DOI: 10.1111/epp.12862

ORIGINAL ARTICLE

Guidelines for improving statistical analyses of validation datasets for plant pest diagnostic tests

Sebastien Massart[1] | **Benedicte Lebas**[1] | **Aude Chabirand**[2] | **Anne-Marie Chappé**[3] | **Tanja Dreo**[4] | **Francesco Faggioli**[5] | **Catherine Harrison**[6] | **Roy Macarthur**[6] | **Natasha Mehle**[4,7] | **Monica Mezzalama**[8] | **Françoise Petter**[9] | **Maja Ravnikar**[4] | **Jean-Philippe Renvoisé**[10] | **Davide Spadaro**[8] | **Laura Tomassoli**[5] | **Jenny Tomlinson**[6] | **Charlotte Trontin**[10] | **René van der Vlugt**[11] | **Ana Vučurović**[4] | **Rebecca Weekes**[8] | **Yves Brostaux**[12]

[1]Gembloux Agro-Bio Tech, TERRA, Laboratory of Plant Pathology, University of Liège, Liège, Belgium

[2]Pests and Tropical Pathogens Unit, ANSES Plant Health Laboratory, Saint Pierre, France

[3]Nematology Unit, ANSES Plant Health Laboratory, Le Rheu, France

[4]National Institute of Biology, Ljubljana, Slovenia

[5]Consiglio per la Ricerca in Agricoltura e l'analisi dell'economia agraria – Centro di Ricerca Difesa e Certificazione, Roma, Italy

[6]Fera Science Ltd, York, UK

[7]School for Viticulture and Enology, University of Nova Gorica, Vipava, Slovenia

[8]Department of Agricultural Forest and Food Sciences and AGROINNOVA – Centre of Competence for the Innovation in the Agro-environmental Sector, University of Torino, Turin, Italy

[9]European and Mediterranean Plant Protection Organization, Paris, France

[10]Quarantine Unit, ANSES, Plant Health Laboratory, Clermont-Ferrand, France

[11]Wageningen University and Research, Wageningen, The Netherlands

[12]Gembloux Agro-Bio Tech, University of Liège, TERRA, Liège, Belgium

Correspondence
Sebastien Massart, University of Liège, Gembloux Agro-Bio Tech, TERRA, Laboratory of Plant Pathology, Liège, Belgium.
Email: sebastien.massart@uliege.be

Abstract

Appropriate statistical analysis of the validation data for diagnostic tests facilitates the evaluation of the performance criteria and increases the confidence in the conclusions drawn from these data. A comprehensive approach to analysing and reporting data from validation studies and inter-laboratory comparisons such as test performance studies is described. The proposed methods, including statistical analyses, presentation and interpretation of the data, are illustrated using a real dataset generated during a test performance study conducted in the framework of the European project, VALITEST. This analytical approach uses, wherever possible and whenever applicable, statistical analyses recommended by international standards illustrating their application to plant health diagnostic tests. The present work is addressed to plant health diagnosticians and researchers interested and/or involved in the validation of plant diagnostic tests, and also aims to convey the necessary information to those without a statistical background. Detailed statistical explanations are provided in the Appendices.

Directives pour améliorer les analyses statistiques de jeux de données de validation pour les tests de diagnostic phytosanitaire

Une analyse statistique appropriée des données de validation des tests de diagnostic facilite l'évaluation des critères de performance et augmente la confiance dans les résultats tirés de ces données. Cet article décrit une approche globale qui consiste à analyser et rapporter les données issues d'études de validation et d'études comparatives inter-laboratoires telles que les études de performance de tests. Les méthodes proposées, notamment les analyses statistiques, la présentation et l'interprétation des données, sont illustrées dans cet article à partir d'un jeu de données réel généré lors d'une étude de performance de test menée dans le

Sebastien Massart and Benedicte Lebas equally contributed to the publication.

Funding information
This article is based upon work from the work package 2 of the project VALITEST (https://www.valit est.eu/), supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no 773139.

cadre du projet européen VALITEST. Cette approche analytique utilise, lorsque cela est possible et applicable, les analyses statistiques recommandées par les normes internationales illustrant ainsi leur application aux tests de diagnostic phytosanitaire. Le présent article s'adresse aux professionnels du diagnostic phytosanitaire et aux chercheurs intéressés et/ou impliqués dans la validation de tests de diagnostic phytosanitaire. Il vise également à transmettre les informations nécessaires à ceux qui n'ont pas de formation statistique. Des explications statistiques détaillées sont fournies en annexes.

Руководство по улучшению статистического анализа валидационных наборов данных для диагностических тестов на наличие вредных организмов по отношению к растениям

Должный статистический анализ валидационных данных для диагностических тестов облегчает оценку критериев эффективности и повышает доверие к выводам, сделанным на основе этих данных. Описан комплексный подход к анализу и представлению данных валидационных исследований и межлабораторных сравнений, таких как исследования эффективности тестов. Предлагаемые методы, включая статистический анализ, представление и интерпретацию данных, проиллюстрированы реальным набором данных, полученных в ходе исследования эффективности испытаний, проведенного в рамках европейского проекта VALITEST. Этот аналитический подход использует, где это возможно и когда применимо, статистические анализы, рекомендованные международными стандартами, иллюстрируюя их применимость в диагностических тестах для защиты здоровья растений. Данная работа адресована специалистам по диагностике защиты растений и исследователям, заинтересованным и/или участвующим в валидации тестов для диагностики в защите растений, а также призвана донести необходимую информацию до тех, кто не имеет статистического образования. Подробные статистические пояснения приведены в приложениях.

1 | INTRODUCTION

The validation of any diagnostic test, i.e. the determination of its performance characteristics, is essential before it can be implemented in plant pest diagnostics. Validation provides essential information on the performance of a test, supports the reliability of the diagnostic activity and assists in the selection of a test that is appropriate for its intended use (Trontin et al., 2021). Test validation can be carried out within a laboratory (intra-laboratory) or in the framework of inter-laboratory comparison studies by several laboratories (such as proficiency testing or test performance study, TPS). The evaluation, carried out on a panel of samples, can include a single test or a comparison of several tests. A scientifically sound evaluation relies on results from a properly designed sample panel and, for a TPS, from enough participating laboratories, allowing the calculation of performance characteristics of the test(s) and their comparison. Several guidelines for the validation of a test and the calculation of its performance characteristics are available for plant health (EPPO, 2019) or specifically for seed testing (ISF, 2020; ISTA, 2019).

Statistical analyses of validation data facilitate the interpretation and comparison of tests and increase the confidence in the conclusions drawn from the validation data. For example, the use of confidence intervals associated with each estimate allows a better interpretation of the value calculated for a performance criterion taking into account the intended use of the test, as shown for the performance evaluation of three RT-PCR protocols for fruit tree virus detection (Massart et al., 2009a, 2009b) or in the frame of an in-depth statistical analysis of TPS results for the molecular detection of phytoplasmas on grapevine (Chabirand et al., 2017). The identification of outliers is also an important objective and usually relies on the expertise of the organizer of the study while it can be supported by proper statistical analysis.

The selection of statistical methods for the evaluation of performance criteria presented in the present

publication was based on (i) the applicability of the statistical method in the context of plant health diagnostic laboratories, (ii) the minimal number of samples and replicates required to correctly perform the statistical method and (iii) the ease of application and interpretation of the results. This work was carried out in the framework of the Europe-funded project VALITEST (https://www.valitest.eu/index), which aimed at improving the validation approaches for diagnostic tests to maximize their usefulness for users (diagnosticians) and decision-makers (at national, European or regional levels) and their use in routine diagnostics.

Statistical analyses can be perceived as complex and difficult to understand, mainly because of the lack of guidance on their use in the context of plant health diagnostics. In this publication, we propose guidelines for the evaluation of an extended range of performance criteria and the application of statistical analyses on validation datasets. These guidelines also provide information on the identification of outliers as well as on the management of inconclusive or missing results. The results generated by these analyses applied to validation datasets allow tests to be selected in an evidence-based way considering their intended use and performance.

2 | DEFINING A SAMPLE PANEL AND THE NUMBER OF PARTICIPATING LABORATORIES

The data generated during intra-laboratory validation studies and inter-laboratory comparison need to be of sufficient quantity to correctly perform the statistical analyses for a proper estimation of performance characteristics of a test and their correct interpretation. Increasing the number of data points is often the best way to improve the confidence concerning the calculated performance characteristics. The magnitude of the uncertainty on a proportion (i.e. performance characteristics such as accuracy) corresponds to $1/\sqrt{n}$, with n equal to the sample size (Newcombe, 1998). Although it is possible to estimate the false positive rate of two tests based on 10 test results, it will not be possible to conclude on the difference between these tests as there will be an uncertainty of 32% associated with each estimate (a confidence interval, CI, of 95%). Overall, the amount of available data depends on the number of samples in the panel, and for comparison studies involving multiple laboratories, the number of participating laboratories. The information in this paper is complemented by a video tutorial: https://www.youtube.com/watch?v=AVxuEDxerGM.

Designing a validation relies on a balance between the available resources (limiting the amount of generated data) and the need for a reliable statistical analysis (requiring more data). The number of samples included in the panel is often limited by the resources of the laboratory to perform the validation study (e.g. cost, time or the availability of reference material, personnel or equipment). For intra-laboratory validation studies, the number of samples included in the panel can usually be larger than that for inter-laboratory comparison studies (e.g. more strains/isolates, more samples with closely related pests or organisms, more dilution levels). For inter-laboratory comparison studies, the number of samples is usually limited owing to the difficulty of preparing reference material in sufficient quantities for many laboratories and the financial resources required to test and prepare the samples for the panel (e.g. availability of personnel, time). Although increasing the number of laboratories participating in an inter-laboratory comparison study will improve the reliability of the calculated performance characteristics, it usually requires more resources (https://www.youtube.com/watch?v=AVxuEDxerGM).

2.1 | Recommended number of participating laboratories

Increasing the number of laboratories participating in an inter-laboratory comparison study positively affects the estimation of the reproducibility as well as the robustness of the calculation of other performance characteristics and of their confidence intervals. For example, the EPPO Standard PM 7/122 (EPPO, 2014) states that test performance studies require a minimum number of participating laboratories (ideally a minimum of 10 valid laboratory datasets). If the statistical analyses are carried out on the results delivered by a small number of laboratories, this leads to an increased uncertainty and wider confidence intervals around estimates of test performance. In addition, the organizers of such a study should present the conclusions of their analyses with caution because the presence of outliers or other kinds of inconsistent results are less likely to be detected when the number of laboratories is small.

2.2 | Recommendations for the sample panel

The composition of the sample panel, i.e. the type (infested or not by the target pest – hereafter called the 'target') and number of samples and the number of biological replicates (several samples prepared each from a biologically distinct example of the same type of biological material, not to be confused with technical replicates, which correspond to the number of reactions prepared from one sample for a test, e.g. duplicate/triplicate reactions in ELISA/PCR tests), is critical. The sample panel depends on the intended use of the test (which determines the performance criteria to be evaluated) and the availability of reference material. It usually includes samples infested with the target (including dilutions) and

samples free from the target (but that may be infested by closely related species).

From a statistical point of view, a serial dilution is not the best way to prepare diluted samples because it introduces a correlation between samples at different dilutions. Theoretically, all of the diluted samples should be prepared independently from each other (but they can be prepared from the same initial sample). However, in practice, serial dilutions are generally carried out when preparing the sample panel to optimize time and limit the complexity of preparation as the bias owing to serial dilution is considered as low.

Figure 1 shows which samples are used to evaluate the different performance characteristics of a test. All of the samples can be used to estimate repeatability (if replicates are included as recommended) and reproducibility. The results obtained from the diluted infested samples are used to estimate the analytical sensitivity. The results from the infested samples and samples free from the target pest (negative controls) are used to estimate the other performance criteria.

The organizers should be experts on the test(s) to be validated and on their intended use. Hence, they should be able to specify the most critical performance criteria to evaluate and how to sufficiently challenge the test so that estimates of performance are robust in practice. This should be addressed by the sample panel composition. For example, if the analytical sensitivity is critical for the intended use of the test and is a potential weakness, then its evaluation will be a priority (i.e. number of dilution points). However, when exclusivity is critical (ability to distinguish the target from another pest/organism), the panel should include samples free from the target pest and infested with closely related species that might be detected, causing a problem of exclusivity of the test.

The type and number of samples, and the number of dilution points and of biological replicates in a sample panel, can influence the determination of the performance characteristics of a test. The sample panel should ideally include samples representing specific difficulties for the detection or identification of the target pest (e.g. low concentration or a range of the genetic variability). Indeed, a sample panel consisting only of infested samples at very high concentrations might result in 100% diagnostic sensitivity for all of the evaluated tests, limiting the ability to discriminate between tests for this criterion.

In the framework of VALITEST, a sample panel was proposed (Supplementary Material S1) after the evaluation of the statistical analyses conducted on 10 datasets generated during test performance studies and with input from their organizers. The proposal is a balance between the statistical power of the analysis and 'practicality' for the study organizers. It should be noted that this proposed sample panel may not always be feasible (e.g. owing to the limited availability of reference material) and should be adapted depending on the performance criteria to evaluate.

An overview of the sample panels used in the test performance studies within the framework of the VALITEST project is provided in the Supplementary material S1b. This overview shows the adaptations made by the organizers depending on the different constraints and the scope of their test performance studies.

Note that the use of reference material is critical for the preparation of the panel of samples. Guidelines have been developed as part of the VALITEST project from which the EPPO Standard PM 7/147 *Guidelines for the production of biological reference material* (2021) has been established.

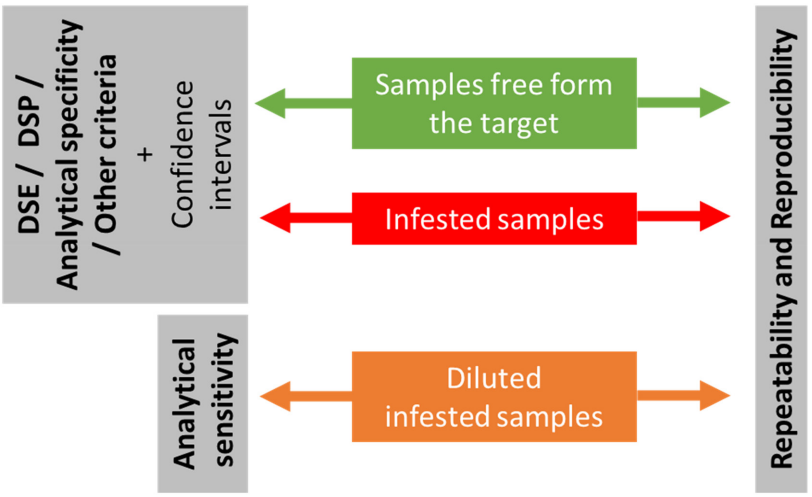

FIGURE 1 Schematic representation indicating the use of the results of each type of sample in the calculation of the performance characteristics of a test. Other performance criteria include accuracy, diagnostic odd ratio, false positive and false negative rates, positive predictive value and negative predictive value, positive and negative likelihood ratios. DSE, Diagnostic sensitivity; DSP, diagnostic specificity

3 | COMPOSITION OF THE SAMPLE PANEL OF THE EXAMPLE DATASET

The statistical analyses proposed in this paper are illustrated using the results from one of the TPSs carried out within the framework of VALITEST. The study involved 16 laboratories in which two tests were compared (here called Test A and Test B) for the detection of a plant virus. The sample panel consisted of 22 samples including samples infected with the target virus, samples free from the target virus (samples free from viruses and samples infected with closely related non-target viruses) and a series of dilutions prepared from a sample infected with the target virus (see Supplementary material S2). All samples containing the target virus and samples free from viruses were tested in duplicate. The values of the performance criteria were calculated from the results obtained with this sample panel. The results are shown graphically and are discussed below in the relevant sections.

4 | INCONCLUSIVE AND MISSING RESULTS

Sometimes results can be missing or inconclusive. A missing result is when the laboratory has not reported it. This may happen when a test could not be performed because of, for example, sample degradation, failure of nucleic acid extraction, failure of controls, or a shortage of samples/specific reagents for a test. A result can be considered inconclusive when it was not possible for the laboratory to assign a positive or a negative result to a sample. This may happen when the amount of the target in the sample is close to the level of detection of the test.

Given their origin, missing results are not representative of the performance of a test and should be excluded from any further analysis. The inclusion in the analysis of inconclusive results can vary between performance criteria, based on their suspected origin and the expertise of the study organizer.

In the case of a missing or inconclusive result, it is recommended that, whenever possible, the participating laboratory should repeat the test to deliver a dataset that is as complete as possible. Repetition of the experiment should be clearly stated in the report.

However, where this is not possible or if missing or inconclusive results occur again, then the missing results are discarded, and decisions must be taken on how to consider inconclusive results. These decisions depend on how the study organizer considers these results based on his/her expertise and knowledge of the pests and the tests. In any case, inconclusive results should be treated with caution to minimize biases to the analyses. The way inconclusive results were addressed in this publication is presented below:

- For the calculation of analytical sensitivity, inconclusive results were considered as negative results, i.e. pest not detected in the diluted sample(s). This is because inconclusive results might occur owing to the very small amount of target in such samples and should therefore be taken into account.
- For the calculation of repeatability and reproducibility, inconclusive results were excluded from the analysis. Inconclusive answers were excluded from the analysis to avoid two inconclusive results given on the same sample contributing positively to the performance criteria (repeatability or reproducibility).
- For the calculation of the other performance criteria described in this publication with their corresponding confidence intervals, inconclusive results were treated as erroneous results (results not showing the real status of the sample, corresponding to false positive result for a healthy sample and false negative result for an infested sample).

Missing and inconclusive results must undergo further examination on a case-by-case basis to better understand the origin of their occurrence.

5 | OUTLIER RESULTS

Outliers may be detected before or during the statistical analysis and interpretation of the results. For example, outliers can be detected by looking at strong deviations among laboratories or samples. Among the analyses proposed in this paper, the following ones can be useful to identify outliers (see Supplementary Material S3 for examples of tables and figures): accordance and concordance per sample (Figure S3.1A); accordance per laboratory (Figure S3.1B); analytical sensitivity per laboratory (Figure S3.4); diagnostic sensitivity; and diagnostic specificity per laboratory (Table S3.3 and Figure S3.5). Note that, even in the case of very high diagnostic sensitivity or diagnostic specificity (as shown in our example for diagnostic sensitivity, those parameters can still be useful to detect outliers).

Different types of outliers can be detected:

1. the results obtained in one laboratory are very different from those from the other participating laboratories;
2. the results obtained by most participating laboratories for a sample are different from the expected results.

The cause of the outliers should be evaluated by a diagnostician who has experience in statistics or working in collaboration with statisticians and a decision made on the exclusion (or not) of those results from the statistical analysis. For example, data may be excluded from the analysis when it can be established that they are the result of contamination, or from a specific deviation

such as the failure of a machine. Communication with the laboratory that has generated potential outliers can help in identifying the cause of these results.

6 | CONFIDENCE INTERVALS

Diagnostic test results can be influenced by numerous random variation sources (e.g. variation between samples, preparation of the samples, execution of the measurement process). Therefore, the performance characteristics calculated on the sample panels are estimates of the true values which remain unknown. The quality of those estimates is important for end-users, as it will affect the confidence in the test results. Confidence intervals provide an estimate of the uncertainty that is associated with the estimate of the value of each performance criteria from a limited number of samples, as compared with an entire population (Hess et al., 2012). Confidence intervals give a range of values that contain the real value of the parameter with a fixed prior probability (usually 95%). For example, if 100 confidence intervals of a performance characteristic are estimated from 100 independent samplings in the whole population, a 95% confidence level means that, on average, 95 confidence intervals contain the real value of the performance characteristics. Thus, the confidence intervals give information on the dispersion of the individual values around the average criteria. The narrower the confidence interval, the less dispersed the values are and the more confident the user is about the estimates. A two-sided confidence interval of 95% is commonly used as it provides an estimate with a 5% risk that the real value of the criteria is outside this confidence interval.

Confidence intervals are useful in several situations: (i) when the numbers of datasets used to calculate and compare the performance characteristics of several tests are different; (ii) when the measurements (associated to the observed values of the compared performance criteria) present a marked variability between tests or laboratories; and (iii) as an indication of the uncertainty of the estimates of performance characteristics of the tests. However, since the comparison of confidence intervals is not equivalent to a statistical test, but only an approximation, an appropriate statistical test should be carried out to accurately evaluate the statistical difference and its associated probability if there was no variation.

The methods used to determine confidence intervals are linked to the statistical analyses used to estimate the performance criteria (see Appendix 1). Confidence intervals for analytical sensitivity, repeatability and reproducibility were not determined because they require complex calculation whereas this study is intended to be accessible to people whatever their knowledge in statistics.

In the example in Supplementary Material S3 (Table S3.3), confidence intervals were calculated to have a better estimation of the accuracy of two diagnostic tests.

7 | REPEATABILITY AND REPRODUCIBILITY

The EPPO Standard PM 7/98 *Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity* (2019) defines repeatability and reproducibility respectively as 'the level of agreement between replicates of a sample tested under the same conditions' and 'the ability of a test to provide consistent results when applied to aliquots of the same sample tested under different conditions (e.g. time, person, equipment, location)'. We propose to estimate repeatability (within a laboratory) and reproducibility (between laboratories) by calculating accordance and concordance, respectively. In the plant sector, accordance and concordance are used by the International Seed Federation (ISF-ISHI-Veg, 2020), the International Seed Testing Association (ISTA, 2019) and ANSES based on the recommendation of the standard ISO 16140 in the 2003 version (Chabirand et al., 2017).

The calculation is based on simple counts of concordant and non-concordant results between replicates (whatever the status of the samples). These measures evaluate the probability of achieving the same test results for identical samples within (accordance) and between (concordance) laboratories (Langton et al., 2002). For the repeatability assessment, each biological replicate must be obtained through an identical but independent process and should not be a repeated measure of the same aliquot. For reproducibility assessment, concordance can be calculated for any reproducibility conditions, for example, the day, the equipment or the operator within a laboratory. For this calculation, inconclusive results are excluded from the analyses.

Accordance can be calculated per test and per sample. At the level of the test, accordance shows the expected agreement between the results from replicates of all samples in each laboratory taken individually. Calculating accordance values for each sample is possible when biological replicates are included in the sample panel. At the level of the sample, accordance is used to identify samples that give discordant results for replicates analysed at the same time under the same conditions in each participating laboratory independently. The replicates should be biological replicates obtained from the same sample through an identical but independent process. Technical replicates corresponding to repeated measures on the same aliquot are not recommended but can be used in the absence of biological replicates. Accordance is calculated per laboratory (for all the replicates received) to identify laboratories with poor repeatability.

Concordance of a test as a performance characteristic will provide information about the test's capacity to give

consistent results across the random between-laboratory variation, as observed between the different laboratories in a TPS. If concordance is smaller than accordance, it indicates that two replicates are more likely to give the same result if they are analysed by the same laboratory than if they are analysed by different laboratories. A test with higher concordance will give more consistent results across the different levels of factors that may vary across laboratories: environment, operator, machine, etc. At the sample level, concordance will help identify samples that give inconsistent results between laboratories.

As accordance and concordance are not totally independent, an additional tool is needed to identify if there is extra variability between laboratories that is not only the result of within-laboratory variation. Accordance and concordance estimates can be used to calculate the concordance odds ratio (COR, by samples and by tests). The magnitude of this ratio provides the relative chance ('odds' in other terms) of getting the same result when two samples are analysed in the same laboratory compared with if they are analysed by different laboratories (Langton et al., 2002). For example, an odds ratio of 2.5 indicates that the samples are 2.5 times more likely to produce the same result (i.e. both positive or both negative) when analysed in the same laboratory than when analysed in different laboratories.

7.1 | Illustration of the accordance, concordance and concordance odds ratio

Accordance, concordance and concordance odds ratio principles (Langton et al., 2002), are illustrated in Supplementary Material S3 using the example dataset. Detailed information on their calculations is described in Appendix 2.

8 | ANALYTICAL SENSITIVITY

The EPPO standard PM 7/98 (EPPO, 2019) defines analytical sensitivity as 'the smallest amount of target that can be detected reliably (also referred to as the limit of detection)' and provides recommendations on how to assess it for different methods and discipline. For example, for the validation of serological and molecular tests, the same standard recommends conducting at least three experiments with a series of dilutions for the estimation of the analytical sensitivity. The definition refers not only to the capacity to detect a small amount of target, but also to the capacity detect it with high certainty. In this context, the use of a probability of detection model based on a binomial generalized linear model as recommended in ISO 16140-2 (2016) is recommended for the determination of the analytical sensitivity. Generalized linear models (McCullagh & Nelder, 1989) are an extension of the classical linear model (analysis of variance, linear regression), adapted for non-normal responses (here, the probability of detection). As the probability of detection ranges from 0 to 1 (pest not present/present), a binomial family response was chosen, which describes a number of successful events (here, detection) on a fixed number of trials (the diagnostic tests). The generalized model provides an estimate of the probability of detection for any value in the range of dilutions of the sample panel, not only for the observed ones (Supplementary Material S3, Figure S3.3).

Unlike ISO 16140-2 (2016) a logit link function ($\log[p/(1-p)]$ where p is the probability of detection) was chosen for the adjustment of the parameters of the model, because the tools to apply the model with this link function are widely available and easy to handle. In addition, the resulting estimates are very close to the complementary log–log link model ($\log[-\log(1-p)]$, where p is the probability of detection) recommended by ISO 16140-2 (2016). In the binomial generalized linear model, the probability (expressed as a percentage) of detecting a target is a function of its concentration as a continuous variable which is presented in a graph, helping the interpretation of the data. The model can be applied on all qualitative methods with binary outputs (i.e. positive/negative answers; Wehling et al., 2011).

The generalized linear model does not require any assumption on the number of technical and/or preferably biological replicates and it can be used when those numbers vary between samples and/or laboratories. However, the model requires a minimum of five dilution points to perform correctly. Each diluted sample should be analysed at least three times from the same positive sample. The model can be used for one or several diagnostic tests (Appendix 3).

The analytical sensitivity calculated using the probability of detection model can be absolute or relative. For some pests, such as bacteria, fungi or nematodes, an absolute level of the analytical sensitivity can be determined where the probability of detection model is expressed, for example, in the number of cells, spores or cysts. In the case of pests for which the concentration cannot be quantified, such as viruses, viroids or phytoplasmas, a relative analytical sensitivity can be determined, where the probability of detection model is expressed as a dilution level. Note that the relative analytical sensitivity may also be used in the case of quantifiable pests.

The probability of detection model (whether absolute or relative) can be used to compare the analytical sensitivity of different tests using fixed levels of detection probability. Two levels that are often referred to in scientific publications are 50% and 95%. A level of 95% probability of detection means that, at the corresponding dilution level, the pest can be detected on average in 95% of the tests carried out (Supplementary Material S3, Figure S3.3 and Table S3.2).

Caution is required when using specific statistical models (such as a binomial generalized linear model)

TABLE 1 Terminology used for the classification of test results when compared with the status of reference samples

	Status of reference sample	
	Target present	**Target absent**
Positive test result	True Positive (TP)	False Positive (FP)
Negative test result	False negative (FN)	True Negative (TN)

with data that have been transformed (such as a logit link between the target concentration expressed in log scale and the detection status) to calculate the analytical sensitivity from data of the diluted samples for each test and each laboratory (see the section on 'Analytical sensitivity'). Those models are based on the assumption that the probability of detection is decreasing as the dilution level increases. This hypothesis can be shown to be incorrect in different cases: for example, when all the samples gave the same result whatever the dilution level or when the observed detection rate shows contradictory behaviour (e.g. a significant decrease then significant increase again or the Hook effect for immunological detection in the presence of an excess amount of target; Schiettecatte et al., 2012).

When this happens, the model cannot be trusted. Hence, the diagnostician should check that the model is fit for purpose before interpreting the calculated analytical sensitivity value. To evaluate the adequacy of those assumptions, the analysis can also be carried out for each laboratory independently, as shown in Figure S4 (Supplementary Material S3).

The probability of detection can be determined for a test across laboratories or within each laboratory, and the detailed results obtained from our dataset are presented in Supplementary Material S3.

The performance criteria gathered in this section have been used in the framework of the VALITEST test performance studies for plant pests (https://www.valitest.eu/work_packages/index#wp1) and correspond to internationally used recommendations (EPPO standard PM 7/122, 2014; EPPO Standard PM 7/98, 2019; ISF, 2020; ISTA, 2019). Some of these performance criteria have also been referred into scientific publications in intra-laboratory validation studies and test performance studies (Chabirand et al., 2017; Franco Ortega et al., 2020, 2021; Massart et al., 2009a, 2009b; Renvoisé et al., 2019).

This section provides a description of each performance criteria so that diagnosticians can decide which ones are more appropriate for their analysis. In addition, the calculation of confidence intervals is proposed for those performance criteria. Further explanations are available in the following video tutorial: https://www.youtube.com/watch?v=otDdi5sY_uU. In addition, details on the calculation of these criteria are given in Table 2.

In our example dataset, the infection status (i.e. presence or absence of the target) of each sample of the panel was used to classify test results into true positive/true negative or false positive/false negative. This terminology is commonly used in validation studies (ISF, 2020; ISTA, 2019; NATA, 2018).

A true positive or a true negative outcome is reported when the result of the test is in agreement with the assigned value of each reference sample. A false positive or a false negative outcome is reported when the result of the test is not in agreement with the assigned value of reference samples (Table 1).

8.1 | Diagnostic sensitivity and diagnostic specificity

Diagnostic sensitivity (DSE) and diagnostic specificity (DSP) are widely used criteria to evaluate the performance of plant pest detection tests (Franco Ortega et al., 2020, 2021; Renvoisé et al., 2019). Diagnostic sensitivity represents the proportion of infested samples that correctly tested positive using a specific test, while diagnostic specificity corresponds to the proportion of samples free from the target pest that correctly tested negative. Diagnostic sensitivity and diagnostic specificity range between 0 and 1 and can be expressed in percentages. The higher the diagnostic sensitivity or diagnostic specificity, the better the performance of the test. A test with a high diagnostic sensitivity indicates a high probability of detecting the target pest when it is present in a sample while a test with a high diagnostic specificity has a high probability of correctly diagnosing the absence of the target pest when it is truly absent. The use of confidence intervals is recommended for diagnostic sensitivity and diagnostic specificity.

It should be noted that the diagnostic sensitivity and the diagnostic specificity are heavily dependent on the choice of samples of the panel (infested samples and samples free from the target pest). In particular, when the numbers of these samples vary between laboratories and tests, it is important to include confidence intervals in the analysis to better compare the estimations made from heterogeneous datasets. In any case, the comparison between tests should always be interpreted with caution.

In diagnostics, two tests are usually performed along with the performance characteristics that allow an informed decision to be made about which tests should be used first and which second, depending on the intended use. For example, the test presenting the highest diagnostic sensitivity may be different from the test presenting the highest diagnostic specificity. By combining tests, the first test could maximize the diagnostic sensitivity (a screening test with a low false negative rate), while the second test could maximize the diagnostic specificity (a confirmatory

TABLE 2 Performance criteria used for measuring the effectiveness of a test in plant health diagnostic laboratories, with their formulae and meaning (source: https://en.wikipedia.org/wiki/Confusion_matrix, accessed 1 February 2021)

Performance criteria	Formula	Note
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Accuracy (syn. trueness) is the proportion of true test results (positive and negative) in the tested samples. The value can range between 0 and 1 and can be expressed as a percentage. The higher the value is, the better the performance of the test Warning: Accuracy estimation can yield misleading results if the dataset is unbalanced between infested and non-infested samples. For example, if the sample panel contains only a small proportion of non-infested samples, a very high accuracy can be obtained despite a very low diagnostic specificity (high frequency of FP among the few non-infested samples) Note: The Accuracy formula is presented in EPPO Standards PM 7/98 (2019) and PM 7/122 (2014) with the terms positive/negative agreement/disagreement
Diagnostic sensitivity	$\frac{TP}{TP+FN}$	Diagnostic sensitivity (DSE) is the proportion of infested samples that correctly tested positive for a test. Diagnostic sensitivity is complementary to the false negative rate (see below). The value can range between 0 and 1 and can be expressed as a percentage. The higher the value is, the better the performance of the test
Diagnostic specificity	$\frac{TN}{TN+FP}$	Diagnostic specificity (DSP) is the proportion of healthy samples correctly tested negative for a test. Diagnostic specificity is complementary to the false positive rate (see below). The value can range between 0 and 1 and can be expressed as a percentage. The higher the value, the better the performance of the test
Diagnostic odd ratio	$\frac{TP/FN}{FP/TN}$	Diagnostic odd ratio (DOR) of a test is the ratio of the odds of positivity in subjects with disease relative to the odds in subjects without disease. The DOR values have no limit but, if they are infinite, they cannot be interpreted properly. The higher the value is, the better the performance of the test
False positive rate	$\frac{FP}{FP+TN} = 1 - DSP$	False positive rate (FPR) is a ratio expressing the probability of false positive detection among healthy samples. In other words, the proportion of healthy samples tested positive. It is linked to the diagnostic specificity (1 − DSP) The value can range between 0 and 1 and can be expressed as percentage. The lower the value is, the lower the number of false positive results
False negative rate	$\frac{FN}{FN+TP} = 1 - DSE$	False negative rate (FNR) is a ratio expressing the probability of false negative detection among infected samples. In other words, the proportion of infected samples tested negative. It is linked to the diagnostic sensitivity (1 − DSE) The value can range between 0 and 1 and can be expressed as percentage. The lower the value is, the lower the number of false negative results
Positive predictive value	$\frac{TP}{TP+FP}$	Positive predictive value (PPV) is the ratio of infected samples among the positive results, e.g. what proportion of the positive results come from an infected sample. The value can range between 0 and 1 or be expressed as percentage. The higher the value is, the better the performance
Negative predictive value	$\frac{TN}{TN+FN}$	Negative predictive value (NPV) is the ratio of healthy samples among the negative results, e.g. what proportion of the negative results come from a healthy sample. The value can range between 0 and 1 or be expressed as percentage. The higher the value is, the better the performance

Abbreviations: DSE, diagnostic sensitivity; DSP, diagnostic specificity; FN, false negative; FP, false positive; TN, true negative; TP, true positive.

test with a low false positive rate) Finally, if diagnostic sensitivity and diagnostic specificity are useful parameters to compare the performance of tests, they cannot be used directly to estimate the probability that a target pest is present or not in a matrix. To estimate the discriminating performance of the test when samples are infected or not, both parameters should be combined into one measure called the likelihood ratios (see below).

8.2 | Accuracy

Accuracy is the proportion of true (positive and negative) test results of the tested samples (EPPO Standard PM 7/122, 2014; EPPO, 2014). Trueness is a synonym of accuracy (NATA, 2018). The value ranges between 0 and 1 or can be expressed as percentage. The higher the value, the better the performance of the test.

Accuracy is usually determined for each test when analysing the performance of tests but can also be estimated for each participating laboratory (in particular when the objective is to evaluate the performance of laboratories in proficiency tests).

It is important to stress that accuracy estimation can be misleading if the dataset is unbalanced between infested samples and samples free of the target pest. For example, if the sample panel contains only a small proportion of samples free from the target pest, a very high accuracy

can be obtained despite a very low diagnostic specificity (high frequency of false positives among the few samples free from the target pest). This underlines the importance of appropriate sample panel design as discussed above.

8.3 | Diagnostics odds ratio

The diagnostics odds ratio (DOR) of a test is the ratio of the odds (in another term, the probability) of positivity in infested samples (i.e. true positive and false negative) relative to the odds in samples free from the target (i.e. false positive and true negative; Glas et al., 2003). The higher the value, the better the performance of a test. Diagnostic odd ratios are useful to compare the performance of tests, but it is impossible to give a cut-off value to consider if the DOR is appropriate or not. An infinite diagnostics odds ratio (e.g. when the value of false negative, false positive or true negative is zero) cannot be interpreted properly. The diagnostic odds ratio is positively correlated to accuracy as both values increase when the number of true positives or true negatives rises. However, poor results (for example low diagnostic specificity) obtained with an unbalanced sample panel (a low proportion of samples free from the target pest) will have a greater impact on the diagnostic odds ratio than on accuracy.

8.4 | False positive rate and false negative rate

The false positive (or negative) rate is the proportion of target-free (or infested) samples tested positive (or negative). The false positive rate is linked to the DSP as its formula is 1 − DSP. On the contrary, the false negative rate is linked to the DSE as its formula is 1 − DSE. The values range between 0 and 1 and can be expressed as percentages. The lower the value, the lower the number of false positive (or negative) results.

8.5 | Positive predictive value and negative predictive value

The positive (or negative) predictive value (PPV – NPV) is the ratio of infested (or target-free) samples among the positive (or negative) results, e.g. what proportion of the positive (or negative) results comes from an infested (or target-free) sample (Šimundić, 2009). The value ranges between 0 and 1 and can be expressed as a percentage. The higher the value is, the better the performance.

8.6 | Positive and negative likelihood ratios

The likelihood ratio for positive test results tells us how much more likely the positive test result is to occur in samples with the target compared with those without the target. The positive likelihood ratio is usually higher than 1 because it is more likely that the positive test result will occur in infested plants than in healthy plants (Šimundić, 2009) with higher values corresponding to better tests, but there is no absolute threshold. An infinite value is obtained if the DSP is 100%.

The likelihood ratio for a negative test result represents how much less likely the negative test result is to occur in a plant with the target than in a plant without the target. The negative likelihood ratio usually ranges between 0 and 1 because it is less likely that a negative test result will occur in plants with disease than in plants without disease (Šimundić, 2009). The lower values correspond to better tests but there is no absolute threshold. For ease of interpretation and to be symmetrical with the positive likelihood ratio, the negative likelihood ratio formula can be inverted (Parikh et al., 2009; see Table 2). In this case, the higher the values of the negative likelihood ratios are, the better the tests are and an infinite value is obtained if the DSE is 100%. As such, they directly link the pre-test and post-test probability of a disease in a specific sample and depend on the values of diagnostic sensitivity and diagnostic specificity (see Table 2).

The likelihood ratios are well suited to assist in the selection of tests when several tests are evaluated for a target pest. Plotting in a graphical form the inverted negative likelihood ratio and the positive likelihood ratio of each test provide a visual representation of how the tests compare with each other in relation to the diagnostic sensitivity and diagnostic specificity.

Figure 2 shows a comparison of 18 tests for the detection of a pest. These 18 tests have been plotted according to their positive likelihood ratio (x-axis) and inverted negative likelihood ratio (y-axis). Test no. 11 has an infinite inverted negative likelihood ratio (e.g. a negative likelihood ratio of zero) while test no. 4 has an infinite positive likelihood ratio (100% of diagnostic specificity). According to the y-axis, e.g. the inverted negative likelihood ratio representing the trust that the diagnostician can have in the obtained negative results, three tests (4, 11 and 15) have higher values than all of the other tests. Tests 4, 11 and 15 show the highest inverted negative likelihood ratio values and could be used for applications requiring high confidence in negative results, e.g. as a first screening test in quarantine for certification diagnostics. Tests 4, 7 and 9 show the highest positive likelihood ratio values and could be used for applications requiring high confidence in positive results, like for a confirmatory test upon a first detection. In both cases, the calculation of confidence intervals (like in Figure S5, Supplementary Material S3) could help identify differences among the best tests. Finally, the tests located in the orange circle correspond to tests that have a relatively low positive likelihood ratio and inverted negative likelihood ratio compared with the above-mentioned tests.

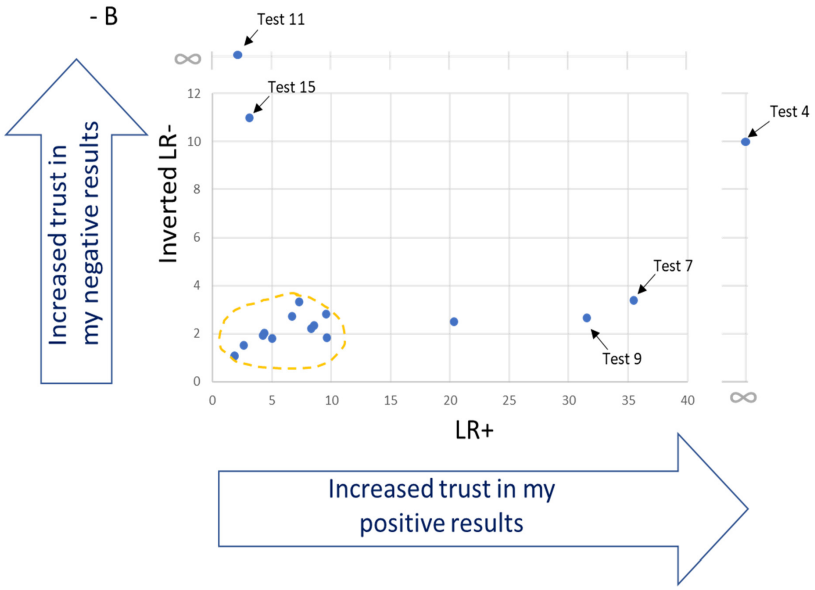

FIGURE 2 Positive likelihood ratio (LR+, *x*-axis) and inverted negative likelihood ratio (LR−, *y*-axis) calculated from the validation data for 18 tests. Each dot represents one test

In this case, confidence intervals may give additional information by providing upper or lower levels of confidence for this ratio.

8.7 | Determination of performance criteria per participating laboratory

All of the performance criteria analysed in the above section can also be determined per participating laboratory. Their calculation per laboratory can be useful to identify laboratories having divergent results compared with the other participating laboratories, i.e. to identify outliers. The results of the calculation of DSE and DSP per laboratory for tests A and B are detailed in Supplementary Material S3.

9 | PRACTICAL APPLICATION OF THESE GUIDELINES

9.1 | Layout of results

All of the results should preferably be compiled in a single document so as to facilitate the statistical analyses. To facilitate data entry, the following recommendations are made:

- Enter each test result, including technical replicate results, in a single line in a table (e.g. MS Excel ou Calc spreadsheet), with all the relevant information reported in different columns.
- Clearly identify each individual biological sample with its target (and non-target) organisms and, when possible, their estimated concentration in appropriate units (e.g. cfu for bacteria, dilution scale for viruses).
- Use simple codes for reporting test results or sample true status, e.g. 0 for absence/negative, 1 for presence/positive, 2 for inconclusive results. Missing results should also be shown in the raw data file by leaving the relevant cell empty in the Results column.

The layout of the dataset used for the case study example is shown in Supplementary Material S4.

9.2 | R script publicly available

R scripts can be used to automate the calculation of the performance criteria and the production of the graphical representations shown in the article, provided that the original data follow the data entry layout described above. Further information on R scripts can be found in Supplementary Material S5.

9.3 | Accessibility of validation reports

Validation reports from several diagnostic tests, in particular for regulated pests, are freely available to save time, costs and resources by avoiding duplicative validation efforts and allowing simpler and faster verification studies for laboratories. Existing resources where validation reports are freely available include: (i) EPPO validation reports, https://dc.eppo.int/validation_data/validationlist; (ii) ISF validation reports, https://www.worldseed.org/our-work/phytosanitary-matters/seed-health/ishi-veg-validation-reports/; and (iii) STA validation reports, https://www.seedtest.org/en/method-validation-reports-_content---1--3459--467.html.

10 | CONCLUSION

The statistical approaches described in this paper can be used to analyse the results of test validation studies (intra-laboratory validation studies or inter-laboratory comparison studies). The recommendations given in this paper aim to help harmonize the analyses of validation studies and whenever possible to foster their comparisons, despite differences in sample panels. The statistical methods proposed in this publication have been used to analyse 10 datasets generated in test performance studies conducted in the framework of the European project, VALITEST (https://www.valitest.eu/).

The calculation and statistical analysis of the performance criteria is needed to assist the reliable selection of tests for specific intended uses (e.g. screening test or confirmatory test) as shown for the likelihood ratio of tests A and B in the case study. Validation data are essential to support discussions between risk assessors and managers and laboratory experts regarding the selection of tests to be used in routine diagnostics.

The statistical analysis of test performance characteristics also provides some indication on how test results should be interpreted because positive and negative results may not be equally informative in all contexts. This is highlighted with the use of, for example, likelihood ratios and positive and negative predictive values.

The quality of the statistical analysis of data generated during validation studies depends on several factors such as the composition of the sample panel, the number of participating laboratories (for inter-laboratory comparison studies), the identification of outliers and the inclusion or not of missing and outliers results in the analysis.

Statistical analysis of validation studies can be challenging. However it is important to undertake such analyses as they facilitate the evaluation of the performance criteria and increase the confidence in the conclusions drawn from these data. Improving statistical analysis is an ongoing process and further statistical tests could be envisioned in the future. For example, a statistical inference test, such as Fisher's exact test (Fisher, 1922), can be performed on some performance criteria to underline if there is a significant difference between the performance characteristics of the tests or the different laboratories.

ACKNOWLEDGEMENTS

We would like to thank C. Eskes, R. Souza Richards and J. Woudenberg of the International Seed Federation, Nyon (Suisse) for their valuable input during the development of this article.

DISCLAIMER

The content of the manuscript represents the opinion of the authors and not of their institutions.

REFERENCES

Agresti A, Coull B A (1998) Approximate is better than "Exact" for interval estimation of binomial proportions. *American Statistician* 52, 119–126.

Chabirand A, Loiseau M, Renaudin I, Poliakoff F (2017) Data processing of qualitative results from an interlaboratory comparison for the detection of 'Flavescence dorée' phytoplasma: how the use of statistics can improve the reliability of the method validation process in plant pathology. *PLoS ONE* 12(4), e0175247, https://doi.org/10.1371/journal.pone.0175247

Collet D (2003) Modelling Binary Data, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

EPPO (2014) PM7/122 (1) Guidelines for the organization of interlaboratory comparisons by plant pest diagnostic laboratories. *Bulletin OEPP/EPPO Bulletin* 44(3), 390–399, https://onlinelibrary.wiley.com/doi/epdf/10.1111/epp.12162.

EPPO (2019) PM 7/98 (4) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. *Bulletin OEPP/EPPO Bulletin* 49(3), 530–563, https://onlinelibrary.wiley.com/doi/epdf/10.1111/epp.12629.

EPPO (2021) PM 7/147 (1) Guidelines for the production of biological reference material. *Bulletin OEPP/EPPO Bulletin* 51(3), 499–506, https://onlinelibrary.wiley.com/doi/full/10.1111/epp.12781.

Fisher R A (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85(1), 87–94, https://doi.org/10.2307/2340521.

Fleiss JL, Levin B, Paik MC (2003) Statistical analyses for rates and proportions. Third Edition. John Wiley & Sons. New York.

Franco Ortega S, del Pilar Bustos López M, Nari L, Boonham N, Gullino ML, Spadaro D (2021) Rapid detection of *Monilinia fructicola* and *Monilinia laxa* on peaches and nectarines using loop-mediated isothermal amplification. *Plant Disease* 103, 2305–2314, https://apsjournals.apsnet.org/doi/pdfplus/10.1094/PDIS-01-19-0035-RE

Franco Ortega S, Prencipe S, Gullino ML, Spadaro D (2020) New molecular tool for a quick and easy detection of apple scab in the field. *Agronomy* 10 (4), 581, https://doi.org/10.3390/agronomy10040581.

Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM (2003) The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 56(11), 1129–1135.

Hess AS, Shardell M, Johnson JK, Thom KA, Strassle P, Netzer G, Harris AD (2012) Methods and recommendations for evaluating and reporting a new diagnostic test. *European Journal of Clinical Microbiology and Infectious Diseases* 31(9), 2111–2116.

ISF (2020) ISHI-Veg guidelines for the validation of seed health methods, version 3, https://www.worldseed.org/wp-content/uploads/2020/12/MVGuidelines_v3_November-2020.pdf.

ISO 16140-2 (2016) Microbiology of food chain – method validation – Part 2: Protocol for the validation of alternative (proprietary) methods against a reference method. International organization for standardization, Geneva, Switzerland.

ISTA (2019) Procedure 'validation methods and organizing and analyzing results of interlaboratory comparative tests (CT)', https://www.seedtest.org/upload/cms/user/TCOM-P-10-ValidatingmethodsandresultsofCTsV1.141.pdf.

Langton SD, Chevennement R, Nagelkerke N, Lombard B (2002) Analysing collaborative trials for qualitative microbiological methods: accordance and concordance. *International Journal of Food Microbiology* 79, 175–181.

McCullagh P, Nelder JA (1989) Generalized Linear Models, 2nd revised edition, London: Chapman & Hall.

Massart S, Brostaux Y, Brabarossa L, Batlle A, César V, Dutrecq O, Fonseca F, Guillem R, Komorowska B, Olmos A, Steyer S, Wetzel T, Kummert J, Jijakli M H (2009a) Inter-laboratory evaluation of

two reverse-transcriptase polymeric chain reaction-based methods for the detection of four fruit tree viruses. *Annals of Applied Biology* 154, 133–141.

Massart S, Brostaux Y, Brabarossa L, César V, Cieslinska M, Dutrecq O, Fonseca F, Guillem R, Laviña A, Olmos A, Steyer S, Wetzel T, Kummert J, Jijakli MH (2009b) Inter-laboratory evaluation of a duplex RT-PCR method using crude extracts for the simultaneous detection of Prune dwarf virus and *Prunus* necrotic ringspot virus. *European Journal of Plant Pathology* 122, 539–547.

NATA (2018) Guidelines for the Validation and Verification of Quantitative and Qualitative Test Methods, in: Technical Note 17. Canberra, Australia, https://www.nata.com.au/phocadownload/gen-accreditation-guidance/Validation-and-Verification-of-Quantitative-and-Qualitative-Test-Methods.pdf.

Newcombe RG (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 17, 857–872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.

Parikh R, Parikh S, Arun E, Thomas R (2009) Likelihood ratios: clinical application in day-to-day practice. *Indian Journal of Ophtalmology* 57(3), 217–221.

Renvoisé JP, Chambon F, Gleize M, Pradeilles N, Garnier S, Rolland M (2019) Selection, optimization and characterization of molecular tests for the detection of Tobacco ringspot virus (TRSV). *Bulletin OEPP/EPPO Bulletin* 49(1), 111–121.

Schiettecatte J, Anckaert E, Smitz J (2012) Interferences in immunoassays. In: *Advances in Immunoassay Technology*. Chiu NHL, Christopoulos TK (eds). IntechOpen, ISBN 978-953-51-0440-7, 10.5772/1967. https://www.intechopen.com/chapters/33740

Simel DL, Samsa GP, Matchar DB (1991) Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology* 44(8), 763–770.

Šimundić, AM (2009). Measures of diagnostic accuracy: basic definitions. *The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine* 19(4), 203–211.

Trontin C, Agstner B, Altenbach D, Anthoine G, Bagińska H, Brittain I, Chabirand A, Chappé AM, Dahlin P, Dreo T, Freye-Minks C, Gianinazzi C, Harrison C, Jones G, Luigi M, Massart S, Mehle N, Mezzalama M, Mouaziz H, Petter F, Ravnikar M, Raaymakers T M, Renvoisé J P, Rolland M, Santos Paiva M, Seddas S, van der Vlugt R, Vučurović A (2021) VALITEST: Validation of diagnostic tests to support plant health. *Bulletin OEPP/EPPO Bulletin* 51(1), 198–206.

Wehling P, LaBudde RA, Brunelle SL, Nelson MT (2011) Probability of detection (POD) as a statistical model for the validation of qualitative methods. *Journal of AOAC International* 94(1), 335–347.

SUPPORTING INFORMATION

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

How to cite this article: Massart, S., Lebas, B., Chabirand, A., Chappé, A-M, Dreo, T. & Faggioli, F. et al. (2022) Guidelines for improving statistical analyses of validation datasets for plant pest diagnostic tests. *EPPO Bulletin*, 52, 419–433. Available from: https://doi.org/10.1111/epp.12862

APPENDIX 1 - CONFIDENCE INTERVALS – ADDITIONAL INFORMATION

The calculation of the confidence intervals can be determined using statistical software such as R software.

Agresti–Coull confidence intervals

Confidence intervals of 95% using the Agresti–Coull method (Agresti & Coull, 1998) were applied for the following performance criteria: accuracy, diagnostic sensitivity, diagnostic specificity, false positive rate and false negative rate, and the rate of true positive and rate of true negative.

It is a general formula for calculating binomial confidence intervals.

Given X successes in n trials, define $\tilde{n} = n + z^2$ and

$$\tilde{p} = \frac{1}{\tilde{n}}\left(X + \frac{z^2}{2}\right)$$

where

$$z = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

is the quantile of a standard normal distribution (for example, a 95% confidence interval requires $\alpha = 0.05$, thereby producing $z = 1.96$).

Then, a confidence interval for p is given by

$$\tilde{p} \pm z\sqrt{\frac{\tilde{p}}{\tilde{n}}(1-\tilde{p})}$$

Confidence interval for the diagnostic odds ratio

The 95% confidence intervals for diagnostic odds ratio can be calculated using Simple OR CI (Fleiss et al., 2003).

The formula is

$$e^{\left[\ln(\mathrm{DOR}) \pm 1.96\sqrt{\frac{1}{\mathrm{TP}}+\frac{1}{\mathrm{FN}}+\frac{1}{\mathrm{FP}}+\frac{1}{\mathrm{TN}}}\right]}$$

The DOR calculation is impossible if the false positive (FP) or false negative (FN) is equal to zero. Confidence interval calculation is impossible if one of the false positive (FP), true positive (TP), false negative (FN) or true negative (TN) is equal to zero.

Confidence intervals for the generalized linear model

Confidence intervals for the analytical sensitivity can be extracted from the parameters of the generalized linear model used to adjust the probability of detection on the diluted series data, but it needs some advanced computation.

TABLE A1 Terminology used for the classification of test results when compared with the status of reference samples

	Status of reference sample	
	Target present	**Target absent**
Positive test result	True positive (TP)	False positive (FP)
Negative test result	False negative (FN)	True negative (TN)

The (log)dilution factor associated with a probability of detection of p ($\hat{x}_p$) can be extracted by the following formula, where g(.) is the link function of the generalized linear model, and $\hat{\beta}_0$ and $\hat{\beta}_1$ its parameters.

$$\hat{x}_p = \frac{\left(g(p) - \hat{\beta}_0\right)}{\hat{\beta}_1} = f\left(\hat{\beta}_0, \hat{\beta}_1\right)$$

The confidence interval associated with this value can be calculated by different computational methods, exact or approximate, that are beyond the scope of this report.

More information can be found in McCullagh and Nelder (1989) or Collet (2003) among others.

Confidence intervals for positive and negative likelihood ratios

The 95% confidence intervals for positive and negative likelihood ratio can be calculated using Simel's method (Simel et al., 1991), using the results of Table A1.

95% confidence interval of positive likelihood ratio:

$$e^{\ln\left(\frac{\mathrm{DSE}}{1-\mathrm{DSP}}\right) \pm 1.96 \times \sqrt{\frac{1-\mathrm{DSE}}{\mathrm{TP}} + \frac{\mathrm{DSP}}{\mathrm{FP}}}}$$

95% confidence interval of negative likelihood ratio:

$$e^{\ln\left(\frac{\mathrm{DSE}}{1-\mathrm{DSP}}\right) \pm 1.96 \times \sqrt{\frac{\mathrm{DSE}}{\mathrm{FN}} + \frac{1-\mathrm{DSP}}{\mathrm{TN}}}}$$

APPENDIX 2 - ACCORDANCE AND CONCORDANCE FOR THE DETERMINATION OF THE REPEATABILITY AND REPRODUCIBILITY AND CONCORDANCE ODDS RATIO – ADDITIONAL INFORMATION

The calculation of the accordance, concordance and concordance odds ratio can be done using a spreadsheet or a programming software such as R software.

Determination of the accordance for the estimation of the repeatability

The accordance is the percentage chance of finding the same result from two replicates of the same sample analysed in the same laboratory (Langton et al., 2002) under repeatability conditions. The accordance is calculated by dividing the number of pairs of equal results between replicates of the samples by the total number of pairs of results between replicates. The true status of the sample (i.e. target absent or target present) is not used in this calculation.

The accordance is bound between 0 and 1, 0 meaning that not a single pair of replicates shows the same result, and 1 that all the replicates have the same result.

For a given sample, if a laboratory performed n replicates and k of these gave identical positive results (note, the number of identical negative results can also be used), then the accordance for that sample is estimated as

$$\frac{k(k-1) + (n-k)(n-k-1)}{n(n-1)}$$

The accordance of a test obtained using the results of a test performance study as a whole is the average (mean) of the accordance values calculated for each laboratory.

Determination of the concordance for the estimation of the reproducibility

The concordance is calculated in the same way, using the results of the same sample measured by different laboratories (in the context of test performance studies) instead of replicates in the same laboratory. One way of calculating this is using the same formulas as accordance but considering all results disregarding laboratory information, then subtracting the number of matching and total pairs within each laboratory (which are linked to the accordance, not concordance). For a given sample, N results were obtained from the different laboratories (including replicates) and K of these gave identical positive results (note, the number of identical negative results can also be used), then the concordance for that sample is estimated by

$$\frac{K(K-1) + (N-K-1) - \sum_i \left[k_i\left(k_i-1\right) + \left(n_i - k_i - 1\right)\right]}{N(N-1) - \sum_i \left[n_i\left(n_i-1\right)\right]}$$

If the accordance is higher than the concordance, it indicates that two identical samples are more likely to give the same result if they are analysed by the same laboratory than if they are analysed by different ones, suggesting that there can be variability in performance between laboratories. A concordance value much lower than the accordance value can suggest that the method is not robust enough to reproduce the same results under different laboratory conditions. The comparison between accordance and concordance can be achieved through the concordance odds ratio evaluation.

Determination of the concordance odds ratio for the estimation of inter-laboratory variation

The concordance odds ratio (COR) is a ratio of the accordance and concordance for the estimation of the degree of inter-laboratory variation. The ratio removes the bias related to the accuracy of the results (i.e. numbers of true positive/negative and of false positive/negative) which are used to calculate the two parameters (i.e. concordance and accordance) taken separately.

The formula of COR is defined as follows:

$$\frac{\text{acc}(1-\text{conc})}{\text{conc}(1-\text{acc})}$$

with acc for accordance and conc for concordance.

As the accordance of a test should normally be superior to its concordance (this is because the repeatability is expected to be higher than the reproducibility of a test), this ratio should show values between 1 and the positive infinite. The higher the COR value is, the greater the variability between laboratories.

However, when there are many accordance values of 1 (meaning that the tests are highly stable with a reproducibility identical to the repeatability), concordance odds ratios are of little help to discriminate the tests, as most of the estimates are either 1 or infinite values. To get meaningful results, the COR estimation can be completed using Fisher's test, which tests the hypothesis that there is a significant variation of the results between laboratories for a particular sample based on the fact that COR values significantly greater than 1 indicate a significant variability of the results between laboratories.

APPENDIX 3 - PROBABILITY OF DETECTION FOR THE DETERMINATION OF THE ANALYTICAL SENSITIVITY – ADDITIONAL INFORMATION

The calculation of the probability of detection can be done using any statistical software capable of adjusting a binomial generalized model (also called logistic regression) such as R software.

For each test, data for the diluted samples were used to adjust binomial generalized linear models (bGLM) with a logit link between the dilution (expressed by the base 10 negative exponent of the corresponding dilution) and the detection status. The number of dilution levels being very limited, the adjustment of bGLM is not always possible as this method requires at least five levels, and the laboratory effect has been neglected. This type of model is easily adjusted in R with the glm() function, using argument family = binomial to account for the binary nature of the result.

An example of an Excel spreadsheet developed by the ISO for the determination of the limit of detection (terms relative level of detection in that spreadsheet) between laboratories during an inter-laboratory comparison can be found at the following link: https://standards.iso.org/iso/16140/-2/ed-1/en. One spreadsheet (RLOD_MCS_clause_5-1-4-2_V3_2015-08-15) is the template used to enter analytical data; the second spreadsheet (RLOD_inter-lab-study_16140-2_AnnexF_ver1_28-06-2017) provides information on the program and the equations through examples.