This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 773139

Grant agreement N. 773139

DELIVERABLE N° 2.1

Title:

Guidelines for the revision of the EPPO Standards PM 7/98 and PM 7/122 for validation studies (including Test Performance Studies)

Validation of diagnostic tests to support plant health

Due date:	Month M35
Actual submission date	27-04-2021 (Month M36)
Start date of the project	01-05-2018
Deliverable lead contractor (organization name)	ULg
Participants (Partners short names)	ANSES, CREA, EPPO, FERA, GIORiN, NIB, ULg, UNITO, WBF, WR
Author(s) in alphabetical order	Brostaux Y., Chabirand A., Lebas B., Massart S.
Contact for queries	sebastien.massart@uliege.be
VALITEST reviewers (in alphabetical order)	Chappé A.-M., Faggioli F., Macarthur R., Mehle N., Mezzalama M., Spadaro D., Petter F., Renvoisé J.-P., Tomlinson J., Trontin C., van der Vlugt R., Vučurović A., Weekes R.
External reviewers (in alphabetical order)	Eskes C., Souza Richards R., Woudenberg J.
Level of dissemination	Public
Type of deliverable	Report

Abstract

These guidelines describe a new and improved approach to analyse and report data of validation studies and inter-laboratory comparisons studies (i.e. test performance studies) for integration in the appropriate EPPO standards (i.e. PM 7/98 (2019) and PM 7/122 (2014)). The proposed methods of analysis including statistical methods, have been tested on the 10 data sets obtained from the test performance studies conducted in the framework of VALITEST (as part of Work package 1). This new approach uses, wherever possible and whenever applicable, statistical methods recommended by international standards with examples of their application to plant health diagnostic tests. This approach is a consensus agreement between VALITEST members.

The analysis, presentation and interpretation of data are presented using data based on a real data set of a test performance study conducted by VALITEST (as part of Work package 1). It has been written in a simple manner for diagnostician with limited knowledge in statistics. More advanced statistical explanation are provided in Annexes.

The content of this deliverable represents the views of the author only and is his/her sole responsibility; it cannot be considered to reflect the views of the European Commission and/or the Research Executive Agency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use that may be made of the information it contains.

TABLE OF CONTENTS

TERMS, ABBREVIATIONS AND DEFINITIONS

Accordance: *"The (percentage) chance that two identical test materials analysed by the same laboratory under repeatability conditions will both be given the same result (i.e. both found positive or both found negative)"* (Langton *et al*., 2002).

Analytical sensitivity: *"The smallest amount of target that can be detected reliably (this is sometimes referred to as the limit of detection)"* (EPPO standard PM 7/76, 2018).

Analytical specificity: It comprises inclusivity and exclusivity.

AOAC: Association of official analytical chemists

Concordance: *"The (percentage) chance that two identical test materials sent to different laboratories will both be given the same result (i.e. both found positive or both found negative)"* (Langton *et al*., 2002).

Concordance odds ratio: *"The relative chance ('odds' in betting terms) of getting the same result when two samples are analysed in the same laboratory compared to if they are sent to the different laboratories"* (Langton *et al*., 2002).

Diagnosis (of a pest): *"The process of detection and identification of a pest (ISPM No. 27, 2006) (i.e. the interpretation of the result of a diagnostic process)"* (EPPO PM standard 7/76, 2018).

Diagnostic sensitivity: *"The proportion of infected/infested samples testing positive compared with results from an alternative test (or combination of tests). Diagnostic sensitivity = true positives / (true positives + false negatives)"* (EPPO standard PM 7/76, 2018). In EPPO standard PM 7/98 (2019), the formula of diagnostic sensitivity is applied to the comparison of a test A with a validated test B and uses the terms positive/negative agreement/disagreement.

Diagnostic specificity: *"The proportion of uninfected/uninfested samples (true negatives) testing negative compared with results from an alternative test (or combination of tests). Diagnostic specificity = true negatives / (true negatives + false positives)"* (EPPO standard PM 7/76, 2018). In EPPO standard PM 7/98 (2019), the formula of diagnostic specificity is applied to the comparison of a test A with a validated test B and uses the terms positive/negative agreement/disagreement.

EPPO: European and Mediterranean Plant Protection Organization.

Exclusivity: *"The performance of a test with regards to cross-reaction with non-targets (e.g. closely related organisms, contaminants)"* (EPPO standard PM 7/76, 2018).

Generalized linear model: Generalized linear model is an extension of the classical linear model (which includes linear regression) that allows to model responses which do not follow a normal distribution. This is the case if the result of a diagnostic test is binary (e.g. detected or not).

Link function: Link functions are used in generalized linear models to transform the linear predictor estimated by the model in the response space (here, a probability). The default link function for binomial models is the logit function, which is symmetrical and give equal weights to true or false responses (presence/absence, detection/non detection).

Inclusivity: *"The performance of a test with a range of target organisms covering genetic diversity, different geographical origin and hosts"* (EPPO standard PM 7/76, 2018; EPPO standard PM 7/122, 2014).

Intra-laboratory comparison: "Organization, performance and evaluation of measurements or tests on the same or similar items within the same laboratory in accordance with pre-determined conditions" (ISO 17025, 2018).

Inter-laboratory comparison: "*Organization, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with pre-determined conditions (i.e. proficiency testing or test performance studies)*" (EPPO standard PM 7/76, 2018).

ISO: International organization for standardization, based in Geneva, Switzerland.

Limit of detection: Synonym of analytical sensitivity which is defined above.

Outlier: "*Member of a set values which is inconsistent with other members of that set*" (ISO 13528, 2015).

Probability of detection: In the context of this deliverable, a statistical model to determine the probability of a target pest being detected.

Repeatability: "*The level of agreement between replicates of a sample tested under the same conditions*" (EPPO standard PM 7/76, 2018).

Replicates: They can be technical or biological depending on the context. Biological replicates correspond to several samples prepared from the same original biological material (and can be included in a sample panel). Technical replicates correspond to the number of reactions prepared from one sample for a test (e.g. duplicate/triplicate reactions in ELISA/PCR tests).

Reproducibility: "*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)*" (EPPO standard PM 7/76, 2018).

Risk analysis: A process of determining the factors influencing a test results by identifying the performance criteria needed to be evaluated and to what extent (EPPO standard PM 7/98, 2019).

Robustness: "*The extent to which altered test conditions (e.g. temperature, volume, change of reagents) affect the established test performance values (e.g. analytical sensitivity, analytical specificity)*" (EPPO standard PM 7/76, 2018).

Selectivity: "*The extent to which variations in the matrix affect the test performance (matrix effect)*" (EPPO standard PM 7/76, 2018).

Screening (for a pest): The process of looking for a pest among a large number of samples (e.g. insects, plants, soil, water) to identify a pest that may be confirmed by a second test.

Test: "*The application of a method to a specific pest and a specific matrix*" (EPPO standard PM 7/76, 2018).

TPS: Test performance study, "*Evaluation of the performance of one or more tests by two or more laboratories using defined samples (evaluation of a test)*" (EPPO standard PM 7/76, 2018).

Validation (of a test): A means of providing "*objective evidence that the test is suitable for the circumstances of use*" (EPPO standard PM 7/98, 2018).

Verification (of a test): A means of providing "*objective evidence that the laboratory is competent to perform a validated test according to the relevant performance characteristics*" (EPPO standard PM 7/98, 2018).

1 Purpose

The proper validation of any diagnostic test, e.g. the determination of its performance characteristics, is a mandatory prerequisite before its application in plant pest diagnostics. Such evaluation can be carried out within the laboratory (intra-laboratory) or in the framework of test performance studies by two or more laboratories. The evaluation, carried out on a panel of reference samples, can include a single test or compare several tests simultaneously.

A scientifically sound evaluation relies on the generation of results dataset from a properly designed sample panel and, for a test performance study, a sufficient number of laboratories, allowing the proper calculation of performance characteristics of the test(s) and their comparison. The sample panel design, relying on the expertise of the organiser, and the number of participating laboratories are two important elements for a proper statistical analysis of intra-laboratory validation studies and inter-laboratory comparison studies.

An appropriate statistical analysis increases the confidence in the conclusions drawn from the validation data. The use of statistics in data processing during intra- and inter-laboratories studies facilitates the interpretation and comparison between tests for a given performance criteria. The clear added value of statistical tools is also to provide confidence intervals and p-values associated with each estimate. For a given test, this allows a better interpretation of each calculated performance criterion for the intended use. This also allows the possibility to statistically compare tests for a given performance characteristic.

The identification of outliers is also a very important focus when analysing the results obtained during an intra-laboratory validation study or inter-laboratory comparison study. This identification relies on the expertise of the organiser of the study, but it can be guided by statistical analysis allowing to highlight divergent results for a sample or a laboratory.

Numerous statistical methodologies have been developed for interpreting the results of the validation of a test. So far, an in-depth analysis of data generated during intra-laboratory validation studies or test performance studies showed a lack of guidance on the methods of calculation of performance criteria and/or their interpretation. The range of performance criteria for the analyses can be extended. The purpose of this document is to provide guidelines for improving the analyses of data generated during diagnostic test validation studies (EPPO standard PM 7/98, 2019) and inter-laboratory comparisons (EPPO standard PM 7/122, 2014) as well as proficiency testing. These guidelines aim to harmonize the analyses of the validation dataset and to foster their comparisons, even between different studies. They aim to provide information in order to facilitate the data interpretation.

2 Scope

This document is applicable to the analysis of diagnostic tests validation data generated during intra-laboratory validation studies or test performance studies. Some statistical analyses outlined in this document can also be applicable to the analysis of data generated during proficiency tests programmes (e.g. diagnostic sensitivity, diagnostic specificity or repeatability).

The guidelines are relevant for plant health diagnostic laboratories that perform validation studies before using a test routinely and organisers of test performance studies and proficiency testing studies. The scope of test includes the detection and/or identification of plant pests (e.g. arthropods, bacteria, fungi, nematodes, invasive plants, protozoan, viroids, viruses or weeds) from any types of matrices (e.g. pure microbial culture, plant tissue, soil, water).

3 Introduction

Diagnostic tests (e.g. commercial and non-commercial tests) need to be appropriately validated before their use as a routine test in a plant health diagnostic laboratory. Test validation (or test verification if using a validated test) is mandatory to be accredited to ISO 17025 standard which will be a requirement for all national reference laboratories in the European Union by 29 April 2022 (EU Regulation 2017/625).

The ISO 17025 (2017) standard recommends that "*the validation shall be as extensive as is necessary to meet the needs of the given application or field of application*". Indeed, as stated in EPPO standard PM 7/76 (2018) "*diagnostic tests have different levels of analytical sensitivity, analytical specificity, speed and cost*" and the laboratory should select tests that are suitable according to the circumstances of use. For example, a test used for screening (e.g. testing for a pest during a surveillance programme) may not have the same validation requirements as a test used for diagnosis. Before performing the validation of a test, a risk analysis (new requirement in the 2017 version of ISO 17025), is conducted in order "*to identify which performance criteria need to be evaluated and to what extent*" (see EPPO standard PM 7/98 (2019) for further details).

The performance criteria described in EPPO standard PM 7/98 (2019) for the validation of a test are analytical sensitivity, analytical specificity which includes inclusivity and exclusivity, selectivity, robustness, repeatability and reproducibility with detailed guidance for each of the diagnostic methods by plant pest field including botany. Diagnostic sensitivity and diagnostic specificity performance criteria are provided in EPPO standard PM 7/76 (2018) and applied to the comparison of a test A with a validated test B in EPPO standard PM 7/98 (2019). The performance characteristics of a test allow a better understanding of the reliability of a test. For test performance studies, the performance criteria of accuracy, and rates of true positive and true negative are provided in EPPO standard PM 7/122 (2014).

VALITEST partners agreed to focus on three priorities:

(i) improvement of the evaluation of the analytical sensitivity,
(ii) more robust evaluation of the repeatability and reproducibility and,
(iii) introduction of the likelihood ratios and other criteria in the context of performance evaluation.

In most of the intended use contexts, the analytical sensitivity, analytical specificity, diagnostic specificity, diagnostic sensitivity, repeatability and reproducibility are considered as the core performance criteria for a test and thus their evaluation requires an appropriate analysis and interpretation of the data collected during validation studies and test performance studies. During the work conducted in the test performance studies conducted in the framework of VALITEST (as part of Work package 1), it appeared that more performance criteria may be valuable and are consequently suggested. These additional performance criteria that have been included in these guidelines are diagnostic odd ratio, false positive and negative rates, rates of true positive and true negative, and positive and negative likelihood ratios.

To establish these guidelines, the choice of the statistical methods for the determination of the performance characteristics was based on the applicability of the method in the context of plant health diagnostic laboratories, the minimum number of samples and replicates required for the statistical method to perform correctly, the ease of application and interpretation of the results.

The statistical tools have been evaluated with 10 datasets from VALITEST test performance studies carried out in Work package 1. This allowed an in-depth analysis on how the statistical tools behave (data not shown) and thus to refine (if required) the recommendations. Furthermore, a range of criteria used by the organisers of the test performance study was assessed to measure the effectiveness of a test (e.g. accuracy, diagnostic odd ratio, false positive and negative rates, rates of true positive and true negative). As a result, this document provides a consensus approach for the analysis of data obtained in validation studies and test performance studies. The application of the statistical tools is presented using some of the data generated during the VALITEST test performance studies to illustrate their usefulness and limitations. In addition, the interpretation of the results is provided. Further details on the statistical tools are provided in Annexes. This document also provides information on how to establish the panel of samples, how to deal with inconclusive and missing results and how to identify and deal with outlier results. Finally, the use of confidence intervals and, for concordance odds ratio, the p-values is also described and proposed as optional tools to assess the statistical confidence in the estimation of the performance characteristics of the tests.

4 Composition of dataset

4.1 Background information

The data generated during validation studies and test performance studies need to be of sufficient quantity for the statistical methods to perform correctly. The amount of available data depends on the number of samples in a panel and, for test performance studies involving multiple laboratories, the number of participating laboratories. Both elements are crucial in the proper estimation of the performance characteristics of the test and their sound interpretation for validating a test. It should be noted that the performance characteristics are properties of the test itself, which depend on the matrix (e.g. plant, soil, water), on the quantity of pest present in the samples, and on the number of replicates used to make a diagnosis.

The composition of the panel plays a crucial role. It usually includes samples infested with the target pest (including serially diluted samples) and samples free from the target pest (but that might be infested by closely related species). These samples should be preferably prepared at least in duplicate in order to properly evaluate the repeatability of the test(s) and to compare to its reproducibility. The number of reference samples included in the panel and the proportion of infested, non-infested and serially diluted samples in the sample panel rely on several criteria.

First, differences between tests can be difficult to assess or can be biased when using a small number of data. The main limitation for an appropriate statistical analysis of data is the need for a minimum number of data points generated during a validation study. Increasing the number of data is often the best way to improve the reliability of the calculated performance characteristics. For example, in theory, the uncertainty of the diagnostic sensitivity or diagnostic specificity of a test based on 10, 100 and 1000 samples is estimated to be around 32%, 10 % and 3.2% respectively [1].

The number of samples included in the panel is limited by the resources of the laboratory to perform the validation study (e.g. cost, time, availability of reference material). A distinction should be made here between intra-laboratory validation studies and test performance studies. Indeed, for intra-laboratory validation studies, the number of samples included in the panel can usually be larger than for test performance studies (e.g. testing more strains/isolates, more samples with closely related pest or targets, possibility of preparing more dilutions…). For test performance studies, the number of samples is usually limited due to the difficulty to get reference materials in sufficient quantity for 10-15 laboratories and the resources to test and prepare the samples for the panel. For a test performance study, raising the number of participating laboratories will also improve the reliability of the calculated performance characteristics but also need more resources.

Designing the sample panel and recruiting a sufficient number of laboratories for a test performance study therefore relies on a balance between the available resources (limiting the number of generated data) and the needs for a reliable statistical analyses (requiring more data). This section provides suggestions on sample panel and number of participating laboratories but the expertise of the organiser of the validation study / test performance study and its knowledge on the strengths and weaknesses of the test(s) to be validated remain crucial.

For further explanations, a tutorial video is available here: https://www.youtube.com/watch?v=AVxuEDxerGM

4.2 Number of participating laboratories

The number of participating laboratories in a test performance study affects the estimation of the reproducibility using the concordance of Langton *et al.* (2002) as well as the robustness of the calculation of other performance characteristics and of their confidence intervals.

To minimize this effect, it is proposed to have a minimum of 10 laboratories considered as proficient for a method as recommended in EPPO standard PM 7/122 (2014). As explained before, the inclusion of more laboratories will increase the reliability of the estimations of the performance characteristics but is constrained by available resources.

If the statistical analyses are carried out on the results delivered by a smaller number of laboratories, users should be aware that the conclusions of their analyses should be presented with caution and appropriate warnings. This is because the resulting performance characteristics will be subject to an increased uncertainty with possible incorrect estimations of the confidence intervals due to the limited number of laboratories.

4.3 Sample panel recommendations

The composition of the sample panel (e.g. the type and number of samples and, the number of replicates) depends on the intended use of the test, on the availability of reference material and on the experience of the organiser of the intra-laboratory validation study / test performance study. Figure 1 shows which samples are used to estimate the different performance characteristics of a test. The results obtained from the diluted infested samples are used to estimate the analytical sensitivity, the repeatability (if replicates are included as recommended) and the reproducibility (see sections 5.4 and 5.5). The results from infested and non-infested

[1] The magnitude of the uncertainty on a proportion (e.g. diagnostic sensitivity and diagnostic specificity) corresponds to $1/\sqrt{n}$, with n equal to the sample size.

samples are used to estimate the repeatability, reproducibility, the diagnostics sensitivity and diagnostics specificity and the related performance criteria (see sections 5.6 and 5.7).

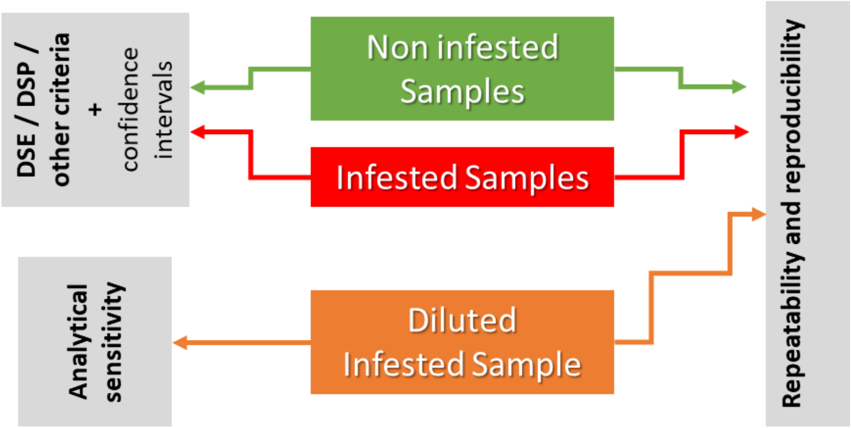
DSE / DSP / other criteria + confidence intervals
Non infested Samples
Infested Samples
Repeatability and reproducibility
Analytical sensitivity
Diluted Infested Sample

Figure 1. Schematic representation indicating the use of the results of each type of sample in the calculation of the performance criteria of a test. Healthy samples can be infested by another (closely related) species. Acronyms stands for: DSE: diagnostic sensitivity and DSP: diagnostic specificity. Other performance criteria are accuracy, false positive and false negative rates, rates of true positive and true negative, positive and negative likelihood ratios.

The organiser of the validation study / test performance study often knows the risks posed by the test(s) and, therefore, the most critical performance criteria to evaluate. This will impact the sample panel composition. For example, if the limit of detection is crucial for the intended use of the test and it is a potential weakness, the evaluation of analytical sensitivity will be a priority. On the other hand, if there is a risk of lack of exclusivity with another pest/organism, the panel should include several samples free from the target pest but infested by closely related species in order to evaluate the analytical specificity.

The type and number of samples, the number of dilution points and the number of replicates in a sample panel can influence the evaluation of the performance characteristics of a test. Indeed, the sample panel should include samples presenting specific difficulties for the pest detection, for example due to low concentration or genetic divergence of the isolate or the presence of a closely related species. For example, a sample panel consisting only of infested samples at very high concentration and “easy” to detect might result in 100% diagnostic sensitivity for all the evaluated tests and would limit the ability to discriminate the tests for this criterion. Some samples should therefore correspond to “worst case scenario” for the pest detection with the test(s) to allow a better comparison of them and, if relevant, their discrimination. Another drawback of having values of 100% for diagnostic sensitivity, diagnostic specificity or repeatability is the infinite values obtained for other performance criteria (see sections 5.6 and 5.7).

In preparation of the VALITEST test performance studies round 2, a sample panel was developed based on the trials of the statistical approaches on the 10 VALITEST datasets and with discussion with the organisers of the test performance studies. The outcome is a balance between statistical power and “practicality” for the organisers of such studies (Table 1). It should be noted that this proposed sample panel may not be feasible due to the limited availability of reference materials and should be adapted depending on the performance criteria to evaluate. The preparation of reference materials is critical and should follow the recommendations of the VALITEST deliverable 3.3 (Guidelines and Standard Operating Procedure (SOP) finalised for the production of the reference materials) (November 2019) and the future EPPO standard on reference materials (under development). It should be noted that the calculation of the performance criteria is done on the basis of expected results (i.e. based on the true status of the samples).

The proposed sample panel takes into consideration the following:

- For the determination of the analytical sensitivity (see section 5.5), a minimum of five different dilution points with a minimum of 3 replicates each, from a sample containing the target pest is required. The range of serial dilution must include the limit of detection determined by the organiser of test performance studies, i.e. the pest will not be consistently detected by the participating laboratories at the highest dilution point. The dilution series may be evaluated during the preliminary's studies in preparation of test performance studies and not included in the panel of samples. Depending on the scope of test performance studies, the evaluation of analytical sensitivity might not be required. For example, if a test is applied on pure culture or on individual pests.

- For the determination of the diagnostic sensitivity and diagnostic specificity (see section 5.6), the additional optional performance criteria (see section 5.7) and their corresponding confidence intervals (see section 5.3), a minimum of three samples free from the target pest (i.e. negative samples) and two samples infested with the target pest (i.e. positive samples) are required. These samples should be independent from each other, e.g. having different origins: not the same isolate (for infested samples) nor the same plant/batch of plants (for infested and non-infested samples). Furthermore, at least one of the two positive samples should have a low concentration of the target pest, i.e. close to the limit of detection estimated by the organiser of the test performance study. It might be important to include species closely related to the target pest in the negative samples to evaluate the analytical specificity of a test. The number of samples may be increased, for example when the dilution series is evaluated in preparation of test performance studies.

- For the determination of the repeatability and reproducibility from a test performance study (see section 5.4), the positive and negative samples described above should be processed at least in two replicates and the dilution series is included in the analysis as well. For the determination of the repeatability and reproducibility from an intra-laboratory validation study, a minimum of three replicates is recommended in EPPO standard PM 7/98 (2019).

Table 1. Panel of 25 samples for the determination of the performance criteria of diagnostic test using the statistical tools presented in sections 5.4 to 5.7.

Type of samples	Number of biological samples	Number of replicates	Dilution	Note
Negative	3	2	Not applicable	Negative samples should be independent from each other (i.e. not an aliquot of the same sample)
Positive	2	2	One positive sample has a medium concentration and the other one has a low concentration of the target pest (i.e. close to the limit of detection)	Positive samples should be independent from each other (i.e. not an aliquot of the same sample)
Serial dilution	1	3	Five dilution points with the highest dilution not consistently detected	Positive sample used for the dilution series should be independent to the other positive samples

4.4 Additional consideration on sample panel composition

The previous section underlines the importance of proper design of the sample panel and proposed an example of sample panel. It is essential to remind that this sample panel is indicative and that any sample panel will be influenced by many constraints: statistical power of the tests, availability of financial resources and reference samples, number of participating laboratories, risk analysis determining critical performance criteria based on preliminary test characteristic evaluation by the organiser of an intra-laboratory validation study / test performance study, intended use of the test, etc. In any case, the inclusion of replicates is strongly recommended and, if evaluating the analytical sensitivity, the determination of five dilution points is recommended, even in duplicates instead of triplicates if the number of samples is limited.

Table 2 shows a summary of sample panels used during the second round of a test performance study in the framework of the VALITEST project. This table exemplifies the adaptations made by the organisers depending on the different constraints and the scope of their test performance studies.

5 Performance criteria analysis

5.1 Structure of the validation dataset used as an example

5.1.1 Composition of the sample panel of the example dataset

The previous section describes the recommended composition of the sample panel, which is particularly important in order to calculate reliable estimates of the performance characteristics of the test(s).

The statistical analyses proposed in these guidelines are illustrated using the results from a reference dataset obtained from a VALITEST test performance study involving 16 laboratories and comparing two tests (called Test A and Test B) for the detection of a viral pest. The sample panel contained 22 samples as presented in Table 3. The panel included samples infected with the target pest, samples free from the target pest and a serial dilution prepared from a sample infected with the target pest. All the samples were processed in duplicates. The different performance criteria have been calculated from the obtained dataset and the results shown graphically (see sections 5.4 to 5.7).

Table 2. Information on the panel composition used during the second round of the VALITEST test performance study.

Pest	Minimum number of participants*	Maximum number of participants*	Number of samples in the panel	Use of biological replicates?	Number of samples infested with the target pest	Number of samples free from the target pest	Among which samples with close relative to the target pest	Dilution points	Replicates per dilution point	Scope of test performance study
Cryphonectria parasitica (Cp)	7	10	16	No	3	2	Yes	2	1 to 3	Detection of Cp in symptomatic and asymptomatic wood material using molecular tests
Plum pox virus (PPV)	9	14	22	Yes	5	4	No	2	2	Detection of PPV in symptomatic and asymptomatic leaves of *prunus* spp. using serological and molecular on-site tests
Tomato brown rugose fruit virus (ToBRFV)	26	34	22	No	2	2	No	5	2 for the least diluted 3 for other dilutions	Detection and identification of ToBRFV in symptomatic and asymptomatic leaves and fruit of tomato and pepper using serological and molecular tests
Tomato spotted wilt virus (TSWV)	13	17	22	Yes	2	7 (depends on the method)	Yes	5	2 for the least diluted 3 for other dilutions (depends on the method)	Detection and identification of TSWV in symptomatic leaves of tomato using serological and molecular tests
Xylophilus ampelinus (Xa)	3	11	22	Yes	2	2	Yes	5	2 for the least diluted 3 for other dilutions	Detection of Xa in symptomatic and asymptomatic stem material using serological and molecular tests

*if several tests were included, the least number of participants and the highest number of participants for a test are provided

Table 3. Details of the sample panel used in the analysis as an example. The status of a sample is either (F, in green) free from the target pest (in our case, a virus) or (I, in red) infected with the target pest at a medium concentration (not too close to the limit of detection). Dilution series (in orange) prepared from an infected sample different to the other infected sample, is expressed as I-10^{-2} for a 10^{-2} dilution and the sample originating from the same source as another sample (i.e. link between samples) is shown with the status (i.e. F or I) followed by the name of the other sample (e.g. F-A2).

Sample ID	Status	**Sample ID**	Status	**Sample ID**	Status	**Sample ID**	Status
A1	F-A2	**A6**	F*	**A11**	I-10^{-5}	**A16**	I-10^{-3}
A2	F-A1	**A7**	F*	**A12**	I-10^{-5}	**A17**	I-10^{-2}
A3	F	**A8**	F*	**A13**	I-10^{-4}	**A18**	I-10^{-2}
A4	F*	**A9**	I-10^{-6}	**A14**	I-10^{-4}	**A19**	I-A20
A5	F*	**A10**	I-10^{-6}	**A15**	I-10^{-3}	**A20**	I-A19
						A21	I-A22
						A22	I-A21

*These samples are free from the target pest but each is infected with a different species closely related to the target pest.

5.1.2 Lay out of results

Results should be gathered in a single document in a manner to facilitate the statistical analyses. For this, it is recommended to:

- Enter each test result, including technical replicate's results, in a single line in an Excel spreadsheet, with all the relevant information reported in different columns;

- Identify clearly each biological sample and when possible, with 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 box empty in the results column. Further details on how to deal with inconclusive and missing results are given in section 5.1.3.

The lay out of the results for the processing of the data of our example dataset is shown in Table 4.

Table 4. Copy of part of the results as an example of the lay out of the data in an Excel file for their processing with the statistical methods proposed in these guidelines.

Sample ID	Test name	Laboratory code	Technical replicate[1]	Test results[2]	True status[3]	Concentration/ quantity/ dilution[4]	Linked sample[5]	Sample info[6]
A4	A	L01	1	1	0			Non-target virus 1
A4	A	L01	2	1	0			Non-target virus 1
A5	A	L01	1	0	0			Non-target virus 2
A5	A	L01	2	0	0			Non-target virus 2
A6	A	L01	1	1	0			Non-target virus 3
A6	A	L01	2	1	0			Non-target virus 3
A1	A	L01	1	0	0		A2	Healthy host 1
A1	A	L01	2	0	0		A2	Healthy host 1
A2	A	L01	1	0	0		A1	Healthy host 1
A2	A	L01	2	0	0		A1	Healthy host 1
A3	A	L01	1	0	0			Healthy host 2
A3	A	L01	2	0	0			Healthy host 2
A7	A	L01	1	0	0			Non-target virus 4
A7	A	L01	2	0	0			Non-target virus 4
A8	A	L01	1	0	0			Non-target virus 5
A8	A	L01	2	0	0			Non-target virus 5
A19	A	L01	1	1	1	10^{-02}	A20	Target virus isolate 1
A19	A	L01	2	1	1	10^{-02}	A20	Target virus isolate 1
A20	A	L01	1	1	1	10^{-02}	A19	Target virus isolate 1
A20	A	L01	2	1	1	10^{-02}	A19	Target virus isolate 1
A21	A	L01	1	1	1	10^{-04}	A22	Target virus isolate 2
A21	A	L01	2	1	1	10^{-04}	A22	Target virus isolate 2
A22	A	L01	1	1	1	10^{-04}	A21	Target virus isolate 2
A22	A	L01	2	1	1	10^{-04}	A21	Target virus isolate 2
A9	A	L01	1	0	1	10^{-06}	A10	Target virus isolate 3
A9	A	L01	2	0	1	10^{-06}	A10	Target virus isolate 3
A10	A	L01	1	0	1	10^{-06}	A11	Target virus isolate 3
A10	A	L01	2	0	1	10^{-06}	A11	Target virus isolate 3
A11	A	L01	1	0	1	10^{-05}	A12	Target virus isolate 3
A11	A	L01	2	0	1	10^{-05}	A12	Target virus isolate 3
A12	A	L01	1	0	1	10^{-05}	A13	Target virus isolate 3
A12	A	L01	2	0	1	10^{-05}	A13	Target virus isolate 3
A13	A	L01	1	1	1	10^{-04}	A14	Target virus isolate 3
A13	A	L01	2	1	1	10^{-04}	A14	Target virus isolate 3
A14	A	L01	1	1	1	10^{-04}	A15	Target virus isolate 3
A14	A	L01	2	1	1	10^{-04}	A15	Target virus isolate 3
A15	A	L01	1	1	1	10^{-03}	A16	Target virus isolate 3

A15	A	L01	2	1	1	10^{-03}	A16	Target virus isolate 3
A16	A	L01	1	1	1	10^{-03}	A17	Target virus isolate 3
A16	A	L01	2	1	1	10^{-03}	A17	Target virus isolate 3
A17	A	L01	1	1	1	10^{-02}	A18	Target virus isolate 3
A17	A	L01	2	1	1	10^{-02}	A18	Target virus isolate 3
A18	A	L01	1	1	1	10^{-02}	A17	Target virus isolate 3
A18	A	L01	2	1	1	10^{-02}	A17	Target virus isolate 3

[1] For technical replicates, add a row for each technical replicate and name the replicate (1, 2, 3, etc…)

[2] Coding of test results: 0: negative result, 1: positive result, 2: inconclusive result and for missing result, leave the cell blank

[3] True status of the samples is coded as 0 for negative result and 1 for positive result

[4] Enter the dilution rate or concentration (e.g. cfu); If the sample was not diluted, then leave the cell blank

[5] Linked sample: Enter the sample reference ID that the sample relates to; If the sample is not linked to any sample, then leave the cell blank

[6] Sample info may be host name, pest name, name of isolate

5.1.3 Inconclusive and missing results

Results of an intra-laboratory validation study or a test performance study can sometimes be inconclusive or missing. A result is missing when the laboratory has not reported it.

Missing results may happen when a test could not be conducted because of, for example, sample degradation, failure of nucleic acid extraction, failure of controls, ran out of samples/specific reagents for a test. A result is considered inconclusive when it was not possible for the laboratory to decide whether the sample should be assigned positive or negative. This may happen when it is closed to the limit of detection of the test (i.e. the grey zone area). For example, when the cycle threshold (Ct) is close to the cut off Ct for a real-time PCR. The participating laboratories can react differently when they observe an inconclusive result: some might just indicate this statement while others might do additional testing to clarify. Whatever the situation, the operation carried out should be clearly stated in the report.

The data analysis of the first round of VALITEST test performance studies reveals that inconclusive and missing results were treated differently by the organisers of these studies. These results were either considered as false results or excluded from the analysis, for example when it was not possible to make a justified interpretation of the results reported as inconclusive. In some cases, a value of 0.5 instead of 0.0 was added when the absence of a category (i.e. false positive/negative, true positive/negative) would prevent the calculation of performance characteristics. This could potentially affect the results of the analysis of test performance studies. The inconclusive results should be treated with caution to minimise biases to the analyses. Thus, some guidelines on how to deal with inconclusive and missing results are proposed here.

For the calculation of the analytical sensitivity, inconclusive results were treated as false results, i.e. pest not detected in the diluted sample(s). The missing results in contrast were not included in the analytical sensitivity analysis. Indeed, inconclusive results might be due to the very low amount of pest in the diluted samples and they should be taken into account while missing result are independent of the dilution level.

For the calculation of the repeatability and reproducibility, both inconclusive and missing results were excluded in the analysis. For the calculation of all other performance criteria with their corresponding confidence intervals, both inconclusive and missing results were included in the analysis. Missing results were treated the same way as inconclusive results (i.e. considered as false positive or false negative), as they would affect the performance of the corresponding test. These decisions on including or not missing or inconclusive results are indicative as the validation or test performance organiser should consider these results based on his/her expertise and knowledge on the pests and the tests.

5.2 Outlier results

5.2.1 *Background*

Before and during the statistical analysis and the interpretation of results, outliers can be identified, for example for a sample or a participating laboratory, and should be checked. This is because their presence in the dataset may introduce bias and affect the analysis of intra-laboratory validation studies or test performance studies. There is a whole range of methods that allow results to be designated as outliers. However, this statistical characterisation is only a marker aimed at drawing the attention of the diagnostician. The outlier identification relies mainly on the expertise of the organiser of an intra-laboratory validation study / test performance study.

Different types of outliers can be identified for further examination and possible exclusion from the statistical analysis including data sets for which:

1) Results for the controls were not as expected,
2) There are missing results,
3) Results very different for one laboratory from the rest of the participating laboratories,
4) Results for one sample different from the expected results for all the participating laboratories.

5.2.2 *Proposed approach for outlier identification*

Outliers can be detected by calculating performance criteria at laboratory and/or sample level and looking for strong individual deviations among them. Those deviations can be investigated by a diagnostician with some insight in statistics, leading to the possible exclusion of the corresponding data from the general analysis, for example if it can be established that they are the consequence of contamination or a specific deviation is identified. Even in the case of very high diagnostic sensitivity or diagnostic specificity (as shown in our example for diagnostic sensitivity, Figure 9), those parameters can still be useful to detect outliers, which show below than average performance event in those situations. A smooth communication with the laboratory having generated these potential outliers can help in identifying the root cause of these results and to avoid elimination of results relevant for the statistical analysis.

As a non-exhaustive list, the following analyses can be useful to identify outliers:

- Accordance and concordance for each sample (see section 5.4, Figures 2 and 4)
- Accordance for each laboratory (see section 5.4, Figure 3)
- Analytical sensitivity for each laboratory (see section 5.5, Figure 7)
- Diagnostic sensitivity and diagnostic specificity for each laboratory (see section 5.6, Figure 9)

5.3 Confidence intervals

5.3.1 *Background*

We propose the optional use of confidence intervals, a statistical tool that assesses the quality of the estimation of a parameter (e.g. performance criteria such as diagnostic sensitivity, diagnostic specificity, repeatability). The estimated confidence intervals give a range of values that will contain the real value of the parameter with a fixed probability (usually a 95% confidence level is used). For example, if 100 confidence intervals of a performance criteria are estimated from 100 independent sampling [2] in the whole population, a 95% confidence level means that, on average, 95 confidence intervals will contain the real value of the performance criteria. Confidence intervals help to visualize the uncertainty of the estimation of a parameter obtained from a limited number of samples, as compared to the population (Hess *et al*., 2012; Erdoğan and Gülhan, 2016).

Confidence intervals are useful:

- When the number of datasets used to calculate and compare a performance characteristic of several tests is different.
- To evaluate if there is a significant difference between 2 observed values for the same performance criteria (for example diagnostic specificity): differences can approximately be interpreted as significant (with an associated risk of 5%) when 95% confidence intervals do not overlap. However, since the comparison of confidence intervals is not strictly equivalent to a statistical test, but only an approximation, an appropriate statistical test should be carried out to strictly evaluate the statistical difference and its associated probability.
- When the measurements (associated to the observed values of the compared performance criteria of interest) present a contrasted variability, with either high or low variability, between tests or laboratories.

A two-sided confidence interval of 95 % is commonly used in the scientific fields and the same confidence level can be used to evaluate the performance of plant health diagnostic tests as it provides an estimate with a risk level kept at a reasonable level (i.e. 100%-95% = 5%).

For further explanations, a tutorial video is available here: https://www.youtube.com/watch?v=9Lq7bZaJ4Mc

5.3.2 Proposed statistical methods for calculating confidence intervals

The methods used to determine confidence intervals are linked to the statistical methods used to calculate the performance criteria as shown in Table 5.

It was decided not to determine the confidence intervals for the analytical sensitivity, repeatability and reproducibility in the context of plant health diagnostic because of the complex calculation required.

[2] In the case of a test performance study, a sampling corresponds to the subset of laboratories participating in that study compared to all the laboratories providing pest diagnostic; for an intra-laboratory validation study, this corresponds to the subset of infested matrix (plant, plant organ, water, soil, etc) or isolates compared to all the infested matrix or the genetic diversity of the pest.

Table 5. Confidence interval's methods for the performance criteria measuring the effectiveness of a diagnostic test.

Performance criteria (EPPO standard PM 7/98, 2019)			
	Method of calculation of the parameter	**Confidence interval's method**	**Note on the calculation of confidence intervals**
Analytical sensitivity	Probability of detection (see section 5.5)	Generalized linear models	Not applied for this deliverable as it is more complex to perform
Diagnostic sensitivity	(see section 5.6)	Agresti-Coull method (Massart *et al*., 2008 and 2009)	Easy to calculate and good coverage for extreme values of ratio
Diagnostic specificity	(see section 5.6)	Agresti-Coull method (Massart *et al*., 2008 and 2009)	Easy to calculate and good coverage for extreme values of ratio
Repeatability	Accordance (see section 5.4)	Derived from bootstrap standard errors (Langton *et al*., 2002)	Not applied for this deliverable as it is more complex to perform
Reproducibility	Concordance (see section 5.4)	Derived from bootstrap standard errors (Langton *et al*., 2002)	Not applied for this deliverable as it is more complex to perform
Other performance criteria (optional)			
	Method of calculation of the parameter	**Confidence interval's method**	**Note on the calculation of confidence intervals**
Accuracy	See Table 10 (see section 5.7)	Agresti-Coull method (Massart *et al*., 2008 and 2009)	Easy to calculate and good coverage for extreme values of ratio
Diagnostic odds ratio	See Table 10 (see section 5.7)	Simple OR CI (Fleiss *et al*., 2003)	Fairly easy to calculate
False positive and false negative rates	See Table 10 (see section 5.7)	Agresti-Coull method (Massart *et al*., 2008 and 2009)	Easy to calculate from the CI of diagnostic sensitivity and diagnostic specificity
Positive and negative predictive values	See Table 10 (see section 5.7)	Agresti-Coull method (Massart *et al*., 2008 and 2009)	Easy to calculate and good coverage for extreme values of ratio
Positive and negative likelihood ratios	See Table 10 (see section 5.7)	Simel method (Simel *et al*., 1991)	Easy to calculate and good coverage for extreme values of ratio

5.3.3 Application on example dataset

Determination of confidence intervals

The confidence intervals obtained, based on the example dataset, for the performance criteria proposed in these guidelines and their interpretation, are presented here below in each criteria's corresponding section. The calculation of the confidence intervals was performed from the positive and negative samples only (not the dilution series, see section 4.3 for details) using the formula described in Annex 1.

Interpretation of confidence intervals

In general terms, 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). Its usefulness for analysing the results of an intra-laboratory validation study or test performance study, between tests and/or laboratories is explained in sections 5.6.3 and 5.7.3.

5.4 Repeatability and reproducibility

5.4.1 Background

The EPPO standard PM 7/98 (2019) provides recommendations for the assessment of repeatability and reproducibility for each type of method per discipline. For example, for the validation of serological and molecular tests used in plant health diagnostic, EPPO standard PM 7/98 (2019) recommends to "*analyse at least three replicates of spiked sample extracts with a low concentration*" for the assessment of the repeatability and the same is recommended for the assessment of the reproducibility "*but with different operators, if possible, on different days and with different equipment when relevant*". For the verification of serological and molecular tests, EPPO standard PM 7/98 (2019) recommends to "perform at least three simultaneous tests on the same material with low levels of target". EPPO standard PM 7/98 (2019) does not provide specific advice on the analysis of the data for the repeatability and reproducibility.

For further explanations, a tutorial video is available here: https://www.youtube.com/watch?v=T3h6Cipteks

5.4.2 Proposed approach for repeatability and reproducibility evaluation

We propose to estimate the repeatability (within a laboratory) and reproducibility (between laboratories) of a detection test by calculating the accordance and concordance, respectively. The calculation is based on simple counts of concordant and non-concordant results between replicates (whatever the true status of the samples) and is easy to calculate. These measures evaluate the probability of achieving the same test results for identical samples within (accordance) and between laboratories (concordance) (Langton *et al.* 2002). In the plant field, accordance and concordance are used by the International seed testing association (ISTA[3]) and ANSES based on the recommendation of the standard ISO 16140-2 in the 2003 version (Chabirand *et al.*, 2017).

Accordance and concordance can be calculated per test, per laboratory or per sample. For this calculation, inconclusive and missing results are excluded from the analyses (see section 5.1.3).

At the level of the test, it represents the repeatability of the test (performance characteristic that is usually reported), e.g. the agreement between the results from replicates of all samples in all laboratories. At the level of the sample: it is used to identify samples that give discordant results for technical and/or biological replicates analysed at the same time under the same conditions in the participating laboratories. At the level of the laboratories: it is used to identify laboratories that produce discordant results between technical and/or biological replicates, whatever the samples. The accordance will be calculated for the test (representing the impact of

[3] ISTA, Seed Health Committee (SHC) Tool box, https://www.seedtest.org/en/shc-tool-box-_content---1--1410--811.html, accessed June 2020

different conditions [i.e. reproducibility] on the agreement of results) and for each sample. The calculation of accordance is based on the agreement of all the results analysed by pair as illustrated in Table 6.

Table 6. Example of data generated by three laboratories (Lab 1, Lab 2, Lab 3) testing two samples (Sample X and Y) in duplicates (Replicate A and B) for one test (Test 1) for the calculation of accordance and concordance. Coding of each result is as follows: **+** or **–** for positive or negative results, **XA1** means results of replicate A for sample X by Laboratory 1.

TEST 1		Lab 1	Lab 2	Lab 3
Sample X	Replicate A	**+** **(XA1)**	**-** **(XA2)**	**+** **(XA3)**
	Replicate B	**+** **(XB1)**	**-** **(XB2)**	**+** **(XB3)**
Sample Y	Replicate A	**-** **(YA1)**	**-** **(YA2)**	**-** **(YA3)**
	Replicate B	**-** **(YB1)**	**-** **(YB2)**	**+** **(YB3)**

Based on our example of Table 6, the pairs for accordance and concordance are as follows:

Accordance per laboratory: pairs of results analysed:

- Lab1: 2 (XA1=XB1 , YA1=YB1)
- Lab 2: 2 (XA2=XB2 , YA2=YB2)
- Lab 3: 2 (XA3=XB3 , YA3 ≠ YB3)

Accordance per sample: pairs of results analysed:

- Sample X: 3 (XA1=XB1 , XA2=XB2 , XA3=XB3)
- Sample Y: 3 (YA1=YB1 , YA2=YB2 , YA3 ≠ YB3)

Concordance for sample X: 15 pairs of results analysed:

- (XA1=XB1 , XA1 ≠ XA2 , XA1 ≠ XB2 , XA1=XA3, XA1=XB3, etc …..)

Concordance for test 1: 30 pairs of results analysed (15 pairs per sample):

The accordance and concordance estimates can also be used to calculate the concordance odds ratio (COR) by samples and by tests for the estimation of the level of variation between laboratories. Concordance odds ratio removes the bias related to the accuracy of the results (i.e. numbers of true positive/negative and of false positive/negative related to the true status of the samples) which are used to calculate the two parameters (i.e. concordance and accordance) taken separately. Thus, the magnitude of the ratio provides the relative chance ('odds' in betting terms) of getting the same result when two samples are analysed in the same laboratory compared to if they are sent to the 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 (both positive or both negative) if they are analysed in the same laboratory than if they are analysed in different laboratories.

As additional information, the p-value (Fisher test) can be determined for the concordance odds ratio (see section 5.4.3) to underline if there is a significant difference between the tests and thus can assist in the selection of a test

5.4.3 Application on example dataset

Accordance, concordance and concordance odds ratio per test

The results of the repeatability and reproducibility of the tests have been determined using the accordance and concordance as well as the concordance odds ratio of Langton *et al*. (2002) (Table 7). Further information on the calculations for the accordance, concordance and concordance odds ratio can be found in Annex 2.

Table 7. Estimates of the accordance, concordance and concordance odds ratio by test for the estimation of the repeatability, reproducibility and the estimation of the degree of variation between laboratories, respectively.

Criteria	Test A	Test B
Accordance	99 %	100 %
Concordance	91 %	92 %
Concordance odds ratio	9.8	Infinite

Tests A and B have very similar values of accordance (repeatability) and concordance (reproducibility).

If concordance is smaller than accordance, it indicates that two replicates are more likely to give the same result if they are analyzed by the same laboratory than if they are analyzed by different ones, suggesting that there can be variability in performance between laboratories. When the variation between the results from different laboratories increases, compared to the variation of the results within each laboratory, the COR value increases. The value of the COR therefore indicates if the test is robust enough to reproduce the same results under different laboratory conditions (the lower the COR, the more robust the test). Note that the accordance, concordance and concordance odds ratio calculation can be refined by sample and by laboratory for the accordance (see below).

The concordance odds ratio value of test B is infinite while the concordance odds ratio value for test A is 9.8. The infinite concordance odds ratio value is a result of the accordance value being 100% and, in this case, this criterion is not very informative to compare the tests and no statistical comparison can be carried out.

Accordance values per sample and per laboratory

Calculating accordance values for each sample is possible when replicates are included in the sample panel. It can help detecting samples with a particularly low or high repeatability of the test results. This parameter can be split by laboratory or by test to give further insight about the factors influencing the results.

Accordance values calculated per sample (Figure 2) shows that the replicates of all the samples provided the same result with test B (i.e. accordance of 100%) while two samples (samples A1 and A7) gave some discordant results between replicates using test A (with accordance values of 94%).

This example highlights the potential occurrence of outlier results (with samples A1 and A7 for test A) that may indicate potential issues with samples or laboratory(ies). An investigation to find out the source of the issue, by for example questioning the laboratories and checking the samples preparation, may be useful, in order to determine how these results should be considered in the analyses (e.g. outliers or not).

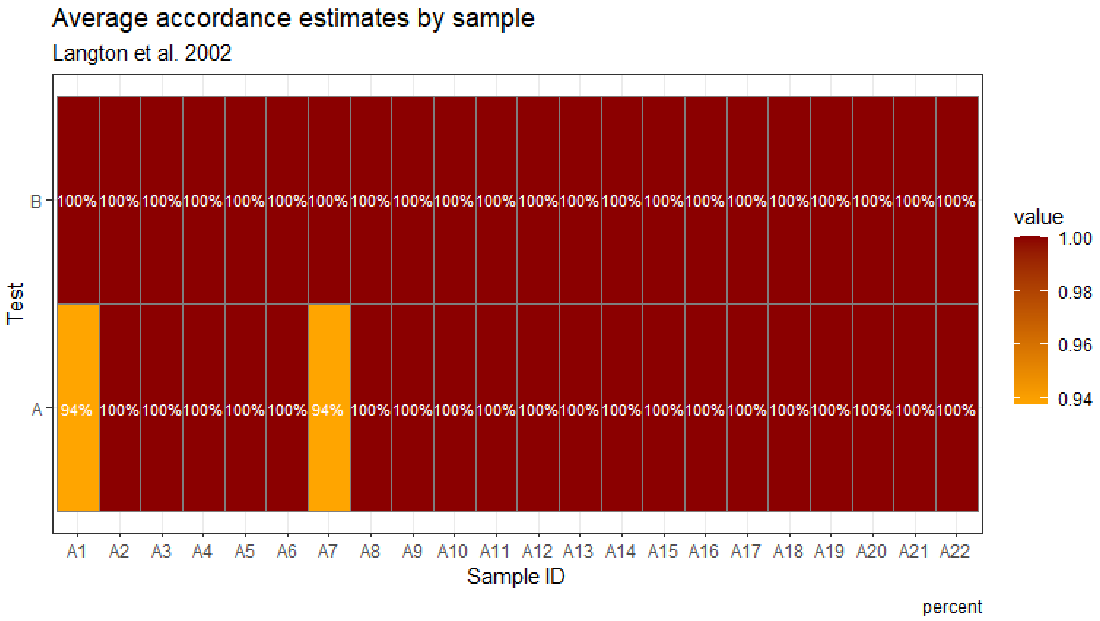

Figure 2. Accordance values by sample considering the results of all participating laboratories for the estimation of the repeatability. Accordance is bound between 0 and 1 and can be expressed in percentage: 0 means that not a single pair of replicates shows the same value, and 1 that all the replicates have the same result.

Figure 3 shows the accordance calculated for each laboratory per test comparing the two replicates of each sample. This figure underlines a very good repeatability (100% of accordance) for all laboratories except a single one, L05 with an accordance of 91%. The analysis of Figures 2 and 3 reveals that only the laboratory number L05 did not get repeatable results with samples A1 and A7, which explains the lower value obtained for test A. This observation adds further insight on the results of the average values of accordance, concordance and concordance odds ratio calculated in Table 7. We can also observe that laboratories L09 and L10 did not analyse replicates of samples for test B, hence does not have any accordance estimate for this test.

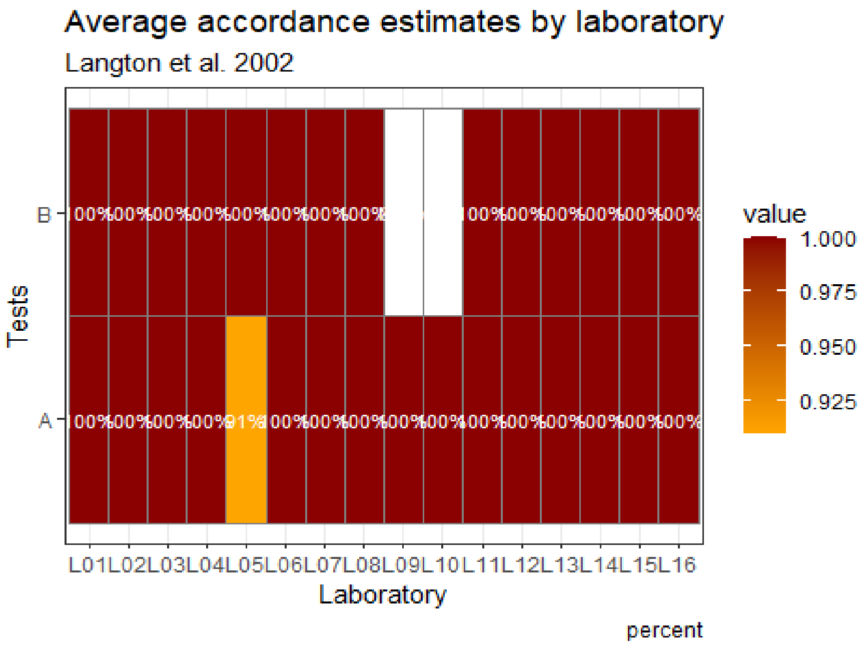

Figure 3. Accordance values by laboratory for the estimation of the repeatability. Accordance is bound between 0 and 1 and can be expressed in percentage: 0 means that not a single pair of replicates shows the same value, and 1 that all the replicates have the same result.

Concordance values per sample

Calculating concordance estimates for each sample based on the results obtained from different laboratories is useful to highlight the samples leading to high discrepancies between laboratories. From Figure 4, we can see that samples A4 and A6 led to higher variability between laboratories for Test A. On average, in our example, more samples provide the same results between laboratories with Test B than Test A, indicating that Test B is slightly more reproducible than Test A (see Table 7).

Similarly, to the accordance per sample, an investigation to find out the source of the issues with samples A4 and A6 for test B, by for example questioning the laboratories and, checking the samples preparation, may be useful, in order to determine how these results should be considered in the analyses (i.e. outliers or not).

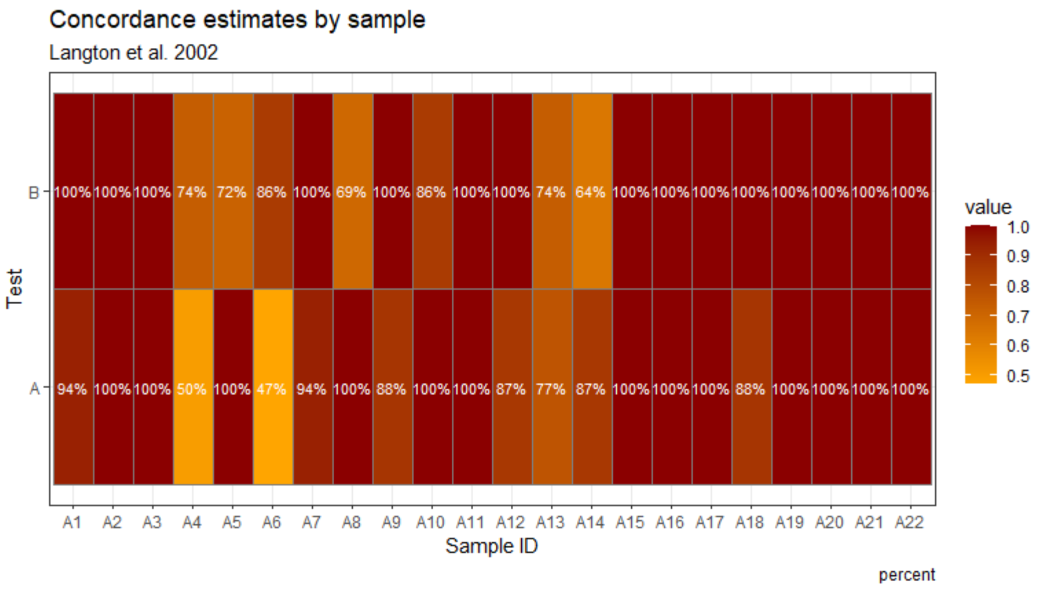

Figure 4. Average concordance values calculated for each sample based on the test results from all the participating laboratories for the estimation of the reproducibility. Concordance is bounded between 0 and 1 and can be expressed in percentage: 0 means that laboratories can provide different values, and 1 that all the laboratories provided the same result.

Concordance odds ratio per sample

The degree of inter-laboratory variation was also assessed by calculating the concordance odds ratio (COR) for each sample (Figure 5).

The larger the concordance odds ratio, the more predominant the inter-laboratory variation. For concordance odds ratio values above 1.00, the Fisher's exact test can be used, to evaluate the statistical significance of the variation between laboratories. A significant result (p-value below 0.05) indicates a significant inter laboratory variation. As an example, we can see here that some samples are associated with increased inter laboratory variability for both tests: A4, A6, A13, A14. For those samples, discordant results between laboratories are more frequent than between intra laboratory replicates.

In this case, the concordance odds ratio is of little help to discriminate all but the least efficient tests, as most of the estimates are either 1 or infinite values.

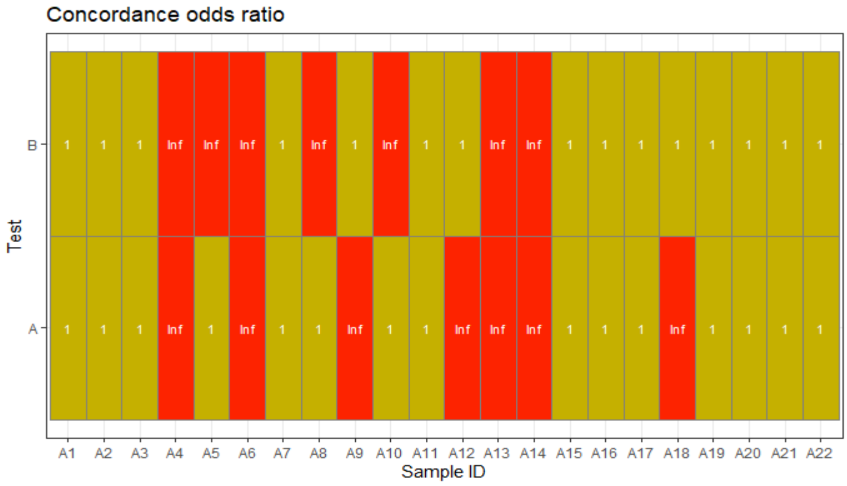

Figure 5. Concordance odds ratio based on the accordance (within laboratories) and concordance (between laboratories) estimates by samples. Red blocks indicate a significant Fisher's test result (p-value < 0.05), hence a significant inter laboratory variability. Light green blocks illustrate p-value >0.05.

5.5 Analytical sensitivity

5.5.1 Background

The EPPO standard PM 7/98 (2019) provides recommendations to assess the analytical sensitivity for different methods and discipline. For example, EPPO standard PM 7/98 (2019) recommends conducting at least three experiments with serial dilutions for the estimation of the analytical sensitivity for the validation of serological and molecular tests. For the verification of serological and molecular tests, EPPO standard PM 7/98 (2019) recommends "*to analyse at least eight samples at the established limit of detection*". In both cases, EPPO standard PM 7/98 (2019) does not advise on data analysis.

The determination of analytical sensitivity in other fields such as analytical chemistry and microbiology, is based on a statistical concept called the 'probability of detection' which is the probability of the target microorganism being detected (Wehling *et al.* 2011, Uhlig and Gowik 2018). The probability of detection model allowed a better discrimination between tests allowing the identification of the least performant. The model developed by Wehling *et al.* (2011) was approved by the Association of Official Agricultural Chemists (AOAC) and included in their guidelines for validation of microbiological methods for food and environmental surfaces (AOAC, 2012). A different probability of detection's model has also been used in ISO 16140-2 (2016), a standard for method validation for microbiology of the food chain, to compare an alternative method against a reference method. This model is also recommended in the method validation guidelines developed by ANSES (2015). Probability of detections is briefly mentioned in EPPO standard PM 7/122 (2014) with reference to Wehling *et al.* (2011)'s model. The generalized linear models have rarely been used in plant health diagnostics (Chabirand *et al*., 2017; Massart *et al*., 2008).

For further explanations, a tutorial video is available here: https://www.youtube.com/watch?v=ibHwKW448io

5.5.2 Proposed approach for analytical sensitivity evaluation

We propose the use of a probability of detection (POD) model called generalized linear model as recommended in ISO 16140-2 (2016), for the determination of the analytical sensitivity, but with a logit link function (log[p/(1-p] where p is the probability of detection). We choose the logit link because the tools to adjust this model are much widely available, and the resulting estimates are very close from the complementary log log link model. In this model, the probability (expressed in percentage) of detecting a target is a function of the target concentration as a continuous variable. The resulting graph helps in 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). Inconclusive results are treated as false results, i.e. counting as false negative or false positive depending on the real status of the sample while missing results are not included in the analysis (see section 5.1.3).

The generalized linear model does not require any assumption on the number of technical and/or 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 (see sections 4.3 and 5.5.3). In order to meet the validation requirements of PM7/98 (2019), the serial dilution should be repeated on the same material at least three times. The model can be used for one or more diagnostic tests.

The limit of detection calculated using the probability of detection model can be absolute or relative. For some pests (e.g. bacteria, fungi, nematode), it is possible to determine an absolute level of the limit of detection with the probability of detection model that may be expressed for example in number of cells, spores, cysts. In the case of pests for which the concentration cannot be quantified (e.g. viruses, viroids, phytoplasmas), a relative limit of detection can be determined, where the probability of detection model is relative to a dilution level but could be used to compare two or more diagnostic tests. The relative limit of detection can also be adopted for quantifiable pests.

The probability of detection model can be used to compare different tests using fixed levels of detection probability. We used here two levels that are often referred in the scientific publication: 50% and 95%. A level of 50% probability of detection means that, at this dilution level, the pest can be detected in 50% of the tests carried out on average.

5.5.3 Application on example dataset

Determination of the probability of detection per test

For each test, results from the serial dilution samples were used to adjust binomial generalized linear models (bGLM) with logit link between the dilutions (regardless of the dilution factors, for example dilution of 10th as shown in Figure 6) which are expressed by the base 10 negative exponent of the corresponding dilution, and the detection status. The inconclusive results were included while the missing results were excluded for this analysis (see section 5.1.3).

A model, corresponding to a function representing the estimated probability of detection depending on the dilution factor, is generated for each test using for example R software (Figure 6), dilutions corresponding to a 50% or 95% probability of detection have been calculated as an example of the possible LOD to report (LOD50 and LOD95 – see Table 8). Further information on the calculations for the probability of detection can be found in Annex 3.

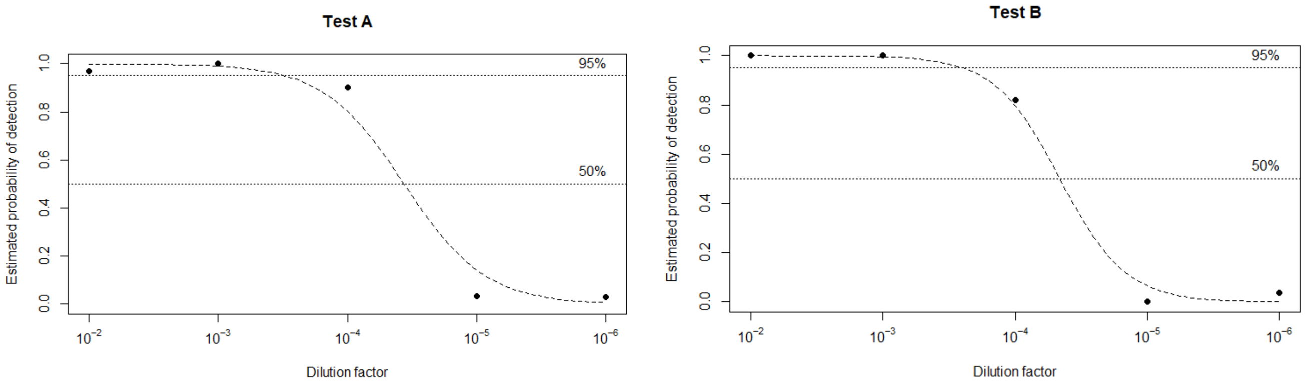

Figure 6. Probability of detection (Y axis) based on the dilution factor (X axis) for the target pest for tests A and B, with data combined from all the laboratories. The dotted lines are the 50% and 95% probabilities of detection.

Table 8. Dilution values (expressed as log dilution factor) for the target pest by the two tests for 50% and 95% of probability of detection.

Limit of detection (LOD)	**Dilution values**	
	Test A	**Test B**
LOD at 50%	$10^{-4.4}$	$10^{-4.3}$
LOD at 95%	$10^{-3.5}$	$10^{-3.6}$

In our example, tests A and B have a very similar levels of detection for the probability of detection rates of 50% and 95%. For comparing several tests with similar probability of detection more accurately, the confidence intervals can be calculated for each test. However, the calculation of the confidence intervals from generalized linear models is more complex and would require more advanced statistical knowledge. Therefore, it was not applied in our example.

In order to evaluate the impact of the laboratory on the model, the analysis can also be carried out for each laboratory independently as shown in Figure 7.

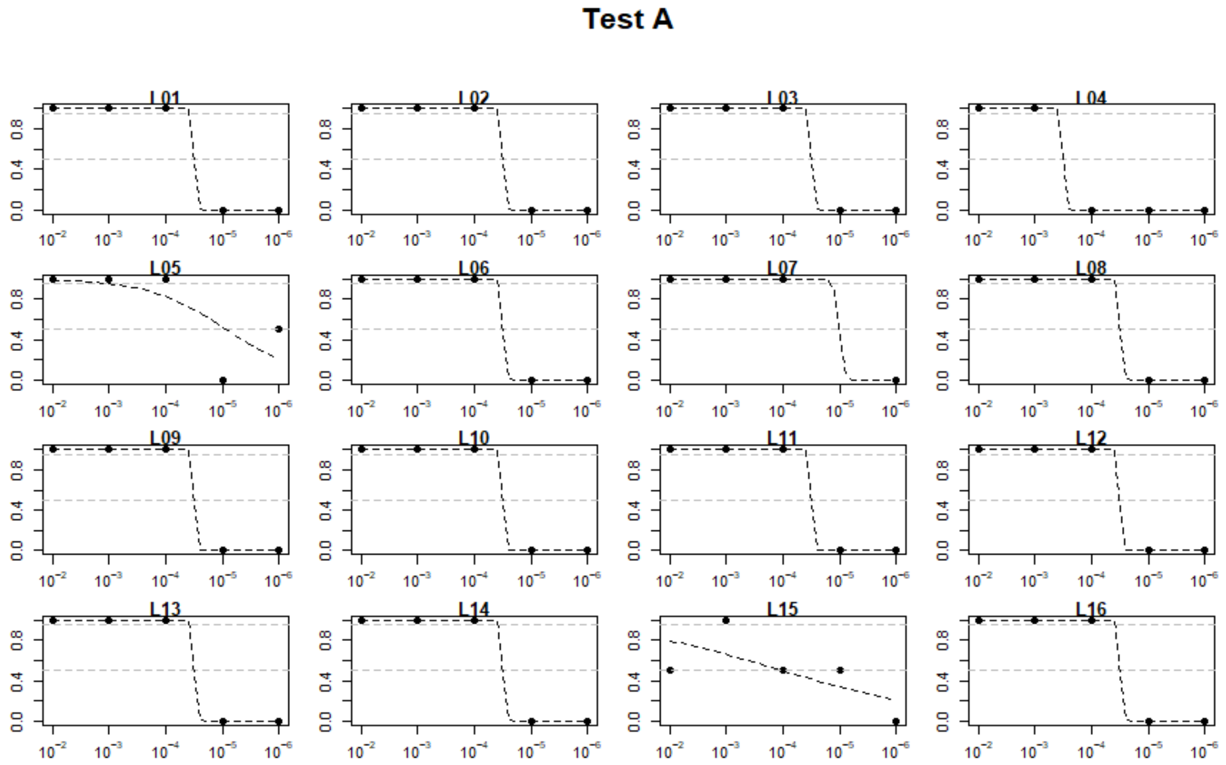

Figure 7. Determination of the probability of detection (Y axis) based on the dilution factor (X axis) for the target pest for Test A per laboratory. The dotted lines are the 50% and 95% probabilities of detection.

Interpretation of the probability of detection

Once the graphs have been built (see Figure 7), the diagnostician can select the most appropriate probability of detection level fitting the application of the diagnostic test. The selection of the threshold of the probability of detection, for example 95% or 99%, impacts the limit of detection of the pest by tolerating the non detection of a pest at this concentration in approximately 5 or 1% of the analysed samples respectively. It should be based on a risk analysis taking in consideration the severity of the pest, the purpose of the test (e.g. detection or identification, pre-screening or confirmatory test), the cost of the test, the probability of arrival, the phytosanitary status, etc… (see EPPO standard PM7/98 (2019) for points that could be considered in the risk analysis).

Caution is required when using a binomial generalized linear models (bGLM) with logit link between the target concentration (expressed in log scale) and the detection status to calculate LOD from data of the diluted samples for each method and each laboratory. Those models are based on the assumption that the probability of detection is decreasing when the dilution level increases. This hypothesis can be violated in different cases:

- When all the samples show the same status whatever the dilution level;
- When the observed detection rate shows contradictory behavior (e.g. decrease then increase again).

When this happens, the model cannot be trusted. Hence, diagnostician should verify that the model is fit before interpreting the calculated limit of detection value. An example of a problematic case is observed with laboratory 05 for Test A in Figure 7. The organizer of test performance studies should pay attention to the other values for performance criteria obtained by the different laboratories. In this case, it was noted that the value for repeatability was lower for laboratory L05 (see section 5.4.3). Another example is the POD result of laboratory L05 which reveals decreases then increases of the detection rate (high detection rate at the highest dilution level with Test A). This laboratory has a 100% repeatability, suggesting that samples may have been swapped around during the preparation of the test or samples may not have been evenly mixed. The results of laboratory L15 did not allow the determination of the limit of detection due to incongruent results (i.e. decrease then increase again). These odd results may come from for example, cross contamination or mixing of samples.

5.6 Diagnostic sensitivity and diagnostic specificity

5.6.1 Background

Diagnostic sensitivity and diagnostic specificity performance criteria are defined in EPPO standard PM 7/76 (2018) and applied to the comparison of a test A with a validated test B in EPPO standard PM 7/98 (2019). These standards do not provide any advice on the interpretation of the analysis of the data.

5.6.2 Proposed approach for diagnostic sensitivity and diagnostic specificity evaluation

We recommend the calculation explained in EPPO standard PM 7/98 (2019) for the determination of the diagnostic sensitivity and the diagnostic specificity for the validation of a new test by comparison with a validated test (see Annexe 4). In addition, we propose the calculation of confidence interval at 95% for these two criteria (see section 5.3).

We also propose to broaden the EPPO approach by including the option for a newly developed test using reference samples when, for example the new test is the only available test for the detection of a pest using reference material identified by another method, such as morphology, biochemical tests. In this case, the following grid is proposed:

		True status of reference sample	
		Target present	Target absent
Test result	Positive	**TP**	FP
	Negative	FN	**TN**

TP: true positive, FP: false positive, FN: false negative, TN: true negative

TP and TN: the result of the test is in agreement with the true status of the reference samples (i.e. positive or negative results)

FP and FN: the result of the test is not in agreement with the true status of reference samples as being infested with the target pest (i.e. false negative or false positive)

In this scenario, the formula for determining of the diagnostic sensitivity and diagnostic specificity is:

Diagnostic sensitivity: $\frac{TP}{TP+FN}$ Diagnostic specificity: $\frac{TN}{TN+FP}$

The confidence intervals are proposed to be determined for the diagnostic sensitivity and diagnostic specificity, as an optional tool for the data analysis. Further details on confidence intervals are explained in section 5.3 and Annex 1.

5.6.3 Application on example dataset

Determination of the diagnostic sensitivity and diagnostic specificity per test

The diagnostic sensitivity and diagnostic specificity of two tests with their corresponding confidence intervals have been calculated according to the true status of the reference material used during the test performance study. For those performance estimations, we excluded the data from the dilution series, keeping only the non-diluted sample of each series, so that the results are estimated on independent samples only (all the samples of a dilution series are connected to each other as they derive from the same biological sample). Both the inconclusive and missing results were included for this analysis (see sections 4.3 and 5.1.3). The results are given together with their respective lower and upper 95% confidence levels (see section 5.3).

In our example, Tests A and B for a target pest have a very similar diagnostic sensitivity with confidence intervals overlapping (Table 9, Figure 8). On the other hand, there is a significant difference between the two tests with regard to the diagnostic specificity. The diagnostic specificity of Test A (with a value of 88%) is more than 1.5 higher than the diagnostic specificity of Test B (with a value of 55%). This means that false positive results are more likely to be obtained with Test B than with Test A. In addition, the two confidence intervals are clearly distinct suggesting that the real value of the diagnostic specificity for the tests are different (Table 9, Figure 8). This can be critical when the intended use of the test in the diagnostic context requires detection test to have high diagnostic specificity.

Table 9. Diagnostic sensitivity (DSE) and diagnostic specificity (DSP) of Test A and Test B for a viral pest with their lower and upper 95% confidence levels (LCL and UCL).

Criteria	**Test A**	**Test B**
DSE	99%	100%
LCL_{DSE}	96%	97%
UCL_{DSE}	100%	100%
DSP	88%	55%
LCL_{DSP}	84%	49%
UCL_{DSP}	92%	62%

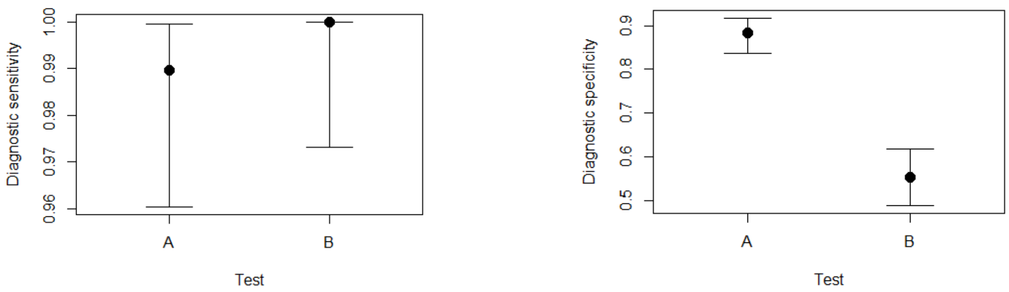

Figure 8. Graphical representation of the 95% confidence intervals for the diagnostic sensitivity and diagnostic specificity of Test A and Test B for a target pest.

It should be noted that the diagnostic sensitivity and the diagnostic specificity are heavily dependent on the choice of the positive and negative reference samples (see section 4). If the number of samples vary between laboratories and tests, it is particularly important to include confidence intervals in the analysis. Furthermore, the comparison between tests should be interpreted with caution.

Determination of the diagnostic sensitivity and diagnostic specificity for each participating laboratory

The calculation of diagnostic sensitivity and diagnostic specificity for each laboratory can be useful to spot laboratory with divergent results compared to the other participating laboratories, i.e. to identify outliers (see section 5.2). For example, analyses of the diagnostic sensitivity using non diluted infested samples revealed that one laboratory (L15) has 83% diagnostic sensitivity with Test A while all other laboratories have 100% diagnostic sensitivity (Figure 9), suggesting that results of laboratory L15 for Test A may be potential outliers. Similarly, analyses of the diagnostic specificity using negative samples could show poor performance of some laboratories, tests or a combination of the two (Figure 9). In this case example, Test B resulted in a lower diagnostic specificity in a majority laboratory as compared to Test A.

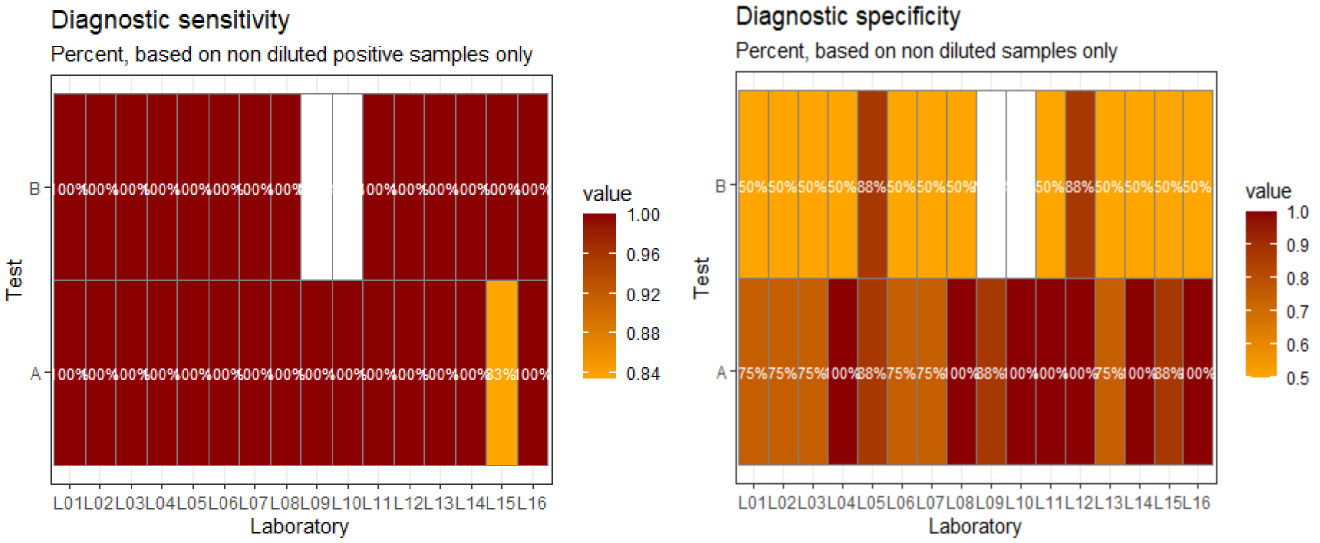

Figure 9. Determination of the diagnostic sensitivity (DSE) and diagnostic specificity (DSP) per laboratory, to assist in identifying outliers. DSE and DSP are bound between 0 and 1 and can be expressed in percentage: 0 means that all the results are wrong (false positive/negative results), and 1 that the expected results are obtained for all the samples.

Interpretation of the diagnostic sensitivity and the diagnostic specificity

A test with a high diagnostic sensitivity strongly indicates a high probability of detecting a pest when it is present in a sample while a test with a high diagnostic specificity has a high probability of diagnosing the absence of a pest when it is truly absent. It should be noted that, when comparing several tests, the test presenting the highest diagnostic sensitivity is not the same as the test presenting the highest diagnostic specificity. In some instances, it may be useful to conduct two successive tests to confirm the results. By doing so, the first round of test could maximise the diagnostic sensitivity and the second round of test could maximise the diagnostic specificity.

Diagnostic sensitivity and diagnostic specificity are useful parameters to compare the test performance but they cannot be used to estimate the probability that a pest is present in a matrix. In order to do that, both parameters can be combined into one measure called the Likelihood ratios which is explained in section 5.7.

5.7 More criteria for measuring the performance of a test

5.7.1 Background

The result analyses of the VALITEST test performance study revealed that some organizers of test performance studies used additional criteria to evaluate the performance of a test. These are:

- Accuracy (already mentioned in EPPO standards PM 7/98 (2019) and PM 7/122 (2014)),
- Diagnostic odd ratio,
- False positive and false negative rates,
- Rates of true positive and true negative (already mentioned in EPPO standard PM 7/122 (2014)) and,
- Positive and negative likelihood ratios.

For further explanations, a tutorial video is available here: https://www.youtube.com/watch?v=otDdi5sY_uU

5.7.2 Proposed additional performance criteria for the analysis of validation data

In addition to accuracy and the rates of true positive and true negative that are already mentioned in EPPO standards PM 7/98 (2019) and/or PM 7/122 (2014), we propose to include in the relevant EPPO standards the following optional additional criteria to evaluate the performance of a test: diagnostic odd ratio, false positive and false negative rates, and positive and negative likelihood ratios. Diagnosticians may decide which performance criteria is more appropriate for their data analysis.

The formula of each parameter is given in Table 10 with a succinct explanation of their meaning and application. In addition, the calculation of their 95% confidence interval is also proposed to be determined for all these additional criteria as an optional tool as it may allow a better comparison of test performances. Further information on confidence intervals is provided in section 5.3.

As for the diagnostic sensitivity and diagnostic specificity, the calculation of these additional criteria has been determined using the true status of the reference material of the non-diluted positive and negative samples (so the dilution series was excluded; see section 4.3 for details).

5.7.3 Application on example dataset

Determination of the other performance criteria

As for the determination of the diagnostic sensitivity and diagnostic specificity the other performance criteria with their corresponding confidence intervals have been calculated according to the true status of the reference material used during the test performance study. For those performance estimations, we only use the data from the non-diluted samples (infested and not infested), so that the results are estimated on independent samples (the dilution series data were excluded because they are connected to each other as they derive from the same biological sample). Both the inconclusive and missing results were included for this analysis (see sections 4.3 and 5.1.3). The results of the different tests performance criteria listed in Table 10, are given in Table 11. Details on the calculation of confidence intervals are given in Annex 1.

Table 10. 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 01 February 2021). Acronyms used in the formulae are defined as per section 5.6 as follows: DSE: diagnostic sensitivity, DSP: diagnostic specificity, TN: true negative, FN: false negative, TP: true positive, FP: false positive.

Performance criteria	Formula	Note
Accuracy or Trueness	$\frac{TP + TN}{TP + TN + FP + FN}$	Accuracy 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 percentage. The higher the value, the better the performance. Warning: Accuracy estimation can yield misleading results if the data set 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 odd ratio (DOR)	$\frac{TP/FN}{FP/TN}$	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 Diagnostic odds ratio values have no limit but, if they are infinite, they could not be interpreted properly. The higher the value is, the better the performance.
False positive rate (FPR)	$\frac{FP}{FP + TN} = 1 - DSP$	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 is.
False negative rate (FNR)	$\frac{FN}{FN + TP} = 1 - DSE$	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 is.
Rate of true positive (RTP)	$\frac{TP}{TP + FP}$	RTP is the ratio of infected samples among the positive results, e.g. which 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, the better the performance.

		Note: The RTP formula is presented in EPPO standard PM 7/122 (2014)
Rate of true negative (RTN)	$\frac{TN}{TN+FN}$	RTN is the ratio of healthy samples among the negative results, e.g. which 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, the better the performance. Note: The RTN formula is presented in EPPO standard PM 7/122 (2014)
Positive likelihood ratio (LR+)	$\frac{DSE}{1-DSP}$	LR+ is how much more likely a sample is infected rather than healthy when the test result is positive. Likelihood ratios use the diagnostic sensitivity and diagnostic specificity of a test to determine whether a test result usefully changes the probability of infection when the result is positive. Note that the LR+ formula contains the false positive rate formula in its denominator (1-DSP). The lower the diagnostic specificity (DSP) is, the lower the LR+ is. The LR+ values range from 0 to infinite. The higher the value, the more likely the probability of a sample is to be infected with the pest when the test result is positive.
Negative likelihood ratio (LR-)	$\frac{DSP}{1-DSE}$	LR- is how much more likely a sample is healthy rather than infected when the test result is negative. Likelihood ratios use the diagnostic sensitivity and diagnostic specificity of a test to determine whether a test result usefully changes the probability of healthy status when the result is negative. Note that the LR+ formula contains the false negative rate formula in its denominator (1-DSE). The lower the diagnostic sensitivity (DSE) is, the lower the LR- is. The LR- values range from 0 to infinite. The higher the value, the more likely the probability of a sample is to be free from the pest when the result is negative.

Table 11. Performance characteristics of Tests A and B for a target pest with their lower and upper confidence levels (LCL and UCL, also represented in a graphical form in Figure 10). The values of accuracy, false positive and negative rates, positive and negative predictive values, range from 0 to 1 and can be expressed as percentage. The values of diagnostic odds ratio and likelihood ratios do not have any limit.

Performance criteria with their confidence levels	**Test A**	**Test B**
Accuracy	93%	74%
$\text{LCL}_{\text{Accuracy}}$	90%	70%
$\text{UCL}_{\text{Accuracy}}$	95%	79%
Diagnostic odds ratio (DOR)	716	Infinite
LCL_{DOR}	169	Not a number
UCL_{DOR}	3034	Infinite
False positive rate (FPR)	12%	45%
LCL_{FPR}	16%	51%
UCL_{FPR}	8%	38%
False negative rate (FNR)	1%	0%
LCL_{FNR}	4%	3%
UCL_{FNR}	0%	0%
Rate of true positive (RTP)	86%	63%
LCL_{RTP}	81%	57%
UCL_{RTP}	90%	68%
Rate of true negative (RTN)	99%	100%
LCL_{RTN}	97%	96%
UCL_{RTN}	100%	100%
Positive likelihood ratio (LR+)	8.4	2.2
$\text{LCL}_{\text{LR+}}$	6.1	1.9
$\text{UCL}_{\text{LR+}}$	11.9	2.6
Negative likelihood ratio (LR-)	84.8	Infinite
$\text{LCL}_{\text{LR-}}$	23.7	24.8
$\text{UCL}_{\text{LR-}}$	308.7	Infinite

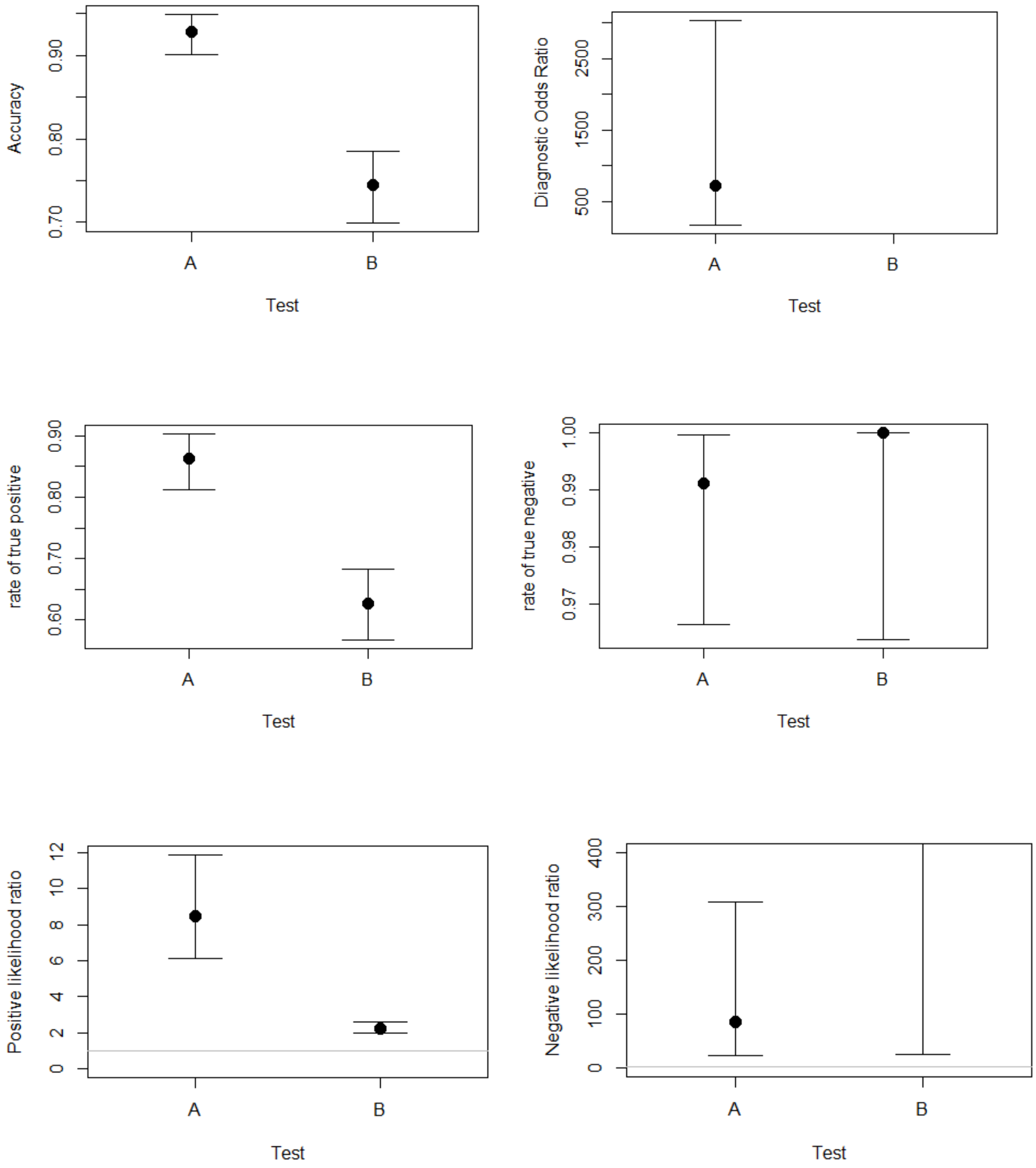

Figure 10. Graphical representation of the 95% confidence intervals for the following performance criteria by test for a target pest: accuracy, diagnostic odds ratio, false positive and negative rates, rates of true positive and true negative and positive and negative likelihood ratios.

Interpretation of accuracy of diagnostic tests

In our example, the accuracy of Test A is higher than for Test B and their confidence interval are not overlapping. This indicates that the results (positive and negative) obtained with Test A are more likely to be true than with Test B. As indicated above, the interpretation of accuracy estimation should be carried out with care when the panel is unbalanced (unequal proportion of infested and non-infested samples).

Interpretation of diagnostic odds ratios

In our example, the diagnostic odds ratio could not be interpreted properly as Test B presented infinite values. The diagnostic odds ratio can complement the performance characteristic of the accuracy. It is worth stating that poor results (for example low diagnostic specificity) obtained with an unbalanced sample panel (low proportion of non-infested samples) will have a greater impact on the results of the diagnostic odds ratio than on the accuracy.

Interpretation of false positive and negative rates

In our example, the false negative rates of Test A and Test B are very similar, 1% and 0% respectively (as these values depend on the diagnostic sensitivity results which were very similar with values of 99% and 100% for Test A and Test B, respectively).

The false positive rate of Test B is higher than with Test A with values of 45% and 12%, respectively with confidence intervals that are not overlapping. This indicates that more false positive results would be obtained with Test B than with Test A.

Interpretation of the rates of true positive and true negative

With regard to the results of the rates of true positive and true negative (and the positive and negative likelihood ratios), the results indicate that the samples tested negative are likely to be free of the pest. However, for samples that tested positives, the probability of the sample being truly infected is much higher with Test A than Test B as it is shown by the confidence intervals of the estimated values that do not overlap between the tests. This is because a higher number of false positive results has been obtained with Test B than with Test A.

Interpretation of the positive and negative likelihood ratios

The likelihood ratios are well suited to assist in the selection of tests when several tests are available for a pest and have been evaluated. Indeed, by plotting on a graph the negative likelihood ratio with the positive likelihood ratio of each test, it provides a visual representation on how the tests compare to each other in relation to the diagnostic sensitivity and diagnostic specificity as these parameters are used for the calculation of the likelihood ratios. However, values of 100% for the diagnostic sensitivity or the diagnostic specificity will act as an artefact in the estimation of likelihood ratios, as they will lead to infinite negative or positive likelihood ratios, masking the effect of the other parameter in the estimate. In this case, confidence intervals may give some additional information by giving upper and lower level of confidences for this ratio (see section 5.3), even sometimes in the case of an infinite estimation.

6 Implementation of the statistical approach

6.1 Inclusion into EPPO standards

The selected (statistical) methods for the assessment of the performance characteristics of diagnostic tests (i.e. analytical sensitivity, diagnostic sensitivity, diagnostic specificity, repeatability and reproducibility) discussed above with the addition of other performance criteria (i.e. false positive and false negative rates, rates of true positive and true negative and, positive and negative likelihood ratios) for measuring the effectiveness of a diagnostic test as optional approaches are proposed to the EPPO expert panel on Diagnostics and Quality assurance for possible inclusion in EPPO standard PM 7/98 (2019): *Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity*, to further improve the data analyses of performance criteria of plant health diagnostic tests and their interpretation. The proposed statistical approach would further support the interpretation of the validation data of a test (e.g. accordance and concordance) and the selection of tests (e.g. probability of detection). Furthermore, the data analyses would be consistent between laboratories with the possibility of comparison of data sets so that a laboratory may only need to verify some performance criteria of a test fully validated with these guidelines.

EPPO standard PM 7/122 (2014): *Guidelines for the organization of inter-laboratory comparisons by plant pest diagnostic laboratories* could also benefit from the addition of the relevant statistical methods for the data analyses of inter-laboratory studies and their interpretation. For example, the evaluation of the performance of a laboratory performance or of a test performance could be improved using the probability of detection model.

6.2 Data sharing

We believe that the Database on Diagnostic expertise – validation data section of the EPPO website is the best place for freely sharing the validation data of diagnostic tests which could be of interest to other laboratories in Europe, in particular for quarantine pests such as *Xylella fastidiosa*. By doing so, it would avoid laboratories having to repeat the complete validation of a test and thus may only require a verification study which would save time, cost and resources.

7 Conclusion and recommendations

This report provides detailed information on how the analyses of performance criteria for the validation of diagnostic tests (or test performance studies) is improved using statistical tools as demonstrated with an example dataset. Other performance criteria for measuring the effectiveness of diagnostic tests are also proposed, as optional, in the data analyses. We believe that the plant health diagnostic laboratories would benefit from such analyses. For example, for the selection of the most appropriate test for the detection of a pest.

It is also recommended that the proposed validation approach be presented to the EPPO diagnostic expert panel on Diagnostics and Quality assurance for their consideration as to whether they could be incorporated into EPPO standards PM 7/98 (2019) and PM 7/122 (2014).

REFERENCES

ANSES (2015) Guide de validation des méthodes d'analyses. ANSES, France, https://www.anses.fr/fr/system/files/ANSES_GuideValidation.pdf

AOAC Standards Development (2012) Guidelines for validation of microbiological methods for food and environmental surfaces. AOAC International., http://aoac.org/aoac_prod_imis/AOAC_Docs/StandardsDevelopment/AOAC_Validation_Guidelines_for_Food_Microbiology-Prepub_version.pdf.

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.

Erdoğan S., Gülhan O.T. (2016) Alternative confidence interval methods used in the diagnostic accuracy studies. *Computational and mathematical methods in medicine*, 7141050, doi: 10.1155/2016/7141050.

European Plant Protection Organization (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.

European Plant Protection Organization (2018). PM7⁄76 (5) - Use of EPPO diagnostic standards. Bulletin OEPP/EPPO Bulletin, 48(3): 373-377, https://onlinelibrary.wiley.com/doi/epdf/10.1111/epp.12506.

European Plant Protection Organization (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.

Fleiss J.L., Levin B., Paik M.C. (2003) Statistical methods for rates and proportions. Third Edition. John Wiley & Sons. New York.

Hess A.S., Shardell M., Johnson J.K., Thom K.A., Strassle P., Netzer G., Harris A.D. (2012) Methods and recommendations for evaluating and reporting a new diagnostic test. *European journal of clinical microbiology and infectious diseases*, 31(9): 2111-2116.

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.

ISO 17025 (2017) General requirements for the competence of testing and calibration laboratories. International organization for standardization, Geneva, Switzerland.

ISTA, Seed Health Committee (SHC) Tool box, Langton accordance and concordance tool https://www.seedtest.org/en/shc-tool-box-_content---1--1410--811.html, accessed February 2019.

Langton S.D., 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 J.A. (1989) Generalized Linear Models, 2nd revised edition, London: Chapman & Hall. 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 M.H. (2008) 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.

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. (2009) 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.

Simel D.L., Samsa G.P., Matchar D.B. (1991) Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology*, 44(8): 763-770.

Uhlig, S., Gowik P. (2018) Efficient estimation of the limit of detection and the relative limit of detection along with their reproducibility in the validation of qualitative microbiological methods by means of generalized linear mixed models. *Journal of consumer protection food safety*, 13: 79-87.

Wehling P., LaBudde R.A., Brunelle S.L., Nelson M.T. (2011) Probability of detection (POD) as a statistical model for the validation of qualitative methods. *Journal of AOAC International*, 94 (1): 335-347.

ANNEX 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

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

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

Diagnostic odds ratio (DOR) 95% confidence interval

Confidence intervals for diagnostic odds ratio can be calculated using Simple OR CI (Fleiss *et al*., 2003).

The formula is

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

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

Confidence intervals for generalized linear model

Confidence intervals for the limit of detection 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.

The (log)dilution factor associated to 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.

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

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.

ANNEX 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

Accordance is the percentage chance of finding the same result from two replicates of the same sample analyzed in the same laboratory (Langton *et al*., 2002), under repeatability conditions. 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.

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

Concordance is calculated 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 laboratories information, then subtracting the number of matching and total pairs within each laboratory (which are linked to 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

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

If accordance is higher than concordance, it indicates that two identical samples are more likely to give the same result if they are analyzed by the same laboratory than if they are analyzed 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

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:

$$\frac{acc\ (1-conc)}{conc\ (1-acc)}$$ 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 variability between laboratories exists.

However, when there are many accordance values of 1 (meaning that the tests are highly stables 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 by a 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 significantly greater than one indicates a significant variability of the results between laboratories.

ANNEX 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 generalised model (also called logistic regression) such as R software.

For each test, data of the diluted samples were used to adjust binomial generalized linear models (bGLM) with logit link between the dilution (expressed by the base 10 negative exponent of the corresponding dilution) and the detection status. The number of dilution level being very limited, the adjustment of bGLM is not always possible as this method require at least 5 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 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.

ANNEX 4 – Diagnostic sensitivity and the diagnostic specificity – additional information

The calculation of the diagnostic sensitivity and diagnostic specificity can be done in an xls spreadsheet type of software.

Determination of the diagnostic sensitivity and the diagnostic specificity for the validation of a new test *by comparison with a validated test*

The formula for the determination of the diagnostic sensitivity and the diagnostic specificity for the validation of a new test (A) by comparison with a validated test (B) is the one provided in EPPO standard PM 7/98 (Appendix 6, 2019), which is:

Diagnostic sensitivity: PA / (PA + ND)

Diagnostic specificity: NA / (NA + PD)

PA and NA: positive and negative agreement, i.e. same result (positive or negative) was obtained for the same sample with the alternative and the reference tests

PD and ND: positive and negative deviation, i.e. different result (positive and negative) was obtained for the same sample with the alternative and the reference tests

The table below adapted from EPPO standard PM 7/98 (2019), illustrates how PA, PD, ND and NA are determined.

Number of samples		Validated test B	
		Positive	Negative
Test A	Positive	**PA**	PD
	Negative	ND	**NA**