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° 4.2

Title: Report on economic impact of priority tests

Validation of diagnostic tests to support plant health

Due date:	Month 40
Actual submission date	30-10-2021 (Month 42)
Start date of the project	01-05-2018
Deliverable lead contractor (organization name)	FERA Science Ltd.
Participants (Partners short names)	FERA, WBF, BIOREBA, EPPO, NIB, ULG, CD, WR, IPADLAB, SEDIAG, GIORIN
Author(s) in alphabetical order	Agstner, B., Macarthur, R., Van den Berg, F.
Contact for queries	Barbara.Agstner@fera.co.uk (economist) Femke.Vandenberg@fera.co.uk (mathematical modeller)
Level of dissemination	Public
Type of deliverable	Report

Abstract

This deliverable describes the development and current form of a mathematical framework, created to support, inter alia, resource allocation for and design of sampling and test programmes in different plant health contexts. The framework is exemplified by using it to compare the optimal sampling programme based on initial estimates of test performance characteristics (diagnostic sensitivity and specificity) to the optimal sampling programme based on the verified test performance characteristics, thereby revealing the 'cost' of not knowing the true performance of the diagnostic tests. The deliverable concludes with suggestions for further framework developments to allow its integration into current work processes of plant health laboratories and decision makers.

Partners involved: FERA, WBF, BIOREBA, EPPO, NIB, ULG, CD, WR, IPADLAB, SEDIAG, GIORIN

Contents

Abbreviations

DSE: Diagnostic sensitivity
DSP: Diagnostic specificity
PCR: Polymerase Chain Reaction
TPS: Test Performance Study

Key definitions in alphabetical order

Please note that this deliverable primarily draws on EPPO definitions, as these are widely recognised in the plant health area. These might differ from definitions used in other subject areas.

Decision rule	Decision on how to combine the results of individual tests to arrive at a diagnosis. E.g. pest presence is established if *all* tests are positive, or *two* out of three, or *any*.
Parallel testing	A set of tests is performed. A decision rule is then applied to *all* of the results to decide presence or absence of the pest.
Proficiency testing	**Evaluation of the competence of the laboratory:** Establishing the competence of a laboratory in analysing defined samples using their established test.
Sampling programme	Describes the number of samples, from where they are taken, what population they represent, what test programme will be used, as well as how diagnostic results will be interpreted (e.g. if all samples provide a negative result the population will be treated as pest free; if three samples give a positive result priority will be given to this population).
Sequential testing (focus of this study)	One test is performed. If applying the decision rule gives "detected" or "not detected" then this is reported and no more tests are done. Otherwise the subsequent tests are performed until a decision is reached.
Test Performance Study (TPS)	**Evaluation of a test:** Evaluation of the performance of one or more tests by two or more laboratories using defined samples. (Sometimes referred to as a ring-test or collaborative trial.)
Test programme	A set of individual physical tests applied in a specific order and using a specific decision rule to establish pest presence.
Validation	Often validation is defined as the exploratory process for **establishing the operational limits and performance characteristics** of a new, modified or otherwise inadequately characterized test (e.g. EPA, 2015). It should, however, be noted that in ISO 17025:2017 and EPPO PM 7/98 (4) validation is defined as providing objective evidence that the test is **suitable for the circumstances of use** (EPPO). This is an important distinction from merely determining performance characteristics. Once a laboratory has performed a validation, it can (optional) produce a validation summary for this particular test and upload it to the EPPO database.
Verification	Verification provides objective evidence that the laboratory is competent to perform a validated test according to the relevant performance characteristics. Verification can also be done by participating in a proficiency test or test performance study, provided that these allow the requirements to be fulfilled. (PM 7/98). Verification can be used as validation for modifications (ISO 17025:2017 3.9).

1. Introduction

It is widely recognised that increased global trade in life plants and plant materials has resulted in heightened pest[1] pressures. As a consequence, testing for plant pests has become a key part of public and private efforts to safeguard agriculture and the wider environment. However, due to the wide range of traded plants species and associated harmful organisms, as well as resource and technology constraints, decisions about what to test for, how to test and how to respond to results, are subject to a number of trade-offs. These complexities are further exacerbated by responsibilities, as well as costs and benefits from testing, being split between different stakeholder groups, including public and private end-users, inspection agencies and laboratories.

The main aim of task 4.2 was to evaluate case studies, selected from tests identified in task 4.1, using a cost-benefit approach, which would include an assessment of different test contexts and costs, market and non-market impacts of pests, as well as trade-offs and responsibilities. Ultimately, the outcomes of this task would therefore provide insights into end-user decision making and market demand. However, as the project evolved a number of additional questions came up, including "how do stakeholder groups communicate" and "what is the value of validation?". In addition, it became clear that a lot of the parameters required to assess case studies as originally defined were either unknown or subject to large uncertainties.

In order to address both the original task and any additional questions, and to provide flexibility to update parameters at a later date, the decision was taken to broaden the case study approach and develop a framework for assessing costs and benefits of testing for plant pests. This deliverable describes how the framework was developed, its current form, as well as initial results. It should be highlighted that while the framework has been applied to specific questions, there is scope to further develop it into a user-friendly tool that can be integrated into ongoing decision processes and workflows of interested stakeholders.

2. Methodology

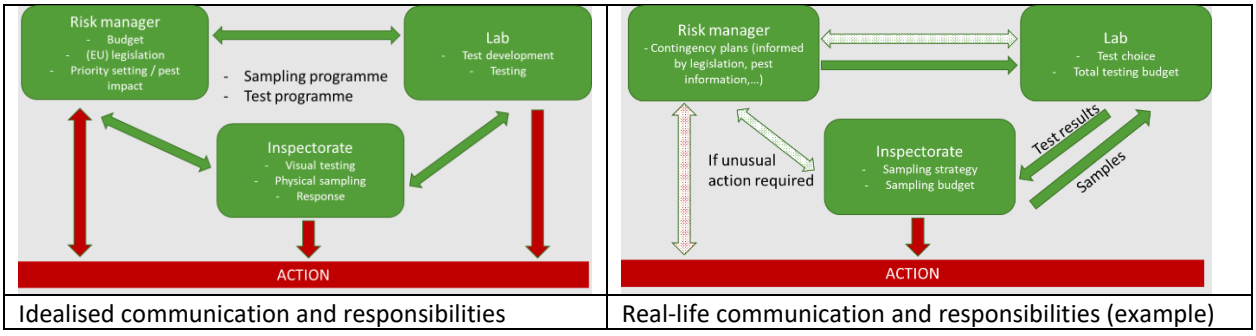
Leading up to the framework development, a number of semi-structured interviews were held with VALITEST partners, laboratory teams and policy staff, to develop an understanding of different testing contexts and stakeholder responsibilities. Questions addressed current testing regimes, laboratory procedures and decision-making processes. Interview outputs were then used to design an online workshop and follow-up exercise for decision makers, to get a wider European view. In parallel, a desk-based exercise was conducted to provide an initial outline of different scenario contexts. The framework itself was constructed by modelling experts, with feedback from laboratory staff and plant health experts, and is coded in R. R has been chosen because it's a key modelling language used by experts in the field and because it allows the future addition of a user-friendly Shiny web-interface if there is interest.

In the following, the framework is described in more detail, including a definition of key terms. The latter is especially important as interviews showed slightly different uses of the same terms across stakeholder groups.

[1] In line with international terminology the term pest will be used to refer to both, pests and pathogens. Consequently, the term infest will be used to refer to both, infestations (generally associated with pests) and infections (generally associated with pathogens).

3. Testing for Plant Pests – Stakeholder Map

On a national level, responsibilities for planning, conducting and responding to plant health tests are usually split between risk managers, inspectorates and laboratories. Risk managers' responsibilities typically include priority setting (e.g. horizon scanning for high risk pests), developing response protocols and contingency plans, feeding into future legislation while ensuring current legislation (national and international) is implemented, as well as high-level budget allocations. Inspectorates are usually responsible for the practical implementation of sampling, which often includes decisions on budget allocations and risk assessments of pathways and plant traders, alongside physical sampling and response actions. Laboratories carry out the actual tests and often are involved in method selection, including budgeting decisions. While in an ideal situation, decisions would be based on close collaboration between all three stakeholder groups (see Figure 1 left image), with risk managers having an oversight of overall resource allocations and how they map to national targets, reality often differs due to a variety of constraints (see Figure 1 right image).



Idealised communication and responsibilities	Real-life communication and responsibilities (example)

Figure 1. Ideal vs real-life communication channels between stakeholders involved in plant health testing

Interviews and the workshop showed that communication between stakeholder groups is often impeded by factors like the complexity of the subject matter, which can be aggravated by time constraints and differences in individual's backgrounds, as well as language barriers, including different associations with terms like sensitivity (e.g. laboratory staff are more likely to think of analytical sensitivity, as it is closely related to their day-to-day work, whereas decision makers are more likely to think of diagnostic sensitivity) and ambiguities over definitions like "validated for a specific purpose". This leads to situations where risk managers might, for example, shy away from being involved in decisions about test methods ("the lab knows best") and resources like validation packs not being shared beyond laboratories.

While all stakeholders are experts in their area, communication is nevertheless crucial, as everyone sees the world from their individual point of view. For example, when asked about the importance of 'sensitivity', risk managers tend to focus on scenarios like 'early detection' and often implicitly include details of the sampling strategy and consequences of false negatives / false positives in their definition; laboratory staff on the other hand tend to focus on biological factors, wherefore sensitivity is considered of particular importance when there is not much DNA in the sample (thus implicitly taking a first visual inspection into account). In other words, if the type of sensitivity is not carefully specified, each group of experts assumes the term relates to the type primarily used in their profession. Similar differences can be seen regarding test speed: while risk managers likely focus on scenarios (e.g. speed is important for perishable goods), laboratory staff are more likely to consider issues around pest biology, like its growth speed; although these views can be complementary, e.g. culturing a pest is

impractical when a fast answer is needed, this might not always be the case. While these differences in viewpoints are natural and important, they need to be recognised in order to make sound decisions in line with national objectives. Supporting the improvement of communication between stakeholder groups by helping to make some of the aforementioned implicit assumptions more visible is thus an important objective of this framework.

4. Mathematical Framework to Assess Test Programmes

4.1. Background and Definitions

Pest presence in plants can be established using tests, where a test describes the combination of pest, matrix (e.g. leaf material) and method (e.g. Polymerase Chain Reaction (PCR)). Although tests are often thought of as individual, self-contained units, the usual *minimum* number of tests applied in a decision process is *two:*

1) a visual assessment for symptoms,
2) a follow-on test which may be applied to symptomatic plants.

Recently, it has become more common practice to use three tests:

1) a visual assessment for symptoms,
2) a rapid low-cost test,
3) a less speedy, more expensive laboratory-based test.

Each of these individual tests will have a different (1) diagnostic sensitivity (DSE), i.e. probability of a positive result if a plant is infested, (2) diagnostic specificity (DSP), i.e. probability of a negative result if a plant is uninfested, and (3) cost of use. It should be noted here that each test will also have a different (4) analytical sensitivity, sometimes referred to as the limit of detection (LoD) for the physical/chemical marker, and (5) analytical specificity, which refers to the ability of e.g. an assay to detect the intended target and its variations (it includes cross-reactivity and interference). As mentioned in the previous section, insufficient specification of terms can lead to misunderstandings; it is therefore important to highlight that this framework uses *diagnostic sensitivity and specificity* to describe test performance (see section 4.2 for mathematical details). Using these characteristics also allows the calculation of the *overall DSE and DSP of multiple tests*, following a widely applied approach used, for example, in radiology, where two tests with different sensitivities and specificities are frequently combined to diagnose a particular disease or condition (Weinstein, Obuchowski et al. 2005); this is explained in more detail below.

Where more than one test is used, they can be performed (1) in parallel, i.e. all tests are performed irrespective of other test results, or (2) sequentially, i.e. the application of subsequent tests is dependent on the result of previous tests. In this study, only sequential testing is considered, as it is a common approach in EU countries and has more scope for realising process efficiencies.

In sequential testing, tests may be undertaken in different *orders* and results of individual tests may be *combined* in different ways (e.g. pest presence is established if all tests are positive, or two out of three, or any = ***decision rule***). Each of these different ways of applying tests, namely the choice of which tests to include × the order of tests × the combination of results, will have its own overall diagnostic sensitivity, diagnostic specificity and cost.

In this deliverable, a set of individual physical tests applied in a specific order and using a specific decision rule to establish pest presence is referred to as a ***test programme***. Test programmes examined in this study allow up to three tests to be performed *on each sample* (an example of a three-test programme may be: visual inspection, on-site testing with a lateral flow device, PCR-based detection in a laboratory, with presence on infestation detected if all three tests are positive). There is scope to extend the number of tests in follow-up work. It is important to distinguish the test programme from the ***sampling programme***, which in this deliverable describes the number of samples, from where they are taken, what population they represent, what test programme will be used, as well as, how diagnostic results are interpreted and acted upon (e.g. if all samples provide a negative result the population will be treated as pest free).

As highlighted, inter alia, by several EPPO guidance documents (e.g. PM 7/76 (5)), it is furthermore important to note that testing for pests can be undertaken for a number of different purposes, including:

- early detection using prospective surveillance,
- defining pest-free areas,
- measuring the prevalence of a pest or pathogen (or possibly some other feature of plants such as species), or
- in support of a programme of eradication.

For each of these purposes the benefits associated with a correct test result are different. Hence, the optimum features of a diagnostic programme (cost, diagnostic sensitivity, diagnostic specificity) may be different across purposes. For the framework development, 'early detection' has been picked as the main scenario to be investigated. In future work, the framework can be adapted / extended to include alternative scenarios; it is suggested that these scenarios are based on stakeholder demand.

Ultimately, the framework has been designed to answer a range of questions, including the following:

- What budget is required to meet a particular goal (e.g. a certain prevalence at first detection) for a test programme conducted for a specific purpose? (= **budget planning support**)
- Which combination of tests (methods, order, …) minimises budget requirements? (**= design support for efficient test programmes**)
- What is the financial cost of inaccurate or imprecise estimates of diagnostic sensitivity and specificity? (**= determine the value of validation**)
- What is the appropriate size of a validation study? (= **resource allocation support for the validation of tests**)

4.2. Test Programme Performance and Test Programme Cost

In the following, methods are summarised and exemplified by a single three-test programme. Appendix provides a detailed description of the derivation of all equations for all test programmes using up to three tests.

The DSE ***of a test*** is defined as the proportion of infested samples that give a positive test result. Thus, the DSE of a test is:

$$DSE = \frac{TP}{TP+FN} \qquad [1]$$

where *TP* and *FN* are the number of true positive and false negative samples, respectively.

The DSP ***of a test*** is defined as the proportion of uninfested samples that give a negative test result. Thus, the DSP of a test is:

$$DSP = \frac{TN}{TN+FP} \qquad [2]$$

where *TN* and *FP* are the number of true negative and false positive samples, respectively.

The DSE ***of a test programme*** is thus defined as the proportion of infested samples which, based on the results of the full test programme (which may comprise multiple tests), is classed as positive. Hence, the test programme DSE and DSP can be calculated according to the total number of TPs, FPs, TNs and FNs found when all samples have been subjected to the full test programme. Figure 2 exemplifies the case where *N* samples are subjected to an A * (B+C) test programme, with test order A then B then C, whereby samples which give a positive result with Test A and also give a positive result with either Test B or C lead to the conclusion that the infestation has been detected.

In summary the test programme A * (B + C) proceeds as follows:

1) **Test A** is applied to all samples; the pest or pathogen is declared absent from samples that give a negative test result.
2) **Test B** is applied to the remaining samples (i.e. those that gave a positive result with Test A); the pest or pathogen is declared present in samples that give a positive result.
3) **Test C** is applied to the remaining samples (i.e. those that gave a negative result with test B); the pest or pathogen is declared present in the samples that give a positive result and absent from those that give a negative result.

This means that in this test programme all samples are tested with Test A; only samples which give a positive result with Test A are tested with test B and only those which gave a positive result with Test A and a negative result with Test B are tested with Test C.

All test results across all potential pathways can then be classed as a TP, FP, TN or FN and the test programme DSE and DSP calculated based on these numbers as per equations [1] and [2].

The average sample testing cost for a programme can be calculated by adding up the probabilities of going along each of the testing pathways. In the example given in Figure 2 all samples are tested with Test A, whereby Test A incurs a cost of c_A; a fraction p_A of these tests will lead to a positive test result and these samples will be tested with Test B, which incurs a cost of c_B per test and finally a fraction $(1 - p_B)$ of these samples will lead to a negative test result and also be tested with Test C, which incurs a cost of c_C per test. Hence, the average sample testing cost for this test programme is:

$$Average\ sample\ testing\ cost = c_A + p_A c_B + p_A(1 - p_B)c_C \qquad [3]$$

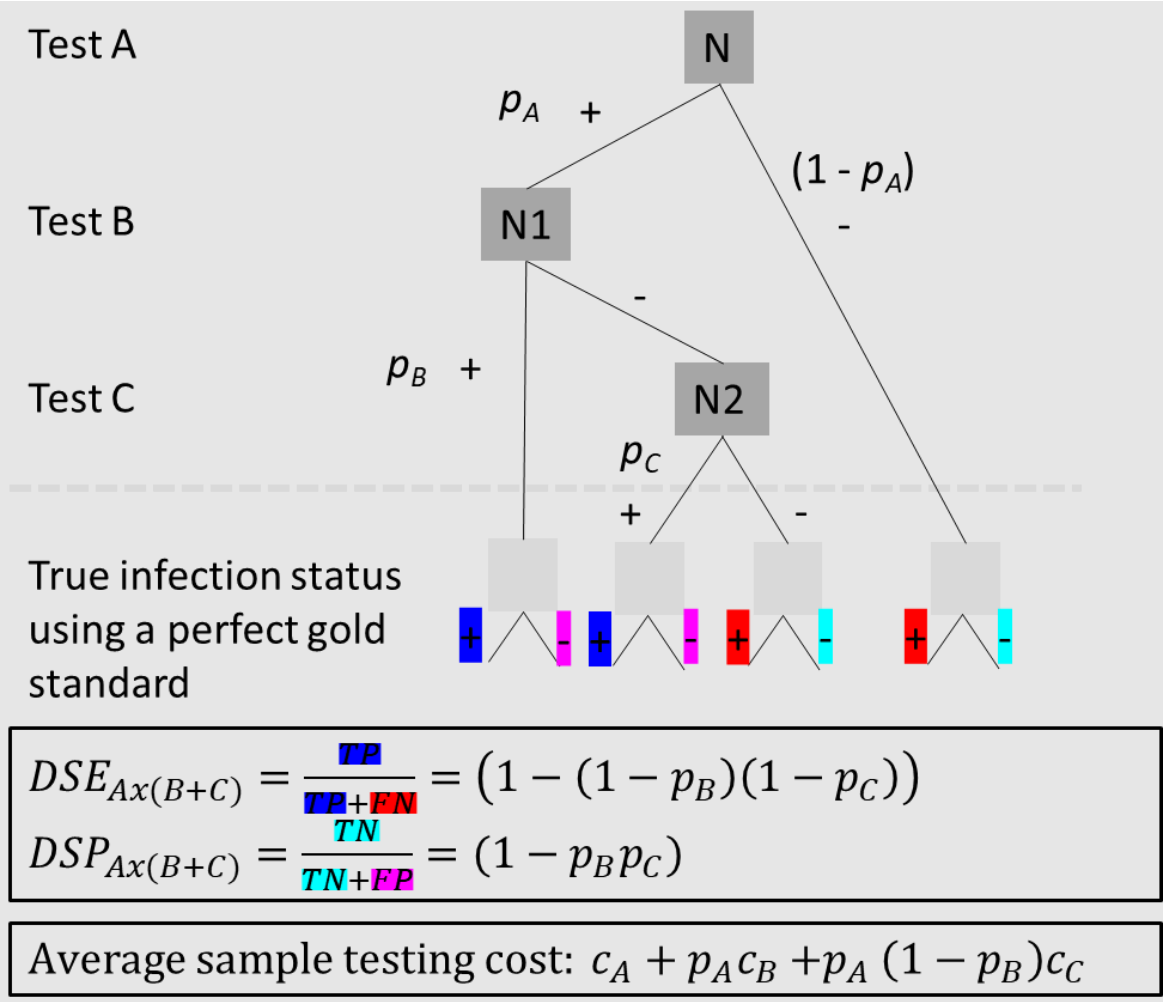


Figure 2. Illustration of the A * (B+C) programme DSE and DSP and average sample testing cost calculations. In the A * (B+C) programme all samples which return a positive result with Test A (N * p_A) will be tested with Test B; after which all samples with a negative test result for Test B (N * $p_A(1 - p_B)$) will be tested with Test C. This means that all samples which result in a positive test result for Test A and a positive test result for either Test B or C are classed as positive samples. Note that the probability of a positive test result is affected by the frequencies of a test returning either a true positive (TP) or a false positive (FP) result, whereas the probability of a negative test result, is affected by the frequencies of a test returning either a true negative (TN) or a false negative (FN) result. The cost of an individual test is represented by c.

4.3. The Cost of Getting it Wrong

The framework considers two types of costs for getting it wrong: (1) costs which are caused by false positive test results and (2) costs which are caused by false negative test results. Please note, that while the costs described below are expected to be generalisable to most scenarios, they are currently based on the 'early detection' scenario. Further feedback is needed to ensure appropriateness for additional scenarios.

False positive test results lead to *excessive management action*. In the case of sampling for early detection, each sample being classed as positive by the test programme would instigate an action, such as confirmatory testing through culturing or further in-situ testing in the area in which the sample classed as positive was found. Hence, false positives result in an unnecessary further spend on confirmatory testing, which should *eventually* reveal that the sample was in fact not infested. In some instances, false positives might also lead to precautionary action, such as the destruction of host

material: e.g. trees in a delineated zone or consignments at the border. Usually the type of response will be outlined in policy documents and contingency plans and is dependent on the risk-rating of the pest in question. Precautionary action is likely more costly than confirmatory testing, though costs might be borne by different stakeholder groups. For example, in case of a positive result some countries give importers the choice between additional, charged-for testing and the destruction of the consignment. Hence, false positive costs are highly context dependent.

In this framework the average cost of a false positive, c_{FP}, is presumed to be constant, but this assumption can easily be adjusted. The expected total cost of false positive test results, $c_{T;FP}$, is composed of the expected number of false positive test results given by the test programme, $E_{FP;k;o}$, and the average cost of dealing with a false positive test result, c_{FP}. This leads to:

$$c_{T;FP} = E_{FP;k} \cdot c_{FP}, \qquad [4]$$

and

$$E_{FP;k} = (1 - prev)(1 - DSP_k)N_k \qquad [5]$$

where *prev* is the pest prevalence, DSP_k the diagnostic specificity of test programme *k* and N_k the number of samples tested with test programme *k*.

False negatives on the other hand lead to *insufficient management action*. For example, not detecting an infestation means the pest or pathogen can spread further, leading - due to the larger initially affected area - to (1) higher management costs upon eventual detection, as well as (2) greater overall pest impacts (e.g. destroyed crops). Depending on the type of pest, impacts can affect market commodities, such as wheat, or non-market commodities, such as amenity trees. While the former primarily impacts industry stakeholders via crop losses or reduced crop quality, the latter impacts the wider public via the loss of ecosystem services. While both types of impacts are quantifiable in principle, ecosystem service values are highly location dependent and currently still difficult to estimate. In this deliverable, false negative impacts are hence expressed in terms of increased prevalence at first detection rather than in monetary terms. Again, this is an element that could be expanded on at a later stage.

5. The Value of Validation

5.1. Introduction

As stated in EPPO (PM 7/76 (5)) “validation data is not available for all tests that are currently widely used in plant pest diagnostic laboratories. Lack of validation data is, in particular, often the case for high-throughput tests such as ELISA.” While there is often a long period of experience of use of such tests, which makes it possible for the laboratory to qualify the reliability of such tests, there is at the moment no way to assess whether this approach leads to hidden costs, and if so of what magnitude.

Even where validation data is available, currently validation studies are only required to include analytical sensitivity and specificity, but not (although they can) diagnostic sensitivity and specificity. In addition, to determining the value of validation for selected case studies, the framework can therefore be used to:

- Inform future discussions on the design and reporting of validation studies, including characteristics like diagnostic sensitivity and specificity;

- Aid the understanding of the difference between analytical and diagnostic sensitivity and specificity and therefore support efficient communication across stakeholder groups.

As mentioned above, the results of a validation study (for definitions see page 1) can, inter alia, provide a prediction of what the diagnostic sensitivity and diagnostic specificity of the method will be in practice. Based on the usual statistical model for variation in analytical performance (e.g. ISO 5725, IUPAC Guide to Single-Laboratory validation), the observed diagnostic sensitivity and diagnostic specificity are expected to vary at random (1) over time when a method is applied in a laboratory (or by a user for on-site tests), and (2) between laboratories or users. Moreover, the long run-mean diagnostic sensitivity and diagnostic specificity observed in a single laboratory is expected to equal the mean across the population of laboratories. Lastly the long run variation within a laboratory is expected to be the same size as the variation between laboratories. Hence, there are two options for estimating the value of diagnostic sensitivity and diagnostic specificity. The first option is in a single laboratory study during which long-run random variation can be expressed in results. The second option is in a multi-laboratories performance study (or Test Performance Study (TPS)). Whether estimates of diagnostic sensitivity and diagnostic specificity are gained from a single-laboratory or multi-laboratories study they are statistical *estimates* of the true values with an associated uncertainty. With some further work, uncertainty in diagnostic sensitivity and diagnostic specificity can be integrated into this framework to elucidate its effect on uncertainty about the cost of test programmes and the expected upper limit for prevalence at first detection provided by sampling programmes.

5.2. Applying the Framework

As a reminder, the scenario examined in this deliverable is 'early detection', where regular sampling and testing of plants is undertaken for pests that have not yet been detected in the country/region in question. The framework will be applied to a case study exemplifying the 'early detection' scenario; however, while parameters (including test characteristics) are based on a real-world example pest and test, they have been abstracted in this deliverable to allow the focus to remain on the mechanics of the framework.

Given that the effectiveness of sampling and testing is estimated by the *prevalence at first detection* (which is a function of the number of samples tested and the sensitivity of the sampling programme that is used to test them) the optimum programme is the one that either:

- provides the lowest estimate for prevalence at first detection for a fixed cost, or
- the lowest cost to achieve a set target for prevalence at first detection.

Choosing an optimal programme for 'early detection':
Following Mastin, van den Bosch et al. (2019), if there is an ongoing sampling programme, a sampling plan in which the product of the number of samples tested and programme diagnostic sensitivity is maximised, provides the lowest estimated prevalence at which a pest is first detected. The goal is to maximise this product which is achieved when the test diagnostic sensitivity divided by the cost per test is maximised.
The framework can then be used to assess the effect of the number of samples tested and the test programme used on the prevalence of first detection, where a test programme with a higher false negative rate will clearly result in a higher prevalence at first detections.
See Appendix for further details and equations.

Irrespective of which approach to determining the optimum programme is taken (cost constraint or 'prevalence at first detection' constraint), the outcome relies on the estimates of test performance characteristics used. However, if the true performance of tests deviates from these estimates, several consequences are possible:

(1) tests are used in-efficiently (sub-optimal programme is chosen) so that, for example,
    i. the cost of false positive results is higher than expected,
    ii. not enough resources are assigned to the sampling and testing to achieve the set target for detecting presence, resulting in a delayed detection of an outbreak
(2) stakeholders falsely believe, for example, that the absence of infestation above a target level has been demonstrated.

All of these consequences are 'costs' of not knowing the true performance of the diagnostic tests and basing decisions on wrong assumptions. Hence, estimates of the relation between the size of these 'costs' and the accuracy and precision of test performance characteristics gained during validation can inform the decision maker about the value obtained by validating performance characteristics.

In the following, the process of estimating the costs of not knowing, and thus the value of validation, is described step-by-step using a case study example. Example parameters are described in the tables below: Table 1 lists illustrative parameters for a potential pest of plant health significance chosen to exemplify the 'early detection' scenario and Table 2 lists the DSE and DSP of the tests used. Please note that this case study has been selected as the original test performance estimates are different to the true performance characteristics identified through a TPS. Reference-names for step-scenario combinations are highlighted in bold and underlined.

Steps to estimate the value of validation:

Step 1 - Optimise sampling programme based on unvalidated test performance characteristics (= **"apparent/unverified"**)

Step 2 - Estimate true cost and maximum prevalence at first detection for three possible scenarios:

a. Scenario 1 (= "t**rue/unverified; fixed budget"**): A scenario where the risk manager follows the apparent/unverified test programme until either the total sample & testing budget has been spent or the planned number of samples has been reached; whichever comes earlier.
b. Scenario 2 (= "**true/unverified; fixed number tests"**): A scenario where the risk manager strictly follows the apparent/unverified test and sampling programme irrespective of the budget.
c. Scenario 3 (= "**true/unverified; fixed prevalence target"):** An illustrative but unrealistic scenario in which a risk manager uses true/validated test performance characteristics to meet a target for prevalence at first detection by increasing the number of samples tested which re-optimises the test programme

Step 3 – Optimise sampling programme based on validated test performance characteristics (= **"true/verified")** – this can also be done based on the three scenarios described above.

Step 4 – Comparison

Summary:

There are three scenarios: **scenario 1)** sampling programme with a fixed budget; **scenario 2)** sampling programme with a fixed number of samples; **scenario 3)** sampling programme with a fixed true prevalence at first detection. There are estimates of performance at three steps: **Apparent/unverified:** apparent performance of a programme optimised using unverified estimates of test performance. **True/unverified:** true performance of a programme optimised using unverified estimates of test performance. **True/verified:** true performance of a programme optimised using verified estimates of test performance.

Table 1: Parameter values describing 'early detection' *scenario*

Parameter	Description	Value
c_C	Average sample collection cost	10
c_{FP}	Average cost immediate action taken following a positive result	500
Δ	Interval over which samples are collected	365 days
x	Confidence level for 'early detection'	95%
r	Epidemic growth rate of target pest	0.0033 day^{-1}

Table 2: Parameter values describing unverified and verified test performance estimates

Test	Parameter	Unverified estimates of performance	Verified estimates of performance
1 Symptoms	DSE	20%	20%
	DSP	95%	95%
	Cost per test	€2	€2
2 ELISA	DSE	95%	95.8%
	DSP	100%	80.0%
	Number of tests for DSE verification	NA	24
	Number of tests for DSP verification	NA	15
	Cost per test	€12	€12
3 PCR	DSE	100	98.1
	DSP	100	95.6
	Number of tests for DSE verification	NA	54
	Number of tests for DSP verification	NA	45
	Cost per test	€14	€14

Ad step 1.) In a first step, an optimal sampling programme (including the test programme) is determined based on apparent/unverified test performance characteristics. For this case study this is the sampling programme **which provided a prevalence at first detection of no more than 1% (at 95% confidence) based on the apparent/unverified diagnostic test performance characteristics.**

Programme details are summarised in Table 3 and total expected costs calculated for this step are depicted by bar 1 in Figure 3.

Ad step 2.) In the first scenario (**true/unverified; fixed budget**) the risk manager follows the **apparent/unverified testing and sampling programme** until the total sample and test program budget has been spent. In the second scenario **(true/unverified; fixed number tests**) the risk manager strictly follows the **apparent/unverified testing and sampling programme** even though the amount spent is larger than expected. The second scenario is representative of a situation where different stakeholders control different elements of development and implementation of the sampling and test programmes, with no immediate feedback loops being in place. For example, risk managers develop the sampling plan, which is passed onto the inspectorate for sample collection after which the samples are passed onto the laboratory staff for testing with the test programme deemed most optimal based on extensive experience (personal communication with risk managers and laboratory staff). In this case the inspectorate will collect a fixed number of samples with the laboratory testing samples according to a predefined test programme, without considering the potentially associated impact costs which may lead to exceedance of the total available budget. Results for the fixed budget and fixed number of tests true/unverified scenarios are shown in Table 3.

Bar 2 of Figure 3 shows the total cost (including the proportion incurred due to false positives) of the third true/unverified scenario (**true/unverified; fixed prevalence target**) in which the target 1% prevalence at first detection is achieved by increasing the number of samples tested but without changing the test programme. This is a sub-optimal use of the information provided by true estimates of test performance, and possibly unrealistic for this reason, but it is included to provide a direct like-for-like cost comparison for sampling programmes that provide the same estimated upper limit for prevalence at first detection.

Ad step 3.) Whether there is a benefit in re-optimising the sampling programme according to the validated performance characteristics is then assessed by comparing the performance of the **true/unverified;** programme with that of the sampling and test programme which has been re-optimised using the true (validated) test performance characteristics. This strategy is referred to as **true/verified**. This illustrates how the optimal test program, the number of samples that ought to be tested and the minimum prevalence at first detection are affected by differences in diagnostic test performance and will provide an evaluation of the cost savings that can be made when the test and sampling programme are re-optimised once validation data is available.

Two true/verified programmes are given. The first is **true/verified (scenario 1)** (Table 3) in which the test programme has been reoptimized and applied in a sampling programme with the original budget used in the apparent/unverified programme. The second is **true/verified (scenario 3)** (Table 3 and bar 3 of Figure 3) in which the test programme has been re-optimised and the sample budget increased to provide an estimated upper limit for prevalence at first detection of 1%.

Ad step 4.) Comparison and summary: Figure 3 shows that for the apparent/unverified test programme the costs are solely associated with testing and that there are no false positive costs given that the DSP is presumed to be 1. Verification of the test performance characteristics has however shown that the DSP is smaller than 1 which means that there is a large cost associated with not optimising the test programme to account for this (red section of bar 2). Firstly, there is a large cost associated with having to go back to the sites for which positive test results were found to perform confirmatory testing to find you were in fact dealing with a false positive. Secondly, the fact that the DSE is slightly lower than expected also means that there is a slightly higher testing cost (blue section

of bar 2) to be able to achieve the same prevalence at first detection threshold. The third bar shows that considerable cost savings can be made while still achieving the target prevalence at first detection by re-optimising the sampling programme according to the validated performance characteristics. The re-optimisation suggests that a small increase testing budget associated with a change in the test programme can lead to a larger reduction in the costs associated with false positives.

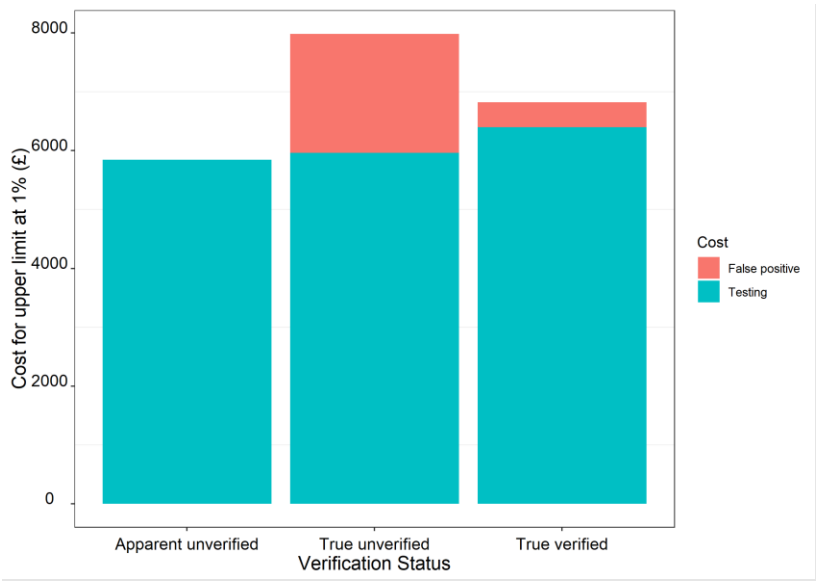


Figure 3. Apparent and true costs of achieving detection at 1% prevalence for sampling programmes based on tests programmes optimised using verified and unverified estimates of test performance

Table 3 gives some more details about the scenarios. Based on the unverified estimates of test performances (see Table 2) it is expected that testing 1807 samples a year using the "A×B; 1,3" test programme (where 1 is symptoms; and 3 is PCR), would provide an upper limit for prevalence at first detection of 1% for an annual budget of €5.847,00. However, the verified, true, performance of the tests is less favourable than the initial estimates used to produce the sampling plan. If these true performance characteristics are used to assess how the "A×B; 1,3" plan will perform then a range of outcomes may arise including primarily a higher-than-expected cost if the planned number of tests is kept fixed, or a higher-than-expected prevalence at first detection if the planned budget is kept fixed (Table 3). The cost performance of the test programme can be improved by introducing an LFD as a third confirmatory test (A×B×C; 1,3,2). Two options for the use of this test programme are shown in Table 3. True outcome (scenario 1) shows a scenario where the budget is fixed as originally planned which provides a prevalence at first detection which is slightly worse than originally planned but is better than was actually being achieved. True outcome (scenario 3) shows a scenario in which the re-optimised test programme is used in a sampling programme which achieves the original target for prevalence at first detection of no more than 1% with 95% confidence.

Table 3: Optimum test programmes and their performance in sampling plans based on apparent and true unverified and verified estimates of test performance

Strategy	Test programme	$prev$	$N_{k;o}$	DSE	DSP	$c_{T;FP}$	$c_{T;S}$	c_T
Note: *Prev* = prevalence; $N_{k;o}$ = Number of samples tested per time interval by a test programme *k* and test order *o*; DSE = diagnostic sensitivity; DSP = diagnostic specificity; $c_{T;FP}$ = total cost of false positive test results; $c_{T;S}$ = total sampling cost; c_T = total cost								
Apparent outcome: programme optimised with **unverified** test parameters	A×B; 1,3	1.00%	1807	0.200	1.000	0	5847	5847
True outcome (scenario 1): programme optimised with **unverified** test parameters (fixed budget)	A×B; 1,3	1.37%	1346	0.196	0.998	1346	4374	5847
True outcome (scenario 2): programme optimised with **unverified** test parameters (fixed number of tests)	A×B; 1,3	1.02%	1807	0.196	0.998	2346	5847	7833
True outcome (scenario 1): programme optimised with **verified** test parameters (fixed budget)	A×B×C; 1,3,2	1.17%	1642	0.188	0.9996	360	5486	5847
True outcome (scenario 3): programme optimised with **verified** test parameters (fixed prevalence target)	A×B×C; 1,3,2	1.00%	1922	0.188	0.9996	423	6397	6819

6. Impact and Next Steps

This deliverable describes the development and current form of a mathematical framework which was created in response to a number of issues/questions that arose during the VALITEST project. Its ultimate goals are to:

- Provide a frame for assessing case studies in a standardised manner
- Facilitate discussions between different stakeholder groups in plant health testing and thus support integrated decision making
- Support decision making in the following areas:
    - Budget planning for sampling / test programmes
    - Design of efficient sampling / test programmes
    - Determining the value of validation
    - Resource allocation for the validation of tests
- Inform discussions on international protocols (e.g. on which performance characteristics should be included in a validation report)

In its current form, the framework allows the assessment of case studies in an 'early detection' scenario. While this report showcases the potential power of the framework using a real-life case study to examine the value of validation, it also highlights areas for future development; these are in no particular order:

- Develop a user-friendly interface together with stakeholders (suggestion: R-based Shiny web-interface)
- Develop and programme additional scenarios. For example, the case of eradication, where one may be looking for symptomatic plants and thus starts off with a visual assessment
- Delve deeper into the 'early detection' scenario and research current processes and terminologies using more case studies. For example, people often mention that they collect asymptomatic plants for testing, which implies that they have already done a visual test. However, they might in fact collect random samples from an area they expect to be completely pest free, wherefore plants are by default asymptomatic
- Account for uncertainty, i.e. integrate uncertainty in diagnostic sensitivity and diagnostic specificity to elucidate its effect on uncertainty about the cost of test programmes and the expected upper limit for prevalence at first detection provided by sampling programmes
- Delve deeper into individual cost elements of excessive or insufficient management action
- Create more case studies for existing pests / tests

It is suggested to prioritise the options above together with potential end-users, to ensure the framework/tool is fit for purpose and can be readily integrated into current workflows and decision processes.

7. References

EPA (2021). [online resource] https://www.epa.gov/sites/default/files/2015-01/documents/final_microbiology_method_guidance_110409.pdf [last accessed October 2021]

EPPO (2018) PM 7/76 (5) Use of EPPO Diagnostic Standards. Bulletin OEPP/EPPO Bulletin, 48 (3), 373-377.

EPPO (2019) PM 7/98 (4) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. Bulletin OEPP/EPPO Bulletin, 49 (3), 530-563.

ISO 17025:2017: General requirements for the competence of testing and calibration laboratories. BSI

ISO 5725:1994: Accuracy (trueness and precision) of measurement methods and results. ISO, Geneva (1994)

Mastin, A. J., van den Bosch, F., van den Berg, F., & Parnell, S. R. (2019). Quantifying the hidden costs of imperfect detection for Early Detection surveillance. Philosophical Transactions of the Royal Society B, 374(1776), 20180261.

Thompson, Michael & Ellison, Stephen & Wood, Roger. (2002). Harmonized guidelines for single-laboratory validation of methods of analysis (IUPAC Technical Report). Pure and Applied Chemistry - PURE APPL CHEM. 74. 835-855. 10.1351/pac200274050835.

Weinstein, S., Obuchowski, N. A., & Lieber, M. L. (2005). Clinical evaluation of diagnostic tests. *American Journal of Roentgenology*, *184*(1), 14-19.

Appendix - Detailed Framework Description

Calculating the overall test programme diagnostic sensitivity and specificity and cost of testing: motivating example

In radiology calculating the diagnostic sensitivity (DSE) and specificity (DSP) of multiple tests is a common statistical problem because frequently two tests (A and B) with different sensitivities and specificities are combined to diagnose a particular disease or condition (Weinstein, Obuchowski et al. 2005). When a test programme consists of multiple tests, they can be interpreted in either an "**and**" or an "**or**" manner.

The probability of a test programme in which two tests (A **and** B) must give a positive response for infestation to be deemed detected (an "A x B" test programme) is[2]:

$$P_{A\times B} = p_A p_B \tag{1}$$

Where, p_A is the probability of test A giving a positive response and p_B is the probability of test B giving a positive response. The probability of this programme giving a positive result is lower than when either test is used alone.

For this test programme, if test A is negative then the second test does not need to be performed because we already know that that test programme will give a negative result. Hence, every time we employ the test programme, we use test A, but we only use test B on a proportion (p_A) of occasions. The average programme cost per plant tested can then be calculated as:

$$c_{A\times B} = c_A + p_A c_B \tag{2}$$

where c_A and c_B are the costs of test A and B, respectively. The cost of this programme is higher than the cost of test A but may be lower than the cost of test B if test A is particularly cheap or test A rarely gives positive results.

We could instead choose to use the two tests in an "**or**" programme, where the infestation is deemed to have been detected if either of the tests gives a positive result. The diagnostic sensitivity of such an A **or** B ("A + B") test programme can be calculated as[3]:

$$p_{A+B} = 1 - (1 - p_A)(1 - p_B). \tag{3}$$

The probability of this two-test "**or**" programme giving a positive result is higher than for either test alone, and the two-test "**and**" programme.

For this test programme if A is positive, then the second test does not need to be performed because we already know that the programme will give a positive result. Hence, the average test programme cost per plant tested can then be calculated as:

$$c_{A+B} = c_A + (1 - p_A)c_B. \tag{4}$$

[2] The programme gives a positive result if two events both occur: A is positive, and B is positive. The probability of two events both occurring is given by multiplying the probability of each event occurring.

[3] The programme gives a positive result if either of the tests is positive. This means that two events have to both occur for the programme to give a negative result: A has to be negative and B has to be negative. Hence, we multiply the probabilities of each test being negative to calculate the probability of the programme giving a negative result and subtract that value from 1 to give us the probability of the programme giving a positive result.

The cost of this programme is higher than the cost of test A, but if p_A is greater than 50% then the cost of this programme is lower than the cost of the two-test "**and**" programme.

If for each test the cost of the test and the probability of the test resulting in a positive response is known, then Equations 1 to 4 can be used to rank the probability of a positive response and cost for both individual tests and the two two-test programmes. See Table 1 for an example, whereby Test A has a lower probability of giving a positive response than Test B, Test A is cheaper than Test B and p_A is greater than 50%.

Table 1. Ranking of the probability of a positive response and cost for the one-test and two-test programmes considered (based on Equations 1-4). In this example Test A has a lower probability of giving a positive response than Test B; Test A is cheaper than Test B and p_A is greater than 50%.

Test programme	Probability of positive result	Average test programme cost
A x B	*LOWEST*	*HIGHEST*
A	*LOWER*	*LOWEST*
B	*HIGHER*	*LOWER*
A + B	*HIGHEST*	*HIGHER*

Hence, in this simple example where there is a single probability of a positive result being observed across plants we have shown how to calculate four probabilities of a positive result being produced and four costs. In any given testing scenario this gives us the information we need to choose the test programme which meets our needs at the lowest cost.

In practice each test is applied to two types of plants: infested plants and uninfested plants. In this case the probability of a positive response depends on the infestation status of the plants and is anticipated to be close to zero for uninfested plants (diagnostic specificity (DSP) close to 100%) and much higher for infested plants (diagnostic sensitivity (DSE) close to 100%). Hence, the performance of test programmes can be evaluated for both types of plant by applying Equations 1 to 4 separately for infested and uninfested plants

For infested plants, $p_A = DSE_A$ and $p_B = DSE_B$.Hence Equations 1 and 2 can be used to calculate $DSE_{A\times B}$ and $c_{infected;A\times B}$ per infested plant tested, whereas Equations 3 and 4 can be used to calculate DSE_{A+B} and $c_{infected;A+B}$ per infested plant tested.

For uninfested plants, $p_A = 1 - DSP_A$ and $p_B = 1 - DSP_B$. Hence in this case Equations 1 and 2 can be used to calculate $1 - DSP_{A\times B}$ and $c_{uninfected;A\times B}$ per uninfested plant tested, whereas Equations 3 and 4 can be used to calculate $1 - DSP_{A+B}$ and $c_{uninfected;A+B}$ per uninfested plant tested.

Then if a test programme is applied to a population in which the prevalence of infested plants is prev, the expected average cost per test for a test programme k is given by

$$cost = c_{uninfected;k} \cdot (1 - prev) + c_{infected;k} \cdot prev \qquad (5)$$

Table 2 shows an example where Test A gives a lower probability of a positive response than test B for both infested and uninfested plants; if Test A p_A is <50% for uninfested plants and >50% for infested plants and as the previous example in Table 1 Test A is cheaper than Test B. This example highlights that typically there are several trade-offs between the test programme cost and both the programme DSE and DSP.

Table 2. Ranking of the probability of a positive (negative) response and cost for the one-test and two-test programmes considered (based on Equations 1-4) and separated for infested and uninfested plants. In this example Test A gives a lower probability of a positive response than test B for both infested and uninfested plants; for Test A p_A is <50% for uninfested plants and >50% for infested plants and Test A is cheaper than Test B.

Test programme	Infested plants		Uninfested plants	
	Probability of a positive result (programme DSE)	Average test programme cost	Probability of a negative result (programme DSP)	Average test programme cost
A x B	*LOWEST*	*HIGHEST*	*HIGHEST*	*HIGHER*
A	*LOWER*	*LOWEST*	*HIGHER*	*LOWEST*
B	*HIGHER*	*LOWER*	*LOWER*	*LOWER*
A + B	*HIGHEST*	*HIGHER*	*LOWEST*	*HIGHEST*

The probability of a positive detection decision and costs for testing a plant for all test programmes consisting of up to three tests is given in Table 3.

The equation for the test programme DSE and the equation for the average cost per infested plant tested are given by replacing p in Table 3 with the DSE of a specific individual test in the equation for the test programme probability of producing a positive result and the equation for the average test programme cost, respectively. The programme DSP can then be derived from the equation for the test programme probability of producing a positive result by substituting p with $(1 - DSP)$, i.e. the false positive rate of each test, and similarly the equation for the average test programme cost for an uninfested plant can be derived from the equation for the average test programme cost by substituting p with $(1 - DSP)$.

Table 3. Equations for the test programme probability of producing a positive result and the average test programme cost for all test programmes containing up to three tests. Note that tests are performed in alphabetic order, i.e. test A will always be performed before test B and C. Each test is associated with its own test specific probability of producing a positive result, denoted by p and cost c. Replacing p with $(1 - DSP)$ in the probability of a positive result and cost equation, converts the equations into equations summarising the test programme DSP and the average test programme cost for an uninfested plant, respectively.

Test programme	Test programme probability of producing a positive result	Average test programme cost
A	p_A	c_A
A x B	$p_A p_B$	$c_A + p_A c_B$
A + B	$1 - (1 - p_A)(1 - p_B)$	$c_A + (1 - p_A)c_B$
A x B x C	$p_A p_B p_C$	$c_A + p_A c_B + p_A p_B c_C$
A x B + C	$1 - (1 - p_A p_B)(1 - p_C)$	$c_A + p_A c_B + (1 - p_A p_B)c_C$
A x C + B	$1 - (1 - p_A p_C)(1 - p_B)$	$c_A + c_B + p_A(1 - p_B)c_C$
A + B x C	$1 - (1 - p_A)(1 - p_B p_C)$	$c_A + (1 - p_A)c_B + (1 - p_A)p_B c_C$
A x (B + C)	$p_A\big(1 - (1 - p_B)(1 - p_C)\big)$	$c_A + p_A c_B + p_A(1 - p_B)c_C$
(A + B) x C	$\big(1 - (1 - p_A)(1 - p_B)\big)p_C$	$c_A + (1 - p_A)c_B + \big(1 - (1 - p_A)(1 - p_B)\big)c_C$
(A + C) x B	$\big(1 - (1 - p_A)(1 - p_C)\big)\, p_B$	$c_A + c_B + (1 - p_A)p_B c_C$
A + B + C	$1 - (1 - p_A)(1 - p_B)(1 - p_C)$	$c_A + (1 - p_A)c_B + (1 - p_A)(1 - p_B)c_C$

Changing the test order

Table 3 shows all of the distinct test programmes for 1, 2 and 3 tests, giving 1, 2 and 8 programmes, respectively. In addition, the order in which tests are applied can change the performance and cost of a test programme, which increases the effective number of potential programmes which can be used: a single test can be done in one order, two tests in two orders, and three tests in six orders. Hence, there is only one way to use a single test to make a diagnostic decision, four ways to use two tests (two programmes in two orders) and 48 different ways of using three tests.

Sample plans and programmes for early detection

The aim of 'early detection' is to detect the presence of a pest or pathogen soon after its establishment into a population while its prevalence is still low. If the prevalence is sufficiently low at first detection, then there may be more options for control or eradication. Following Mastin, van den Bosch et al. (2019), if there is an ongoing sampling programme in which N samples per time interval Δ are tested by a test programme *k* and test order *o*, with an overall a diagnostic sensitivity $DSE_{k;o}$ for a pest or pathogen which has not yet been detected but which may be present and growing exponentially with rate constant *r*, then to ensure (with confidence *x*) that the pathogen is present in a proportion of the population no higher than *prev* when it is first detected we require that:

$$N_{k;o;\Delta} \geq -\ln(1-x)\left(\frac{r\Delta}{DSE_{k;o}\cdot prev}\right) \quad (6)$$

Expressing the rate at which samples are tested per unit time:

$$N_{k;o} \geq -\ln(1-x)\left(\frac{r}{DSE_{k;o}\cdot prev}\right) \quad (7)$$

For 95% confidence this can be approximated by:

$$N_{k;o} \geq \frac{3.r}{DSE_{k;o}\cdot prev}. \quad (8)$$

And

$$prev \leq \frac{3.r}{DSE_{k;o}\cdot N_{k;o}}. \quad (9)$$

Hence, a sample plan in which the product of the number of samples tested and programme sensitivity is maximised provides the lowest estimated upper limit for prevalence at which a pest or pathogen is first detected. The goal for designing a test plan is to maximise this product which is achieved when the test sensitivity / cost per test is maximised.

Two different test programmes applied in a survey such the product $DSE_{k;o} \cdot N_{k;o}$ is the same are equally effective for 'early detection', having the same estimated upper limit for prevalence. And both lead to same estimated prevalence at first detection as a survey in which Neff plants are tested per interval using a method with 100% sensitivity where:

$$N_{eff} = DSE_{k;o} \cdot N_{k;o} \quad (10)$$

And

$$prev \leq \frac{3.r}{N_{eff}} \quad (11)$$

Hence, the value of a survey undertaken for 'early detection' is defined by the effective number of samples N_{eff} and a measure of cost for a test programme is cost per effective sample. The best test programme is the one that minimises the cost per effective sample.

The total cost of the sampling plan, c_T, is composed of both the total sampling cost, c_S, and the total cost of false positive test results (i.e. additional work done to determine that an apparent positive result is a false positive result), c_{FP}, such that:

$$c_T = c_{T;S} + c_{T;FP}. \tag{12}$$

The total sampling cost is made up of the survey location visit costs, and the cost of all the tests performed on the samples collected, leading to:

$$c_{T;S} = N_{k;o} \cdot \left(c_c + c_{uninfected;k;o} \cdot (1 - prev) + c_{infected;k;o} \cdot prev\right) \tag{13}$$

Where $N_{k;o}$, is the number of samples taken and tested using test programme k and test order o, c_C is the sample collection cost, and $c_{uninfected;k;o}$ and $c_{infected;k;o}$ are the average testing cost for an uninfested and infested sample, respectively.

The expected cost of false positive test results, $c_{T;FP}$, is composed of the expected number of false positive test results given by the test programme, $E_{FP;k;o}$, and the average cost of dealing with a false positive test result, c_{FP}. The cost of dealing with a false positive result represents the cost of unnecessary confirmatory testing through culturing or in situ testing at the area in which the sample classed as positive was found to eventually identify that the sample was a false positive rather than a true positive. This leads to:

$$c_{T;FP} = E_{FP;k;o} \cdot c_{FP}, \tag{14}$$

And

$$E_{FP;k;o} = (1 - prev)\left(1 - DSP_{k;o}\right)N_{k;o} \tag{15}$$

with $DSP_{k;o}$ the diagnostic specificity of test programme *k* and test order o.

Hence for a total budget B the number of samples that can be tested (when taken from across v locations) is:

$$N_{k;o} = \frac{B - v.c_v}{c + \left(1 - DSP_{k;o}\right).c_{FP}} \tag{16}$$

For 'early detection', the test programme which maximises the product of $N_{k;o}$ (Equation 16) and $DSE_{k;o}$ (using formulae in Table 3) will provide the lowest prevalence at first detection (Equation 9). The product of $N_{k;o}$ and $DSE_{k;o}$ gives *the effective number of samples tested for 'early detection'.* That is the number of samples tested with 100% sensitivity that would give the same maximum prevalence at first detection. Hence, given three tests each with a sensitivity, specificity and cost we can choose the programme (out of the 51 possible programmes where three tests may be employed) that is most fit for that specific purpose by calculating the total cost of employing each programme across a range of prevalences and the effective number of samples that are tested.