Science of the Total Environment 964 (2025) 178573

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Comparison of statistical evaluation approaches for log-removal validation according to European water reuse regulations

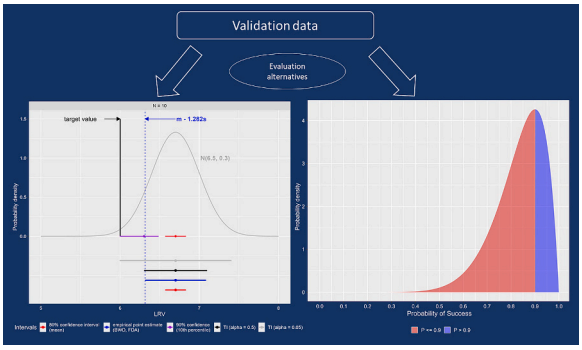
Wolfgang Seis [a,b,*], Michael Stapf [a], Ulf Miehe [a], Thomas Wintgens [c]

[a] Kompetenzzentrum Wasser Berlin gGmbH, Grunewaldstraße 60/61, 10825 Berlin, Germany
[b] Department of Water Management, Faculty of Civil Engineering & Geosciences, Delft University of Technology, 2600 GA Delft, the Netherlands
[c] Institute of Environmental Engineering, RWTH Aachen University, Mies-van-der-Rohe Strasse 1, 52066 Aachen, Germany

HIGHLIGHTS

- Statistical tolerance intervals (TI) used for log_{10}-removal validation in water reuse
- TI allow for robust system validation at lower sample sizes at large effect-sizes
- TI may be more resource-effective as minimum number of samples not required
- TI considered more flexible in comparison to existing percentile-based approaches
- Binomial approach shown to require large sample sizes to be statistically valid

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Ouyang Wei

Keywords:
Water reuse
Process validation
Tolerance intervals

ABSTRACT

In 2020, the European Union published ordinance EU 2020/741, establishing minimum requirements for water reuse in agriculture. The ordinance differentiates between several water quality classes. For the highest water quality class (Class A), the ordinance mandates analytical validation of the treatment performance of new water reuse treatment plants (WRTP) related to the removal of microbial indicators for viral, bacterial, and parasitic pathogens. While the ordinance clearly defines the numeric target values for the required log_{10}-reduction values (LRV), it provides limited to no guidance on the necessary sample sizes and statistical evaluation approaches. The main requirement is that at least 90 % of the validation samples should meet the requirements. However, the interpretation of this 90 % validation target can significantly impact the required sample size, efforts necessary, and the risk of misclassifying WRTPs in practice.

The present study compares different statistical evaluation approaches that might be considered applicable for LRV validation monitoring. Special emphasis is placed on the use of tolerance intervals, which combine percentile estimations with sample size-based uncertainty and confidence regions. Tolerance interval-based approaches are compared with alternative methods, including a) a binomial evaluation and b) the calculation of empirical percentiles. The latter are already used in existing European and U.S. regulations for bathing water and irrigation water quality.

* Corresponding author at: Kompetenzzentrum Wasser Berlin gGmbH, Grunewaldstraße 60/61, 10825 Berlin, Germany.
E-mail address: wolfgang.seis@kompetenz-wasser.de (W. Seis).

https://doi.org/10.1016/j.scitotenv.2025.178573
Received 10 September 2024; Received in revised form 12 December 2024; Accepted 16 January 2025
Available online 23 January 2025

Our study demonstrates that using tolerance intervals allows for the reliable validation of WRTPs that achieve high LRVs relative to regulatory targets with comparatively smaller sample sizes compared to the other two approaches, while reducing the risk of misclassification. Additionally, we show that simplified approaches, such as a "9 out of 10" approach, pose a substantial risk of misclassification and should not be applied. We illustrate the behavior of these different approaches through simulation experiments and application to real data collected in 2022 and 2023 at a large WRTP in Germany.

1. Introduction

In 2020, the European Union published Regulation EU 2020/741 on the minimal requirements for water reuse in agricultural irrigation (EU, 2020). The Regulation aims at ensuring microbial safety by formulating requirements for water reuse in agriculture as a combination of complementing measures, which are specified in Annex I of the Regulation. These measures include a.) the definition of specific water quality classes (Class A – Class D), which subsequently restrict irrigation to specific agricultural products, b.) the definition of minimal treatment technology standards, which have to be implemented specific to each water quality class, c.) a set of routine monitoring parameters, which have to be regularly or continuously monitored during operation, and d.) the validation of the treatment performance of the system regarding the removal of indicator organisms for viral, bacterial and parasitic pathogens. The latter is necessary only for the highest water quality class (Class A), and should be considered a short, but intensive monitoring endeavor, which must be conducted before an envisaged water reuse treatment plant (WRTP) starts operating, or after major changes to the WRTP have been implemented. Performance targets for microbial parameters are commonly defined as $\log_{10}$-reduction values (LRV) and are applied in drinking water quality management (Teunis et al., 2009), meta-analyses for assessing the performance of single treatment technologies (Branch et al., 2021), and water reuse applications (Verbyla et al., 2023). By demanding site specific validation of LRVs (Table 1), the EU Regulation 2020/741 follows existing and proposed regulatory frameworks in the field of water reuse (Verbyla et al., 2023). For the U. S., Verbyla et al. (2023) underlined that standard practices for local validation studies are still lacking, and elaborated on recommendations for conducting validation monitoring procedures. Also, in Europe, the published Regulation does not specify or reference any methodological approach for system validation, beyond the mere definition of numerical target values for the required LRVs (Table 1) and the statement that *"at least 90% of the validation samples should fulfill the requirements"*. No information is provided regarding data requirements, sample sizes or which statistical approach to apply. However, such choices may have relevant effects on the efforts necessary to perform validation monitoring as well as on the reliability and robustness of the obtained results. From a statistical perspective it is crucial to avoid both Type I (false positive) and Type II (false negative) errors, meaning to avoid situations in which a WRTP is declared to be able to meet performance targets, while in reality it is not, and vice versa. In this context, Verbyla et al. (2023) underlined the importance of accounting for the existing parameter uncertainty of the target parameter used for validation, in their case a 5th percentile. The authors highlight that the obtained precision of the estimated percentile not only depends on the sample size but also, as they mention, on the mean and standard deviation of LRVs. However, the effect of sample size, mean and variance on the uncertainty of the estimated parameter is not summarized in a systematic manner. Furthermore, confidence intervals of a 5th percentile are only assessed for the case of normally distributed LRV data. For the non-normal case reference is made to non-parametric approaches, which, however, demand large sample sizes to validate a 5th percentile with sufficient confidence. Alternative parametric approaches which are able to cope with non-normality not considered.

In a European setting, even the statistical target parameter for validation is not well-defined. Against this background the present study elaborates on and compares different statistical evaluation approaches for validating a 90 % threshold, and illustrates its effect on the final validation results. It does so by a.), introducing each method theoretically, b.) illustrating its relevant characteristics by means of statistical simulation experiments, and c.) applying and evaluating each approach to a set of real-word data collected at a large scale WRTP.

Our study includes methods, which are currently used in existing European and international regulations for bathing water and irrigation water quality for estimating high percentiles and complements it with alternative approaches, specifically with the use of a binomial evaluation approach and the use of statistical tolerance intervals. Tolerance intervals are rarely used but may provide beneficial properties to the validation process, by reducing both the risk of type I and type II errors in the validation process.

Table 1
Overview of LRV targets necessary for treatment validation.

Indicator	$\log_{10}$-removal target
E. coli	>5.0
Coliphages/f-specific, Coliphages/somatic Coliphages/total	>6.0
Clostridium perfringens-spores/spore-forming sulfate-reducing bacteria	>4.0 (*Clostridium perfringens*-Spores) >5.0 (spore-forming sulfate-reducing bacteria)

2. Methodology

The present study presents different candidate approaches for validating LRV performance targets based on a 90 % threshold (Section 2.3). It illustrates the behavior of the various approaches both by simulation as well as by applying them to real-world data collected during 2022 and 2023 (Section 2.1).

2.1. Data collection

During May 2022 and August 2023, a total of 24 paired influent and effluent samples were collected at a large wastewater treatment plant in Germany. The WRTP consisted of a full scale activated sludge treatment, followed by additional ozonation, rapid sand filtration and UV disinfection at pilot scale. During the operation period ozone dosage was varied in two phases to achieve control targets defined by % removal of UVT_{254} (Phase 1: Δ UVT = 34 %, Phase 2: Δ UVT = 47 %). For the study, samples collected as 24-composite samples using refrigerated automated samplers in the influent of the WRTP and in the effluent of the UV disinfection were used. Samples were analyzed for *E. coli*, somatic

Table 2
Analytic methods used for analysis.

Parameter	Method	Sample size during operating conditions
Clostridium perfringens spores	DIN ISO 141189:2016	10 paired samples at Δ UVT = 34 %,
E. coli	DIN ISO 9308-2	14 paired samples at Δ UVT = 47 %.

coliphages and *Clostridium perfringens* spp. using standard laboratory methods (Table 2). For the analysis of effluent data, we ensured that the lower limit of quantification (LOQ) was set to <1 per unit volume, which eases the statistical evaluation, as discrete observations <1 per unit volume can be set to zero.

2.2. Calculation of log_{10}-reduction values (LRV)

The log_{10}-reduction is not directly measurable but is a function of the measured influent and effluent concentration. It is calculated by:

$$log_{10} - reduction = f(c(influent), c(effluent)) = log_{10}\frac{c(influent)}{c(effluent)} \quad (3\text{-}1)$$

A common point of discussion is the handling of values below the LOQ in the effluent of the WRTP. Common approaches for analyzing microbiological data below the LOQ include the substitution with the numerical value of the LOQ, half of LOQ, or to model them as censored data (Chik et al., 2018), which are common approaches also for chemical parameters. However, microbial data are discrete in character. Thus, if results are reported as <1 per analyzed volume, the results can be interpreted as 0 per analyzed volume. However, an observation of 0 does not necessarily guarantee that the true underlying average particle concentration is truly <1, as an observation of 0 can also have been caused by random sampling error. Thus, the data are also not truly censored (Chik et al., 2018). In the present study, we applied analytical methods with a lower detection limit of <1 per unit volume for all analyzed microbial indicators in the effluent. Values below the LOQ, reported as "< 1" were set to a value of 0. Since 0 cannot be used in the denominator of Eq. (3-1), we calculate the validated LRV ($LRV_{validated}$), which represents an observed removal limited by the concentration measured in the influent of the WRTP according to (3–2). While we are aware that removing the denominator leads to the same result as using the lower LOQ of 1 in Eq. (3-1), we prefer the phenomenological rational over the purely mathematical or conventional justification. Moreover, setting effluent concentration below the LOQ to 0 has the advantage that no further assumptions are necessary when fitting a negative-binomial distribution to the data (see Section 2.3.3.2) because the distribution covers a value of 0, in contrast to e.g. a lognormal distribution.

$$LRV_{validated} = log_{10}(c(influent)) \quad (3\text{-}2)$$

2.3. Statistical evaluation approaches for LRV validation

According to Regulation EU 2020/741 the validation task requires validating that *"at least 90% of the validation samples should fulfill the requirements"*. Since no further information is provided these "90 %" may be subject to alternative interpretations.

The present section describes different alternatives to interpreting these "90 %" and evaluates how they may influence process validation. The approaches differ in the target parameter (success rate vs. percentile) as well as in whether the collected influent and effluent data is evaluated under a "paired, correlated" or a "unpaired, independent" assumption. While the paired evaluation pairs influent and effluent samples by date and calculates pairwise LRV values, the unpaired approach fits distributions to the influent and effluent data, and uses Monte Carlo Simulation to generate distributions of LRV based on random sampling from the influent and effluent distributions.

2.3.1. Classical and Bayesian approaches to statistics

In the present study *classical* or *frequentist* approaches to statistics are presented next to Bayesian methods. Both approaches fundamentally differ in how probability is interpreted. While classical statistics interprets probability as a result of repeated events, like rolling a dice or flipping a coin, Bayesian statistics uses probability to express one's uncertainty regarding certain quantities of interest. That distinction leads to different interpretation about how classical confidence intervals or Bayesian credible intervals are interpreted. A classical confidence interval describes a procedure which if applied to a hypothetically repeated experiment generates intervals, out of which a certain proportion contains the true value in the long run (Morey et al., 2016). It does not make any statement about the probability that any single interval includes the true value or not. Thus, if the challenge is to construct an interval which contains the true value with a certain probability given the data and further external (prior) information, a Bayesian approach is required, as only Bayesian credible intervals are interpreted that way. Therefore, Bayesian methods are considered the preferred choice. However, for standard normal and binomial distributions Bayesian methods lead to very similar credible limits as classical confidence intervals if flat prior distributions are used. Thus, results from classical approaches can be interpreted as resulting from Bayesian methods with flat priors, thus allowing for a Bayesian interpretation. While we generally favor Bayesian techniques, we included also classical calculation approaches as they are a.) often easier to compute, partly allowing for an analytical solution (at least for simple applications), and b.) are more commonly applied and may be more applicable in practice.

2.3.2. Binomial evaluation of the success rate of calculated LRV

The binomial model is based on the estimation of the success rate p as a target parameter. The approach is included into this study since it might be considered the most intuitive evaluation method, given the question of validating a rate parameter, like the required rate of 90 %. For example, the former European bathing water directive from 1976 (76/160/EEC) assessed water quality compliance based on a binomial approach (EEC, 1976). For LRV validation the binomial evaluation approach is based on LRV values which are calculated from paired influent and effluent samples. If the calculated LRV is above the target value the observation is labelled as "success" otherwise as "failure". The point estimate for the most likely success rate p is calculated by:

$$p = \frac{N(success)}{N(success) + N(failure)} \quad (3\text{-}3)$$

For validation that $p \geq 0.9$, the null-hypothesis H_0 which has to be rejected is set to:

H_0: $p < 0.9$ (the WRTP is not able to meet the requirements)
H (alternative): $p \geq 0.9$ (the WRTP is able to meet the requirements)

In the present study we apply a required confidence level of 95 %, which accepts a type I error rate of 5 %. While confidence levels are arbitrary to a certainty degree, we assume a level 95 % to be widely accepted in the scientific literature. Thus, we calculate 1-sided lower confidence intervals (CI), both by applying a Wilson classical approach (3–4), as well as by simulating a Bayesian credible interval based on a Beta-Binomial model (3–5), with a uniform prior distribution, parameterized as a Beta-distribution with $\alpha = \beta = 1$.

$$CI = \frac{1}{1+\frac{z^2}{n}}\left(p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}\right) \quad (3\text{-}4)$$

$$p(p\,|data) \sim Beta(\alpha + n(success), \beta + n(failure)) \quad (3\text{-}5)$$

with:

CI: The confidence interval for the population proportion p
p: the sample proportion
n: the sample size, i.e., the total number of trials, successes or failures.
z: z-score from the standard normal distribution corresponding to the desired confidence level, here 1.645 for the 1-sided lower confidence limit.

2.3.3. Approaches based on percentile evaluations of obtained LRV values

As a second set of evaluation approaches, methods are considered which are derived from the distribution of numerical LRV values. If 90 % of the evaluated LRV values should be above a pre-defined performance target it essentially means that the 10th percentile of the distribution of LRVs should be located above the performance target. Thus, the 10th percentile is considered the target parameter for validation. High percentiles are also used in the U.S. for LRV validation (Verbyla et al., 2023).

Percentiles can be calculated in multiple ways (Hunter, 2002). The current European Bathing Water directive (2006/7/EC, 2006), uses parametric 90th and 95th percentiles of an assumed lognormal distribution of FIB data (see Section 2.3.3.1) for long-term bathing water quality assessment. The same parametric approach is used by the US-FDA to assess the microbial quality of irrigation water (FDA, 2013). As these approaches are already used and accepted in practice, we included them into the present study. One of the major drawbacks regarding this approach is that it does not explicitly account for sample-size based uncertainty but only calculates the point estimate of the target percentiles from the sample at hand. No level of confidence is explicitly provided, and no generalization from sample to population is explicitly considered. Thus, a hypothetical null-hypothesis, that the true 10th percentile exceeds the defined target value, is rejected if the point estimate of the 10th percentile calculated from the sample at hand exceeds the pre-defined target. In order to ensure a certain degree of precision both regulations define minimum number of samples to be collected (12–16 for the European Bathing Water Directive, 20 for the US-FDA).

Eventually, we included two additional approaches based on the calculation of tolerance intervals (see Section 2.3.3.2) *TI (P, α)*. In contrast to the approach of the European Bathing Water Directive tolerance intervals also account for sample-size-based uncertainty, and allow for adding a desired level of confidence (1-α), regarding the estimate above which a certain proportion (P) of the population distribution is expected.

2.3.3.1. Calculation of point estimates for the 10th percentile. The 10th percentile is an estimator which, if perfectly known, describes a value above which 90 % of the values of a population distribution will fall. The parametric 10th percentile can be calculated by:

$$10th\ percentile = m - 1.282\ s \tag{3-6}$$

where m represents the sample mean and s the sample standard deviation. 1.282 represents the interval factor k, which equals the z-score of the standard normal distribution for the quantile of interest (here 1.282 for the 10th percentile).

2.3.3.2. Calculation of tolerance intervals based on paired and unpaired evaluations. According to the National Institute for Standards and Technology (NIST), tolerance intervals describe an interval which cover a certain proportion of the population with a stated level of confidence α, and allow for answering the question about what interval guarantees with a certain level of confidence that p percent of the population will not fall below a certain lower limit (Heckert et al. (2002), chapter 7.2.6.3). The answer to this question leads to a 1-sided lower tolerance interval, whose endpoint is called *lower tolerance limit*. According to (Hahn et al. (1991), chapter 2.4.2), a lower 1-sided tolerance limit, is equivalent to the lower limit of a 1-sided confidence interval for a given quantile, e.g. a 10th percentile. Using a lower tolerance limit with $P = 0.9$ and $\alpha = 0.05$ is thus equivalent to ensuring that the 1-sided 95 % confidence limit (5 %–100 %) of the 10th percentile is larger than the pre-defined target value.

In the context of LRV validation we are interested in the question about an interval, whose lower limits guarantees that 90 % of the validation measurements will not fall below the target value defined by the European Reuse Regulation 2020/741, given the observed data. Therefore, tolerance intervals may seem an appropriate approach to that question.

2.3.3.2.1. Classical tolerance interval calculations of normally distributed LRV data derived from paired data. Microbiological data can often be well described by using a lognormal distribution, i.e. fitting a normal distribution to the log-transformed observations. Against this background, normally distributed LRV values can be expected in cases when a.) both influent and effluent data follow a lognormal distribution, since the difference of two normal distributions, will again be normally distributed, and b.) when all effluent concentrations are constant, since manipulations of a normal distribution with a constant value, again lead to a normal distribution. While the first case can be expected when all effluent data are well above the limit of quantification, the second case might be observed if all data are below the limit of quantification, i.e. having a constant value. In both cases, normal tolerance intervals and the lower tolerance limits (TL_{Lower}) with $P = 0.9$ and $\alpha = 0.05$ can be calculated by:

$$k = \frac{1}{\sqrt{n}} t_{n-1;1-\alpha}\left(\sqrt{n}\, z_P\right) \tag{3-7}$$

$$TL_{Lower} = \mathrm{m} - k\ \mathrm{s} \tag{3-8}$$

where m and s represent the sample mean and standard deviation. The interval factor k now derives from the non-central t-distribution $t_{(n-1,\ 1-\alpha)}$ and depends on the level of confidence α, the available sample size n, and the z-score of the quantile of interest z_p. In the present study we used the R-package *tolerance (*Derek, 2010*)* and the implemented function *normtol.int()* for calculations.

2.3.3.2.2. Bayesian tolerance intervals based on unpaired evaluation. Deviating from classical normally distributed tolerance intervals derived from paired influent and effluent data, can be motivated by several reasons. First, the resulting LRV values may not be normally distributed. The most likely case for that observation is that only a certain proportion of the effluent data is below the limit of quantification, leading to a right-skewed distribution of LRV data.

A second motivation may be that there are different numbers of observation available for influent and effluent of the WRTP. In that case, part of the information might get lost during a paired evaluation. In such cases, an unpaired evaluation might be preferred. For both cases, we constructed a 1-sided lower Bayesian tolerance limit based on an unpaired evaluation of the influent and effluent data, following the following steps:

1. Fitting of a lognormal distribution to the influent data, leading to 10,000 correlated Markov Chain Monte Carlo (MCMC) samples for μ and σ
2. Fitting of a negative-binomial distribution to the effluent data, leading to 10,000 correlated MCMC samples of the location and scale parameters.
3. Construction of 10,000 influent and 10,000 effluent distributions based on a row-wise correlated MCMC samples for the individual distributions
4. Simulation of 10,000 distributions of LRV values using the generated influent and effluent distribution in Eqs. (3-1) and (3-2) (if simulation trials of the negative binomial are equal to 0), respectively.
5. Calculation of the 10th percentile of each of the 10,000 LRV distributions (content of the tolerance intervals), leading to 10,000 simulations for the 10th percentile
6. Calculation of the α-% quantile of the distribution of 10th percentile (confidence – level of 95 %)

The approach of constructing tolerance intervals from row-wise evaluation of the MCMC samples of the marginal distributions of distribution parameters is adapted from Stoudt et al. (2021), who nicely illustrate the differences between different probabilistic intervals. An

example in R (R Development Core Team, 2008) illustrating of the simulation process based on simulated data is provided in the SI. Bayesian analyses are conducted using the R-package *brms* (Bürkner, 2017).

2.4. Model comparison

To illustrate relevant aspects of the behavior of the different approaches we

a.) conduct a simulation-based analysis for both the binomial and distributional approaches, and
b.) apply both approaches to the collected real-world data

2.4.1. Simulation analysis

2.4.1.1. Binomial analysis. For illustrating the behavior of the binomial model, we assume the number of failures to increase from 0 to 5. For each number of failures, we increase the number of successes until the resulting lower 1-sided 95 % confidence/credible limit exceed a value of 0.9. Moreover, for the Bayesian approach we additionally compute P ($p > 0.9|$data), the probability that the "true" success rate is >0.9 for each combination of successes and failures. To reproduce results, we provide the complete R-Code in the supplementary information (SI).

2.4.1.2. Percentile analysis. As can be seen from Eqs. (3-6) and (3-8) the calculation of empirical percentiles and 1-sided lower tolerance limits is structurally identical, except from the numerical value of the interval factor k. For tolerance intervals this factor depends on the sample size N and the desired level of confidence (1-α). For illustrating the behavior of evaluation approaches, which are derived from the distribution of LRV values, we compute the interval factor k for different levels of confidence (1-α) ranging from 50 % to 100 %, and sample sizes ranging from 5 to 50. Additionally, we illustrate the behavior of the different approaches and intervals by generating simulated datasets of sample sizes of 5, 10, 20, and 50, which have exactly the same sample mean *($m = 6.5$)*, and sample standard deviation *($s = 0.3$)*. For each dataset, we compute:

- The 80 % confidence interval of the population mean μ.
- The 90 % confidence interval of the 10th percentile (lower 1-sided tolerance interval with $P = 0.9$ and $\alpha = 0.05$ and $\alpha = 0.95$)
- The empirical 10th percentile according to Eq. (3-6) with $k = 1.282$
- The 1-sided tolerance interval with $P = 0.9$ and $\alpha = 0.05$ (confidence level 95 %)
- The 1-sided tolerance interval with $P = 0.9$ and $\alpha = 0.5$ (confidence level 50 %),

and compare results to a hypothetical target value of 6 LRV. All simulations were conducted in R. Tolerance intervals and confidence intervals of the 10th percentile were calculated using the *normtol.int()* function from the R-package *tolerance*. To reproduce results, we provide the complete R-Code in the Supplementary information (SI).

2.4.2. Application to real dataset

2.4.2.1. Paired evaluation. For the paired evaluation, samples taken on the same day from the influent and effluent were evaluated. Incomplete datasets were removed. The evaluation is limited to the operating condition of ΔUVT = 47 %, which was shown favorable for disinfection. Although the influent values of the treatment plant are independent of the operating condition, the influent values for the operating condition ΔUVT = 34 % are discarded in the paired evaluation as there is no compatible effluent value. For the operating condition ΔUVT = 47 %, a total of 14 pairs of values were available. For each pair of values, LRVs were calculated according to Section 2.2. The obtained LRVs were further evaluated as follows:

1. Calculation of the success rate p and the associated parameter uncertainty using the Bayesian Beta-Binomial method (Eq. (3-5)).
2. Calculation of the one-sided lower tolerance limit with ($P = 0.9$) and $\alpha = 0.05$.
3. Calculation of the point estimator for the 10th percentile (Section 2.3.3.1).

For the binomial approach, the probability that the true success rate is >0.9 was also calculated. For better visualization of uncertainty intervals of the entire population distribution, the 90th percentile and the upper tolerance limits were additionally calculated complementing the 10th percentile and the lower tolerance limit. To illustrate the dependency of the calculated uncertainty intervals on the sample size N, all evaluations were carried out chronologically ascending from $N = 3$ to $N = 14$.

2.4.2.2. Unpaired evaluation. In addition to the paired evaluation, the lower and upper tolerance limits were determined using a Bayesian evaluation according to Section 2.3.3.2.2. In a first approach, as for the paired evaluation, the evaluation was also limited to the operating condition ΔUVT = 47 %. In a second approach (unpaired, all data), the influent data of the test series with ΔUVT = 34 % were also included for determining the distribution of the influent values. In this way, 10 additional measurements are available for the influent. As for the paired evaluation, the evaluation was carried out for N = 3 to = 14. For the evaluation in which all values are used, the evaluation starts with 13 influent and 3 effluent values and increases to 24 influent and 14 effluent values.

3. Results

3.1. Illustrating intervals

3.1.1. Binomial approach

The classical binomial and Bayesian beta-binomial model are illustrated in Fig. 1. Fig. 1 (top) illustrates the 1-sided lower 95 % confidence limits of the Bayesian and Wilson confidence intervals. The figure shows that for the case that no failure event is detected, i.e. all validation samples fulfill the requirements, the Wilson method requires 25 successes until the lower confidence limit exceeds the target value of 0.9, whereas the Bayesian approach requires 28 samples. In general, both approaches behave quite similar. As the number of failures increases the difference between both approaches stays relatively stable between 2 and 3 samples. Fig. 1 (top), moreover, shows that each failure has to be compensated for by increasing the number of successes. Thus, a single failure may have a large effect on the effort necessary for log-credit validation. Fig. 1 (bottom) illustrates the probability the true success rate is larger than the target value of 0.9, calculated by the Bayesian beta-binomial approach. The figure illustrates that for the combination of 1 failure out of 10 trials, i.e. 9 successes, the probability that the true success rate is larger than target value of 0.9, is only 30 %. For the case, of 5 failures and 45 successes, the probability that $p > 90$ % is still only 40 %. The results of this simulation experiment show that in cases where the point estimate is located at the target value of 0.9, the probability that the system does not fulfill the regulatory requirements is larger than the probability that it does. Therefore, the result shows that simplified, naive approaches like 1 out of 10 imply the risk of falsely validating a system which might actually not fulfill the regulatory requirements. A real-world example, where such misclassification happened is shown for the case of *Clostridium perfringens* in Section 3.2.

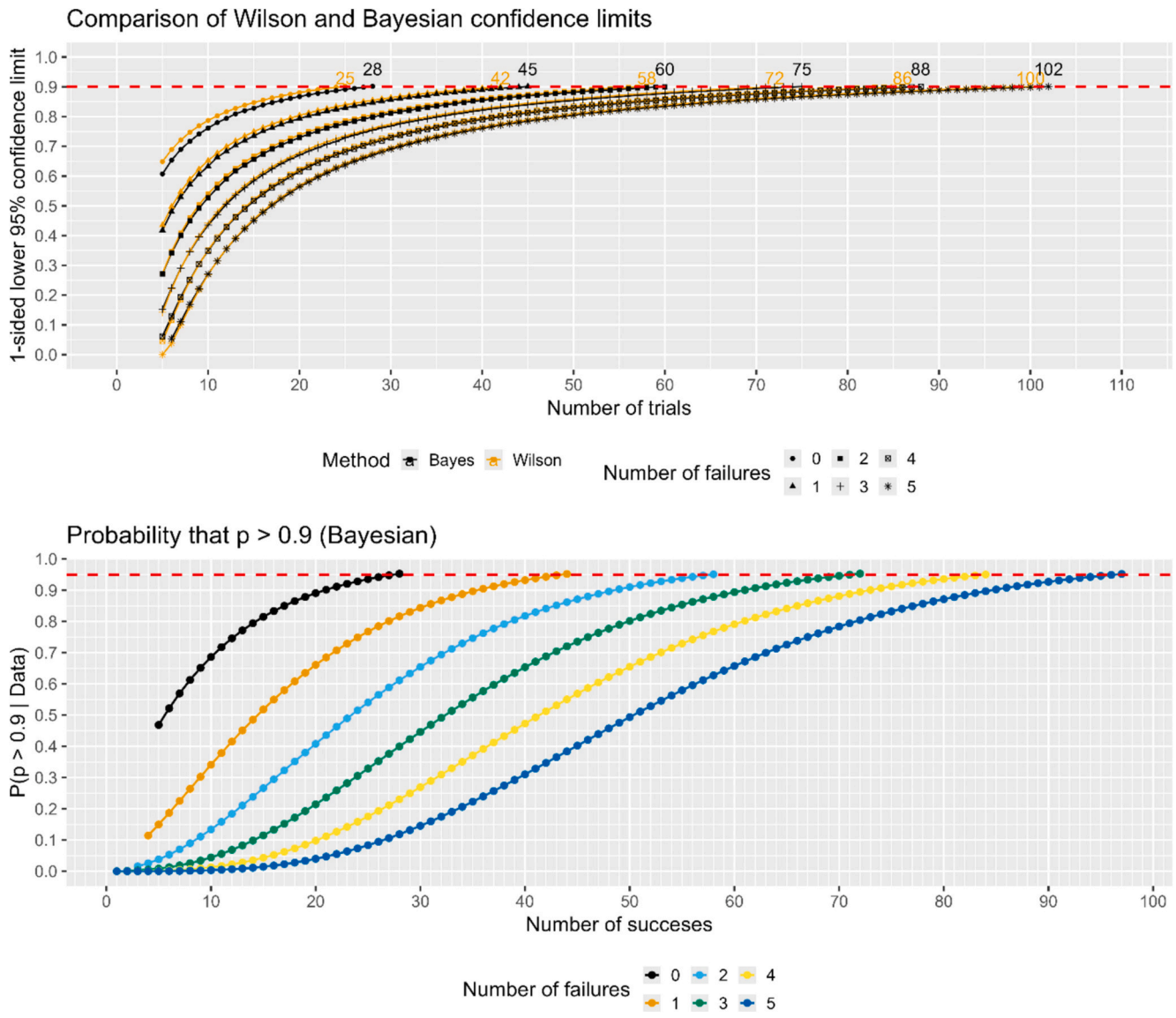


Fig. 1. Comparison of Bayesian and Wilson confidence/credible intervals (top) and the probability P, that success rate $p > 0.9$ based on the Bayesian evaluation (bottom).

3.1.2. *Approaches based on LRV distribution*

The differences in interval factors k of tolerance intervals with $P = 0.9$ and the empirical 10th percentile are illustrated in Fig. 2. The figure shows that for a level of confidence of $\alpha = 0.5$, i.e. 50 % confidence, the interval factor k of the tolerance interval rapidly converges to the z-score of the empirical percentile. At a sample size of 5 the difference is approximately 0.1 decreasing to levels below 0.02 at $N = 20$. Thus, the lower 1-sided tolerance limit with $P = 0.9$ and $\alpha = 0.5$ can be approximately set equal the point estimate of the 10th percentile. If the required confidence level (1- α) increases to larger values, e.g. 90 %, the k factor of the tolerance limits become wider. While all tolerance factors will eventually converge to the z-score if sample sizes become very large, such large sample sizes are not realistic in practice. Thus, in practice, tolerance intervals with confidence levels >50 % will be wider in comparison to intervals covering empirical percentiles. An illustrative example is provided in Fig. 3. Since all datasets deliver the same sample mean and sample standard deviation, the obtained empirical estimate for the 10th percentile is the same for all sample sizes between 5 and 50, and fulfills the hypothetical target value of 6 even for small sample sizes. In contrast, the lower 1-sided tolerance limit falls below the target value of 6 at low sample sizes of 5, and above the target value at sample sizes >10. This illustration underlines how tolerance intervals account for sample-size based uncertainty. From Figs. 2, 3 and (3–8) it also becomes apparent that the width of the interval, which covers the relevant uncertainty, is determined by N, α, and the sample standard deviation (s). Given the same width, it is essentially the difference between the sample mean (m) and the target value, i.e. the effect size in statistical terms, which determines whether the interval fulfills performance requirements or not. The larger that difference between m and the target value, i.e. the effect size, the more uncertainty can be tolerated. Fig. 4 illustrates the relationship between sample size, effect size and standard deviation required for successful validation based on a normal tolerance interval TI ($P = 0.9$, $\alpha = 0.05$).

3.2. *Application to real-world data*

The results obtained from applying the different evaluation approaches to real world data are shown in Fig. 5. Focusing first on the evaluation of uncertainty intervals derived from 14 paired samples it can be seen that all uncertainty intervals are completely above the target values for *E. coli*, but not for *C. perfringens*. Therefore, the intervals agree in their final assessment after 14 samples that the system can be

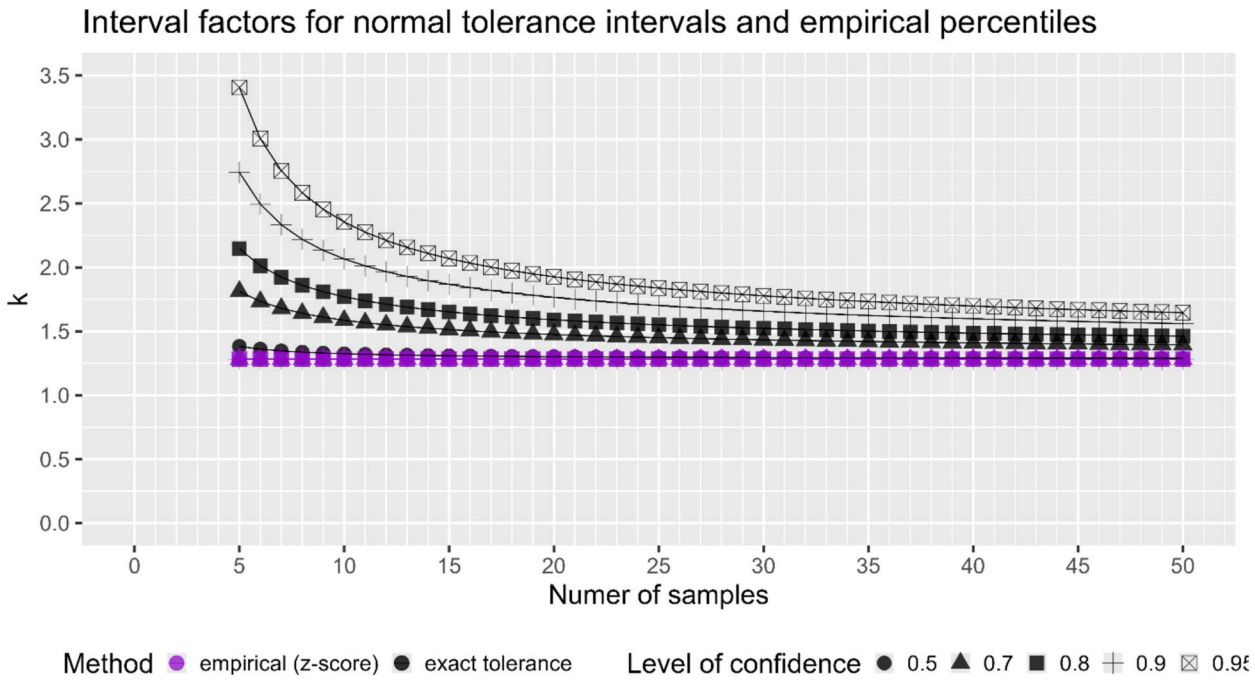


Fig. 2. Comparison of interval factors k for normal tolerance intervals with P = 0.9 and the empirical 10th percentile against sample size and desired level of confidence.

considered *validated* for *E. coli* but not for *C. perfringens*. While intervals agree in their final assessment it can be seen that the intervals behave quite differently between the minimum sample size of 3 and the maximum of 14. Since the empirical parametric evaluation approach does not consider sample size-based uncertainty, the empirical uncertainty intervals are comparatively narrow from the beginning ($N = 3$) in comparison to the approaches based on tolerance intervals. Indeed, for the empirical parametric approach applied to *C. perfringens* data, it takes up to a sample size of 9 until the lower limit of the constructed interval falls below the regulatory target of 4 LRV, indicating that the system might not comply with regulatory targets. Thus, the approach changes from a positive to a negative assessment of the system, raising the question about whether to trust an assessment based on low sample sizes.

In contrast, all approaches based on tolerance intervals are consistent in their negative assessment for *C. perfringens* from N = 3 to $N = 14$. At low sample sizes tolerance intervals are very wide due to the large parameter uncertainty, that these intervals do account for. As the sample size increases, intervals become narrower and eventually stabilize. At very low sample sizes <5 the classical tolerance interval is wider than the Bayesian one. At larger values differences are rather small with sometimes the Bayesian and sometimes the classical approach being slightly narrower or wider.

The figure also illustrates that the difference between the observed average treatment performance in relation to the regulatory target value, i.e. the effect size, and the width of the uncertainty interval are essential for a successful validation. For example, in the case of *E. coli* the average $\log_{10}$-removal lies at approximately 7 LRV. Thus, the effect size (difference to the regulatory target (5 LRV)) is 2 orders of magnitude, which corresponds to a factor of 100. Since the average lies two orders of magnitude above the target, tolerance intervals based on a paired evaluation already fall completely above the target value at comparatively low sample sizes of $N = 5$–6, since the large difference between average and target value also tolerates comparatively large statistical uncertainty. That shows that if the difference between the average treatment performance and regulatory target is large, a validation based on tolerance intervals can be achieved with comparatively low samples sizes. Since sample sized based uncertainty is accounted for results from a tolerance interval-based approach based on low sample sizes are more reliable, which is a major advantage over the empirical approach.

The figure moreover shows some potential shortcomings and disadvantages of the binomial approach, especially how naïve approaches like 9 successes out of 10 trials can lead to a false positive validation and thus potential health risk. As can be seen for the case of *C. perfringens* after 14 LRV evaluations only 10 fulfill the regulatory requirements. This corresponds to a success rate of $p = 71$ %, and a probability that $p >$ 0.9 of approximately 1 %. Thus, regulatory requirements are not achieved. However, after 10 sampling events 9 out of 10 evaluations indeed were above the target. If a "9 out of 10" approach would be chosen it would have led to a false validation of the system (false positive, Type I error). From Fig. 1 it can be deduced that >80 successes would be necessary after 4 failures to achieve 95 % confidence that p is truly larger than 90 %. Another disadvantage of the binomial approach is that it requires larger sample sizes as interval-based approaches. Since a binomial approach requires a minimum of 25–28 success for validation if all data fulfill the requirements the 14 successful trials for *E. coli* are not sufficient for validation (Fig. 1).

The larger number of required samples of the binomial approach in comparison to percentile-based approaches is caused by the fact, that the binomial approach ignores the information about the difference between the observed $\log_{10}$-removal and the regulatory target value. For example, when applying the binomial approach for validating *E. coli* removal, an observed LRV of 5.01 influences the results in the same way as an observed LRV of 7. Both values enter the model as "*success*". The information that an observed LRV of 7 overfulfills the regulatory target by a factor of 100, while the other barely achieves it is ignored and is, thus, not accounted for.

Tolerance intervals for the unpaired case based on all data (green tolerance interval), show that validation efforts can be further reduced if all available data are included into the statistical evaluation. Such situations may occur if after a first validation attempt, results indicate that process parameters need to be further adjusted. In such cases, influent data from this first validation attempt, which are independent from the adjustment of process parameters can be included into a second validation phase and reduce validation efforts by providing complementing information on influent water quality.

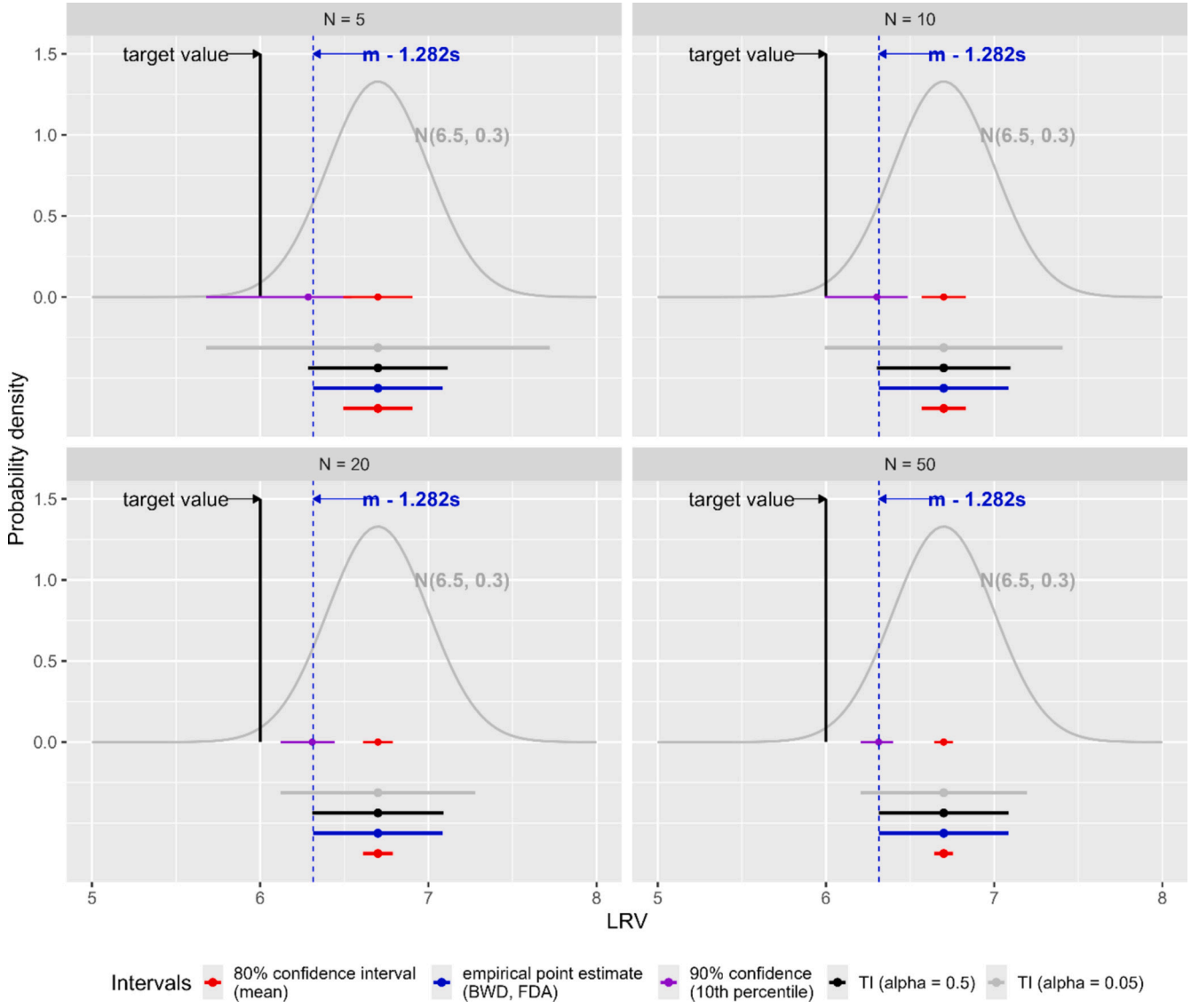


Fig. 3. Illustrating the effect of sample size on confidence intervals, tolerance intervals and empirical percentile intervals. Example datasets with $N = 5, 10, 20, 50$ were generated leading to the exact same sample mean and sample standard deviations. Tolerance intervals (TI) below the distribution always show the lower 1-sided TI (P, α) and upper 1-sided TI (1-P, α) tolerance limits with corresponding values for α. Empirical intervals show point estimates for the 10th and 90th percentile.

4. Discussion

In the present study we evaluated different statistical methods for validating a 90 % threshold, with a strong reference to $\log_{10}$-removal validation according to the most recent water reuse regulation EU 2020/741 in Europe. We selected this example since the regulation currently lacks detailed recommendations on which statistical approach to apply. Our study introduced tolerance intervals as a rarely used potential candidate approach and illustrated its advantages over a binomial evaluation approach and the calculation of empirical parametric percentiles. The latter are used in existing regulations in Europe and the U.S. for the assessment of bathing water and irrigation water quality, respectively.

We showed that tolerance intervals extend existing percentile-based approaches by complementing the target percentile, e.g. 10th percentile, with an associated level of confidence (1-α), e.g. 95 %. We also showed that in contrast to empirical parametric percentile calculation, tolerance intervals do account for sample-size based parameter uncertainty. By doing so, tolerance intervals generalize from sample to population and do not rely on a minimum number of data points to ensure a certain level of statistical precision. As tolerance intervals are very wide at low sample sizes, a positive validation at low sample sizes can be regarded as valid, since it will only be achieved if the true average treatment performance is much higher than the regulatory target, i.e. if the effect size is very large. The opportunity to conduct a reliable validation even at low sample sizes is a major difference and practical advantage in comparison to the use of empirical parametric percentiles, where low sample sizes leads to a high risk of type I error. The use of parametric percentiles in combination with defining a minimum number of samples to ensure a certain level of precision, as implemented by existing U.S. and European regulations, seems therefore less flexible and potentially demands unnecessary high numbers of samples in some cases.

In the context of irrigation water standard in the U.S., Gerdes et al. (2022) reported that the number of 20 samples is often criticized to put an disproportionate economic burden especially on end-users with very limited resources, who have to carry the costs for elaborating microbial water quality profiles (MWQP). Therefore, Gerdes et al. (2022) conducted a Monte Carlo analysis to assess whether percentile calculation based on fewer samples might have led to the same assessment result as percentile calculations based on a complete dataset of 20. Similar to our

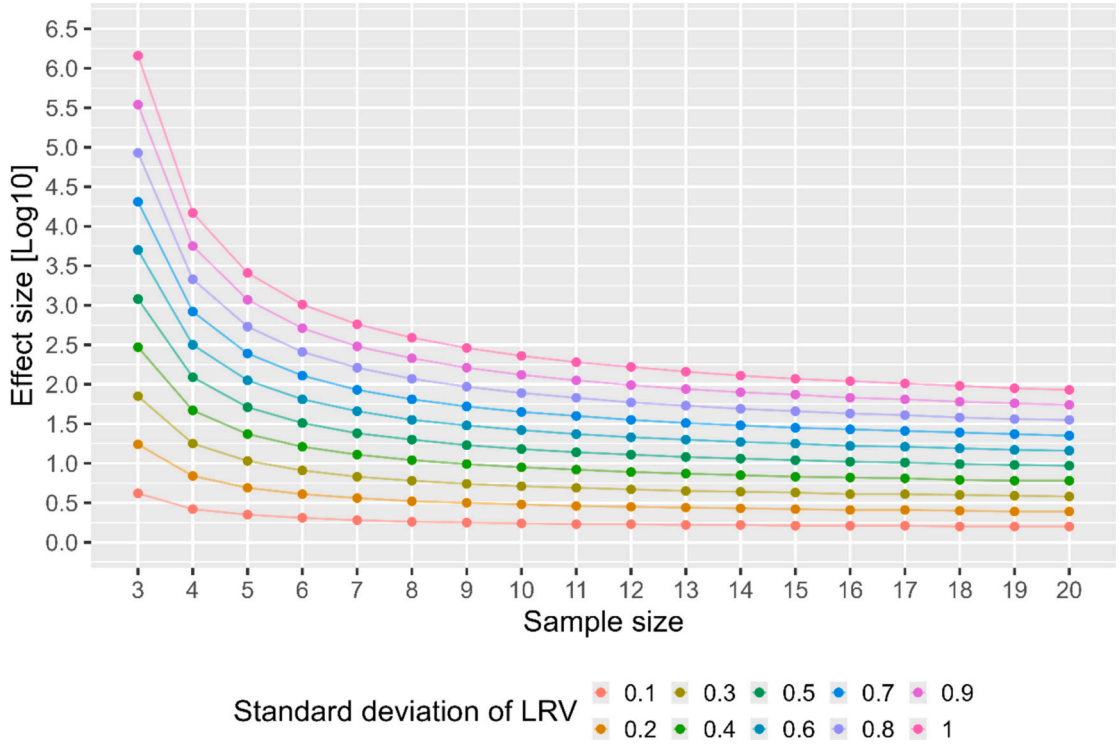


Fig. 4. Relation between sample size, standard deviation and effect size necessary for successful validation based on a normal tolerance interval with P = 0.9 and α = 0.05.

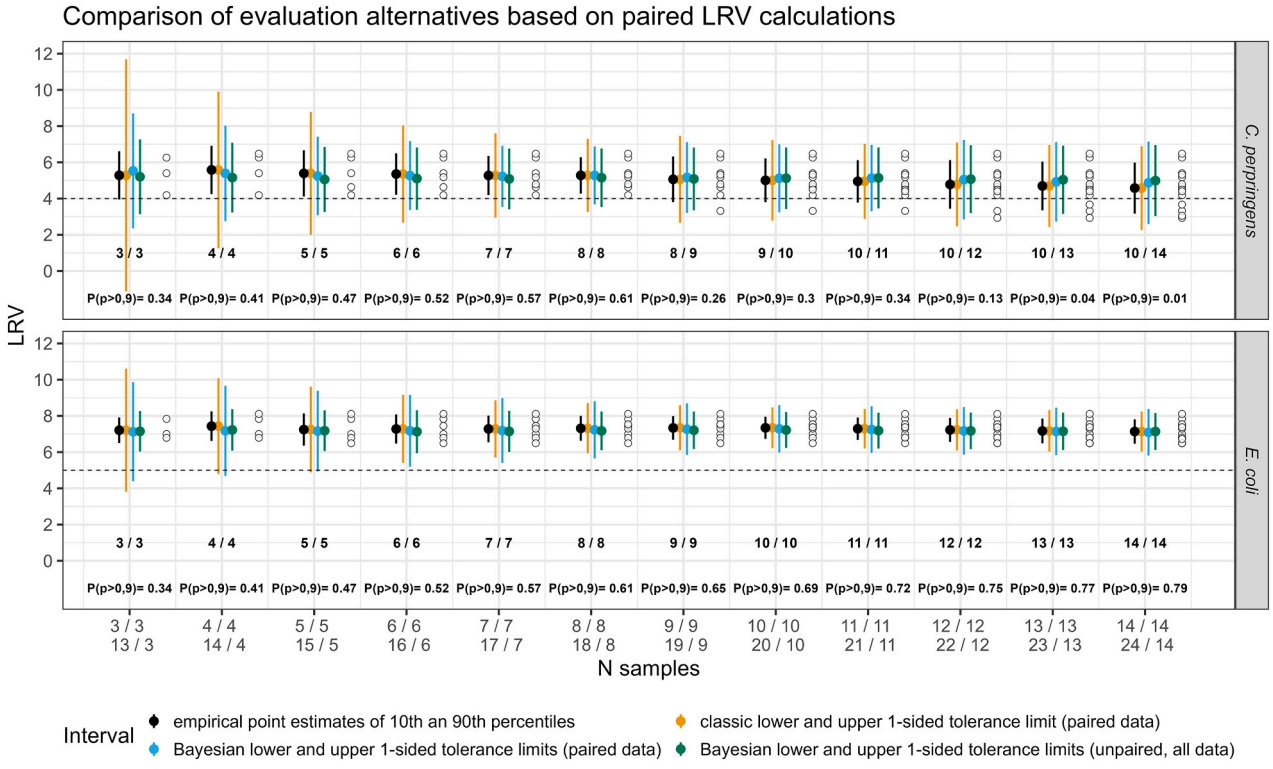


Fig. 5. Comparison of different evaluation approaches based on paired LRV calculations. Horizontal dashed lines indicate target LRVs for the two species. Numbers below the uncertainty intervals refer to the binomial evaluation represented as the number of successes and the total number of trials. The calculated probabilities P ($p > 0.9$) refer to the probability, evaluated by the Bayesian beta-binomial model, that the true success rate is larger than 0.9 given the available number of successes and failures. White circles indicate the LRV values from which intervals and success rates are inferred from. The two rows of tick marks on the x-axis refer to the number of influent and effluent samples evaluated for the scenarios "paired data only", and "unpaired evaluation of all data".

results they showed that if the underlying water quality is far from the regulatory target (below or above), percentile calculations based on fewer samples may lead to the same assessment as calculations based on a dataset of 20. Thus, the authors conclude that a reduced number of samples might be justifiable in some cases. However, the remaining problem when relying on empirical percentiles is that there is no way to assess whether a reduced number of samples would have been sufficient in a specific case *beforehand*. In contrast, the application tolerance intervals would allow such evaluations, and thus may provide a suitable way to make water quality validation less resource intensive and effective, while ensuring microbial safety.

The problem that the parameter uncertainty of estimated percentiles increases with the distance of the estimated percentile from the average has been described before (Berthouex and Hau, 1991). Such uncertainties may lead to uncertainties in the decision-making process and the question regarding which level of precision is needed. Verbyla et al. (2023) addressed this issue by underlining that the level precision of reported percentile estimates should reflect the existing parameter uncertainty. Our study provides an alternative approach by recommending that the lower 1-sided tolerance interval should exceed regulatory target values. While there is certainly more than one valid approach, we regard the use of tolerance intervals as one particularly transparent and methodologically valid approach, since it accounts for existing parameter uncertainty and can be compared the other kinds of statistical significance testing. The latter provide an indication about whether observed differences are likely caused by some real effect or are simply the result of random sampling error. Therefore, the proposed approach aligns well with good scientific practice. To the best of our knowledge, the use of tolerance intervals has not been suggested in the context of water quality management before, but seems to be a potentially valuable amendment to existing and future regulations, where percentile thresholds are applied. Examples include application in bathing water and irrigation water quality.

Our present study, however, only focused on the validation of technical systems under routine operating conditions. In such cases reduced sample sizes may seem to be more appropriate since system validation should take place independently for different operating conditions (e.g. rain weather, dry weather flow), while the management of extraordinary event conditions should be part of complementary risk management plans. Therefore, water quality during process validation can be regarded as relatively stable in comparison to the monitoring of natural waters, like bathing waters or surface waters for irrigation. In the latter cases reduced sample sizes may also reduce probability of detecting of peak events of fecal contaminations. Thus, in other contexts, additional aspects may have to be considered and additional research may be needed before transferring the proposed approach to other contexts.

In order to exploit the practical benefits of potentially fewer samples in practice, data evaluations could be conducted incrementally, following pre-defined protocols, e.g. after increments of 3 or 5 samples. A potential downside might be the additional level of complexity to calculate the 1-sided lower tolerance limit in comparison to empirical percentiles. However, the only difference in comparison to the approach for calculating parametric empirical percentiles is that the interval factor k changes with sample size. Since changes in k become rather small above a sample size of 20, 20 samples seem to be a reasonable upper limit, and specific values for k could easily be provided for the selected increments, e.g. for sample sizes of 5, 10, 15, and 20. Such procedures and protocols should be further developed and validated by applying the proposed approach to additional comprehensive test datasets.

A potential limitation of exploiting the described advantages may be that the observable effect size, i.e. the difference between observed $\log_{10}$-reduction and target value can be limited by low influent concentrations of fecal indicator organisms. For example, an average inflow *E. coli* concentration of 10^6/100 mL, limits the observable LRV to a value of 6 and the observable effect size to a difference of 1, since the target value is 5. In such cases, increasing the sample volumes in the effluent might potentially allow for increasing the observable effect sizes, but such methods may have to be further developed, since additional factors, like recovery rates of enrichment methods have to be accounted for, which in turn have to be validated in practice.

5. Conclusion

- Statistical tolerance intervals account for sample-size-based uncertainty and thus allow for validating performance treatment targets even at lower sample sizes if effect sizes are large.
- Bayesian tolerance intervals based on unpaired evaluation of influent and effluent data are considered the most flexible approach for handling different challenges regarding data evaluation, like unequal sample sizes and various proportions of data below the LOQ.
- Regulations, currently relying on the calculation of parametric empirical percentiles, like bathing water and irrigation water regulations might benefit from re-evaluating potential benefits tolerance intervals provide,
- Further comprehensive test data sets are required to further validate the application of the proposed approach, and refine validation protocols.

CRediT authorship contribution statement

Wolfgang Seis: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Michael Stapf:** Data curation. **Ulf Miehe:** Writing – review & editing. **Thomas Wintgens:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research part of the research project FlexTreat which was funded by the Federal Ministry of Education and Research (BMBF) under the reference number: 02WV1561A-L. We gratefully acknowledge Xylem for providing the pilot treatment plant essential for this research. We also thank the Wastewater Association Braunschweig for granting access to their facilities and to support our experiments, especially Dr. Franziska Gromadecki and Janina Heinze.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2025.178573.

Data availability

Data will be made available on request.

References

2006/7/EC 2006 Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing. Community, E. (ed).

Berthouex, P.M., Hau, I., 1991. Difficulties related to using extreme percentiles for water quality regulations. Research Journal of the Water Pollution Control Federation 63 (6), 873–879.

Branch, A., Leslie, G., Le-Clech, P., 2021. A statistical review of pathogen and indicator log removal values from membrane bioreactor literature. Crit. Rev. Environ. Sci. Technol. 51 (16), 1866–1890.

Bürkner, P.-C., 2017. brms: An R Package for Bayesian Multilevel Models Using Stan, 80 (1), p. 28.

Chik, A.H.S., Schmidt, P.J., Emelko, M.B., 2018. Learning something from nothing: the critical importance of rethinking microbial non-detects. Front. Microbiol. 9.
Derek, S.Y., 2010. tolerance: an R package for estimating tolerance intervals. J. Stat. Softw. 36 (5), 1–39.
EEC 1976 Council Directive 76/160/EEC of 8 December 1975 concerning the quality of bathing water.
EU 2020 REGULATION (EU) 2020/741 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on minimum requirements for water reuse.
FDA 2013 FDA Food Safety Modernization Act.
Gerdes, M.E., Cruz-Cano, R., Solaiman, S., Ammons, S., Allard, S.M., Sapkota, A.R., Micallef, S.A., Goldstein, R.E.R., 2022. Impact of irrigation water type and sampling frequency on microbial water quality profiles required for compliance with U.S. food safety modernization act produce safety rule standards. Environ. Res. 205, 112480.
Hahn, G.J., Meeker, W.Q., Escobar, L.A., 1991. Statistical Intervals: A Guide for Practitioners.
Heckert, N., Filliben, J., Croarkin, C., Hembree, B., Guthrie, W., Tobias, P., Prinz, J., 2002. Handbook 151: NIST/SEMATECH e-Handbook of Statistical Methods, NIST Interagency/Internal Report (NISTIR). National Institute of Standards and Technology, Gaithersburg, MD.
Hunter, P.R., 2002. Does calculation of the 95th percentile of microbiological results offer any advantage over percentage exceedence in determining compliance with bathing water quality standards? Lett. Appl. Microbiol. 34 (4), 283–286.
Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E.-J., 2016. The fallacy of placing confidence in confidence intervals. Psychon. Bull. Rev. 23 (1), 103–123.
R Development Core Team 2008 A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
Stoudt, S., Pintar, A., Possolo, A., 2021. Coverage Intervals. Journal of research of the National Institute of Standards and Technology 126, 1–22.
Teunis, P.F.M., Rutjes, S.A., Westrell, T., de Roda Husman, A.M., 2009. Characterization of drinking water treatment for virus risk assessment. Water Res. 43 (2), 395–404.
Verbyla, M., Fani, M., Walker, B., 2023. Pathogen Removal Credits for Wastewater Treatment: Guidance for Study Plans and Reporting (The Water Research Foundation).