Parametric estimation of $P(X > Y)$ for normal distributions in the context of probabilistic environmental risk assessment

Rianne Jacobs[1], Andriëtte A. Bekker[2], Hilko van der Voet[1] and Cajo J.F. ter Braak[1]

[1] Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands
[2] Department of Statistics, University of Pretoria, Pretoria, South Africa

ABSTRACT

Estimating the risk, $P(X > Y)$, in probabilistic environmental risk assessment of nanoparticles is a problem when confronted by potentially small risks and small sample sizes of the exposure concentration X and/or the effect concentration Y. This is illustrated in the motivating case study of aquatic risk assessment of nano-Ag. A non-parametric estimator based on data alone is not sufficient as it is limited by sample size. In this paper, we investigate the maximum gain possible when making strong parametric assumptions as opposed to making no parametric assumptions at all. We compare maximum likelihood and Bayesian estimators with the non-parametric estimator and study the influence of sample size and risk on the (interval) estimators via simulation. We found that the parametric estimators enable us to estimate and bound the risk for smaller sample sizes and small risks. Also, the Bayesian estimator outperforms the maximum likelihood estimators in terms of coverage and interval lengths and is, therefore, preferred in our motivating case study.

Subjects Environmental Sciences, Toxicology, Statistics
Keywords Bayesian estimator, Maximum likelihood estimator, Risk assessment, Simulation, Nanoparticle

Submitted 1 June 2015
Accepted 17 July 2015
Published 18 August 2015

Corresponding author
Rianne Jacobs,
rianne.jacobs@wur.nl

Academic editor
Budiman Minasny

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj.1164**

INTRODUCTION

Like all novel materials, engineered nanoparticle (ENPs) have no history of safe use. A risk assessment is important for the societal acceptance and safe use of ENPs. In order to perform a proper risk assessment, one needs knowledge and data on the properties of nanoparticles. These properties can be different in nanoparticles compared to conventional chemicals in areas such as physicochemical properties, life cycle, toxicokinetics and environmental fate. This information is hard to come by because of lack of knowledge and technical limitations, resulting in no or only small datasets for effect concentrations of ENPs. In the EU, environmental risk assessment is regulated by the European Chemicals Agency (ECHA) and probabilistic risk assessment is level 3 of their tiered risk assessment approach (*ECHA, 2012*).

In the motivating case study of aquatic risk assessment of nano-Ag (*Gottschalk, Kost & Nowack, 2013*), we are confronted with such a small dataset of effect concentrations.

How to cite this article Jacobs et al. (2015), Parametric estimation of $P(X > Y)$ for normal distributions in the context of probabilistic environmental risk assessment. **PeerJ 3:e1164; DOI 10.7717/peerj.1164**

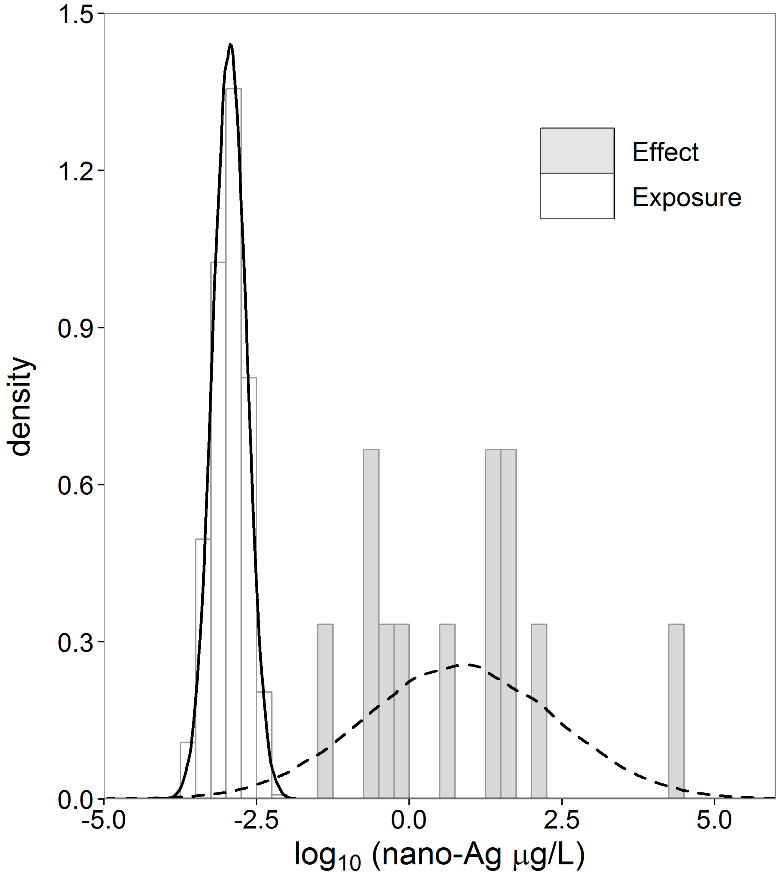

Figure 1 Histograms and normal density curves of exposure (n_x = 1,000) and effect (n_y = 12) concentration nano-Ag (µg/L). Data taken from *Gottschalk, Kost & Nowack (2013)*.

Gottschalk, Kost & Nowack (2013) modeled the exposure of nano-Ag from surface water with a probabilistic material-flow model (*Gottschalk, Scholz & Nowack, 2010*) to obtain a distribution of exposure concentration values. They collected the effect concentration data from available toxicity studies found in the literature. These effect concentration data consist of toxic endpoints (eq. LC_{50}, EC_{50}, NOEC) for 12 aquatic species. For some of these species there were more than one data point. We averaged these to obtain one value for each species. Histograms and normal density curves of the exposure and effect concentration data are given in Fig. 1.

In probabilistic risk assessment, the variability of environmental exposure due to natural variation in concentration values over various environments is modelled by an exposure concentration (X) distribution (ECD). Similarly, the variability in effect concentration values due to natural variation among species in their sensitivity to nanoparticles is modelled by a species sensitivity distribution (SSD) or effect concentration (Y) distribution. Probabilistic risk estimation is based on the overlapping of the ECD and the SSD (*ECHA, 2012*). The risk, $R = P(X > Y)$, is the area under the curve obtained by multiplying the probability density function (pdf) of the ECD with the cumulative distribution function (cdf) of the SSD. *Verdonck et al. (2003)* critically discuss this approach to risk assessment.

In the ecotoxicological risk assessment literature, $R = P(X > Y)$ as a definition for risk was first developed by *Suter, Vaughan & Gardner (1983)*. This concept was further developed by *Van Straalen (2002)* as ecological risk (ER). A similar concept, known as expected total risk, was developed by the Water Environment Research Foundation (WERF) (*Cardwell et al., 1993*; *Warren-Hicks, Parkhurst & Butcher, 2002*). For a visual representation of risk, the exceedance function (1-cdf) of the ECD is plotted against the cdf of the SSD. Such a plot is called a joint probability curve (JPC) (*ECOFRAM, 1999*; *Solomon, Giesy & Jones, 2000*; *Solomon & Takacs, 2002*). The area under the JPC is the risk (*Solomon & Takacs, 2002*). *Aldenberg, Jaworska & Traas (2002)* showed that the area under the JPC, ER and expected total risk are mathematically identical. In this paper, we refer to the probability, $P(X > Y)$, as the risk, R.

When we consider the case study data, there is no overlap between the effect and the exposure histograms (Fig. 1). There is no exposure concentration that is greater than an effect concentration, and, therefore, the empirical estimate of $R = P(X > Y)$ is zero for these datasets. To conclude, however, that the true R is zero based on a small sample is rather imprudent. The denial of a probability of zero is referred to as Cromwell's rule by Lindley (*1971*, p. 105–106; *2006*, p. 90–91). Several possibilities exist to address the zero problem empirically. This is further discussed in Section 'Non-parametric estimator'. The zero problem can also be solved by fitting a parametric distribution to the data. When considering the normal density curves in Fig. 1, we note that there is some overlap between the exposure and effect concentration distributions and, therefore, some non-zero probability of exposure values exceeding effect values.

It is common to assume independent log-normal distributions for the exposure distribution and the species sensitivity distribution (SSD) (*Aldenberg, Jaworska & Traas, 2002*; *Verdonck et al., 2003*; *Wagner & Løkke, 1991*). This is the same as assuming normal distributions on the log-transformed exposure and effect concentrations. This normal–normal model was developed in some detail by *Aldenberg, Jaworska & Traas (2002)* and allows an analytic expression for the risk when parameter values are known.

Estimation of $R = P(X > Y)$ is also of interest in other areas such as engineering and medical applications. In engineering, $R = P(X > Y)$ is known as the reliability in stress–strength models. This is a well-known concept and has been studied extensively for the normal–normal model (*Barbiero, 2011*; *Church & Harris, 1970*; *Downtown, 1973*; *Enis & Geisser, 1971*; *Govidarajulu, 1967*; *Nandi & Aich, 1996*; *Voinov, 1986*; *Weerahandi & Johnson, 1992*) as well as for other distributions (*Kundu & Gupta, 2006*; *Mokhlis, 2005*; *Nadar, Kızılaslan & Papadopoulos, 2014*). None of these papers give sufficiently general theory for obtaining trustworthy interval estimates in the case study. In receiver operating characteristic (ROC) analysis such as used in medical applications, $P(X > Y)$ is known as the area under the ROC curve (*Li & Ma, 2011*). Although usually used for categorical data, the area under the ROC curve can also be obtained for continuous data in both a non-parametric way and for the normal–normal model (*Krzanowski & Hand, 2009*).

In this paper, we will investigate the influence of sample size on the estimation of $R = P(X > Y)$, with special attention to the sample size of effect concentrations. We also investigate the behaviour of the estimators of $R = P(X > Y)$ for small risks. We consider one non-parametric estimator and three parametric estimators, namely, the maximum likelihood estimator (MLE), quasi maximum likelihood estimator (QMLE) and Bayesian estimator with noninformative prior for the normal–normal model. In comparing the parametric estimators with the non-parametric one, we investigate the maximum gain possible when making strong parametric assumptions as opposed to making no parametric assumptions at all. This is done in a simulation study in which we also assess the accuracy and precision of the estimators and compare them for various combinations of sample sizes and risks.

In 'Theory and Methods', we derive the estimators and provide the simulation structure. In 'Simulation Results', the simulation results will be given and discussed. In 'Case Study', the application is discussed in the context of the simulation results. 'Discussion and Conclusion' provides some general discussion, conclusions and recommendations for further study.

THEORY AND METHODS

In this section, we describe the theory and methodology of our approach. We start by deriving the risk for the normal–normal model. Next we discuss the four estimation methods, provide the simulation structure and discuss the performance measures that we used.

Risk

Let X be the $\log_{10}$ exposure concentration random variable and Y be the $\log_{10}$ effect (or $\log_{10}$ sensitivity) concentration random variable.

In the normal–normal model, the distributions are given by

$$X \sim N(\mu_x, \sigma_x)$$
$$Y \sim N(\mu_y, \sigma_y).$$

Due to the additive property of the normal distribution we have

$$X - Y \sim N(\mu_x - \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2}).$$

The risk for the normal–normal model is given by

$$\begin{aligned} R &= P(X > Y) \\ &= P(X - Y > 0) \\ &= 1 - \Phi\left(\frac{-(\mu_x - \mu_y)}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right) \\ &= \Phi\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right) \qquad (1) \end{aligned}$$

where $\Phi(\cdot)$ denotes the standard normal distribution. Equation (1) is a well-known result (*Reiser & Guttman, 1986*).

We note location-scale invariance in Eq. (1). The value of R is determined only by the difference of the expected values and the sum of the variances. The absolute value of the individual parameters is not relevant.

Point estimation

In the following sections, we derive the MLE, QMLE, Bayesian estimator and non-parametric estimator for the risk, R. We let $(x_1, x_2, \ldots, x_{n_x})$ be a random sample of size n_x of log exposure concentrations and $(y_1, y_2, \ldots, y_{n_y})$ an independent random sample of size n_y of log effect concentrations.

Maximum likelihood estimator

The most straightforward way of estimating R is by means of maximum likelihood estimation. The estimator obtained in this way is denoted as $\hat{R}_{\text{MLE}}$. From the invariance property of MLEs (*Bain & Engelhardt, 1992*, p. 296), we obtain $\hat{R}_{\text{MLE}}$ by substituting the MLEs of μ_x, μ_y, σ_x^2 and σ_y^2 in Eq. (1). These MLEs are given by

Parameter	Maximum likelihood estimator
μ_x	$\bar{x} = \frac{1}{n_x}\sum_{i=1}^{n_x} x_i$
μ_y	$\bar{y} = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i$
σ_x^2	$\hat{\sigma}_x^2 = \frac{1}{n_x}\sum_{i=1}^{n_x} (x_i - \bar{x})^2$
σ_y^2	$\hat{\sigma}_y^2 = \frac{1}{n_y}\sum_{i=1}^{n_y} (y_i - \bar{y})^2$

Equation (1) then becomes

$$\hat{R}_{\text{MLE}} = \Phi\left(\frac{\bar{x} - \bar{y}}{\sqrt{\hat{\sigma}_x^2 + \hat{\sigma}_y^2}}\right). \tag{2}$$

Note that $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ are the MLEs of the variance, which are not unbiased.

Quasi maximum likelihood estimator

The QMLE is similar to the MLE, differing only in the use of unbiased estimators for σ_x^2 and σ_y^2 instead of the MLEs. We then obtain

$$\hat{R}_{\text{QMLE}} = \Phi\left(\frac{\bar{x} - \bar{y}}{\sqrt{s_x^2 + s_y^2}}\right) \tag{3}$$

with $s_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \bar{x})^2$ and $s_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \bar{y})^2$.

Bayesian estimator

Our third way of estimating R is Bayesian. Whereas maximum likelihood estimation uses the data only, Bayesian estimation combines prior knowledge about the parameter(s) with the data. The prior knowledge is specified by a prior distribution and the information in the data by the likelihood. The prior distribution and the likelihood are then combined into what is called the posterior distribution of the parameter (*Gelman et al., 2014*). We will derive the joint posterior distribution of the parameters μ_x, μ_y, σ_x^2 and σ_y^2. This distribution together with Eq. (1) will provide us with the posterior distribution, $f_R(r)$, of R.

Unfortunately we have often very little prior knowledge. Therefore, we derive $\hat{R}_{\text{Bayes}}$ assuming a non-informative prior distribution for the parameters, namely $p(\mu_x,\sigma_x^2) \propto \frac{1}{\sigma_x^2}$ for the joint prior distribution of μ_x,σ_x^2 and $p(\mu_y,\sigma_y^2) \propto \frac{1}{\sigma_y^2}$ for μ_y,σ_y^2 (*Gelman et al., 2014*, p. 64).

The posterior distributions are then given by (*Gelman et al., 2014*, p. 65)

- $\mu_x \mid \sigma_x^2 \sim N\left(\bar{x}, \frac{\sigma_x}{\sqrt{n_x}}\right)$
- $\mu_y \mid \sigma_y^2 \sim N\left(\bar{y}, \frac{\sigma_y}{\sqrt{n_y}}\right)$
- $\sigma_x^2 \sim \text{Inverse-gamma}\left(\frac{n_x-1}{2}, \frac{(n_x-1)s_x^2}{2}\right)$
- $\sigma_y^2 \sim \text{Inverse-gamma}\left(\frac{n_y-1}{2}, \frac{(n_y-1)s_y^2}{2}\right)$.

From this, we obtain the conditional posterior distribution of $\frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}}$ (*Weerahandi & Johnson, 1992*)

$$\mu_x - \mu_y \mid \sigma_x^2, \sigma_y^2 \sim N\left(\bar{x}-\bar{y}, \sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}\right)$$

$$\therefore \frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}} \mid \sigma_x^2, \sigma_y^2 \sim N\left(\frac{\bar{x}-\bar{y}}{\sqrt{\sigma_x^2+\sigma_y^2}}, \sqrt{\frac{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}{\sigma_x^2+\sigma_y^2}}\right).$$

Using the variable transformation method and integrating σ_x^2 and σ_y^2 out of the joint pdf, $f(r,\sigma_x^2,\sigma_y^2)$, we obtain the marginal pdf, $f_R(r)$ (see Appendix Result A.2 for details). This marginal posterior pdf of R (Eq. (A.4)) can be evaluated using numerical integration.

Alternatively, we can use Monte Carlo sampling to approximate the marginal posterior pdf of R. We used the Method of Composition (*Lesaffre & Lawson, 2012*, p. 93–94) in which we sample from the known posterior distributions of σ_x^2 and σ_y^2, then μ_x and μ_y from their known posterior conditional distributions and then apply Eq. (1) to obtain the corresponding R value. Figure 2 shows histograms of samples drawn from the marginal posterior distribution of R (sample size of 10,000) together with the marginal pdf computed by numerical integration using Eq. (A.4) for different sample sizes and R values. It can be seen that the Monte Carlo method gives a very good approximation to the theoretical posterior pdf evaluated by numerical integration.

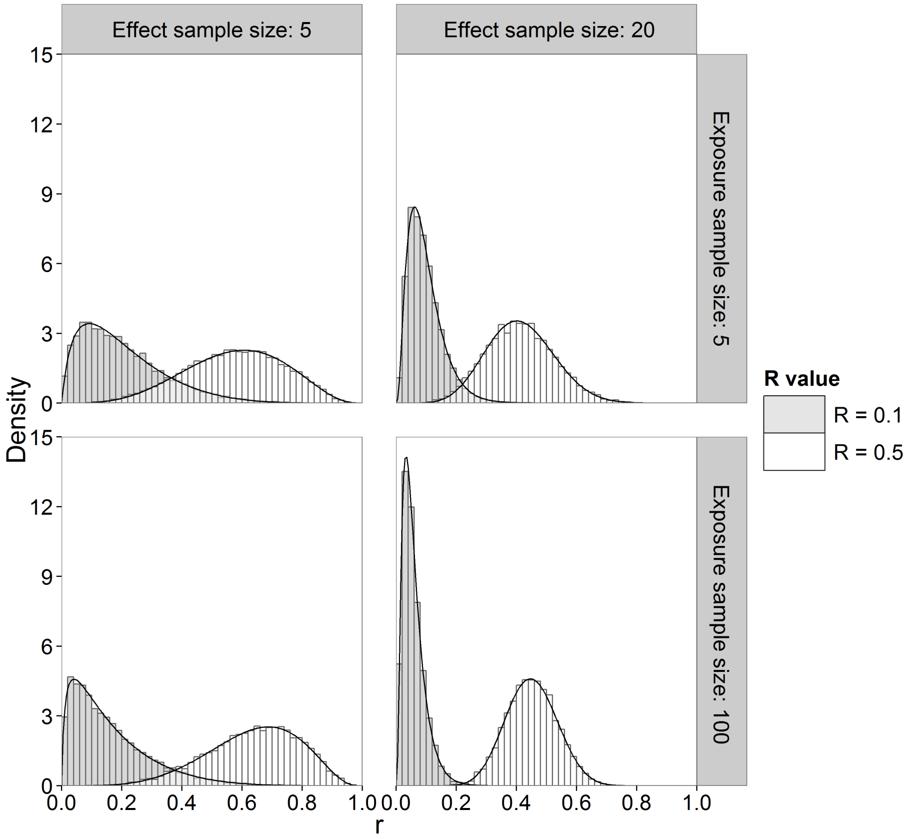

Figure 2 Histogram and theoretical posterior pdf (black solid line) of *R*. Sample sizes of 5 and 20 for effect concentrations, sample sizes of 5 and 100 for exposure concentrations and R value of 0.1 and 0.5.

In this paper, we obtained the posterior distribution of R by sampling, because it required less computing time than numerical integration in our implementation. The posterior mean is often taken as the Bayesian point estimator but we also investigated the posterior median and mode as point estimators of R.

Non-parametric estimator

As a benchmark comparison for the parametric estimators, we included a basic non-parametric estimator. This estimator, $\hat{R}_{\mathrm{NP}}$, is calculated from the data without any distributional assumptions by

$$\hat{R}_{\mathrm{NP}_0} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \left[I(x_i > y_j) + \frac{1}{2} I(x_i = y_j) \right] \tag{4}$$

where $I(\mathcal{S}) = 1$ if $\mathcal{S}$ is true and 0 otherwise (*Krzanowski & Hand, 2009*, p. 65). Alternatively, Eq. (4) can be written as

$$\hat{R}_{\mathrm{NP}_0} = \frac{U}{n_x n_y}$$

where U is the Mann–Whitney statistic (*Gibbons & Chakraborti, 2011*). Equation (4) is also known as the area under the ROC curve and its equivalence to the Mann–Whitney statistic has been shown (*Bamber, 1975*).

The non-parametric estimator of R is related to estimating the success probability, p, in a binomial experiment. As noted in the Introduction, we encounter the zero problem. One possible solution is making use of Laplace's Law of Succession (*Zabell, 1989*). This law states that given k successes in n trials of a binomial experiment, the probability of a success on the next trial is $\frac{k+1}{n+2}$. The validity of this expression has a Bayesian basis. On assuming a uniform prior for p, the posterior distribution of p is a $Beta(k+1, n-k+1)$ distribution (*Gelman et al., 2014*, p. 30), so that the posterior mean is $\frac{k+1}{n+2}$. This expression, denoted $\hat{R}_{\mathrm{NP_{LLS}}}$, is then used instead of the estimator in Eq. (4). Note that the posterior mean is equal to the predictive probability of a success on the next trial.

An alternative solution is to replace the zero with some non-zero value. One option is to estimate the probability of an outcome outside the range of the data as $\frac{1}{2n_x n_y}$. This method is used by Matlab and Genstat to compute quantiles. Another alternative is to use $\frac{1}{n_x n_y+1}$ which is used by Minitab and SPSS.

Interval estimation

We propose interval estimators by calculating credible intervals for Bayesian methods and confidence intervals for others. For the Bayesian estimator, we calculated 90% two-sided highest posterior density (HPD) credible intervals (*Box & Tiao, 1973*, p. 123). These intervals are obtained by finding the interval of the posterior distribution with the highest density, for which we used the 'HPDinterval' function in the 'coda' package in R (*Plummer et al., 2006*). HPD intervals produce the shortest intervals on a chosen scale, e.g., R or a transformation thereof, but are not transformation invariant. To estimate an upper credible bound of the risk, we also calculated the 95% percentile of the posterior. The upper bound of the 90% two-sided HPD intervals is not necessarily equal to the 95% percentile as the probabilities to the left and right of the two-sided HPD interval can be unequal.

For the non-Bayesian estimators, we calculated 90% Bias corrected and accelerated (BCa) parametric bootstrap confidence intervals using the 'boot' package in R (*Canty & Ripley, 2013*; *Davison & Hinkley, 1997*) with 1,000 bootstrap samples. For the non-parametric estimator, the BCa interval algorithm did not converge for small R values and also had some difficulty with the small sample sizes. For these cases we calculated percentile confidence intervals. The percentile method obtains a symmetric $100(1-\alpha)\%$ confidence interval by calculating the $\left(\frac{\alpha}{2}\right)$th and $\left(1-\frac{\alpha}{2}\right)$th percentiles of the bootstrap sample. In the BCa method, these percentiles are adjusted to correct for bias and skewness. For symmetric distributions, the percentile and BCa intervals are equal. Both intervals are also transformation invariant (*Efron & Tibshirani, 1993*, p. 175, 187). When calculating the confidence intervals for small risks, all bootstrap values may be equal, resulting in a zero interval length. For the BCa and percentile interval, the upper bound of a 90% two-sided interval is equal to the upper bound of a 95% one-sided interval

Table 1 Substances, sample sizes and estimated risks in an environmental risk assessment performed by *Gottschalk, Kost & Nowack (2013)*. Sample size of effect concentration data are given for aquatic and soil toxicity. Risks are given for four environmental compartments.

Substance	Sample size		Risks			
	Aquatic	Soil	1	2	3	4
Ag	12	1	0.007	0.397	0	0
CNT	9	2	0	0	0	0
TiO_2	18	2	7.2e−13	0.187	0	1.2e−7
ZnO	17	2	0	0.011	0	0
Fullerenes	4		0	0	0	0

(*Carpenter & Bithell, 2000*). The upper 95% confidence bound is, therefore, trivially obtained from the 90% two-sided interval.

For the MLE-like estimators, we also calculated confidence intervals based on the noncentral t distribution (*Reiser & Guttman, 1986*). In this method, the sum of the two variances (s_x^2 and s_y^2) are approximated with a chi-squared distribution.

The upper confidence (credible) bounds are of special interest in the context of managing risks, as they indicate (with some certainty) that the risk will not be higher than the upper bound.

Simulation study

In this section, we discuss the design of the Monte Carlo (MC) simulation study following the guidelines provided in Burton et al. (2006).

Simulation setup

The simulation study is performed in R (*R Core Team, 2013*). We use the built-in `rnorm` function to sample from a normal distribution using the Mersenne-Twister pseudo-random number generator (*Matsumoto & Nishimura, 1998*). Starting seeds for the different scenarios were drawn from a discrete uniform distribution to produce independent samples for each sample size scenario. The four estimators are calculated on the same sample, thereby avoiding differences among the estimators due to sampling.

To make the simulation as realistic as possible, we chose scenarios that are in line with recent studies of environmental risk assessment. When exposures are measured, it is common to have small sample sizes (*Johnson et al., 2011*; *Westerhoff et al., 2011*), whereas any number of exposure values can be obtained when they are modeled (*Gottschalk, Kost & Nowack, 2013*). For the exposure sample size, therefore, we chose two scenarios: the case of a small number of exposures ($n_x = 5$) and the case of a (relatively) large number of exposures ($n_x = 100$). We chose sample size of effect concentrations and risks loosely suggested by data from *Gottschalk, Kost & Nowack (2013)*. From this data (Table 1), we chose the following scenarios:

- Sample sizes for effect concentrations (n_y): 2, 5, 12, 20, 100
- risks: a grid of values from 1e−14 to 0.5.

The sample size of 100 for effect concentrations was added to study the influence of a large sample size. A risk of 0.5 is obtained when $\mu_x = \mu_y$. We, therefore, chose increasing values of $\mu_x - \mu_y$ to obtain the required range of risks. Considering the standard deviations, we note that the standard deviation of the effect concentration data in the case study is 5.6 times larger than that of the exposure concentration data. Based on this, we chose three scenarios: $\sigma_y = \sigma_x$, $\sigma_y = \frac{1}{5}\sigma_x$ and $\sigma_y = 5\sigma_x$.

The number of simulations was determined by running a pilot simulation (1,000 simulations) for the MLE. From this pilot, we obtained the median empirical standard deviation ($sd = 0.0719496$) of $\hat{R}_{\text{MLE}}$ and the median absolute bias ($\delta = 0.002107476$) in $\hat{R}_{\text{MLE}}$ over all scenarios. These were used to calculate the number of simulations, B, according to *Burton et al. (2006)*

$$\begin{aligned} B &= \left(\frac{z_{0.95} sd}{\delta}\right)^2 \\ &= \left(\frac{1.96 \cdot 0.0719496}{0.002107476}\right)^2 \\ &= 4477.58 \approx 4{,}500. \end{aligned}$$

For each of the 4,500 MC simulations, we calculated $\hat{R}_{\text{MLE}}$, $\hat{R}_{\text{QMLE}}$, $\hat{R}_{\text{Bayes}}$ and $\hat{R}_{\text{NP}}$. Due to the skewness of their sampling distributions, especially for small R values, we decided to use a transformation. Due to the nature of the analytical expression for R, (see Eq. (1)), a probit (inverse standard normal cdf) transformation is a natural choice. Some comparisons between the original scale and the probit transformation are further discussed for the Bayesian case in Section 'Comparison of Bayesian point estimators'.

Simulations were run on a HP desktop computer running Microsoft Windows 7 with processor specification Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz, 3401 MHz, 4 Core(s), 8 Logical Processor(s). Three R-sessions were running the three cases $\sigma_y = \sigma_x$, $\sigma_y = \frac{1}{5}\sigma_x$ and $\sigma_y = 5\sigma_x$ simultaneously. The $\sigma_y = \sigma_x$ case took the longest with the following time (in hh: mm:sec) for each of the four estimators:

- MLE: 03:43:06.18 (bootstrap); 00:10:21.82 (noncentral t)
- QMLE: 03:35:36.44 (bootstrap); 00:10:13.30 (noncentral t)
- Bayes: 03:12:56 (sample size 10,000); 00:41:02.9 (sample size 1,000)
- Non-parametric: 19:13:17.12

The bootstrap of the MLE, QMLE and non-parametric estimator was the cause of the longer runtime. The runtime for the Bayesian estimator is directly related to the size of the posterior sample. All further results are given for the large sample case.

Performance measures

We calculated various performance measures to evaluate the performance of the four point estimators. We calculated the performance measures on the probit scale, so as to be able to highlight differences among methods for small values of R:

- mean: $\text{probit}(\bar{\hat{R}}) = \sum_{i=1}^{4{,}500} \frac{\text{probit}(\hat{R}_i)}{4{,}500}$
- absolute bias: $\text{bias} = \text{probit}(\bar{\hat{R}}) - \text{probit}(R)$
- empirical (or MC) standard deviation: $SD = \sqrt{\sum_{i=1}^{4{,}500} \frac{(\text{probit}(\hat{R}_i) - \text{probit}(\bar{\hat{R}}))^2}{4{,}500}}$
- root mean squared error: $\text{RMSE} = \sqrt{\text{bias}^2 + SD^2}$.

The quality of the interval estimators will be assessed by calculating the coverage probability for each scenario. Note that the confidence intervals we calculated are approximate and do not claim to deliver the correct coverage. In addition, Bayesian credible intervals also do not claim a coverage frequency. For each of the 4,500 simulations, we calculated the confidence (credible) intervals and calculated the proportion of intervals that contained the true R value. We also investigated lengths of confidence (credible) interval over the 4,500 simulations for each scenario and R value. Coverages and lengths of confidence intervals for the different non-parametric estimators were similar. We, therefore, only consider the estimator based on Laplace's Law of Succession.

SIMULATION RESULTS

All results given and discussed are for the scenario that most resembles the case study ($\sigma_y = 5\sigma_x$ and sample sizes $n_x = 100$ and $n_y = 12$) unless explicitly stated otherwise. Graphs and tables for the other scenarios are given in the Supplemental Information. All sampling distribution graphs plot the estimated probit($\hat{R}$) (or $\hat{R}$) value (y-axis) against the true probit(R) (or R) value (x-axis). A diagonal 1–1 line is drawn to indicate where $\text{probit}(\hat{R}) = \text{probit}(R)$ (or $\hat{R} = R$). A logarithmic scale is used when R is plotted.

Comparison of Bayesian point estimators

For Bayesian estimation, we considered three point estimators, namely, the posterior mean, median and mode. We summarize the sampling distribution of each by way of three quantiles (0.5 or median, 0.025 and 0.975). In Fig. 3 these quantiles are plotted as a function of the true value. Figures 3A and 3C show the median and Fig. 3B and 3D show the 0.025 and 0.975 quantiles.

In Figs. 3A and 3B, the quantiles are calculated from the sampling distribution of the estimators on the original scale ($\hat{R}$) and plotted on $\log_{10}$-scale. The lines for the posterior mean are above the 1:1 line, so indicating large positive bias. The lines for the posterior mode go to $\log_{10}(0) = -\infty$, due to the very skew posterior distributions for smaller risks (on original scale of R, as already illustrated in Fig. 2), so indicating large negative bias. The lines for the posterior median are in between and closer to the 1:1 line.

In Fig. 3C and 3D, the quantiles are calculated similarly but on the probit-transformed $\hat{R}$ and plotted on the probit scale as well. Here we see that quantiles of the posterior mean, median and mode almost coincide, indicating that the skewness problem is solved. Simulations for very small sample size of effect concentrations ($n_y = 2$) showed that the mean has a slight advantage because of narrower intervals between the 0.025 and 0.975 quantiles than that of the median and the mode. This difference, however, is very quickly lost for higher sample sizes ($n_y \geqslant 5$) as shown in Fig. S23.

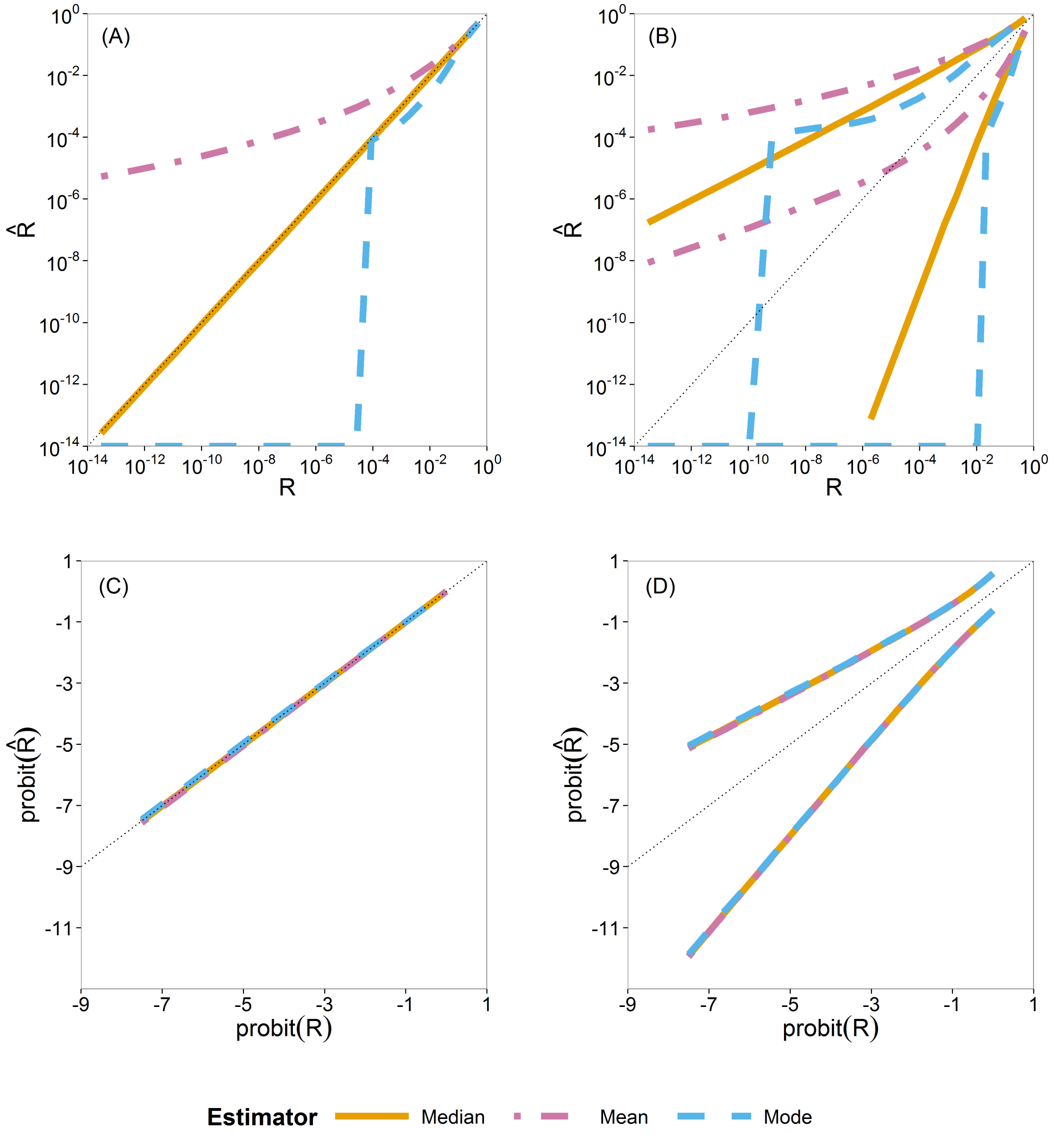

Figure 3 Quantiles of the sampling distribution of the three Bayesian point estimators (mean, median and mode). The median (A, C) and 0.025 and 0.975 quantiles (B, D) of the sampling distribution of the three Bayesian point estimators (mean, median and mode) calculated on the original scale (A, B) and on the probit scale (C, D). When plotted on $\log_{10}$-scale, a zero mode becomes $-\infty$. The diagonal dotted line represents the values where (probit($\hat{R}$) = probit(R) (or $\hat{R} = R$).

From this study of Bayesian point estimators on different scales, we see the advantage of the use of the probit scale for the Bayesian case. Moreover, for ease and its transformation invariance, we chose the posterior median as the Bayesian point estimator of R. The probit scale stretches out small values of R, making possible differences between methods more clearly visible for small R. On this basis, we decided to perform, for all estimators, all further calculations in the simulation study on the probit scale.

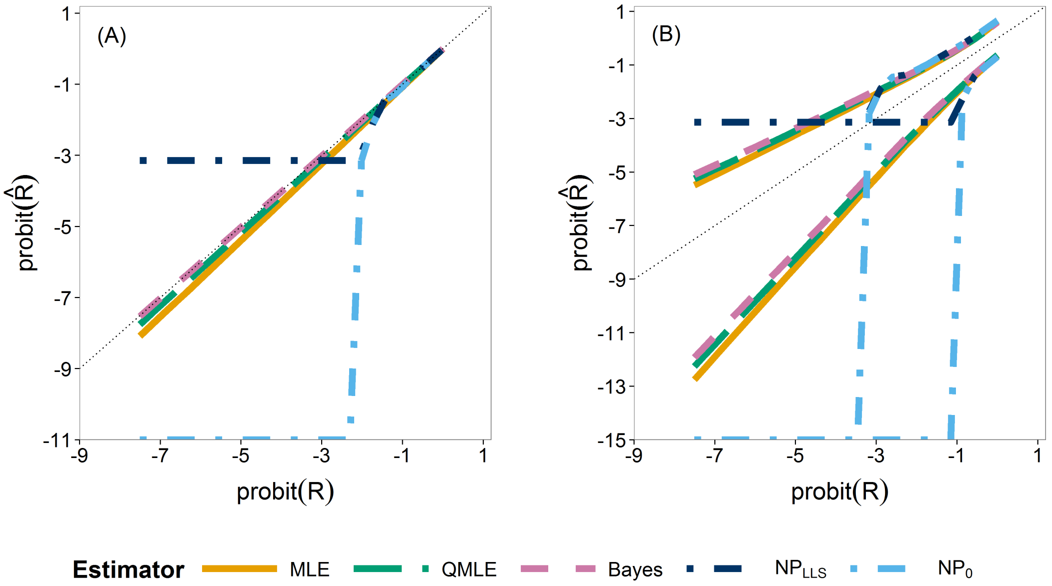

Figure 4 Quantiles of the sampling distribution of the point estimators. The median (A) and 0.025 and 0.975 quantiles (B) of the sampling distribution of the point estimators, $\hat{R}_{\text{MLE}}$, $\hat{R}_{\text{QMLE}}$, $\hat{R}_{\text{Bayes}}$, $\hat{R}_{\text{NP}_{\text{LLS}}}$ and $\hat{R}_{\text{NP}_0}$ calculated on the probit scale. The diagonal dotted line represents the values where $\text{probit}(\hat{R}) = \text{probit}(R)$.

Comparison of the four point and interval estimators

In this section, we show the simulation results for the four estimators.

We first compare the sampling distributions of the four estimators. Figure 4 illustrates the median (A) and 0.025 and 0.975 quantiles (B) of the sampling distributions of $\hat{R}_{\text{MLE}}$, $\hat{R}_{\text{QMLE}}$, $\hat{R}_{\text{Bayes}}$ and $\hat{R}_{\text{NP}}$. For $\hat{R}_{\text{NP}}$, we plotted both the standard estimator, $\hat{R}_{\text{NP}_0}$, (Eq. (4)) which goes to minus infinity on the probit scale and the Laplace version, $\hat{R}_{\text{NP}_{\text{LLS}}}$. These provide the extreme endpoints of the different solutions in solving the zero problem in the non-parametric estimator.

The median of the Bayesian estimator lies closest to the true R (Fig. 4A). This is especially apparent in scenarios with $n_y \leq 12$ (Fig. S24). The non-parametric estimators are clearly not able to estimate R for smaller values as they very quickly jump to their lower bound of either $\text{probit}\left(\frac{1}{n_x x n_y + 2}\right)$ or minus infinity, indicated by horizontal and vertical dash–dot lines respectively. As the sample sizes increase, the three parametric estimators converge (Figs. S24 and S25). The non-parametric estimators remain the worst estimators for all sample sizes when estimating small R values. In our further study, we consider the Laplace version only.

Next, we study the coverage and interval lengths of the two-sided 90% confidence (credible) intervals of the estimators on the probit scale. For each of the 4,500 simulations, we calculated interval lengths and then obtained the median interval length for each combination of estimator, sample sizes and risk value combinations. In order to compare the median interval lengths across different risk values in a single graph, we standardized each one by dividing by the true $\left|\text{probit}(R)\right|$ value to obtain the relative median interval length. Figure 5 plots the relative median interval length (y-axis) against the coverage probabilities (x-axis) for $n_x = 5$ and $n_x = 100$, all investigated sample sizes for effect concentrations, and all R values. Coverage probabilities of less than 0.5 were plotted at 0.5. The vertical line indicates a coverage probability of 90%. A good interval estimator gives points lying on this line with short interval length. This translates to good coverage and narrow intervals. Points corresponding to $n_y = 12$ are indicated by an open black circle.

We found that the MLE (not shown) and the QMLE had a similar pattern for both the bootstrap and noncentral t intervals, with the QMLE consistently having better coverage. Figure 5 shows that the bootstrap intervals have liberal coverage compared to the noncentral t intervals for small sample sizes of effect concentrations. As the sample sizes for effect concentrations increase (bigger dots), the estimators have better coverage. Very small sample size for effect concentrations ($n_y = 2$) gives the worst coverage. For the Bayesian estimator, the sample size has a lesser influence. For small exposure sample size ($n_x = 5$), the coverage of the Bayesian interval estimator tends to be too high. This problem largely disappears for $n_x = 100$, although there is some under-coverage for the $n_y = 2$ case. The parametric estimators have shorter interval lengths when $n_x = 100$ (right column) as compared to $n_x = 5$. For the non-parametric estimator, sample size has a slightly less systematic influence on the coverage.

Compared to the other estimators, the Bayesian interval estimator best maintains the nominal coverage without having larger median interval length and despite the fact of often having a higher than nominal coverage (Fig. 5). The non-Bayesian estimators have smaller than nominal coverage for small sample size of effect concentrations with the non-parametric estimator being the worst. For better comparison of interval lengths among the estimators, the reader is referred to Fig. S26.

In risk assessment, one is often interested in an upper bound on the risk. We studied the coverages and interval lengths of the upper 95% confidence (credible) bounds of the estimators on the probit scale. The interval lengths were quantified as the difference between the upper bound and the true probit(R) value. The median of the 4,500 differences was obtained. In order to compare the median differences across different risk values in a single graph, each median difference was divided by the true $|\text{probit}(R)|$ value being estimated to obtain the relative median difference. Figure 6 plots the relative median difference (y-axis) against the coverage probabilities (x-axis) for $n_x = 5$ and $n_x = 100$, all investigated sample sizes for effect concentrations, and all R values. Coverage probabilities of less than 0.5 were plotted at 0.5. The vertical line specifies a coverage probability of 95%. A good upper bound estimator gives points lying on this vertical line and being close to the horizontal 0 line. This translates to good coverage and an upper bound close to the true R. Points corresponding to $n_y = 12$ are indicated by an open black circle.

We see similar patterns as in the case of the two-sided intervals, with the bootstrap intervals being too liberal. The Bayes estimator gives higher than nominal coverage for small exposure sample size (left column). For large exposure sample size, the Bayesian estimator clearly outperforms the other estimators with good coverage for all R values and

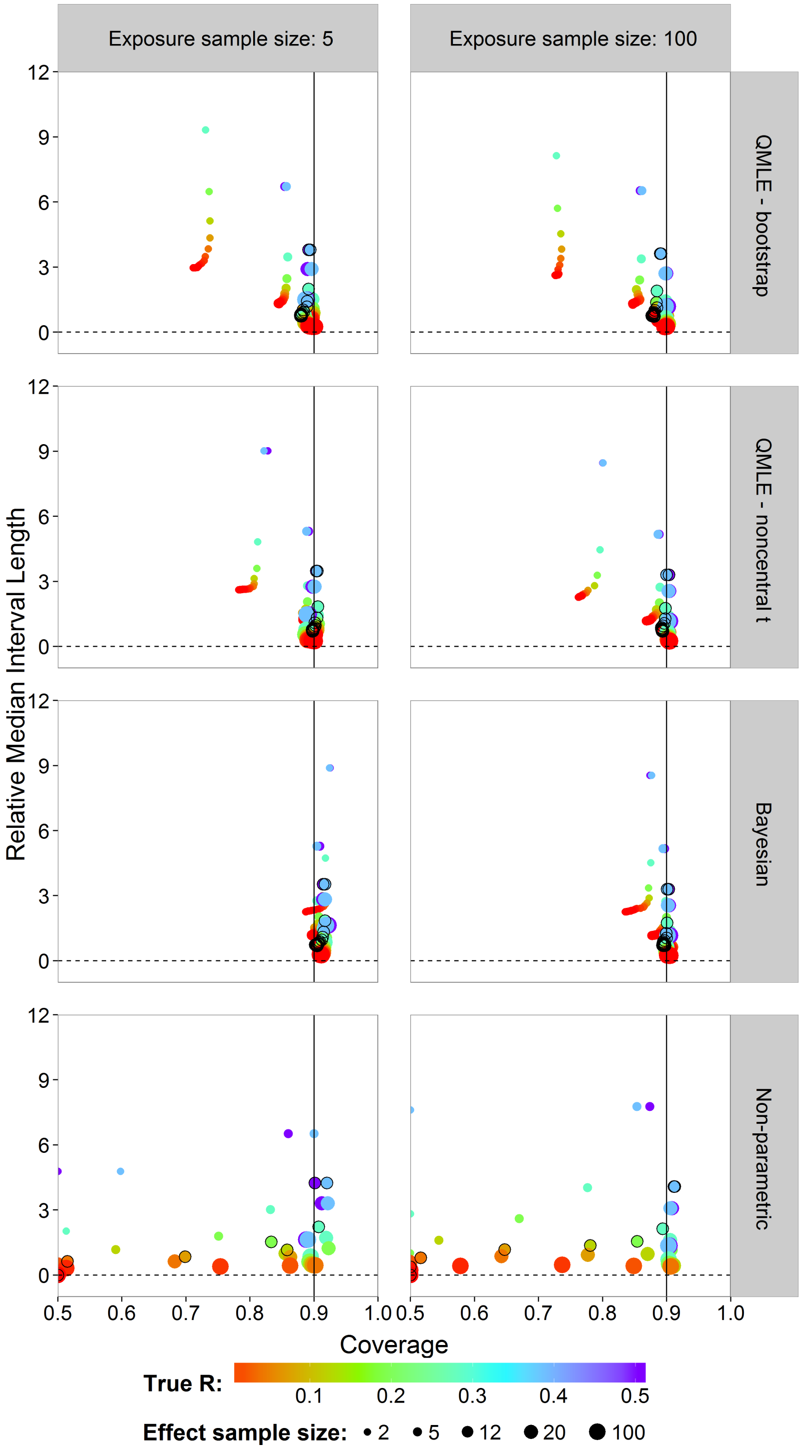

Figure 5 Scatterplots of the 90% two-sided coverage probabilities against the relative median interval length calculated on the probit scale. The value of the true *R* value is illustrated by the color scale. The size of the dots corresponds to the size of the sample size of effect concentrations. A vertical reference line is drawn at a coverage probability of 90%. The points corresponding to $n_y = 12$ are indicated by an open black circle.

sample sizes for effect concentrations without having larger median interval difference. The non-parametric estimator has a severe coverage problem.

The results for the performance measures of the different estimators are given in Tables S2–S5. Due to the lower bound of the non-parametric estimator, the SD, bias and RMSE are not reliable for small R values. Only for a few cases where $R = 0.5$, the non-parametric estimator has slightly lower SD and bias than the parametric estimators. The various graphs have also shown the inability of the non-parametric estimator to estimate small R values.

Among the parametric estimators, the Bayesian estimator as the smallest SD, bias and RMSE on probit scale for all sample sizes and R values. This confirms that the Bayesian estimator is better than the non-Bayesian estimators as also seen for the interval estimator case. The QMLE has smaller SD, bias and RMSE than the MLE for all sample sizes and R values. This also confirms the results of the interval estimators where QMLE has better coverage than MLE.

The Bayesian estimator was in general the best estimator and specifically so for the scenario that is closest to the case study. The Bayesian point estimator was less biased than the MLE and the QMLE in all cases ($\sigma_y = \sigma_x$ (Figs. S8 and S9), $\sigma_y = \frac{1}{5}\sigma_x$ (Figs. S16 and S17) and $\sigma_y = 5\sigma_x$ (Figs. S24 and S25)), and this was especially apparent for small sample sizes for effect concentration ($n_y \leq 12$) and small R values. The Bayesian interval estimator (90% two-sided) had better coverage, with even higher than nominal coverage for exposure sample size $n_x = 5$. This is also seen for the case $\sigma_y = \sigma_x$ (Fig. S10). For the case $\sigma_y = \frac{1}{5}\sigma_x$ (Fig. S18), the higher coverage is only seen for small sample size for effect concentrations as seen by the small dots. When considering the 95% upper bound of the Bayesian estimator compared to MLE and QMLE, we also see better coverage with similar higher than nominal coverage for exposure sample size $n_x = 5$. This is similar for the $\sigma_y = \sigma_x$ case (Fig. S11) and slightly more pronounced for the $\sigma_y = \frac{1}{5}\sigma_x$ case (Fig. S19). Considering the performance measures, the Bayesian estimator performs better (lower values), also for the $\sigma_y = \sigma_x$ case (Tables S2–S4) and the $\sigma_y = \frac{1}{5}\sigma_x$ case (Table S6–S8). For the corresponding case study scenario of $\sigma_y = 5\sigma_x$ and $n_x = 100, n_y = 12$, the Bayesian estimator clearly outperformed the MLE and QMLE. It was less biased and maintained the nominal coverage in both the two-sided and one-sided cases. For better comparison of interval lengths among the estimators, the reader is referred to Fig. S27.

CASE STUDY

In this section, we evaluate the case study results on the basis of the simulation study results. In the case study, we have a sample size of 1,000 of exposure concentrations and a sample size of 12 of effect concentrations. We note that the exposure concentrations come from a simulation model, so that it is possible to generate an arbitrary large sample exposure concentrations. We treat the size of 1,000 exposures as being effectively of size 100.

First, we verify that the normal–normal model is not in conflict with the data. Visually, the normal distribution fits the concentration data quite well (Fig. 1). The small sample size of the effect concentrations gives low power to any formal normality tests, where

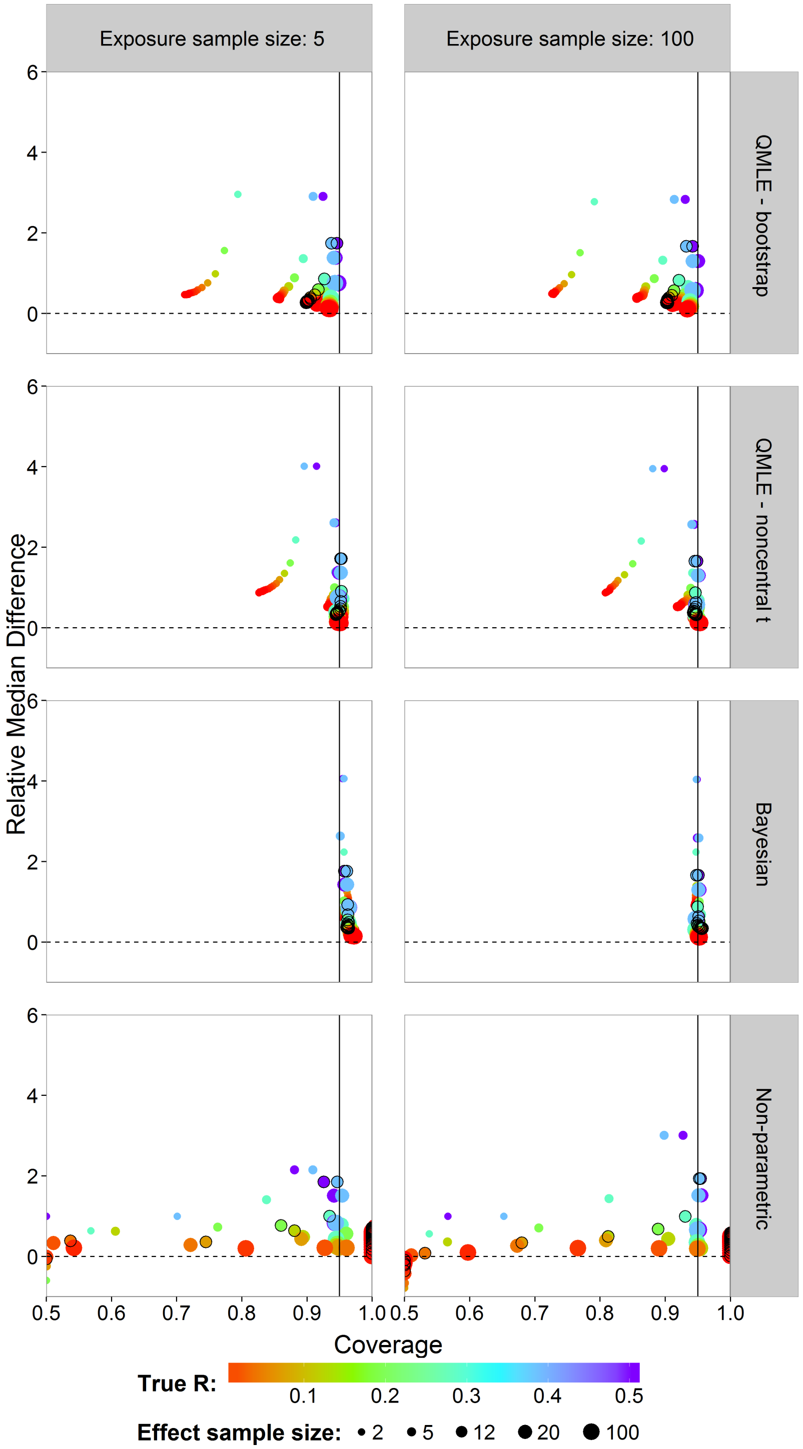

Figure 6 Scatterplots of the 95% one-sided coverage probabilities against the relative median difference calculated on the probit scale. The value of the true R value is illustrated by the color scale. The size of the dots corresponds to the size of the sample size of effect. A vertical reference line is drawn at a coverage probability of 95%. The points corresponding to $n_y = 12$ are indicated by a black circle.

Table 2 Estimated risks ($\hat{R}$), 90% two-sided confidence (credible) intervals (CI) and 95% upper confidence (credible) bounds (CB) for the MLE, QMLE (bootstrap and noncentral *t*), Bayesian and non-parametric estimator.

Estimator	$\hat{R}$	90% 2-sided CI	95% upper CB
MLE	0.0068	0.0003–0.0684	0.0684
QMLE (noncentral *t*)	0.0090	0.0006–0.0784	0.0784
QMLE (bootstrap)	0.0090	0.0002–0.0571	0.0571
Bayes	0.0108	0.0006–0.0776	0.0806
Empirical	0.0001	0.0001–0.0001	0.0001

the large sample size of exposure concentrations gives high power, so that even small deviations from normality are detected. Even so, we cannot reject the null hypothesis of normality (see Table S1) for either the effect or the exposure samples at a 5% significance level. In the exposure concentration data, there is some indication for non-normality evident from two of the normality tests which are only just not significant (*p*-values of 0.0564 and 0.0538). Even so, we take the normal–normal model as a useful model.

Next, we consider the estimates of the risk (Table 2). The estimates and intervals were calculated on the probit-scale and then transformed back to the original scale so as to be able to evaluate the case study results in the light of the simulation study results. For the MLE we calculated the interval estimators based on the noncentral *t* distribution and for the QMLE, the noncentral *t* and parametric bootstrap .

The non-parametric estimator was calculated using Laplace's Law of Succession. For the sample sizes of this case study, $\hat{R}_{\mathrm{NP_{LLS}}}$ then becomes $\frac{0+1}{1{,}000\cdot 12+2} = \frac{1}{12{,}002} = 0.000083$. We note, however, that this value is very much dependent on the sample size. A larger exposure sample size will decrease the estimate. As seen in the simulation study results, it is impossible to draw any meaningful conclusions for small risks based on the non-parametric estimator.

The three parametric estimates are similar. The bootstrap 90% confidence interval of the QMLE is clearly narrower. From the simulation study, however, the bootstrap intervals showed liberal coverage and are, therefore, less trustworthy. Considering the 95% upper confidence bound, we note that the Bayesian bound is slightly higher than that of the MLE and QMLE and higher as well than the upper bound of the Bayesian 90% credible interval. Investigating these aspects in the simulation results, we found that these differences are to be expected (see Figs. S1–S3), although the difference between the Bayesian 95% upper bound and the upper bound of its 90% credible interval is not so typical. The distances between the 95% upper bound and $\hat{R}$ as well as the ratio of the Bayesian upper bound to both the QMLE upper bound and the Bayesian upper bound of the two-sided interval fall within the respective sampling distributions as obtained in the simulations.

Based on the simulation results we, therefore, conclude that the Bayesian estimate is the most appropriate. The upper bound (0.0806) is most reliable as it has the best coverage (compared to MLE and QMLE). This is clearly seen by the black circles in the Bayesian panel in the right column of Fig. 6. This case corresponds most closely to the case study data. Based on the model and the data used, we state with 95% confidence, that the risk will not be greater than 0.0806.

DISCUSSION AND CONCLUSION

In this paper we studied the problem of estimating the risk for the case of small sample size for effect concentrations and small R values. The case study data showed discrepancies between the parametric and non-parametric estimators which we investigated via a simulation study. We derived and compared three parametric estimators and one non-parametric estimator for the risk. This was done under the assumption of normality for both the exposure and effect concentration data. We found that, overall, the parametric estimators have better performance than the non-parametric estimator, and the Bayesian estimator outperformed the maximum likelihood-based ones.

The Bayesian estimator in this paper was based on a non-informative prior on the underlying parameters. This resulted in a prior tendency of R toward 0.5. For small sample sizes, there was not enough data to counter this prior tendency and this resulted in an overestimation of R by the posterior mean estimator calculated on the original R scale (Figs. S4, S5, S12, S13, S20 and S21). To overcome this problem, it was needed to switch to the posterior median estimator or to switch to the probit(R) scale. We used both the probit-scale and the posterior median resulting in an estimator that outperformed its parametric counterparts. More benefit can presumably be obtained from the Bayesian estimator if we can use an informative prior, at least when the prior is not in conflict with the data. In addition, the use of probability matching priors (*Datta & Sweeting, 2005*) may also improve on the coverage of the credible intervals. *Ventura & Racugno (2011)* used a strong matching prior for Bayesian estimation of $P(X > Y)$ based on a profile-likelihood approach.

Using the probit(R) scale in the simulation study enabled us to more easily compare the estimators for small R values. Despite giving nice statistical properties, the probit scale may not directly address a specific risk assessment question.

Comparing the parametric bootstrap and noncentral t interval estimators for the MLE and the QMLE, we found the noncentral t intervals to have better coverage. The bootstrap intervals, although a good alternative, are liberal in coverage (i.e., resulted in smaller than nominal coverage) for small sample sizes of effect concentrations. This was also found by *Tian (2008)*.

It was clearly seen that the non-parametric estimator was not able to estimate the risk for small sample sizes and small R values. For R values above the lower bound of probit$\left(\frac{1}{n_x n_y+2}\right)$, the non-parametric estimator had performance measures similar to that of the MLE. As seen in our case study, however, the non-parametric estimator failed completely. The bootstrap cannot provide any variability of outcome with which to provide an interval for the estimate. In the simulation study, we also found that for small sample sizes, there was often too little variability in the data for the bootstrap to be able to quantify it (Fig. 4B). Although this translated to 0 coverage in Figs. 5 and 6, it really shows that the non-parametric estimator completely fails in these cases. For small sample sizes and small R values, therefore, we advise to use parametric estimators.

Considering the computation times of the simulation study, we note that, in addition to the Bayesian estimator being the best estimator, it can also require shorter computation time compared to the bootstrap alternatives depending on the posterior sample size. The larger posterior sample size (10,000) tends to result in slightly narrower estimates of the posterior distribution than those based on the smaller sample size (1,000). Nevertheless, the main results in Figs. 5 and 6 remain basically unchanged. In the case study, the credible interval becomes slightly wider for the larger sample size and the upper bound is slightly lower. Even so, not much is lost by taking the smaller sample size and this drastically reduces the computation time. The maximum likelihood based estimators have shorter computation time when calculating the interval estimators based on the noncentral t distribution than the bayesian estimator. The non-parametric estimator is by far the most computationally demanding due to the bootstrapping and the calculation of Eq. (4).

Assuming normality in the case and simulation study may seem as a strict assumption and going non-parametric is a way to avoid strict assumptions. For many situations in statistics the normal distribution is considered to have too light tails. In our case with very little data, going non-parametric leads to zero tails outside the range of the data. The usual area-under-the-curve-based non-parametric method can then severely underestimate the risk (often resulting in zero risk), whereas the estimate based on Laplace's Law of Succession overestimates the risk for small true risks. To be able to draw any sensible conclusion, one has to use a parametric method. Our comparison of methods shows the advantage of using parametric methods in this case.

We conclude that making parametric assumptions, enabled us to estimate the risk for smaller sample sizes and small risks in the case the data is in fact normally distributed. Further research is needed to investigate the robustness of the parametric methods on non-normal data. We need to investigate whether semi-parametric methods and methods based on the extreme value distribution are able to estimate the tails of distributions sufficiently well from small data sets, so that they outperform the parametric methods used in this paper.

ACKNOWLEDGEMENTS

We thank Fadri Gottschalk for providing the data used in the application and Peter Craig and an anomymous reviewer for their comments that improved the manuscript.

APPENDIX

Result A.1 Transformation method for obtaining the pdf of a function, $R = \Phi(\theta)$, from the pdf of θ.

Let $\theta \sim N(\mu,\sigma)$.

Then the pdf of R is given by $f_R(r) = f_\theta(r)|J(\theta \to R)|$. We first obtain the Jacobian, $J(\theta \to R)$:

$$\begin{aligned} J(\theta \to R) &= \frac{d\theta}{dR} \\ &= \left[\left. \frac{dR}{d\theta} \right|_{\theta = \Phi^{-1}(r)} \right]^{-1} \\ &= \left[\Phi'(\theta) \big|_{\theta = \Phi^{-1}(r)} \right]^{-1} \\ &= \left[\phi(\Phi^{-1}(r)) \right]^{-1} \\ &= \frac{1}{\phi\left(\Phi^{-1}(r)\right)} \end{aligned} \tag{A.1}$$

where ϕ denotes the pdf of the standard normal distribution.

We then obtain $f_R(r)$:

$$\begin{aligned} f_R(r) &= f_\theta(r)|J(\theta \to R)| \\ &= f_\theta(\Phi^{-1}(r)) \frac{1}{\phi\left(\Phi^{-1}(r)\right)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(\Phi^{-1}(r) - \mu\right)^2}{2\sigma^2} \right] \frac{1}{\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\left(\Phi^{-1}(r)\right)^2}{2} \right]} \\ &= \frac{1}{\sqrt{\sigma^2}} \exp\left[-\frac{\left(\Phi^{-1}(r) - \mu\right)^2}{2\sigma^2} + \frac{\left(\Phi^{-1}(r)\right)^2}{2} \right]. \end{aligned} \tag{A.2}$$

Result A.2 From the conditional posterior distribution of $\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}$ given by

$$\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}} \mid \sigma_x^2, \sigma_y^2 \sim N\left(\frac{\bar{x} - \bar{y}}{\sqrt{\sigma_x^2 + \sigma_y^2}}, \sqrt{\frac{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}{\sigma_x^2 + \sigma_y^2}} \right),$$

we obtain the conditional posterior distribution of R (using Result A.1):

$$f_{R|\sigma_x^2,\sigma_y^2}(r \mid \sigma_x^2, \sigma_y^2) = \left[\frac{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}{\sigma_x^2 + \sigma_y^2} \right]^{-\frac{1}{2}} \exp\left[-\frac{\left(\Phi^{-1}(r) - \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_x^2 + \sigma_y^2}} \right)^2}{2 \frac{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}{\sigma_x^2 + \sigma_y^2}} + \frac{(\Phi^{-1}(r))^2}{2} \right]. \tag{A.3}$$

To obtain the marginal posterior density, $f_R(r)$, we integrate σ_x^2 and σ_y^2 out of the joint pdf, $f(r,\sigma_x^2,\sigma_y^2)$, and obtain the required result.

$$\begin{aligned}
f_R(r) &= \int_0^\infty \int_0^\infty f(r,\sigma_x^2,\sigma_y^2)\,d\sigma_x^2\,d\sigma_y^2\\
&= \int_0^\infty \int_0^\infty f(r \mid \sigma_x^2,\sigma_y^2) f(\sigma_x^2) f(\sigma_y^2)\,d\sigma_x^2\,d\sigma_y^2\\
&= \int_0^\infty \int_0^\infty \left[\frac{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}{\sigma_x^2+\sigma_y^2}\right]^{-\frac{1}{2}} \exp\left[-\frac{\left(\Phi^{-1}(r)-\frac{\bar{x}-\bar{y}}{\sqrt{\sigma_x^2+\sigma_y^2}}\right)^2}{2\frac{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}{\sigma_x^2+\sigma_y^2}}+\frac{(\Phi^{-1}(r))^2}{2}\right]\\
&\quad\times \frac{b_x^{a_x}}{\Gamma(a_x)}(\sigma_x^2)^{-a_x-1}\exp\left[-\frac{b_x}{\sigma_x^2}\right]\frac{b_y^{a_y}}{\Gamma(a_y)}(\sigma_y^2)^{-a_y-1}\exp\left[-\frac{b_y}{\sigma_y^2}\right]d\sigma_x^2\,d\sigma_y^2 \qquad \text{(A.4)}
\end{aligned}$$

with

$$\begin{aligned}
a_x &= \frac{n_x-1}{2}\\
a_y &= \frac{n_y-1}{2}\\
b_x &= \frac{(n_x-1)s_x^2}{2}\\
b_y &= \frac{(n_y-1)s_y^2}{2}.
\end{aligned}$$

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work is supported by NanoNextNL, a micro and nanotechnology consortium of the Government of the Netherlands and 130 partners. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
NanoNextNL.

Competing Interests

Cajo J.F. ter Braak is an Academic Editor for PeerJ.

Author Contributions

- Rianne Jacobs analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Andriëtte A. Bekker and Hilko van der Voet contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.

- Cajo J.F. ter Braak analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.1164#supplemental-information.

REFERENCES

Aldenberg T, Jaworska JS, Traas TP. 2002. Normal species sensitivity distributions and probabilistic ecological risk assessment. In: *Species sensitivity distributions in ecotoxicology*. London: CRC Press, 49–102.

Bain LJ, Engelhardt M. 1992. *Introduction to probability and mathematical statistics*. Second edition. Belmont, CA: Duxbury Press.

Bamber D. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**:387–415 DOI 10.1016/0022-2496(75)90001-2.

Barbiero A. 2011. Confidence intervals for reliability of stress-strength models in the normal case. *Communications in Statistics—Simulation and Computation* **40(6)**:907–925 DOI 10.1080/03610918.2011.560728.

Box GEP, Tiao GC. 1973. *Bayesian inference in statistical analysis*. New York: John Wiley and Sons.

Burton A, Altman DG, Royston P, Holder RL. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine* **25**:4279–4292 DOI 10.1002/sim.2673.

Canty A, Ripley B. 2013. *Boot: bootstrap R (S-plus) functions*. R package version 1.3-17.

Cardwell RD, Parkhurst BR, Warren-Hicks W, Volosin J. 1993. Aquatic ecological risk. *Water Environment and Technology* **5**:47–51.

Carpenter J, Bithell J. 2000. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine* **19(August 1999)**:1141–1164 DOI 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.

Church J, Harris B. 1970. The estimation of reliability from stress-strength relationships. *Technometrics* **12(1)**:49–54 DOI 10.1080/00401706.1970.10488633.

Datta GS, Sweeting TJ. 2005. Probability matching priors. *Handbook of Statistics* **25(252)**:91–114.

Davison A, Hinkley D. 1997. *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.

Downtown F. 1973. The estimation of $Pr(Y > X)$ in the normal case. *Technometrics* **15(3)**:551–558.

ECHA. 2012. Guidance on information requirements and chemical safety assessment Chapter R.19: uncertainty analysis. Technical Report. European Chemicals Agency.

ECOFRAM. 1999. Ecological committee on FIFRA risk assessment methods aquatic report, peer review draft. Technical Report 508. US Environmental Protection Agency, Washington, D.C., USA.

Efron B, Tibshirani RJ. 1993. *An introduction to the bootstrap*. New York: Chapman and Hall.

Enis P, Geisser S. 1971. Estimation of the probability that $Y < X$. *Journal of the American Dietetic Association* **66(333)**:162–168 DOI 10.1080/01621459.1971.10482238.

Gelman A, Carlin JB, Stern HS, Dunson DD, Vehtari A, Rubin DB. 2014. *Bayesian data analysis*. Third edition. Boca Raton: CRC Press.

Gibbons JD, Chakraborti S. 2011. *Nonparametric statistical inference*. Fifth edition. Boca Raton: CRC Press.

Gottschalk F, Kost E, Nowack B. 2013. Engineered nanomaterials in water and soils: a risk quantification based on probabilistic exposure and effect modeling. *Environmental Toxicology and Chemistry* **32(6)**:1278–1287 DOI 10.1002/etc.2177.

Gottschalk F, Scholz R, Nowack B. 2010. Probabilistic material flow modeling for assessing the environmental exposure to compounds: methodology and an application to engineered nano-TiO_2 particles. *Environmental Modelling and Software* **25(3)**:320–332 DOI 10.1016/j.envsoft.2009.08.011.

Govidarajulu Z. 1967. Two-Sided confidence limits for $P(X < Y)$ based on normal samples of X and Y. *Sankhya: The Indian Journal of Statistics, Series B* **29**:35–40.

Johnson AC, Bowes MJ, Crossley A, Jarvie HP, Jurkschat K, Jürgens MD, Lawlor AJ, Park B, Rowland P, Spurgeon D, Svendsen C, Thompson IP, Barnes RJ, Williams RJ, Xu N. 2011. An assessment of the fate, behaviour and environmental risk associated with sunscreen TiO_2 nanoparticles in UK field scenarios. *Science of the Total Environment* **409(13)**:2503–2510 DOI 10.1016/j.scitotenv.2011.03.040.

Krzanowski WJ, Hand DJ. 2009. *ROC curves for continuous data*. Boca Raton: CRC Press.

Kundu D, Gupta RD. 2006. Estimation of $P[Y < X]$ for weibull distribution. *IEEE Transactions on Reliability* **55(2)**:270–280 DOI 10.1109/TR.2006.874918.

Lesaffre E, Lawson AB. 2012. *Bayesian biostatistics*. Chichester: John Wiley and Sons.

Li J, Ma S. 2011. Time-dependent ROC analysis under diverse censoring patterns. *Statistics in Medicine* **30(11)**:1266–1277 DOI 10.1002/sim.4033.

Lindley DV. 1971. *Making decisions*. London: Wiley.

Lindley DV. 2006. *Understanding uncertainty*. Hoboken: John Wiley and Sons.

Matsumoto M, Nishimura T. 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* **8(1)**:3–30 DOI 10.1145/272991.272995.

Mokhlis N. 2005. Reliability of a stress-strength model with burr type III distributions. *Communications in Statistics—Theory and Methods* **34(7)**:1643–1657 DOI 10.1081/STA-200063183.

Nadar M, Kızılaslan F, Papadopoulos A. 2014. Classical and Bayesian estimation of $P(Y < X)$ for Kumaraswamy's distribution. *Journal of Statistical Computation and Simulation* **84(7)**:1505–1529 DOI 10.1080/00949655.2012.750658.

Nandi S, Aich A. 1996. Hypothesis-test for reliability in a stress–strength model, with prior information. *IEEE Transactions on Reliability* **45(1)**:129–131 DOI 10.1109/24.488929.

Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**:7–10.

R Core Team. 2013. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. *Available at http://www.R-project.org/*.

Reiser B, Guttman I. 1986. Statistical inference for $Pr(Y < X)$: the normal case. *Technometrics* **28(3)**:253–257.

Solomon K, Giesy J, Jones P. 2000. Probabilistic risk assessment of agrochemicals in the environment. *Crop Protection* **19(8–10)**:649–655 DOI 10.1016/S0261-2194(00)00086-7.

Solomon K, Takacs P. 2002. Probabilistic risk assessment using species sensitivity distributions. In: *Species sensitivity distributions in ecotoxicology, Environmental and ecological risk assessment.* Boca Raton: CRC Press, 285–313.

Suter GW, Vaughan DS, Gardner RH. 1983. Risk assessment by analysis of extrapolation error: a demonstration for effects of pollutants on fish. *Environmental Toxicology and Chemistry* **2**(**3**):369–378 DOI 10.1002/etc.5620020313.

Tian L. 2008. Confidence intervals for $P(Y_1 < X_2)$ with normal outcomes in linear models. *Statistics in Medicine* **27**:4221–4237 DOI 10.1002/sim.3290.

Van Straalen N. 2002. Theory of ecological risk assessment based on species sensitivity distributions. In: *Species sensitivity distributions in ecotoxicology*, *Environmental and ecological risk assessment.* Boca Raton: CRC Press, 37–48.

Ventura L, Racugno W. 2011. Recent advances on Bayesian inference for $P(X < Y)$. *Bayesian Analysis* **6**(**3**):411–428 DOI 10.1214/ba/1339616470.

Verdonck FAM, Aldenberg T, Jaworska J, Vanrolleghem PA. 2003. Limitations of current risk characterization methods in probabilistic environmental risk assessment. *Environmental Toxicology and Chemistry* **22**(**9**):2209–2213 DOI 10.1897/02-435.

Voinov V. 1986. Unbiased estimation of $P(Y < X)$ in the normal case. *Journal of Soviet Mathematics* **33**(**1**):701–706 DOI 10.1007/BF01091435.

Wagner C, Lø kke H. 1991. Estimation of ecotoxicological protection levels from NOEC toxicity data. *Water Research* **25**(**10**):1237–1242 DOI 10.1016/0043-1354(91)90062-U.

Warren-Hicks W, Parkhurst B, Butcher J. 2002. Methodology for aquatic ecological risk assessment. In: *Species sensitivity distributions in ecotoxicology*, *Environmental and ecological risk assessment.* Boca Raton: CRC Press, 345–382.

Weerahandi S, Johnson A. 1992. Testing reliability in a stress-strength model when X and Y are normally distributed. *Technometrics* **34**(**1**):83–91 DOI 10.2307/1269555.

Westerhoff P, Song G, Hristovski K, Kiser MA. 2011. Occurrence and removal of titanium at full scale wastewater treatment plants: implications for TiO2 nanomaterials. *Journal of Environmental Monitoring* **13**(**5**):1195–1203 DOI 10.1039/c1em10017c.

Zabell SL. 1989. The rule of succession. *Erkenntnis* **31**:283–321 DOI 10.1007/BF01236567.