Cost of Speed: A Practical Approach to Evaluate a Screening Method From a Bayesian Perspective

Susie Dai^{a, b, *}, Huiyan Sang^c, Kyung-Min Lee^a, and Timothy Herrman^a

^a Office of the Texas State Chemist, Texas A&M AgriLife Research, Texas A&M University System, College Station, Texas 77843, United States.

^b Department of Veterinary Physiology and Pharmacology, College of Veterinary Medicine, Texas A&M University, College Station, Texas 77843, United States.

^c Department of Statistics, Texas A&M University, College Station, TX 77843, United States.

*Correspondent author: [telephone (979)845-1121; fax (979)845-1389; e-mail
Susie@otsc.tamu.edu]

ABSTRACT

Utilization of rapid screening method presents an investment decision of laboratory assets.

Screening methods are filtering procedures implemented to enable analysis of a large volume of samples in a rapid fashion. The principle of screening includes establishment of a threshold value as a selection criteria for the follow up confirmative analysis. Selection of a quality screening method determines the optimized utilization of laboratory resources. Two mathematical models are used to evaluate a screening method by directly comparing to a gold standard reference method. The first model utilizes the traditional receiver operating characteristic (ROC) curve model to evaluate the screening method. A new model is proposed to access the accuracy and reliability of the screening method based on the gold standard reference method. The two models are extremely practical for a single laboratory to evaluate the reliability of a screening method when limited resources are available.

Keywords: screening method, analytical reliability, gold standard method, ROC model

1. Introduction

Production of high quality data is the principle objective for analytical laboratories. To ensure high quality data, validated methods should be developed and the uncertainty of the developed method needs to be estimated. The uncertainty of an analytical method correlates to many other laboratory analytical parameters such as reliability, robustness, selectivity, specificity, sensitivity and representativeness [1]. On the other hand, productivity related features such as speed, cost and risks are also important parameters to be considered for analytical methods. However, productivity related features are of less concern in method validation in the traditional metrology approaches. Even though this may pose serious limitations in practice, the priority for method development is toward production of high quality data with reasonable productivity. Nevertheless, “fitness-for-purpose” is the ultimate goal for analytical method validation to provide methods that are useful for real applications.

Rapid, reliable and low cost are the highly desirable features for a screening method in terms of “fitness-for-purpose”. In laboratory practice, when comes to the selection of a screening method for analysis, it is desirable to establish the performance characteristic and value the uncertainty of the screening method [2, 3]. Qualitative test, in other words, methods based on binary yes/no responses are commonly used in practice. AOAC has developed guideline for qualitative chemistry method [4]. Probability of detection (POD) has been used as one of the statistical models to validate qualitative methods [5-7]. The POD model enables the comparison between the new method and reference method and has led to various applications in botanical methods and chemical contaminant methods [8-10]. The binary “yes or no” answer is typically determined based on comparison of the method readout to a reference value. Such binary tests typically follow certain statistical hypothesis testing procedures (e.g., paired t-test) to determine

threshold concentrations based on continuous data and to calculate the associated uncertainties of testing results [5, 6]. In a previous study [11], Steliopoulos developed a nonparametric testing method to establish the cut-off value for a screening method. Important performance characteristics of a binary yes/no qualitative method include false negative rates, false positive rates, sensitivity and specificity. Even though POD has been a commonly accepted model that is used to validate qualitative method, Receiver Operating Characteristic (ROC) curve has also been widely used in modeling binary test results. The name of ROC curve rises from the usage of the curve in signal detection theory and has been generally extended to characterizing the behavior of a classifier over its potential operational ranges [12]. ROC curves have seen widely applications in characterizing diagnostic tests. If a test is based on variables on a continuous or graded scale, ROC curves can then be used to assess the overall test value and predict the accuracy of the diagnostic test. As such, the ROC curve has been used commonly to evaluate screening methods in medical tests and clinical chemistry [12]. Relevant to chemical testing, the probability-concentration graph is constructed to reveal the probability of obtaining a positive response dependence on the chemical concentration. The ROC curve can be used to assist selection of the critical concentration for the binary test threshold. Meanwhile, the curve can be used to estimate the unreliability of the screening method [13]. Besides frequentist methods, the Bayesian approach provides an alternative model-based method to estimate the reliability profile for analytical method over a wide range of concentrations [14].

As aforementioned, the screening method results have to be compared with “true values”. In the majority of the available models, the “true value” can be achieved by adding chemicals of known quantities into a blank sample. Spiking has been a commonly used practice in method validation, in which the addition of the known value is used as the reference value for evaluation [15]. The

spiking experiments are reliable experiment designs to represent the true value for the mathematical model; however, spiking may not represent the ideal scenario for an analytical method, as the extraction of the target analyte from the spiked matrices might be easier than the material that naturally contains the target analyte. Toward this end, reference materials that contains endogenous analytical target are of great value in method validation. Certified reference materials (CRMs) are produced to assist method validation and is defined by ISO as “reference material characterized by a metrologically valid procedure for one or more specified properties, accompanied by a certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability”[16]. Besides the reference material, comparison of the analytical results from the method under validation to a “gold standard” method is also used for evaluation purpose. Furthermore, as the most preferred model to evaluate an analytical study, inter-laboratory studies are designed based on multiple laboratories utilizing the same method to analyze both spiked samples and natural samples. Practically speaking, for a single laboratory validation of a screening method, it is one of the common practices to compare the analytical results from the screening method and a well-established reference method in house when limited resources are available.

Nevertheless, all the aforementioned practices involves comparing the testing results from the method under study to a value that is produced by either spiking (no uncertainty assigned), or a value determined by other means of measurement (with uncertainty assigned). However, most of the aforementioned statistical testing models, when using the spiking values or the other measurements as the reference value for reliability testing and method validation, are based on the assumption that the reference values are true values without any measurement errors. The current study presents the direct evaluation of the reliability of the screening method based on

comparison to a reference method and the uncertainty data generated for the reference method from an inter-laboratory study that has been previously published. Since it is usually resource prohibitive for a single laboratory to perform an inter-laboratory study for a new screening method, the study here presents a valuable and practical approach to evaluate the reliability of a screening method. Assessing the performance characteristics of a new method can be achieved through comparison to an existing reference method. Several mathematical models have been established to analyze the measurement errors in the method validation regime of comparing the new analytical method to an established method. Carefully addressed are the measurement errors associated with the validation studies [17, 18]. Particularly, regression models have been used to evaluate the bias in method comparisons and differences have been observed with various regression models [19-22]. Besides the regression models, the current model is the first Bayesian model that takes into account of measurement uncertainty of the reference method and estimates the uncertainty budget of the new screening method under evaluation. The model can then be used as a general guide to estimate the reliability of a screening method, when limited resource is used to evaluate a screening method. Typically, for a laboratory to select a screening method, the laboratory has established a reference method and has the comprehensive data for the performance characteristics of the reference method. The model in this study reveals how to utilize the available data from the reference method and estimate the performance characteristics of the screening method based on a simple experimental design.

The objective of the current work is to use a case study by investigating a rapid chemical method to quantify a contaminant (e.g. aflatoxin) in the corn matrix by directly comparing to a reference method. The reference method has been validated by eleven laboratories and the method performance is reported in the previous literature. The case study illustrated the most practical

means for a single laboratory to evaluate the screening method for accuracy and reliability profile based on straightforward and simple laboratory design.

2. Experimental

2.1. Standards and reagents

Naturally occurring aflatoxin contaminated corn samples were collected by Texas Feed and Fertilizer Control Service regulatory officials. The collected corn samples were grounded through a commercial grinder (Retsch Ultra Centrifugal Mill ZM 200, Retsch, Haan, German) using a 0.75 mm screen. Certified aflatoxin B₁, B₂, G₁ and G₂ (AFB₁, AFB₂, AFG₁, and AFG₂) were purchased from Romer Lab Inc-Biopure (Tulln, Australia). All solutions are made with HPLC grade solvents and reagent grade materials unless otherwise noted. The concentration of AFB₁ and AFG₁ standard is 2 µg/mL in 5 mL acetonitrile. The concentration of AFB₂ and AFG₂ standard is 0.5 µg/mL in 5 mL acetonitrile. The Neogen Veratox aflatoxin test kit is obtained from Neogen Corporation (Lansing, MI, USA). In the final evaluation, the total aflatoxin is evaluated (concentration calculated is based on the addition of AFB₁, AFB₂, AFG₁, and AFG₂ concentration together).

2.2 Screening procedure

The Veratox aflatoxin test kit is used for the screening procedure in the current study. Fifty grams of corn samples was extracted with 250 mL of 70% methanol/ water and shaken for 1 hr on a shaker at 200 rpm. The extract was filtered and the pH of the solution was adjusted to fall into the range of 6-8. One hundred µL of conjugate (supplied by the kit) was pipetted into red-

colored well provided by the kit. One hundred μL of the calibration standard solution and the sample extract were transferred into the red-colored well and mixed with the conjugate. After mixing, 100 μL of the mixed solution was transferred into the antibody coated well. The solution was incubated in the antibody coated well for 2 minutes and then rinsed out with deionized water. One hundred μL of the substrate solution was added into the antibody coated well after all the rinsing water has been moved out of the well. The solution was incubated for 3 minutes before the 100 μL of the stop solution (provided by the kit) was added into the well. The mixed solution was then put into the reader for immediate reading.

2.3 Reference method

The high performance liquid chromatography (HPLC)/post column derivatization/fluorescence method for aflatoxin analysis is used as the reference method [23] . Briefly, 50 grams of the ground corn samples were extracted with 250 mL of methanol/water (70:30, v/v) by mechanical shaker for 1 hour at 200 rpm. A 15 mL extract was filtrated through a folded filter paper (Whatman #1), and 5 mL of the filtered extract was diluted with water with a dilution factor of 5 and 1 gram of sodium chloride was added into the diluted solution. After filtration, 2 mL of the solution was loaded onto an immunoaffinity column (AflatestTM affinity column, Vicam #12020). The solution was pushed through the column. After washing the column with 5 mL of water twice, 1 mL of methanol was used to elute the aflatoxins out from the column. Eluate was diluted with 1 mL of HPLC water. The mixture was then filtered through a 0.2 μm syringe filter prior to HPLC/Fluorescence analysis.

2.4 Mathematical models

The ROC curve is constructed using the logistic protocol and ROC output in SAS 9.0 (Cary, NC, USA). The binary results are determined as: 1 means the screening test result is bigger than the cut-off concentration and 0 means the screening test result is smaller than the cut-off concentration. When the HPLC results are higher than the cut-off concentration, a confirmation analysis will happen. The numerical values from the screening test results enabled generating yes/no binary conclusions and the evaluation of screening method prediction accuracy at different concentration levels (pre-set cut-off concentrations).

Linear regression model is used to estimate an empirical equation that relates the measurement of the reference method to each targeted spiking levels. The estimated equation can be used to adjust the measurement errors of the reference method, and hence it is often referred to as the calibration curve. Given a set of experimental data, a calibration curve is obtained by fitting an appropriate equation to the data. The most common choice is a linear equation with an unknown intercept and slope. Estimation of the linear regression equation is obtained using the function `lm` in R software based on the maximum likelihood estimation procedure.

An unbalanced one-way analysis of variance (ANOVA) (Shaw and Mitchell-Olds, 1993) is used to test the reference method and estimate its uncertainty (Equation 1). Assume the I populations are normally distributed with common variance, and independent random samples of sizes $J_1, J_2, \dots, J_I$ are collected from each population group. Let R_{ij} denote the j th measurement from the i th population, and let μ_i denote the population mean of the i th group, the one-way ANOVA model is as follows,

$$R_{ij} = \mu_i + \varepsilon_{ij}, i = 1, \dots, I, j = 1, \dots, J_i, \varepsilon_{ij} \sim N(0, \sigma_R^2). \quad (1)$$

where ε_{ij} denote the residual error that follows the normal distribution with mean zero and unknown variance σ^2 . Bayesian ANOVA method [24] is used to simulate posterior samples to estimate mean and 95% credible intervals (analogous to confidence intervals in frequentist statistics) of the unknown parameters in model equation (1). Implementation of the Bayesian ANOVA method is done using the function `lmBF` in the R package `BayesFactor`.

Finally, a Bayesian measurement error model [25] is introduced to estimate the linear calibration curve between the screening method and the reference method while accounting for the uncertainty of the reference method.

It consists of two model equations

$$S_{ik} = \beta_0^s + \beta_1^s \mu_i + \eta_{ik}, i = 1, \dots, I, k = 1, \dots, K_i, \eta_{ik} \sim N(0, \sigma_s^2) \quad (2)$$

$$R_{ij,bc} = \mu_i + \varepsilon_{ij}, i = 1, \dots, I, j = 1, \dots, J_i, \varepsilon_{ij} \sim N(0, \sigma_{R,bc}^2) \quad (3)$$

where S_{ik} denotes the k th measurement from the k th concentration level group obtained from the screening method, μ_i denotes the population mean of the i th concentration level in the reference method measurement for $i=1, \dots, I$; and η_{ik} denotes the variance specific to the i th measurement, which is the residual error that follows the normal distribution. $R_{ij,bc}$ denotes the j th measurement from the i th concentration group obtained from the calibrated reference method, and μ_i denotes the population mean of the i th concentration level in the reference method measurement for $i=1, \dots, I$; and ε_{ij} denotes the variance specific to the i th measurement, which is the residual error that follows the normal distribution.

Equation (2) is established to estimate the linear relationship between the screening method and the true concentration, which is similar to the linear calibration method for the reference method. In the analysis for the calibration of the reference method, concentration measured from the spiked data is assumed to be ground-truth without error. However, the true concentration level in (2) is unknown and can only be measured via a proxy, in this case, the reference method. Classical linear regression estimation method is no longer applicable to such a case in which explanatory variable is not observed. It motivates us to add model equation (3) to take into account measurement error of concentration. We assume that calibrated reference measurement is centered at true concentration level with a constant measurement error variance $\sigma_{R,bc}^2$, and the error term ε_{ij} is independent of the true covariate μ_i and the screening measurement S_{ik} .

We adopt a Bayesian method since it allows us to naturally incorporate prior knowledge for the measurement error variance $\sigma_{R,bc}^2$ obtained from the reliability analysis of the reference method. The two-equation model casts the inference problem under a Bayesian hierarchical modeling framework [26] and sets prior specifications of model parameters. In this case, the standard method is followed to assign flat priors to β_0^s and β_1^s and a vague inverse Gamma prior to σ_s^2 , to reflect the assumption that no prior knowledge is given for these parameters. A strong informative prior is then assigned to $\sigma_{R,bc}^2$. Specifically, a conjugate inverse gamma prior IG(a, b) is used, in which hyper-parameters a and b are obtained from the posterior estimations of $\sigma_{R,bc}^2$ from the Bayesian one-way unbalanced ANOVA analysis of the calibrated reference data in (1).

2.5 Statistical analysis

A performance of the screening method and the reference method was evaluated using the SAS Proc LOGISTICS procedure with a ROC and ROCCONTRAST options for the nonparametric approach to compare the empirical ROC curves of specified models based on the area under the ROC curve along with its standard error and confidence interval for the models. An unbalanced one-way ANOVA was applied for evaluation of the reference method using SAS Proc GLM procedure. A direct comparison between the screening method and the reference method was also made by means of a paired sample t-test, Pearson's correlation coefficient, and coefficient of determination. All statistical analyses and modeling were conducted using SAS software v. 9.3 (SAS Institute, Cary, NC) or Microsoft® Excel (Microsoft, Redmond, WA).

3. Results and discussion

3.1. Determination of reliability of the reference method from inter-laboratory study

Evaluation of the reference method is based on the inter-laboratory study from ten laboratories for both spiked samples and naturally contaminated samples based on the previous study [23]. In this study, reliability is evaluated through the confidence interval at each spiking concentration level in the inter-laboratory study design. Basically, the reference method was used by ten laboratories to analyze seven prepared samples from the organizing laboratory. There are a total of 5 samples at four concentration levels, which are 0, 10, 20 and 30 µg/kg of the target analyte. There was one sample at the three concentrations of 0, 10, and 20 µg/kg, and two samples at the concentration of 30 µg/kg. Meanwhile, two naturally contaminated samples were included in this study. The data used in this current study is included in Table 1. Following the suggestions in [23], we used the Grubbs outlier test to detect and exclude outliers from our study. In addition,

since our current study focus on evaluating the reliability of the screening method with nonzero concentrations, we also exclude the sample at 0 concentration level.

An unbalanced one-way ANOVA as in (1) is used to test the reference method and estimate its uncertainty (Equation 1). Specifically, it is assumed that the three populations corresponding to 3 concentration levels (including two naturally contaminated samples) are normally distributed with equal variance, and that we have independent random samples of sizes $J_1=J_2=10$ and $J_3=20$ from the respective concentration. Indeed, the p-value from the Shapiro-Wilk test is 0.10, which supports normality assumption. The p-value from the Levene's test for equal variances is 0.11 and that from the Bartlett test is 0.04 which is slightly smaller than 0.05. Considering the small sample size of this study, we adopt equal variance assumption in this study. The parameter estimations including mean and 95% credible intervals are given in Table 2. Based on the ANOVA analysis, the reference method seems to underestimate true concentration levels especially as the concentration level increases when the spiking study is performed (Table 2). This motivates us to seek a calibration method correcting such biases in the reference method. Specifically, the concentration levels were treated as numerical variables and a regression model is run as below,

$$R_{ij} = \beta_0^{(r)} + \beta_1^{(r)} \mu_i + \epsilon_{ij}, i = 1, \dots, 3, j = 1, \dots, J_i, \epsilon_{ij} \sim N(0, \sigma_{R,bc}^2) \quad (4)$$

Since the p-value for the estimated intercept is 0.38, the intercept is deleted to run a regression model with slope only. The estimated slope parameter is 0.83 with a small standard error 0.02.

The estimated slope is used to obtain the bias-corrected reference value $R_{ij,bc} = R_{ij} / 0.83$, and the ANOVA model is rerun in equation (1). The results are presented in Table 2. It is clear that the means of the bias-corrected reference values are closer to the spiking value and the 95% credible

intervals after bias-correction are now covering the spike values for all 3 concentration levels. For example, in the ANOVA analysis of the gold reference method, if the bias of the gold reference method is not accounted for, then the value predicted by the reference method measurement of the 20 µg/kg spiking sample is in the range of 15.4 µg/kg to 18.7 µg/kg (Table 2). After the bias correction, the new ANOVA model adjusts to center to the spiking value (e.g. true value) and the bias adjusted model calculated that at the 95% confidence interval, if the reference method is going to produce a value of 20 µg/kg, then the true value of the sample is between 18.6 µg/kg to 22.5 µg/kg. Based on the bias adjusted model, we can then further calculate the confidence interval of the screening method based on the result produced by the screening method.

3.2 Determination of the reliability of the screening method based on direct comparison to reference method.

In real practice, a quantitative screening method is used based on a “cut-off” concentration, by which when the screening method reports a bigger value than the cut-off concentration, the sample will be subjected to a confirmation method for further analysis. In this study, the screening method is used to analyze the same sample that has been measured by the reference method for direct comparison and evaluating the accuracy of the screening method in a specific concentration range. In summary, a total of 419 samples were subjected to the two analyses for comparison. The concentration is between 8 µg/kg to 50 µg/kg. Based on the reference values from the reference method, several receiver operation characteristic (ROC) curves were constructed using different cut-off concentrations. Basically, when setting a cut-off concentration, the true positive response is determined when the screening method reports a bigger value compared to cut-off concentration and the reference method also reports a bigger value than the

cut-off concentration. Meanwhile, the false positive response is defined when the screening method reports a bigger value compared to the cut-off concentration and the reference method reports a smaller value than the cut-off concentration. The probability of true positive response is plotted versus the 1- probability of false positive response to construct the ROC curve. The area under the ROC curve can be used as an index to determine the accuracy of the screening method. By changing the cut-off concentration, the screening method will achieve different false positive rates and false negative rates. By using the reference method as a gold standard method and not considering the uncertainty/unreliability of the reference method, the ROC curve is an ideal model to evaluate a screening method.

Before using the ROC model to evaluate the data, a direct comparison is done for the screening method and the reference method using linear regression. The correlation data is shown in Figure 1. The correlation coefficient is 0.79 for this data set, which suggest that about 64% of the data can be explained by the linear regression model. As the two methods are different in nature, a two-sample t-test assuming unequal variances is run to analyze the data. The 2-tailed P value of 0.02 suggests that the results from the two methods are statistically different.

In order to evaluate the screening method for the least false positive and best positive rate at varied concentration levels, three different cut-off concentration levels were selected. With the current data set, the cut-off concentration is set at 15 $\mu\text{g/kg}$, 20 $\mu\text{g/kg}$ and 25 $\mu\text{g/kg}$. The treatment of the screening test data set enables the binary interpretation of screening decision, 1 means yes for confirmation test and 0 means no for conformation test as judged at different concentration levels. If taking the value from the reference method as the absolute “true” value, then the area under the ROC curve can be used as an index to evaluate the accuracy of the screening method. The areas under the three ROC curves are 0.9409, 0.9469 and 0.9449 for the

cut-off concentration at 15, 20 and 25 $\mu\text{g/kg}$, individually (Figure 2). Apparently, when setting the cut-off concentration at 20 $\mu\text{g/kg}$, the kit measurement will give the best true positive response and the lowest false negative response.

Generally speaking, under the circumstances of a chemical test, the area under the ROC curve represents the probability [27] that when the screening test result is in conformity with the reference method to judge the random sample contains the analyte target at the concentration that is above or below the cut-off concentration. In using the reference method to assess the screening method, the bigger value of the area under the ROC curve, the higher the probability that the screening method will be in agreement with the reference method. In reviewing the total samples analyzed by our lab on the annual basis, in 2011, the laboratory has analyzed a total of 2,175 corn samples for aflatoxin concentration. By using the reference method, a total of 1,110 samples contain less or equal to 20 $\mu\text{g/kg}$ aflatoxin and 1,065 samples contain more than 20 $\mu\text{g/kg}$ aflatoxin. Literally, if using the screening method to screen all the samples and setting 20 $\mu\text{g/kg}$ as the cut-off value, the probability of the sample to contain less or more than 20 $\mu\text{g/kg}$ and confirmed by the reference method is 94.69%. In other words, the chance is that about 115 samples out of the 2,175 samples will be incorrectly assigned for containing less or more than 20 $\mu\text{g/kg}$ when measured by the reference method. However, since the reference method also has its own variance and bias, the above results based on ROC curves may not be accurate. Furthermore, the ROC model can only calculate the probability that the screening method result correctly assign the value above or below the threshold value or not. It is impossible to evaluate the reliability of the screening method over a wider concentration range. Thus more advanced mathematical model is needed to evaluate the accuracy and reliability of the screening method.

3.3 Reliability profile of the screening method.

A Bayesian hierarchical two-equation measurement error model as in (2) and (3) is proposed to estimate the reliability of the screening method based on the paired study, while adjusting the bias of the screening method and accounting for the unreliability of the reference method. The same estimated slope is used as in Section 3.1 to obtain the bias-corrected reference value $R_{ij,bc} = R_{ij}/0.83$ in (3).

In the Bayesian framework, inference on the parameters of the measurement error model is performed through the posterior distribution, which can be computed according to the Bayes theorem.

$$P(\beta_0^{(s)}, \beta_1^{(s)}, \{\mu_i\}, \sigma_s^2, \sigma_{R,bc}^2 \mid \{R_{ij,bc}\}, \{S_{ik}\}) \\ \propto P(\{R_{ij,bc}\}, \{S_{ik}\} \mid \beta_0^{(s)}, \beta_1^{(s)}, \{\mu_i\}, \sigma_s^2, \sigma_{R,bc}^2) P(\beta_0^{(s)}, \beta_1^{(s)}, \{\mu_i\}, \sigma_s^2, \sigma_{R,bc}^2)$$

The Markov Chain Monte Carlo (MCMC) [26] method is used to draw posterior samples of model parameters. The bias adjustment parameters β_0^s and β_1^s as well as μ_i are updated iteratively from their conditional posterior distributions based on a Gibbs sampler. Finally, posterior inferences such as posterior means and credible intervals of model parameters are made by computing summaries of the posterior samples. The posterior samples of μ_i can also be used to conveniently estimate the posterior reliability probability.

The estimated parameter values are $\beta_0^s = 11.23$ with standard error to 0.41, $\beta_1^s = 0.89$ with standard error to 0.02.

The reliability probability (π) is defined as “ the expected probability of the screening method to provide analytical results (X) within predefined acceptance limits ($\pm\lambda$) around their reference estimated concentration values (μ_i) over a defined concentration range and experimental conditions”.

A previous model to estimate the reliability of a quantitative method is presented by Rozet et al.

[14]. The study applied a Bayesian approach to evaluate a method in a single laboratory validation and spiking study data. In the reported study, the method under the study has not been compared to any other reference method. In order to illustrate the difference between the current model and a previous Bayesian model to evaluate the reliability of the single validated method [14], the two probability profiles were presented in Figure 3. Based on the previously reported model, the probability profile of the reference method is presented in dotted line. The probability profile of screening method based on the current study is presented in continuous line. In this figure, the acceptance limits is set at $\pm 30\%$ around the reference values in the profile of the reference method,

Table 3 shows the valid concentration ranges corresponding a set of combinations of minimum reliability (70%, 80%, 90%) and acceptance limits ($\lambda = \pm 15\%, 20\%, 25\%, 30\%$). For example, when the minimum reliability criterion is set at 70 %, it suggests that the minimum probability of obtaining future measurements falling the acceptance limits ($\pm 30\%$ around the reference values) in the daily analysis is 70 %. Using the 70% of minimum reliability criterion and the 30% of the specification limits, based on our proposed Bayesian model, the valid concentration range is then calculated to be higher than 18.22 $\mu\text{g/kg}$ for the screening method. Ultimately, the model is applicable to include any concentration levels that have been included in the reference method validation and covered by the paired comparison study.

In the current study, the valid range is 18.22 $\mu\text{g/kg}$ and higher which suggests that if by using the reference method, the predicted aflatoxin concentration of the testing sample is 18.22 $\mu\text{g/kg}$ or higher, then the probability of the screening method to produce a value that is within 70% to 130% of 18.22 $\mu\text{g/kg}$ (between 12.75 $\mu\text{g/kg}$ to 23.69 $\mu\text{g/kg}$) value is 70% or higher (Figure 3). However, if the reference method is going to predict the sample containing less than 18.22 $\mu\text{g/kg}$,

then the probability of using the screening method to produce the a value that is within 70% to 130% of the reference predicted value is less than 70%.

4. Conclusions

Method validation is critically important to assure the accuracy of testing results. Generally speaking, a multi-laboratory collaborative study is desirable to evaluate the method performance characteristics. However, it is time consuming and resource limiting to obtain the inter laboratory collaborative study data. In our current study, two mathematical models are used to evaluate the performance characteristics of a screening method. First, the ROC model is applied to the evaluation of a screening method and determines the cut-off concentration when the screening method and the gold reference method produce the highest conformity. In our proposed new model, a two equation model is utilized to predict the confidence interval of the screening method resulting value based on the inter-laboratory collaborative study validation data. The two models used in this research can be applied universally to evaluate any chemical concentration measurement method, offering the users the choice to 1) determine the probability of the screening method to correctly assign a “yes/no” value and 2) estimate the confidence interval of a reported value from the screening method. In a resource limiting environment that prohibits sponsoring an inter-laboratory study, the two models thus provide practical means for a single laboratory to perform a reliability test of the screening method by using published inter-laboratory study data.

Acknowledgements

The HPLC and kit measurements conducted by Cindy McCormick are deeply appreciated. The research of Huiyan Sang is partially supported by NSF Grant CNS-1343155.

References

- [1] S. Cardenas, M. Valcarcel, Analytical features in qualitative analysis, *Trac-Trend Anal Chem*, 24 (2005) 477-487.
- [2] B.L. Mil'man, L.A. Konopel'ko, Uncertainty of qualitative chemical analysis: General methodology and binary test methods, *J Anal Chem+*, 59 (2004) 1128-1141.
- [3] R. Song, P.C. Schlecht, K. Ashley, Field screening test methods: performance criteria and performance characteristics, *J Hazard Mater*, 83 (2001) 29-39.
- [4] AOAC, Appendix N: ISPAM Guidelines for Validation of Qualitative Binary Chemistry Methods, AOAC, <http://www.eoma.aoac.org/appendices.asp>, web accessed March 2016.
- [5] P. Wehling, R.A. LaBudde, S.L. Brunelle, M.T. Nelson, Probability of Detection (POD) as a statistical model for the validation of qualitative methods, *J AOAC Int*, 94 (2011) 335-347.
- [6] EU, Commission Regulation (EU) No 519/2014 of 16 May 2014 amending Regulation (EC) No 401/2006 as regards methods of sampling of large lots, spices, and food supplements, performance criteria for T-2 and HT-2 toxin and citrinin and screening methods of analysis. L147/29., (2006).
- [7] R.A. LaBudde, Statistical Analysis of Interlaboratory Studies, XX, Measuring the Performance of a Qualitative Test Method, TR290, Least Cost Formulations, Ltd, Virginia Beach, VA, 2008.
- [8] C. von Holst, J. Stroka, Performance criteria for rapid screening methods to detect mycotoxins, *World Mycotoxin J*, 7 (2014) 439-447.
- [9] R.A. LaBudde, J.M. Harnly, Probability of Identification: A Statistical Model for the Validation of Qualitative Botanical Identification Methods, *Journal of Aoac International*, 95 (2012) 273-285.
- [10] S. Uhlig, S. Krugener, P. Gowik, A new profile likelihood confidence interval for the mean probability of detection in collaborative studies of binary test methods, *Accredit Qual Assur*, 18 (2013) 367-372.
- [11] P. Steliopoulos, Establishing the cut-off value of a screening method, *Accredit Qual Assur*, 17 (2012) 543-547.
- [12] W.J. Krzanowski, D.J. Hand, *ROC Curves for Continuous Data*, CRC Press Taylor & Francis Group, Boca Raton, FL, 2009.
- [13] B.M. Simonet, A. Rios, M. Valcarcel, Unreliability of screening methods, *Anal Chim Acta*, 516 (2004) 67-74.
- [14] E. Rozet, B. Govaerts, P. Lebrun, K. Michail, E. Ziemons, R. Wintersteiger, S. Rudaz, B. Boulanger, P. Hubert, Evaluating the reliability of analytical results using a probability criterion: A Bayesian perspective, *Anal Chim Acta*, 705 (2011) 193-206.
- [15] R. Galarini, R. Buratti, L. Fioroni, L. Contiero, F. Lega, Development, validation and data quality assurance of screening methods: A case study, *Anal Chim Acta*, 700 (2011) 2-10.
- [16] ISO Guide 30:1992/Amd 1:2008.
- [17] J.P. Buonaccorsi, *Measurement Error: Models, Methods, and Applications*, CRC Press Taylor & Francis Group, Boca Raton, FL 2010.
- [18] W.A. Fuller, *Measurement Error Models*, John Wiley & Sons, Inc, New York, 1987.
- [19] J. Riu, F.X. Rius, Method comparison using regression with uncertainties in both axes, *Trac-Trend Anal Chem*, 16 (1997) 211-216.
- [20] C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, D.L. Massart, Detection of bias in method comparison by regression analysis, *Anal Chim Acta*, 338 (1997) 19-40.
- [21] A. Martinez, F.J. del Rio, J. Riu, F.X. Rius, Detecting proportional and constant bias in method comparison studies by using linear regression with errors in both axes, *Chemometr Intell Lab*, 49 (1999) 179-193.
- [22] A. Martinez, J. Riu, F.X. Rius, Evaluating bias in method comparison studies using linear regression with errors in both axes, *J Chemometr*, 16 (2002) 41-53.
- [23] M.W. Trucksess, M.E. Stack, S. Nesheim, S.W. Page, R.H. Albert, T.J. Hansen, K.F. Donahue, Immunoaffinity column coupled with solution fluorometry or liquid chromatography postcolumn

- derivatization for determination of aflatoxins in corn, peanuts, and peanut butter: collaborative study, J Assoc Off Anal Chem, 74 (1991) 81-88.
- [24] R.G. Shaw, T. Mitchell-Olds, Anova for Unbalanced Data - an Overview, Ecology, 74 (1993) 1638-1645.
- [25] S. Muff, A. Riebler, L. Held, H. Rue, P. Saner, Bayesian analysis of measurement error models using integrated nested Laplace approximations, J R Stat Soc C-Appl, 64 (2015) 231-252.
- [26] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, Bayesian Data Analysis (Vol. 2), Chapman & Hall/ CRC, London, 2014.
- [27] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics, 44 (1988) 837-845.

Table 1

Inter-laboratory collective study data for the gold standard reference method

Lab	Total Aflatoxin Concentration Corn (ng/g)						
	0 ^a	10	20	30	30	NC ^b	NC
1	0.6	5.8	16.3	52.0 ^o	28.9	32.6	20.3
2	0.6	1.4 ^o	16.6	24.3	25.1	19.8	17.7
3	1.3	8.7	8.7 ^o	20.8	19.2	17.3	20.4
4	0.0	10.6	20.7	28.4	30.2	22.3	23.2
5	1.6	9.8	16.5	25.0	23.7	20.8	20.0
6	0.1	9.2	15.2	26.1	19.9	16.5	13.2
7	1.2	8.5	16.7	27.8	24.9	17.7	25.9
8	0.0	0.0 ^o	17.9	25.5	24.1	25.9	14.6
9	0.0	8.8	16.2	22.9	22.3	18.2	17.8
10	0.0	9.2	17.2	24.7	25.0	21.8	19.5

^a Spiked sample concentration^b Naturally contaminated samples.^o Outliers

Table 2

AVONA model for the reference method data

Result	10	20	30	σ_R^2
Mean				
(95% CI	8.9	17.0	24.6	23.1
before model	(7.2,10.8)	(15.4,18.7)	(23.5,25.8)	(20.5,27.2)
correction)				
Mean				
(95% CI	10.8	20.5	29.7	29.4
after model	(8.7,12.9)	(18.6,22.5)	(28.3,31.0)	(25.6,35.6)
correction)				

Table 3

Reliability profile of the screening method

Acceptance limit	Minimum reliability (p)		
(λ)	70%	80%	90%
.15	36.44	45.04	57.76
.20	27.33	33.73	43.34
.25	21.82	27.03	34.63
.30	18.22	22.52	28.93

Figure legends

Figure 1	Correlation of screening method results and the gold standard reference method results.
Figure 2	Receiver operating characteristic curve for screening method. The curves are constructed based on different cut-off values. Filled triangle: 15 µg/kg; filled circle: 20 µg/kg; and filled square: 25 µg/kg.
Figure 3	Probability profile for the screening method. The solid curve is generated based on the acceptable limits within 70-130% of the predicted value set by the reference method and the minimum probability is set at 70% (model proposed by the current study). The dotted curve is generated based on the acceptable limits within 70-130% of the true value set by the spiking study and the minimum probability is set at 70% [14].

Figure 1.

Figure 2.

Figure 3.