A simple approach would be the type of data-dependent priors that are used by default for variance components in common software packages (e.g., Bürkner, 2017).

[20] See also Stan Development Team (2025) for this duality of perspectives: "Historically, a prior on the scale parameter with a long right tail has been considered 'conservative' in that it allows for large values of the scale parameter which in turn correspond to minimal pooling. But from a modern point of view, minimal pooling is not a default, and a statistical method that underpools can be thought of as overreacting to noise and thus 'anti-conservative.' "

[21] In principle, other (joint) non-informative priors than our prior P5 would be possible in the BHM, and they might lead to different results (see Hamaguchi et al., 2021 for a study in the meta-analysis context). Gelman (2006) criticizes some common choices of non-informative priors assigned on the variance scale.

They estimate a scale parameter from the sample data and plug it in as the scale hyperparameter of the prior distribution, possibly truncated at a lower limit, maintaining a minimum amount of vagueness.[22] Arguably, this has some nice features (most notably an automatic adaptation to the measurement scale), but it makes the prior sensitive to random sampling variation and violates the principle of assigning a prior before seeing the data. A more exploratory approach could start with a large scale hyperparameter and potentially reduce it after inspecting simulations from the joint prior or prior predictive distribution in case "too much" support for unrealistic values is obtained (a strategy employed by McElreath, 2020 for regression modeling). We also point to the principled procedure for choosing weakly informative priors for smoothing parameters, i.e., the random effects variance in our case, developed by Simpson et al. (2017). Their penalized complexity priors are simple to elicit (subjective input is only needed on one scale parameter) and yet satisfy many appealing properties.

Instead, we have used the approach of setting some scientifically informed constraints on an interpretable scale (e.g., an upper quantile of $\%CV_{IP}$, a prior expectation of ρ), and then worked our way back to the required hyperparameters (similar to considerations by Spiegelhalter et al., 2004, sec. 5.7.3). The constraints should be soft in order to maintain the weakly informative property. A collection of possible guiding questions is presented by Röver et al. (2021, p. 458), intended for medical meta-analysis, but with more general applicability. For analytical assays, broad considerations about the scale of measurement units, the realistic range for individual variabilities (as done by Novick et al., 2021) or the total variability, or the typical ratio of within-run to between-run variance could be helpful. The latter can be informed by considerations about the actual environmental factors that will vary across runs of the assay in the current lab (e.g., will the group of lab analysts be more or less constant?).

We did not include a weakly informative prior for the QC sample mean μ, but this would be a further possibility; in particular if concentrations of QC samples have been established as a reference in the same lab previously. Our normal prior for μ with standard deviation 100 on the $\log_{10}$ scale was highly diffused. Instead, for example, Wang & Cheng (2022, p. 199) center their normal prior for the mean in their Bayesian model at 100, as for "a typical bioassay" on the untransformed scale, and assign it a standard deviation of 20. Subsequently, they conduct prior sensitivity analysis.

Prior P4 is specified in terms of independent priors for intermediate precision variance and intra-run correlation, which is a special case of the hierarchical decomposition priors proposed by Fuglstad et al. (2020). We believe that this approach has some advantages over the others. It can make the elicitation of weakly informative (and also informative, see Section 5.2.2) priors easier, because the two quantities most intuitively interpreted are controlled directly. This gave us a simple way to express true ignorance about ρ with a flat (Beta) density, bearing some resemblance to the idea of putting a uniform prior on the shrinkage factor (Spiegelhalter et al., 2004). As one result of this, only our prior P4 implies that a value of zero is impossible for the intermediate precision standard deviation (Figure 1), in line with the plausible assumption that there is no variability-free clinical assay. More technically, the different specification allows the user to induce prior correlation between the two variance components, while simultaneously controlling the marginal total variance. Independence of variance parameters (as in priors P1, P2, P3, and P5) is typically not a realistic assumption for hierarchical models (see also posterior correlations in Appendix A.7).

Also in terms of simulation results, prior P4 occupies a special place In scenarios with high heterogene-

[22] By default, `brms` (Bürkner, 2017) specifies a half-t distribution with 3 degrees of freedom and minimum scale 2.5 for the standard deviation parameters in the BHM.

ity between runs, but not in scenarios with small or moderate heterogeneity, it produces markedly different performance from all other priors. We explain this with stronger shrinkage of variances towards a common value (Section 4.2.2): Although all our weakly informative priors imply the prior expectation $\mathbb{E}(\rho) = 0.5$ (equal magnitude of variance components), prior P4 does so at lower prior variance. In the investigated designs, there is typically more information in the data on within-run than on between-run variability. Therefore, prior P4 tends to shrink between-run variability towards within-run variability, rather than vice versa. This is least desirable when between-run variability is large ($\rho = 0.8$), as we will underestimate total variability. Consequently, when the amount of empirical information about the two variances is most balanced (designs "Standard", "Unbalanced 2")[23], the reduction in average prediction coverage for prior P4 at high correlation is smallest. Conveniently, prior P4 can be modified easily via the parameters of the Beta density $p(\rho)$ in order to tailor the strength of this tendency to the actual design of baseline data collection – even before any data analysis. In addition, the induced prior correlation $\mathrm{Cor}(\sigma^2, \tau^2)$, which will largely depend on the design (see Searle et al., 1992, p. 85, and Appendix A.7) can be used as an indicator of the suitability of the hyperparameters. Therefore, we also recommend prior P4 for practical usage in assay QC range construction, with the suggestion to test the operating characteristics of other hyperparameter choices, which could vaguely acknowledge more prior information (number and size of runs, expectation about which is the dominant variance component), where available. For example, if there are few runs and one expects between-run to exceed within-run variance, putting $> 50\%$ of prior probability density above $\rho = 0.5$ (e.g., via $\mathrm{Beta}(1, 0.5)$) will likely dampen the underestimation of between-run and total variance we observed for higher values of ρ. The asymmetric prior will also improve prediction coverage, while still being only weakly informative. Even if prior expectation $\mathbb{E}(\rho) = 0.5$ shall be maintained, increasing the variance of the prior on ρ relative to the uniform density (e.g., via $\mathrm{Beta}(0.5, 0.5)$) will likely dampen the underestimation in settings with few, highly heterogeneous runs.

5.2.2 Informative priors

All our priors P1 to P4 can relatively easily be adapted to be more informative. This is relevant for vaccine assay quality control. Due to various factors, the baseline data for the QC sample used for QC range construction could be unreliable. Its variability might not be representative of true assay variability, for example because only results from very few runs, or runs obtained under few similar conditions, could be compiled. In this case, QC range estimation could be enhanced with expert "guesstimates" about the precision of the specific assay at hand. Alternatively, there might be results available for the same or a closely related assay, obtained in other experiments. These could precision experiments, which are conducted during assay validation and provide estimates of assay precision at various sample concentrations (Dessy et al., 2024). Such estimates could then be encoded in informative prior distributions and combined with the assay's baseline data from the QC samples in the BHM. As Wang & Cheng (2022) describe their approach to informative prior specification: "The method variability, e.g., intermediate precision (IP), is often available from the corresponding qualification or validation study of the same or similar compounds." The application of informative priors for specific vaccine assays, or subtypes of vaccine assays (e.g., ELISA vs. functional assays), has been deemed valuable by GSK, but its detailed investigation was beyond the scope of this study.

Nevertheless, we think that the parametrization developed for weakly informative prior P4 is by far

[23] A crude way to approximate the amount of information on the different variance components from the design of data collection alone could be via the ANOVA degrees of freedom $\mathrm{df}_{\text{between}} = m - 1$ and $\mathrm{df}_{\text{within}} = N - m$ (Searle et al., 1992, p. 72). In all our simulated designs, we have $\mathrm{df}_{\text{between}} < \mathrm{df}_{\text{within}}$.

best suited for setting informative priors. For example, the estimate of $\%CV_{IP}$ from the precision experiment could easily be set as prior mode, median, or mean using the Gamma density (see Section 3.2.2.3), with the second parameter giving control over the precise degree of informativeness (or, the plausible range for the specific vaccine assay *after* having observed its precision experiments). If also separate estimates of the variance components can be gleaned from the precision experiment, only under the P4 parametrization can this knowledge be incorporated directly (see also arguments in Fuglstad et al., 2020). A Beta density provides excellent flexibility to express such information directly on the ρ scale.

Information on variability from related clinical assays could also be included in a more structured fashion by extending the model with further hierarchical priors. Imagine measurement data from other assays (with other QC samples) are available for modeling. The assumption that assays' intermediate precisions are drawn from a common, larger population of precisions leads to a Bayesian location-scale multilevel model as proposed and studied by Schach et al. (2025). The authors recommend the model for quantifying biopharmaceutical process variance of data-scarce or new products. Transferring this model to our application, fitting and QC range computation would be possible with the same software and methods that we used for the simpler hierarchical models here.

5.3 Limitations

Although we studied 7 combinations of model and estimation approach, the type of interval used as QC range was never varied in our study. For example, the nominal level was always held at $\beta = 0.99$. This is a relatively high value, although not uncommon in QC applications (e.g., Lewis & Hudson-Curtis, 2022; Novick et al., 2021). At lower values of β, we would probably observe less ceiling effects for prediction coverage (described in Section 4.2.2), as well as less sensitivity of the Bayesian QC range to the tail behavior of the variance priors. Less extreme quantiles are also sampled more efficiently with MCMC (Vehtari et al., 2021). In addition, the Bayesian prediction intervals we examined were based on quantiles of the PPD. Highest posterior density intervals, used by Novick et al. (2021) as bioassay QC limits, provide an alternative way to define a QC range from the PPD at the same probability level, with the possible advantage of being narrower. We did not examine their performance.[24] Similarly, we did not look at the performance of the methods in predicting future run means instead of individual observations. Such predictions would be readily available based on the joint posterior, and they are relevant in other applications of hierarchical models (e.g., meta-analyses, see Hamaguchi et al., 2021), but less so in the application studied here. We neither investigated type II tolerance intervals (Patel, 1986), although they are commonly used in quality control and can be very naturally obtained in a Baysian framework (Lewis & Hudson-Curtis, 2022).

We applied a limited number of interval performance metrics. Further, we did so in isolation, which makes the overall picture somewhat hard to grasp. A fruitful approach could be to analyze interval scores for prediction intervals and predictive quantiles that combine both coverage and width information in a single metric (Gneiting & and Raftery, 2007). Due to focus on prediction intervals as QC ranges, our simulation study did not systematically quantify or compare the bias and variance of different hierarchical model-based point estimators, both frequentist and Bayesian (as in, for example, Browne & Draper, 2006). Supposedly, such an analysis would have provided an even better picture of why and when methods differ in their prediction performance. Nevertheless, we report descriptive

[24] Note that the highest density property does not transfer to the original measurement scale when back-transforming the QC range to the original measurement scale, while the quantile property does.

summaries of parameter estimates (Section 4.2.3, Appendix A.6).

Naturally, the parameter space explored in a simulation study impacts conclusions. We covered a relatively broad range across 30 different scenarios. Still, it would have been interesting to observe performances across scenarios where one variance component is held at constant magnitude, while the other one is varied. We did not include such scenarios. The interpretation of results (and variability of results) for the unbalanced designs is somewhat obfuscated by the fact that the design itself was random. Hence, data sets in the respective scenarios differed not only randomly in terms of observed measurements, but also in the distribution of cluster sizes (Appendix A.1). Further insights could possibly be gained by correlating QC range performance metrics against measures of realized unbalancedness. This is possible with the results from our simulation study, but has so far not been conducted systematically.

6 Conclusion

In this study, we have shown that a Bayesian hierarchical modeling approach can be employed for the problem of QC range construction for clinical assays. Different ways to specify the prior distribution were proposed and implemented with state-of-the-art Bayesian software packages. Joint posterior distributions were obtained efficiently with the MCMC technique and could be used to infer predictive intervals for quality control.

Based on our results, a model that ignores clustering of observations within assay runs cannot safely be used for setting control limits in many realistic settings. In contrast, the prediction interval derived from a hierarchical model estimated via REML often provides a suitable QC range. Yet, since it is not immune to undercoverage in sparse-data and high-correlation settings, and lacks a way to incorporate prior information – a highly valuable feature in CMC applications – we advocate for the use of predictive distributions from Bayesian hierarchical models as a viable alternative. Although they led to notable overcoverage of future measurements relative to the nominal level in some settings, this tendency is sensitive to prior choice. As we showed and discussed, it is reasonable to use weakly informative priors for the variance components in the assay quality control application, rather than non-informative priors. The extension to informative priors is straightforward. We have outlined possible routes for further refinement of prior specification.

As an outlook, there are further advantages of the Bayesian QC range approach. It is easy to relax distributional assumptions, for example about the population from which runs are sampled. Also, a log-normal likelihood could be included directly in the Stan program, avoiding the somewhat *ad hoc* log transformation of the data. In addition, the Bayesian approach is perfectly suited for a dynamic updating of the QC range after observing new runs over the course of clinical assay development, as also noted by Lebrun & Rozet (2020, p. 387): "when obtaining each new data point of the control chart, the control limits can be updated, providing a more precise estimation of the control limits of the assay." Hence, Bayesian predictive distributions can be a step towards a more continuous characterization of a bioassay's performance (Novick et al., 2021).

Finally, any method makes errors. When monitoring clinical assay quality, these errors could be false acceptances or false rejections of new measurements and runs. In our study, different methods, as well as different prior distributions, made different types of errors more likely. The choice of method should be guided by considerations of the costs these errors imply for the user and, ultimately, the risks they pose to patients. This can only be judged with a holistic view on the product's purpose, its development process, and its "real-life" usage for medical decision-making.

7 References

Betancourt, M. (2016). *Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo*. https://arxiv.org/abs/1604.00695

Boulanger, B., & Mutsvari, T. (2020). Product Development and Manufacturing. In *Bayesian Methods in Pharmaceutical Research*. Chapman and Hall/CRC.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514. https://doi.org/10.1214/06-BA117

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., Gabry, J., Kay, M., & Vehtari, A. (2025). *Posterior: Tools for working with posterior distributions*.

Canchola, J. A., Tang, S., Hemyari, P., Paxinos, E., & Marins, E. (2017). Correct use of percent coefficient of variation (%CV) formula for log-transformed data. *MOJ Proteomics & Bioinformatics*, *6*(4), 316–317. https://doi.org/10.15406/mojpb.2017.06.00200

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury Press.

Chaloner, K. (1987). A Bayesian Approach to the Estimation of Variance Components for the Unbalanced One-Way Random Model. *Technometrics*, *29*(3), 323–337. https://doi.org/10.1080/00401706.1987.10488242

DeSilva, B., Smith, W., Weiner, R., Kelley, M., Smolec, J., Lee, B., Khan, M., Tacey, R., Hill, H., & Celniker, A. (2003). Recommendations for the Bioanalytical Method Validation of Ligand-Binding Assays to Support Pharmacokinetic Assessments of Macromolecules. *Pharmaceutical Research*, *20*(11), 1885–1900. https://doi.org/10.1023/B:PHAM.0000003390.51761.3d

Dessy, F., Sonderegger, Wagner, Buoninfante, Wadhwa, Agnes, Aksyuk, Baclin, Bonhomme, Cloney-Clark, Corsaro, Neto, Fries, Gagnon, Garofolo, Giardina, Green, Guimera, Harris, … and Zhu, M. (2024). Harmonization of Vaccine Ligand Binding Assays Validation. *Bioanalysis*, *16*(19-20), 1067–1091. https://doi.org/10.1080/17576180.2024.2411925

Faes, C., Molenberghs, G., Aerts, M., Verbeke, G., & Kenward, M. G. (2009). The Effective Sample Size and an Alternative Small-Sample Degrees-of-Freedom Method. *The American Statistician*, *63*(4), 389–399. https://www.jstor.org/stable/25652320

Faya, P., & Pourmohamad, T. (2022a). Introduction. In *Case Studies in Bayesian Methods for Biopharmaceutical CMC* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003255093-1

Faya, P., & Pourmohamad, T. (2022b). *Case Studies in Bayesian Methods for Biopharmaceutical CMC* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003255093

FDA. (2011). *Guidance for Industry: Process Validation: General Principles and Practices*. U.S. Food and Drug Administration.

Francq, B. G., Lin, D., & Hoyer, W. (2019). Confidence, prediction, and tolerance in linear mixed models. *Statistics in Medicine*, *38*(30), 5603–5622. https://doi.org/10.1002/sim.8386

Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., & Riebler, A. (2020). Intuitive Joint Priors for Variance Parameters. *Bayesian Analysis*, *15*(4), 1109–1137. https://doi.org/10.1214/19-BA1185

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2025). *Cmdstanr: R Interface to 'CmdStan'*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A. (2009). Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics. *Statistical Science*, *24*(2), 176–178. https://doi.org/10.1214/09-STS284D

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gneiting, T., & and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Hamaguchi, Y., Noma, H., Nagashima, K., Yamada, T., & Furukawa, T. A. (2021). Frequentist performances of Bayesian prediction intervals for random-effects meta-analysis. *Biometrical Journal*, *63*(2), 394–405. https://doi.org/10.1002/bimj.201900351

ICH. (2009). *ICH Guideline: Pharmaceutical Development Q8(R2)*. International Conference on Harmonisation.

ICH. (2023). *ICH Guideline: Quality Risk Management Q9(R1)*. International Council for Harmonisation.

Junker, B., Zablackis, E., Verch, T., Schofield, T., & Douette, P. (2015). Quality-by-Design: As Related to Analytical Concepts, Control and Qualification. In B. K. Nunnally, V. E. Turula, & R. D. Sitrin (Eds.), *Vaccine Analysis: Strategies, Principles, and Control* (pp. 479–520). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-45024-6_12

Kiyani, S., Pappas, G., & Hassani, H. (2024). Conformal prediction with learned features. *Proceedings of the 41st International Conference on Machine Learning*.

Lebrun, P., & Rozet, E. (2020). Analytical Method and Assay. In *Bayesian Methods in Pharmaceutical Research* (1st ed.). Chapman and Hall/CRC.

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian Biostatistics*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119942412

Lewis, R., & Hudson-Curtis, B. (2022). Calculating Statistical Tolerance Intervals Using SAS. In *Case Studies in Bayesian Methods for Biopharmaceutical CMC* (1st ed.). Chapman and Hall/CRC.

Lewontin, R. C. (1966). On the Measurement of Relative Variability. *Systematic Biology*, *15*(2), 141–142. https://doi.org/10.2307/sysbio/15.2.141

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

Molenberghs, G., & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer. https://doi.org/10.1007/0-387-28980-1_14

Novick, S. J., Christian, E., Farmer, E., & Tejada, M. (2021). A Bayesian Statistical Approach to Continuous Qualification of a Bioassay. *PDA Journal of Pharmaceutical Science and Technology*, *75*(1), 8–23. https://doi.org/10.5731/pdajpst.2019.011221

Patel, J. K. (1986). Tolerance limits - a review. *Communications in Statistics - Theory and Methods*, *15*(9), 2719–2762. https://doi.org/10.1080/03610928608829278

Peterson, J. J. (2020). Process Development and Validation. In *Bayesian Methods in Pharmaceutical Research* (1st ed.). Chapman and Hall/CRC.

Pick, J. L., Kasper, C., Allegue, H., Dingemanse, N. J., Dochtermann, N. A., Laskowski, K. L., Lima, M. R., Schielzeth, H., Westneat, D. F., Wright, J., & Araya-Ajoy, Y. G. (2023). Describing posterior distributions of variance components: Problems and the use of null distributions to aid interpretation. *Methods in Ecology and Evolution*, *14*(10), 2557–2574. https://doi.org/10.1111/2041-210X.14200

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11.

Polson, N. G., & Scott, J. G. (2012). On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis*, *7*(4), 887–902. https://doi.org/10.1214/12-BA730

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S., & Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*, *12*(4), 448–474. https://doi.org/10.1002/jrsm.1475

Schach, S., Eilert, T., Presser, B., & Kunzelmann, M. (2025). Bayesian Hierarchical Modeling for Variance Estimation in Biopharmaceutical Processes. *Bioengineering*, *12*(2), 193. https://doi.org/10.3390/bioengineering12020193

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. Wiley-Interscience. https://doi.org/10.1002/9780470316856

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, *32*(1), 1–28. https://doi.org/10.1214/16-STS576

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470092602

Stan Development Team. (2024). *Stan Modeling Language Users Guide and Reference Manual* (Version 2.36).

Stan Development Team. (2025). Prior Choice Recommendations. In *GitHub*. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Tian, Q., Nordman, D. J., & Meeker, W. Q. (2022). Methods to Compute Prediction Intervals: A Review and New Results. *Statistical Science*, *37*(4), 580–597. https://doi.org/10.1214/21-STS842

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved R-hat for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer.

Wang, K., & Cheng, A. (2022). Bayesian Evaluation and Monitoring of Process Comparability. In *Case Studies in Bayesian Methods for Biopharmaceutical CMC* (1st ed.). Chapman and Hall/CRC.

West, R. M. (2022). Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry*, *59*(3), 162–165. https://doi.org/10.1177/00045632211050531

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC.

A Appendix

A.1 Unbalanced design generation

Algorithm A1 produces the unbalanced run sizes that form part of some designs in our simulations (Section 3.1.2). Since the order of run sizes is not important, the algorithm can be viewed as generating partitions of n that consist of m positive integer parts. The algorithm has properties which align well with real-world scenarios: Extreme partitions (one very large run) are less probable than somewhat more balanced partitions, but completely uniform run sizes are very unlikely (in line with the goal to create unbalanced run sizes). Indeed, no uniform run size distribution occurred by chance in our simulated data sets. Table A1 summarizes the run sizes that were randomly generated by Algorithm A1 for all scenarios based on the two unbalanced designs. The most frequent ordered run size vectors for designs "Unbalanced 1" and "Unbalanced 2" were (1, 2, 2, 3, 4, 4, 5, 6, 6, 7) and (1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4), with frequencies 25 and 310, respectively.

Algorithm A1: Generate vector of run sizes.

Input: $n \geq m \geq 1$ (total number of observations n, number of runs m)

Output: Random partition of n into m positive parts

1 Sample $m-1$ cut points independently from $\{0, 1, \ldots, n-m\}$ with equal probability;

2 Sort the cut points and add boundaries 0 and $n-m$;

3 Compute gaps between consecutive points, add 1 to each gap;

Table A1: Frequency of run sizes generated in the two designs with unbalanced run sizes. Design 'Unbalanced 2' contains double the number of runs as design 'Unbalanced 1', but the same total sample size.

	Scenario	
Run size	Unbalanced 1	Unbalanced 2
1	7,145	33,750
2	9,897	34,469
3	7,388	13,841
4	5,606	5,172
5	4,289	1,886
6	3,223	617
7	2,235	183
8	1,655	61
9	1,165	11
10	813	9
11	540	1
12	353	0
13	259	0
14	171	0
15	107	0
16	65	0
17	43	0
18	21	0
19	11	0
20	8	0
21	3	0
22	2	0
23	1	0
All	45,000	90,000

A.2 Expected prediction coverage estimates

Table A2: Monte Carlo estimates of expected prediction coverage of QC range with Monte Carlo standard error (MCSE). SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

				Estimated expected prediction coverage and MCSE						
Design	Scenario	$\%CV_{IP}$	ρ	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	0.9897 (0.0003)	0.9903 (0.0003)	0.9931 (0.0003)	0.9930 (0.0003)	0.9928 (0.0003)	0.9930 (0.0003)	0.9926 (0.0003)
Standard	2	10	0.5	0.9877 (0.0004)	0.9894 (0.0004)	0.9922 (0.0003)	0.9921 (0.0003)	0.9919 (0.0003)	0.9916 (0.0003)	0.9919 (0.0003)
Standard	3	10	0.8	0.9858 (0.0006)	0.9895 (0.0005)	0.9919 (0.0004)	0.9918 (0.0004)	0.9916 (0.0004)	0.9902 (0.0004)	0.9919 (0.0004)
Standard	4	40	0.2	0.9899 (0.0003)	0.9905 (0.0003)	0.9932 (0.0003)	0.9923 (0.0003)	0.9915 (0.0003)	0.9924 (0.0003)	0.9927 (0.0003)
Standard	5	40	0.5	0.9877 (0.0004)	0.9894 (0.0004)	0.9922 (0.0003)	0.9911 (0.0003)	0.9901 (0.0004)	0.9908 (0.0003)	0.9919 (0.0003)
Standard	6	40	0.8	0.9860 (0.0006)	0.9897 (0.0005)	0.9921 (0.0004)	0.9907 (0.0004)	0.9898 (0.0005)	0.9892 (0.0005)	0.9920 (0.0004)
Unbalanced 1	7	10	0.2	0.9871 (0.0004)	0.9895 (0.0004)	0.9947 (0.0002)	0.9946 (0.0002)	0.9943 (0.0002)	0.9949 (0.0002)	0.9943 (0.0003)
Unbalanced 1	8	10	0.5	0.9816 (0.0007)	0.9889 (0.0006)	0.9945 (0.0003)	0.9944 (0.0004)	0.9940 (0.0004)	0.9937 (0.0003)	0.9943 (0.0004)
Unbalanced 1	9	10	0.8	0.9682 (0.0013)	0.9874 (0.0009)	0.9935 (0.0006)	0.9932 (0.0006)	0.9927 (0.0006)	0.9898 (0.0007)	0.9934 (0.0006)
Unbalanced 1	10	40	0.2	0.9874 (0.0004)	0.9895 (0.0004)	0.9945 (0.0003)	0.9932 (0.0003)	0.9921 (0.0003)	0.9939 (0.0003)	0.9942 (0.0003)
Unbalanced 1	11	40	0.5	0.9818 (0.0007)	0.9892 (0.0006)	0.9946 (0.0004)	0.9927 (0.0004)	0.9914 (0.0005)	0.9923 (0.0004)	0.9944 (0.0004)
Unbalanced 1	12	40	0.8	0.9681 (0.0014)	0.9879 (0.0008)	0.9935 (0.0006)	0.9909 (0.0006)	0.9894 (0.0007)	0.9878 (0.0007)	0.9936 (0.0006)
Sparse 1	13	10	0.2	0.9873 (0.0007)	0.9901 (0.0006)	0.9990 (0.0002)	0.9987 (0.0002)	0.9983 (0.0002)	0.9983 (0.0002)	0.9988 (0.0002)
Sparse 1	14	10	0.5	0.9791 (0.0010)	0.9869 (0.0009)	0.9986 (0.0002)	0.9983 (0.0003)	0.9977 (0.0003)	0.9971 (0.0003)	0.9985 (0.0002)
Sparse 1	15	10	0.8	0.9550 (0.0023)	0.9800 (0.0018)	0.9969 (0.0006)	0.9964 (0.0006)	0.9956 (0.0007)	0.9911 (0.0009)	0.9969 (0.0006)
Sparse 1	16	40	0.2	0.9879 (0.0006)	0.9909 (0.0005)	0.9988 (0.0001)	0.9970 (0.0002)	0.9957 (0.0003)	0.9972 (0.0002)	0.9992 (0.0001)
Sparse 1	17	40	0.5	0.9802 (0.0010)	0.9882 (0.0009)	0.9981 (0.0002)	0.9957 (0.0004)	0.9941 (0.0005)	0.9951 (0.0004)	0.9988 (0.0002)
Sparse 1	18	40	0.8	0.9607 (0.0020)	0.9846 (0.0014)	0.9970 (0.0005)	0.9932 (0.0007)	0.9915 (0.0008)	0.9887 (0.0009)	0.9981 (0.0004)
Unbalanced 2	19	10	0.2	0.9886 (0.0004)	0.9895 (0.0004)	0.9927 (0.0003)	0.9926 (0.0003)	0.9924 (0.0003)	0.9927 (0.0003)	0.9922 (0.0003)
Unbalanced 2	20	10	0.5	0.9876 (0.0005)	0.9900 (0.0004)	0.9927 (0.0004)	0.9927 (0.0004)	0.9924 (0.0004)	0.9922 (0.0004)	0.9925 (0.0004)
Unbalanced 2	21	10	0.8	0.9838 (0.0007)	0.9893 (0.0005)	0.9917 (0.0004)	0.9916 (0.0004)	0.9914 (0.0004)	0.9898 (0.0005)	0.9916 (0.0005)
Unbalanced 2	22	40	0.2	0.9895 (0.0004)	0.9903 (0.0004)	0.9933 (0.0003)	0.9924 (0.0003)	0.9915 (0.0003)	0.9926 (0.0003)	0.9928 (0.0003)
Unbalanced 2	23	40	0.5	0.9876 (0.0005)	0.9899 (0.0005)	0.9926 (0.0004)	0.9916 (0.0004)	0.9905 (0.0004)	0.9913 (0.0004)	0.9924 (0.0004)
Unbalanced 2	24	40	0.8	0.9843 (0.0007)	0.9897 (0.0005)	0.9919 (0.0004)	0.9906 (0.0005)	0.9896 (0.0005)	0.9890 (0.0005)	0.9918 (0.0005)
Sparse 2	25	10	0.2	0.9865 (0.0005)	0.9895 (0.0004)	0.9982 (0.0002)	0.9979 (0.0002)	0.9975 (0.0002)	0.9979 (0.0001)	0.9982 (0.0002)
Sparse 2	26	10	0.5	0.9747 (0.0010)	0.9856 (0.0009)	0.9981 (0.0003)	0.9977 (0.0003)	0.9972 (0.0003)	0.9964 (0.0003)	0.9981 (0.0003)
Sparse 2	27	10	0.8	0.9537 (0.0021)	0.9836 (0.0015)	0.9980 (0.0004)	0.9977 (0.0004)	0.9969 (0.0005)	0.9932 (0.0007)	0.9981 (0.0003)
Sparse 2	28	40	0.2	0.9867 (0.0005)	0.9898 (0.0004)	0.9978 (0.0002)	0.9959 (0.0002)	0.9947 (0.0003)	0.9967 (0.0002)	0.9983 (0.0001)
Sparse 2	29	40	0.5	0.9772 (0.0009)	0.9874 (0.0007)	0.9975 (0.0003)	0.9952 (0.0004)	0.9939 (0.0004)	0.9949 (0.0003)	0.9983 (0.0002)
Sparse 2	30	40	0.8	0.9518 (0.0022)	0.9821 (0.0016)	0.9962 (0.0006)	0.9926 (0.0008)	0.9911 (0.0009)	0.9879 (0.0010)	0.9973 (0.0005)

A.3 Expected interval width estimates

Table A3: Monte Carlo estimates of expected interval width of QC range with Monte Carlo standard error (MCSE). SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

				Estimated expected interval width and MCSE						
Design	Scenario	$\%CV_{IP}$	ρ	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	0.236 (0.001)	0.238 (0.001)	0.250 (0.001)	0.249 (0.001)	0.248 (0.001)	0.249 (0.001)	0.247 (0.001)
Standard	2	10	0.5	0.233 (0.001)	0.239 (0.001)	0.250 (0.001)	0.250 (0.001)	0.249 (0.001)	0.247 (0.001)	0.249 (0.001)
Standard	3	10	0.8	0.232 (0.001)	0.245 (0.001)	0.255 (0.001)	0.254 (0.001)	0.253 (0.001)	0.246 (0.001)	0.255 (0.001)
Standard	4	40	0.2	0.910 (0.004)	0.919 (0.004)	0.964 (0.004)	0.945 (0.004)	0.932 (0.004)	0.948 (0.004)	0.955 (0.004)
Standard	5	40	0.5	0.900 (0.004)	0.925 (0.005)	0.968 (0.005)	0.945 (0.005)	0.930 (0.004)	0.939 (0.004)	0.963 (0.005)
Standard	6	40	0.8	0.896 (0.005)	0.947 (0.005)	0.985 (0.005)	0.954 (0.005)	0.943 (0.005)	0.932 (0.005)	0.984 (0.005)
Unbalanced 1	7	10	0.2	0.230 (0.001)	0.239 (0.001)	0.264 (0.001)	0.263 (0.001)	0.261 (0.001)	0.262 (0.001)	0.262 (0.001)
Unbalanced 1	8	10	0.5	0.226 (0.001)	0.253 (0.002)	0.284 (0.002)	0.282 (0.002)	0.277 (0.002)	0.269 (0.002)	0.283 (0.002)
Unbalanced 1	9	10	0.8	0.217 (0.002)	0.266 (0.002)	0.297 (0.002)	0.294 (0.002)	0.287 (0.002)	0.266 (0.002)	0.297 (0.002)
Unbalanced 1	10	40	0.2	0.897 (0.004)	0.934 (0.005)	1.028 (0.005)	0.986 (0.005)	0.965 (0.005)	0.997 (0.005)	1.025 (0.006)
Unbalanced 1	11	40	0.5	0.882 (0.005)	0.987 (0.007)	1.089 (0.007)	1.022 (0.006)	1.003 (0.006)	1.006 (0.006)	1.104 (0.008)
Unbalanced 1	12	40	0.8	0.848 (0.007)	1.044 (0.009)	1.125 (0.008)	1.039 (0.007)	1.035 (0.008)	0.987 (0.007)	1.164 (0.010)
Sparse 1	13	10	0.2	0.246 (0.002)	0.266 (0.002)	0.397 (0.004)	0.369 (0.003)	0.345 (0.003)	0.333 (0.002)	0.402 (0.004)
Sparse 1	14	10	0.5	0.238 (0.002)	0.292 (0.004)	0.457 (0.005)	0.411 (0.004)	0.380 (0.004)	0.337 (0.003)	0.475 (0.006)
Sparse 1	15	10	0.8	0.221 (0.002)	0.329 (0.005)	0.503 (0.006)	0.444 (0.005)	0.407 (0.004)	0.327 (0.003)	0.533 (0.007)
Sparse 1	16	40	0.2	0.959 (0.006)	1.036 (0.008)	1.285 (0.007)	1.143 (0.006)	1.121 (0.007)	1.161 (0.007)	1.571 (0.016)
Sparse 1	17	40	0.5	0.931 (0.007)	1.154 (0.013)	1.334 (0.009)	1.180 (0.008)	1.206 (0.011)	1.150 (0.008)	1.887 (0.024)
Sparse 1	18	40	0.8	0.885 (0.009)	1.330 (0.019)	1.356 (0.009)	1.201 (0.009)	1.296 (0.015)	1.106 (0.009)	2.154 (0.029)
Unbalanced 2	19	10	0.2	0.232 (0.001)	0.236 (0.001)	0.248 (0.001)	0.248 (0.001)	0.247 (0.001)	0.248 (0.001)	0.246 (0.001)
Unbalanced 2	20	10	0.5	0.233 (0.001)	0.242 (0.001)	0.253 (0.001)	0.253 (0.001)	0.251 (0.001)	0.249 (0.001)	0.252 (0.001)
Unbalanced 2	21	10	0.8	0.230 (0.001)	0.245 (0.001)	0.254 (0.001)	0.253 (0.001)	0.252 (0.001)	0.245 (0.001)	0.254 (0.001)
Unbalanced 2	22	40	0.2	0.911 (0.004)	0.925 (0.004)	0.973 (0.004)	0.952 (0.004)	0.938 (0.004)	0.956 (0.004)	0.965 (0.004)
Unbalanced 2	23	40	0.5	0.904 (0.004)	0.939 (0.005)	0.981 (0.005)	0.956 (0.005)	0.941 (0.005)	0.950 (0.004)	0.978 (0.005)
Unbalanced 2	24	40	0.8	0.897 (0.005)	0.956 (0.005)	0.992 (0.006)	0.960 (0.005)	0.949 (0.005)	0.937 (0.005)	0.992 (0.006)
Sparse 2	25	10	0.2	0.231 (0.001)	0.247 (0.002)	0.352 (0.003)	0.332 (0.003)	0.315 (0.002)	0.306 (0.002)	0.357 (0.004)
Sparse 2	26	10	0.5	0.221 (0.001)	0.274 (0.003)	0.439 (0.005)	0.398 (0.004)	0.369 (0.004)	0.322 (0.003)	0.455 (0.006)
Sparse 2	27	10	0.8	0.214 (0.002)	0.339 (0.005)	0.527 (0.006)	0.462 (0.005)	0.424 (0.004)	0.337 (0.003)	0.558 (0.007)
Sparse 2	28	40	0.2	0.895 (0.004)	0.956 (0.006)	1.180 (0.007)	1.072 (0.006)	1.050 (0.006)	1.091 (0.005)	1.382 (0.014)
Sparse 2	29	40	0.5	0.860 (0.005)	1.068 (0.011)	1.292 (0.008)	1.146 (0.008)	1.167 (0.010)	1.110 (0.007)	1.783 (0.021)
Sparse 2	30	40	0.8	0.823 (0.008)	1.301 (0.018)	1.353 (0.009)	1.200 (0.009)	1.295 (0.014)	1.098 (0.009)	2.148 (0.030)

A.4 RMSCE estimates

Table A4: Monte Carlo estimates of root mean squared coverage error (RMSCE) of QC range. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

				Estimated RMSCE						
Design	Scenario	$\%CV_{IP}$	ρ	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	0.009	0.009	0.008	0.008	0.008	0.008	0.008
Standard	2	10	0.5	0.012	0.011	0.009	0.009	0.009	0.009	0.009
Standard	3	10	0.8	0.016	0.013	0.011	0.011	0.011	0.012	0.011
Standard	4	40	0.2	0.009	0.009	0.008	0.008	0.008	0.008	0.008
Standard	5	40	0.5	0.012	0.011	0.009	0.010	0.010	0.010	0.009
Standard	6	40	0.8	0.016	0.014	0.012	0.012	0.013	0.013	0.012
Unbalanced 1	7	10	0.2	0.012	0.010	0.008	0.008	0.008	0.008	0.008
Unbalanced 1	8	10	0.5	0.020	0.015	0.010	0.011	0.011	0.010	0.011
Unbalanced 1	9	10	0.8	0.042	0.024	0.016	0.016	0.016	0.018	0.016
Unbalanced 1	10	40	0.2	0.012	0.011	0.009	0.009	0.009	0.008	0.009
Unbalanced 1	11	40	0.5	0.021	0.016	0.011	0.012	0.013	0.012	0.012
Unbalanced 1	12	40	0.8	0.043	0.023	0.016	0.017	0.020	0.020	0.016
Sparse 1	13	10	0.2	0.018	0.016	0.010	0.010	0.010	0.010	0.010
Sparse 1	14	10	0.5	0.031	0.025	0.011	0.011	0.011	0.011	0.011
Sparse 1	15	10	0.8	0.071	0.050	0.018	0.018	0.019	0.025	0.018
Sparse 1	16	40	0.2	0.017	0.015	0.009	0.009	0.010	0.009	0.010
Sparse 1	17	40	0.5	0.030	0.024	0.010	0.011	0.014	0.012	0.010
Sparse 1	18	40	0.8	0.063	0.039	0.015	0.019	0.023	0.024	0.015
Unbalanced 2	19	10	0.2	0.011	0.010	0.009	0.009	0.008	0.008	0.009
Unbalanced 2	20	10	0.5	0.013	0.012	0.010	0.010	0.010	0.010	0.010
Unbalanced 2	21	10	0.8	0.019	0.014	0.012	0.012	0.012	0.013	0.012
Unbalanced 2	22	40	0.2	0.010	0.010	0.009	0.009	0.009	0.009	0.009
Unbalanced 2	23	40	0.5	0.013	0.012	0.010	0.011	0.011	0.010	0.011
Unbalanced 2	24	40	0.8	0.020	0.015	0.012	0.013	0.014	0.014	0.013
Sparse 2	25	10	0.2	0.013	0.012	0.009	0.009	0.009	0.009	0.009
Sparse 2	26	10	0.5	0.031	0.024	0.011	0.011	0.011	0.011	0.011
Sparse 2	27	10	0.8	0.068	0.040	0.013	0.013	0.014	0.019	0.012
Sparse 2	28	40	0.2	0.013	0.012	0.009	0.008	0.009	0.008	0.009
Sparse 2	29	40	0.5	0.027	0.021	0.010	0.011	0.013	0.010	0.011
Sparse 2	30	40	0.8	0.071	0.045	0.017	0.021	0.025	0.026	0.016

A.5 Distribution of observed prediction coverages

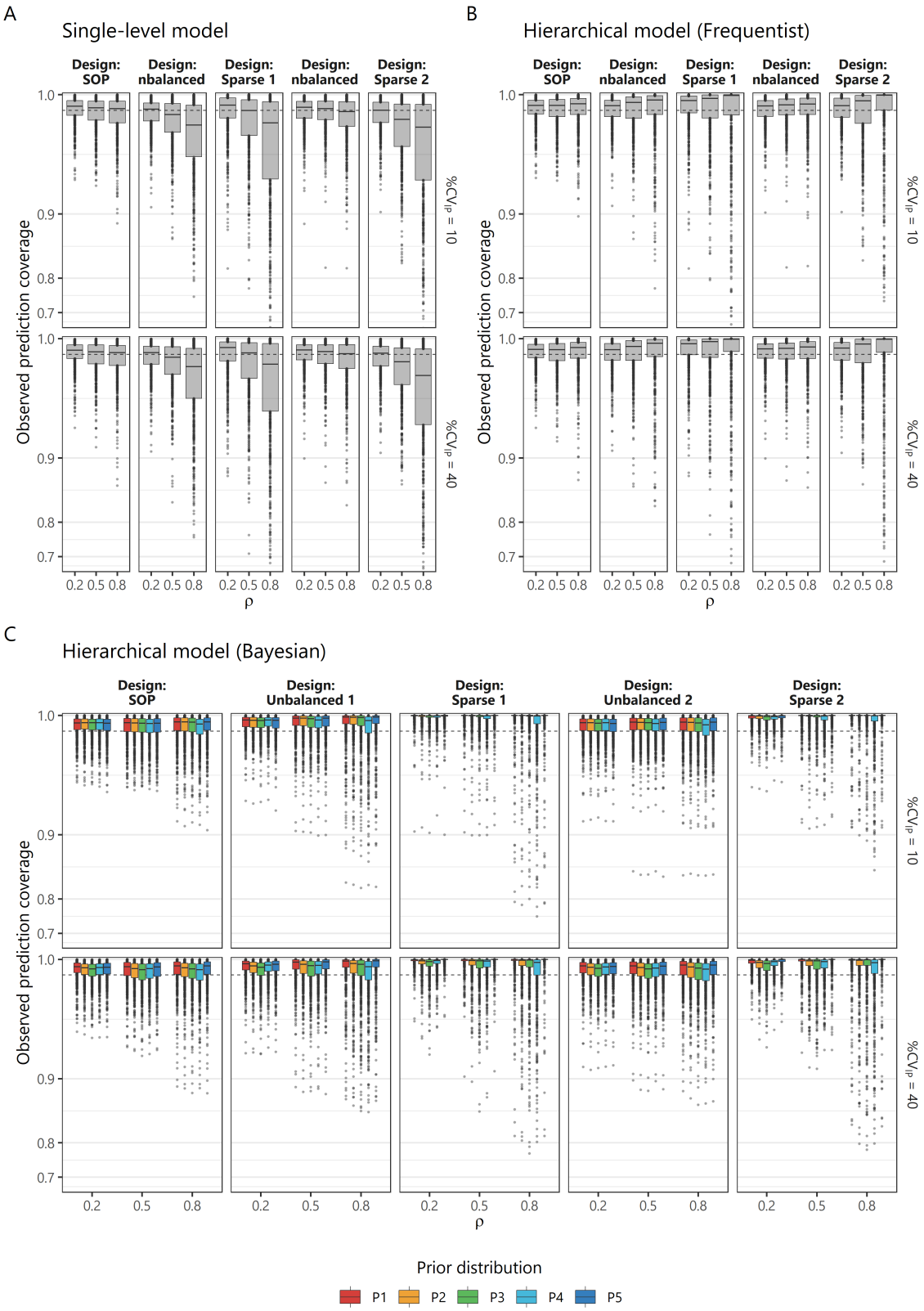

Figure 1: Observed prediction coverages of the QC ranges for all simulation repetitions and all scenarios based on (A) the single-level model, (B) the frequentist hierarchical model, and (C) the Bayesian hierarchical model, with its five different prior distributions. The vertical axis is power scaled for readability.

A.6 Median parameter estimates

Table A5: Median point estimate for parameter μ under different methods, compared to the true parameter value used during data generation. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

					Median point estimate of μ						
Design	Scenario	$\%CV_{IP}$	ρ	True μ	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	**1.998**	1.998	1.998	1.998	1.998	1.998	1.999	1.999
Standard	2	10	0.5	**1.998**	1.998	1.998	1.998	1.998	1.998	1.998	1.998
Standard	3	10	0.8	**1.998**	1.998	1.998	1.998	1.998	1.998	1.998	1.998
Standard	4	40	0.2	**1.968**	1.969	1.969	1.969	1.969	1.969	1.969	1.969
Standard	5	40	0.5	**1.968**	1.968	1.968	1.968	1.968	1.968	1.968	1.968
Standard	6	40	0.8	**1.968**	1.966	1.966	1.967	1.967	1.966	1.967	1.966
Unbalanced 1	7	10	0.2	**1.998**	1.997	1.997	1.998	1.998	1.998	1.998	1.998
Unbalanced 1	8	10	0.5	**1.998**	1.998	1.998	1.998	1.998	1.998	1.998	1.998
Unbalanced 1	9	10	0.8	**1.998**	1.998	1.998	1.998	1.998	1.998	1.998	1.998
Unbalanced 1	10	40	0.2	**1.968**	1.968	1.969	1.970	1.969	1.969	1.969	1.969
Unbalanced 1	11	40	0.5	**1.968**	1.965	1.966	1.966	1.966	1.966	1.967	1.966
Unbalanced 1	12	40	0.8	**1.968**	1.969	1.969	1.969	1.970	1.969	1.969	1.969
Sparse 1	13	10	0.2	**1.998**	1.997	1.997	1.997	1.998	1.997	1.997	1.997
Sparse 1	14	10	0.5	**1.998**	1.999	1.999	1.999	1.999	1.999	1.999	1.999
Sparse 1	15	10	0.8	**1.998**	1.999	1.999	1.999	1.999	1.999	1.999	1.999
Sparse 1	16	40	0.2	**1.968**	1.967	1.967	1.967	1.967	1.967	1.967	1.967
Sparse 1	17	40	0.5	**1.968**	1.970	1.970	1.970	1.970	1.970	1.970	1.969
Sparse 1	18	40	0.8	**1.968**	1.971	1.971	1.971	1.971	1.971	1.971	1.971
Unbalanced 2	19	10	0.2	**1.998**	1.998	1.998	1.998	1.998	1.998	1.998	1.998
Unbalanced 2	20	10	0.5	**1.998**	1.998	1.998	1.998	1.998	1.998	1.998	1.998
Unbalanced 2	21	10	0.8	**1.998**	1.998	1.998	1.998	1.998	1.998	1.998	1.998
Unbalanced 2	22	40	0.2	**1.968**	1.968	1.968	1.968	1.968	1.968	1.967	1.967
Unbalanced 2	23	40	0.5	**1.968**	1.968	1.967	1.967	1.966	1.967	1.966	1.967
Unbalanced 2	24	40	0.8	**1.968**	1.970	1.968	1.968	1.968	1.968	1.968	1.968
Sparse 2	25	10	0.2	**1.998**	1.998	1.998	1.998	1.998	1.998	1.998	1.998
Sparse 2	26	10	0.5	**1.998**	1.997	1.997	1.997	1.997	1.997	1.997	1.997
Sparse 2	27	10	0.8	**1.998**	1.997	1.997	1.997	1.997	1.997	1.997	1.997
Sparse 2	28	40	0.2	**1.968**	1.966	1.966	1.965	1.965	1.965	1.966	1.965
Sparse 2	29	40	0.5	**1.968**	1.963	1.963	1.963	1.964	1.963	1.963	1.963
Sparse 2	30	40	0.8	**1.968**	1.969	1.969	1.969	1.969	1.969	1.970	1.969

Table A6: Median point estimate for parameter σ under different methods, compared to the true parameter value used during data generation. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

					Median point estimate of σ						
Design	Scenario	$\%CV_{IP}$	ρ	True σ	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	**0.0387**	-	0.0379	0.0394	0.0395	0.0394	0.0381	0.0386
Standard	2	10	0.5	**0.0306**	-	0.0297	0.0315	0.0315	0.0315	0.0310	0.0307
Standard	3	10	0.8	**0.0194**	-	0.0191	0.0200	0.0200	0.0200	0.0203	0.0195
Standard	4	40	0.2	**0.1496**	-	0.1463	0.1522	0.1507	0.1508	0.1455	0.1486
Standard	5	40	0.5	**0.1183**	-	0.1169	0.1242	0.1235	0.1231	0.1210	0.1208
Standard	6	40	0.8	**0.0748**	-	0.0743	0.0779	0.0778	0.0771	0.0787	0.0758
Unbalanced 1	7	10	0.2	**0.0387**	-	0.0381	0.0392	0.0392	0.0391	0.0384	0.0386
Unbalanced 1	8	10	0.5	**0.0306**	-	0.0302	0.0312	0.0312	0.0312	0.0309	0.0306
Unbalanced 1	9	10	0.8	**0.0194**	-	0.0191	0.0197	0.0197	0.0197	0.0198	0.0193
Unbalanced 1	10	40	0.2	**0.1496**	-	0.1472	0.1513	0.1498	0.1493	0.1470	0.1489
Unbalanced 1	11	40	0.5	**0.1183**	-	0.1176	0.1215	0.1207	0.1201	0.1199	0.1194
Unbalanced 1	12	40	0.8	**0.0748**	-	0.0730	0.0754	0.0752	0.0748	0.0757	0.0740
Sparse 1	13	10	0.2	**0.0387**	-	0.0373	0.0395	0.0395	0.0394	0.0385	0.0382
Sparse 1	14	10	0.5	**0.0306**	-	0.0299	0.0320	0.0320	0.0318	0.0315	0.0309
Sparse 1	15	10	0.8	**0.0194**	-	0.0189	0.0201	0.0201	0.0200	0.0202	0.0193
Sparse 1	16	40	0.2	**0.1496**	-	0.1450	0.1541	0.1508	0.1494	0.1465	0.1486
Sparse 1	17	40	0.5	**0.1183**	-	0.1143	0.1223	0.1206	0.1188	0.1184	0.1175
Sparse 1	18	40	0.8	**0.0748**	-	0.0741	0.0788	0.0783	0.0774	0.0791	0.0761
Unbalanced 2	19	10	0.2	**0.0387**	-	0.0377	0.0390	0.0390	0.0390	0.0377	0.0381
Unbalanced 2	20	10	0.5	**0.0306**	-	0.0300	0.0317	0.0317	0.0316	0.0312	0.0309
Unbalanced 2	21	10	0.8	**0.0194**	-	0.0191	0.0200	0.0200	0.0200	0.0203	0.0195
Unbalanced 2	22	40	0.2	**0.1496**	-	0.1467	0.1518	0.1504	0.1504	0.1459	0.1484
Unbalanced 2	23	40	0.5	**0.1183**	-	0.1184	0.1247	0.1239	0.1236	0.1218	0.1213
Unbalanced 2	24	40	0.8	**0.0748**	-	0.0742	0.0782	0.0780	0.0773	0.0789	0.0760
Sparse 2	25	10	0.2	**0.0387**	-	0.0383	0.0393	0.0393	0.0392	0.0388	0.0388
Sparse 2	26	10	0.5	**0.0306**	-	0.0302	0.0310	0.0310	0.0310	0.0308	0.0305
Sparse 2	27	10	0.8	**0.0194**	-	0.0192	0.0197	0.0197	0.0197	0.0198	0.0194
Sparse 2	28	40	0.2	**0.1496**	-	0.1479	0.1519	0.1505	0.1498	0.1486	0.1497
Sparse 2	29	40	0.5	**0.1183**	-	0.1166	0.1198	0.1192	0.1184	0.1184	0.1181
Sparse 2	30	40	0.8	**0.0748**	-	0.0742	0.0762	0.0759	0.0755	0.0763	0.0751

Table A7: Median point estimate for parameter τ under different methods, compared to the true parameter value used during data generation. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

					Median point estimate of τ						
Design	Scenario	$\%CV_{IP}$	ρ	True τ	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	**0.0194**	-	0.0191	0.0180	0.0179	0.0178	0.0217	0.0189
Standard	2	10	0.5	**0.0306**	-	0.0297	0.0295	0.0294	0.0292	0.0295	0.0301
Standard	3	10	0.8	**0.0387**	-	0.0380	0.0395	0.0393	0.0391	0.0378	0.0395
Standard	4	40	0.2	**0.0748**	-	0.0701	0.0676	0.0658	0.0593	0.0813	0.0712
Standard	5	40	0.5	**0.1183**	-	0.1138	0.1130	0.1101	0.1058	0.1118	0.1152
Standard	6	40	0.8	**0.1496**	-	0.1456	0.1508	0.1469	0.1445	0.1420	0.1516
Unbalanced 1	7	10	0.2	**0.0194**	-	0.0164	0.0181	0.0181	0.0178	0.0215	0.0185
Unbalanced 1	8	10	0.5	**0.0306**	-	0.0286	0.0316	0.0314	0.0310	0.0305	0.0318
Unbalanced 1	9	10	0.8	**0.0387**	-	0.0366	0.0405	0.0402	0.0396	0.0372	0.0405
Unbalanced 1	10	40	0.2	**0.0748**	-	0.0647	0.0715	0.0677	0.0619	0.0814	0.0734
Unbalanced 1	11	40	0.5	**0.1183**	-	0.1125	0.1229	0.1158	0.1100	0.1153	0.1244
Unbalanced 1	12	40	0.8	**0.1496**	-	0.1439	0.1591	0.1502	0.1460	0.1410	0.1597
Sparse 1	13	10	0.2	**0.0194**	-	0.0160	0.0224	0.0220	0.0211	0.0255	0.0229
Sparse 1	14	10	0.5	**0.0306**	-	0.0271	0.0343	0.0335	0.0323	0.0311	0.0348
Sparse 1	15	10	0.8	**0.0387**	-	0.0340	0.0439	0.0428	0.0411	0.0357	0.0441
Sparse 1	16	40	0.2	**0.0748**	-	0.0656	0.0868	0.0754	0.0650	0.0929	0.0910
Sparse 1	17	40	0.5	**0.1183**	-	0.1090	0.1326	0.1162	0.1064	0.1151	0.1397
Sparse 1	18	40	0.8	**0.1496**	-	0.1354	0.1674	0.1463	0.1399	0.1315	0.1762
Unbalanced 2	19	10	0.2	**0.0194**	-	0.0184	0.0181	0.0181	0.0180	0.0218	0.0191
Unbalanced 2	20	10	0.5	**0.0306**	-	0.0300	0.0300	0.0300	0.0299	0.0302	0.0307
Unbalanced 2	21	10	0.8	**0.0387**	-	0.0375	0.0388	0.0386	0.0384	0.0372	0.0390
Unbalanced 2	22	40	0.2	**0.0748**	-	0.0725	0.0708	0.0682	0.0626	0.0828	0.0744
Unbalanced 2	23	40	0.5	**0.1183**	-	0.1153	0.1153	0.1121	0.1083	0.1133	0.1179
Unbalanced 2	24	40	0.8	**0.1496**	-	0.1481	0.1533	0.1493	0.1469	0.1439	0.1541
Sparse 2	25	10	0.2	**0.0194**	-	0.0164	0.0211	0.0209	0.0204	0.0234	0.0212
Sparse 2	26	10	0.5	**0.0306**	-	0.0275	0.0359	0.0352	0.0340	0.0320	0.0359
Sparse 2	27	10	0.8	**0.0387**	-	0.0367	0.0479	0.0466	0.0446	0.0388	0.0478
Sparse 2	28	40	0.2	**0.0748**	-	0.0640	0.0812	0.0732	0.0650	0.0861	0.0825
Sparse 2	29	40	0.5	**0.1183**	-	0.1087	0.1377	0.1221	0.1142	0.1178	0.1414
Sparse 2	30	40	0.8	**0.1496**	-	0.1393	0.1731	0.1519	0.1455	0.1368	0.1819

Table A8: Median point estimate for parameter $\%CV_{IP}$ under different methods, compared to the true parameter value used during data generation. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

					Median point estimate of $\%CV_{IP}$						
Design	Scenario	$\%CV_{IP}$	ρ	True $\%CV_{IP}$	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	**10.0**	9.9	9.9	10.4	10.4	10.4	10.4	10.4
Standard	2	10	0.5	**10.0**	9.7	9.8	10.3	10.3	10.2	10.2	10.2
Standard	3	10	0.8	**10.0**	9.7	9.9	10.3	10.3	10.3	10.0	10.3
Standard	4	40	0.2	**40.0**	39.5	39.6	41.9	41.2	40.6	41.2	41.4
Standard	5	40	0.5	**40.0**	39.1	39.4	41.5	40.8	40.1	40.5	41.2
Standard	6	40	0.8	**40.0**	38.9	39.3	41.3	40.4	39.8	39.4	41.2
Unbalanced 1	7	10	0.2	**10.0**	9.7	9.8	10.5	10.5	10.5	10.6	10.4
Unbalanced 1	8	10	0.5	**10.0**	9.3	9.7	10.5	10.5	10.4	10.3	10.5
Unbalanced 1	9	10	0.8	**10.0**	8.9	9.5	10.5	10.4	10.3	9.8	10.4
Unbalanced 1	10	40	0.2	**40.0**	38.8	39.2	42.3	41.4	40.5	41.7	41.9
Unbalanced 1	11	40	0.5	**40.0**	37.6	39.0	42.5	41.1	40.0	40.7	42.2
Unbalanced 1	12	40	0.8	**40.0**	35.7	38.4	42.4	40.3	39.2	38.2	42.3
Sparse 1	13	10	0.2	**10.0**	9.5	9.7	11.4	11.3	11.2	11.2	11.2
Sparse 1	14	10	0.5	**10.0**	9.1	9.4	11.4	11.3	11.0	10.7	11.3
Sparse 1	15	10	0.8	**10.0**	8.4	8.9	11.3	11.1	10.7	9.6	11.2
Sparse 1	16	40	0.2	**40.0**	38.8	39.6	46.7	43.9	42.3	44.1	46.1
Sparse 1	17	40	0.5	**40.0**	37.2	38.5	46.3	42.6	41.0	41.9	46.8
Sparse 1	18	40	0.8	**40.0**	34.6	37.2	45.6	40.7	39.1	37.8	47.2
Unbalanced 2	19	10	0.2	**10.0**	9.8	9.8	10.4	10.4	10.3	10.3	10.3
Unbalanced 2	20	10	0.5	**10.0**	9.7	9.8	10.3	10.3	10.3	10.2	10.3
Unbalanced 2	21	10	0.8	**10.0**	9.5	9.7	10.2	10.2	10.1	9.9	10.2
Unbalanced 2	22	40	0.2	**40.0**	39.6	39.9	42.2	41.4	40.8	41.5	41.8
Unbalanced 2	23	40	0.5	**40.0**	39.2	39.7	41.8	41.0	40.3	40.6	41.5
Unbalanced 2	24	40	0.8	**40.0**	38.8	39.7	41.6	40.6	40.0	39.7	41.6
Sparse 2	25	10	0.2	**10.0**	9.6	9.7	10.9	10.8	10.7	10.8	10.8
Sparse 2	26	10	0.5	**10.0**	9.1	9.5	11.2	11.0	10.8	10.5	11.1
Sparse 2	27	10	0.8	**10.0**	8.9	9.6	12.0	11.8	11.4	10.1	12.0
Sparse 2	28	40	0.2	**40.0**	38.6	39.5	44.4	42.6	41.5	43.0	44.1
Sparse 2	29	40	0.5	**40.0**	36.5	38.1	44.4	41.5	40.1	40.8	45.1
Sparse 2	30	40	0.8	**40.0**	34.5	37.5	45.4	40.7	39.3	37.3	47.7

Table A9: Median point estimate for parameter ρ under different methods, compared to the true parameter value used during data generation. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

					Median point estimate of ρ						
Design	Scenario	$\%CV_{IP}$	ρ	True ρ	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	**0.20**	-	0.21	0.17	0.17	0.17	0.25	0.19
Standard	2	10	0.5	**0.50**	-	0.51	0.48	0.48	0.48	0.49	0.50
Standard	3	10	0.8	**0.80**	-	0.80	0.80	0.80	0.80	0.78	0.81
Standard	4	40	0.2	**0.20**	-	0.19	0.16	0.15	0.13	0.23	0.18
Standard	5	40	0.5	**0.50**	-	0.49	0.46	0.45	0.43	0.46	0.49
Standard	6	40	0.8	**0.80**	-	0.79	0.79	0.79	0.78	0.77	0.80
Unbalanced 1	7	10	0.2	**0.20**	-	0.16	0.18	0.18	0.18	0.24	0.19
Unbalanced 1	8	10	0.5	**0.50**	-	0.48	0.51	0.51	0.51	0.50	0.52
Unbalanced 1	9	10	0.8	**0.80**	-	0.79	0.81	0.81	0.80	0.78	0.82
Unbalanced 1	10	40	0.2	**0.20**	-	0.17	0.18	0.17	0.14	0.23	0.19
Unbalanced 1	11	40	0.5	**0.50**	-	0.49	0.52	0.49	0.47	0.49	0.53
Unbalanced 1	12	40	0.8	**0.80**	-	0.79	0.82	0.80	0.79	0.77	0.82
Sparse 1	13	10	0.2	**0.20**	-	0.16	0.23	0.23	0.22	0.30	0.25
Sparse 1	14	10	0.5	**0.50**	-	0.46	0.54	0.53	0.51	0.50	0.57
Sparse 1	15	10	0.8	**0.80**	-	0.76	0.83	0.82	0.81	0.76	0.84
Sparse 1	16	40	0.2	**0.20**	-	0.17	0.24	0.20	0.15	0.28	0.27
Sparse 1	17	40	0.5	**0.50**	-	0.48	0.55	0.49	0.46	0.49	0.60
Sparse 1	18	40	0.8	**0.80**	-	0.77	0.82	0.78	0.77	0.74	0.85
Unbalanced 2	19	10	0.2	**0.20**	-	0.20	0.18	0.18	0.17	0.25	0.20
Unbalanced 2	20	10	0.5	**0.50**	-	0.50	0.48	0.48	0.48	0.48	0.50
Unbalanced 2	21	10	0.8	**0.80**	-	0.79	0.79	0.79	0.79	0.77	0.80
Unbalanced 2	22	40	0.2	**0.20**	-	0.20	0.18	0.17	0.15	0.25	0.20
Unbalanced 2	23	40	0.5	**0.50**	-	0.49	0.47	0.46	0.44	0.47	0.50
Unbalanced 2	24	40	0.8	**0.80**	-	0.79	0.79	0.78	0.78	0.77	0.80
Sparse 2	25	10	0.2	**0.20**	-	0.16	0.23	0.22	0.22	0.28	0.24
Sparse 2	26	10	0.5	**0.50**	-	0.44	0.56	0.55	0.54	0.51	0.57
Sparse 2	27	10	0.8	**0.80**	-	0.78	0.85	0.85	0.83	0.79	0.86
Sparse 2	28	40	0.2	**0.20**	-	0.16	0.22	0.19	0.16	0.25	0.24
Sparse 2	29	40	0.5	**0.50**	-	0.47	0.57	0.51	0.49	0.50	0.60
Sparse 2	30	40	0.8	**0.80**	-	0.78	0.83	0.80	0.79	0.76	0.85

A.7 Correlation between variance components

Table A10: Median point estimate for correlation between variance components' estimators (for the FHM), or for the correlation between variance components' posteriors (for the BHM). under different methods. SLM is the single-level model; FHM is the frequentist hierarchical model; BHM is the Bayesian hierarchical model; P1 to P5 refer to the different prior distributions for variance components.

				Median point estimate of correlation between estimators/posteriors of σ^2 and τ^2						
Design	Scenario	$\%CV_{IP}$	ρ	SLM	FHM	BHM (P1)	BHM (P2)	BHM (P3)	BHM (P4)	BHM (P5)
Standard	1	10	0.2	-	-0.453	-0.298	-0.292	-0.301	-0.210	-0.294
Standard	2	10	0.5	-	-0.302	-0.325	-0.328	-0.333	-0.240	-0.303
Standard	3	10	0.8	-	-0.106	-0.119	-0.121	-0.123	-0.112	-0.108
Standard	4	40	0.2	-	-0.453	-0.282	-0.297	-0.307	-0.211	-0.282
Standard	5	40	0.5	-	-0.310	-0.326	-0.348	-0.372	-0.250	-0.306
Standard	6	40	0.8	-	-0.111	-0.125	-0.139	-0.144	-0.124	-0.112
Unbalanced 1	7	10	0.2	-	-0.210	-0.091	-0.094	-0.098	-0.028	-0.087
Unbalanced 1	8	10	0.5	-	-0.132	-0.083	-0.086	-0.092	-0.042	-0.077
Unbalanced 1	9	10	0.8	-	-0.045	-0.028	-0.031	-0.033	-0.016	-0.026
Unbalanced 1	10	40	0.2	-	-0.215	-0.094	-0.115	-0.126	-0.032	-0.084
Unbalanced 1	11	40	0.5	-	-0.121	-0.087	-0.104	-0.115	-0.047	-0.071
Unbalanced 1	12	40	0.8	-	-0.045	-0.034	-0.041	-0.043	-0.022	-0.027
Sparse 1	13	10	0.2	-	-0.146	-0.006	-0.015	-0.018	0.062	-0.003
Sparse 1	14	10	0.5	-	-0.105	-0.025	-0.033	-0.037	0.023	-0.010
Sparse 1	15	10	0.8	-	-0.037	-0.015	-0.018	-0.018	0.014	-0.006
Sparse 1	16	40	0.2	-	-0.173	-0.039	-0.059	-0.050	0.046	-0.004
Sparse 1	17	40	0.5	-	-0.099	-0.070	-0.087	-0.070	0.006	-0.009
Sparse 1	18	40	0.8	-	-0.035	-0.037	-0.043	-0.033	0.007	-0.006
Unbalanced 2	19	10	0.2	-	-0.408	-0.261	-0.262	-0.267	-0.182	-0.258
Unbalanced 2	20	10	0.5	-	-0.304	-0.288	-0.286	-0.292	-0.210	-0.271
Unbalanced 2	21	10	0.8	-	-0.128	-0.131	-0.133	-0.135	-0.123	-0.122
Unbalanced 2	22	40	0.2	-	-0.424	-0.262	-0.280	-0.295	-0.190	-0.259
Unbalanced 2	23	40	0.5	-	-0.309	-0.294	-0.313	-0.335	-0.226	-0.277
Unbalanced 2	24	40	0.8	-	-0.125	-0.128	-0.142	-0.151	-0.131	-0.117
Sparse 2	25	10	0.2	-	-0.097	-0.013	-0.018	-0.021	0.028	-0.006
Sparse 2	26	10	0.5	-	-0.044	-0.012	-0.015	-0.017	0.020	-0.005
Sparse 2	27	10	0.8	-	-0.011	-0.003	-0.004	-0.005	0.015	-0.002
Sparse 2	28	40	0.2	-	-0.101	-0.038	-0.049	-0.044	0.023	-0.007
Sparse 2	29	40	0.5	-	-0.041	-0.027	-0.033	-0.029	0.018	-0.005
Sparse 2	30	40	0.8	-	-0.012	-0.010	-0.012	-0.009	0.017	-0.002

A.8 Software code

The software code for both R and Stan is provided in a separate zip file.