Opinion Paper

Andrea Rita Horvath*, Patrick M.M. Bossuyt, Sverre Sandberg, Andrew St John, Phillip J. Monaghan, Wilma D.J. Verhagen-Kamerbeek, Lieselotte Lennartz, Christa M. Cobbaert, Christoph Ebert and Sarah J. Lord, for the Test Evaluation Working Group of the European Federation of Clinical Chemistry and Laboratory Medicine

Setting analytical performance specifications based on outcome studies – is it possible?

DOI 10.1515/cclm-2015-0214
Received March 2, 2015; accepted April 10, 2015

Abstract: The 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine proposed a simplified hierarchy for setting analytical performance specifications (APS). The top two levels of the 1999 Stockholm hierarchy, i.e., evaluation of the effect of analytical performance on clinical outcomes and clinical decisions have been proposed to be replaced by one outcome-based model. This model can be supported by: (1a) direct outcome studies; and (1b) indirect outcome studies investigating the impact of analytical performance of the test on clinical classifications or decisions and thereby on the probability of patient relevant clinical outcomes.

This paper reviews the need for outcome-based specifications, the most relevant types of outcomes to be considered, and the challenges and limitations faced when setting outcome-based APS. The methods of Model 1a and b are discussed and examples are provided for how outcome data can be translated to APS using the linked evidence and simulation or decision analytic techniques.

Outcome-based APS should primarily reflect the clinical needs of patients; should be tailored to the purpose, role and significance of the test in a well defined clinical pathway; and should be defined at a level that achieves net health benefit for patients at reasonable costs. Whilst it is acknowledged that direct evaluations are difficult and may not be possible for all measurands, all other forms of setting APS should be weighed against that standard, and regarded as approximations. Better definition of the relationship between the analytical performance of tests and health outcomes can be used to set analytical performance criteria that aim to improve the clinical and cost-effectiveness of laboratory tests.

Keywords: analytical performance; clinical effectiveness; evidence-based laboratory medicine; outcome.

Introduction

The IFCC-IUPAC Stockholm Consensus Conference defined a scientific approach for setting analytical performance goals for laboratory tests in 1999. In their landmark paper the authors presented a hierarchy of models with the “evaluation of the effect of analytical performance on clinical outcomes” representing the top of the hierarchy,
followed by the “evaluation of the effect of analytical performance on clinical decisions” [1].

Fraser noted that the outcome- and decision-based model must be related to well-characterized and accepted medical strategies directly guided by the test results. It was also acknowledged that these approaches are sometimes difficult to apply because most laboratory tests are used in many different ways by clinicians depending on the clinical scenario and clinical pathway [2].

In the 15 years since the Stockholm Conference, very few papers have been published about outcome-based analytical performance criteria. One may therefore question whether it is worth keeping this model at the top of the hierarchy and, if so, what the profession should do to ensure that outcome-based analytical performance specifications (APS) are developed and implemented in the practice of laboratories, in vitro diagnostic (IVD) manufacturers, quality assurance and regulatory bodies.

This report reflects the views and opinions of the Test Evaluation Working Group (WG-TE) of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). This working group includes representatives of laboratory medicine, the IVD industry and experts in epidemiology, evidence-based diagnostics and health technology assessment. The group reviews the impact of analytical performance characteristics of laboratory tests on health and other patient relevant outcomes in the context of their intended use in clinical practice. The same principles apply for laboratory tests performed for different purposes, (i.e., screening, diagnosis, prognosis, risk assessment, treatment selection or monitoring) and in different roles in a given clinical pathway (e.g., as a replacement of another test, as an add-on test in a multiple biomarker strategy, or as a triage test in a diagnostic algorithm) [3].

We have previously defined analytical performance as one of the essential components of test evaluation, alongside the clinical performance; clinical effectiveness; cost-effectiveness; and the broader impact of testing on organizational, societal and other consequences [3]. We describe test evaluation as a dynamic cycle that integrates the above components, each linked by the core component, i.e., the outcome-driven clinical pathway in which the test will be used (Figure 1). Using this framework, we recognize that tests can impact on patient health and broader stakeholder outcomes through multiple pathways as described elsewhere [4, 5]. Given the main function of testing is to provide information that guides clinical decisions about patient management, the major pathway is through the impact of test results on clinical decisions about further testing and treatment and thereby health outcomes – the “testing-management-outcomes” pathway. We acknowledge that other pathways may work through improved safety of the test procedure, e.g., with the introduction of a less invasive test; through improved turn-around-time (TAT) of testing to support fast medical decisions, e.g., with point-of-care testing in acute care; through improved patient convenience of the procedure leading to improved patient uptake and thereby clinical outcomes; and through reduced costs leading to improved societal benefits.

When setting APS, however, our main focus is on the impact of variations in analytical performance on clinical performance, such as diagnostic or prognostic accuracy, and thereby clinical effectiveness, through the testing-management-outcomes pathway. More accurate tests improve patient outcomes only if the analytically improved test increases the rate of appropriate diagnoses (i.e., the diagnostic yield of a test) and appropriate treatment selection and uptake (i.e., the therapeutic yield) [5]. Other attributes of the test that impact on health outcomes through other pathways, such as test safety, speed,
convenience and costs are secondary considerations when setting APS.

**Do we need outcome-based analytical performance specifications?**

APSs are a set of criteria that specify, in quantitative terms, the quality required for analytical performance to deliver laboratory test information that would achieve the best possible health outcomes for patients. The ultimate goal of any medical intervention, including laboratory testing, is to provide health benefits to patients without causing harm. It follows that APS for tests should seek to define the level of reliability required to achieve a favorable balance between health benefits and harms at acceptable costs.

Testing itself rarely leads to health outcomes. As described above, testing guides the actions of clinicians and patients, and these actions may be responsible for changes in health outcomes. Thus, laboratory tests are almost always linked to outcomes in an “indirect” manner by applying test information to clinical decisions. This makes measuring the effectiveness of testing methodologically complicated, which may explain, at least partially, the scarcity of outcome studies in laboratory medicine [6]. We believe that the lack of direct outcome studies, and the practical difficulties in designing such studies, however, does not justify the removal of outcome-based APS from the Stockholm hierarchy. Outcome studies allow an assessment of the impact of the analytical performance of a test on health outcomes together with any other unique attributes that may affect the benefit-harm balance, such as improved patient compliance, convenience or satisfaction. This information is needed to achieve meaningful evidence-based APS and can also guide the development of better tests.

It is well established that poor analytical performance – in particular bias around clinical practice guideline-driven decision limits – can lead to poor diagnostic yield with misclassification and subsequent mismanagement of patients [7]. For example, there is a wide variation in the analytical performance of intact parathyroid hormone (iPTH) immunoassays between laboratories. In one study a 16% difference was observed in the number of dialysis patients achieving guideline-driven therapeutic targets for secondary hyperparathyroidism, depending on which iPTH assay was used [8]. After harmonizing the analytical performance of different iPTH tests, about one fifth of patients would have undergone different clinical management [8, 9].

We emphasize the need for considering the consequences of testing and thus clinical outcomes because assumptions about the relationship between analytical performance and health outcomes have not always proven to be correct. High quality analytical performance does not guarantee high quality clinical action or patient compliance or that the chosen treatment will be effective. The opposite is also true; poor analytical performance of a test that plays a small part in a complex clinical pathway may not necessarily lead to adverse or unfavorable outcomes. Below we describe an example which demonstrates that a reduction in the analytical performance of certain laboratory tests does not necessarily lead to patient harm. A systematic review of several analytical performance studies has shown that prothrombin time/INR measurement by point-of-care testing is less precise and accurate than by laboratory-based coagulation methods [10]. It was assumed that such variations in results could lead to inappropriate warfarin dosing decisions. Several meta-analyses of randomized controlled trials (RCTs) comparing point-of-care and standard monitoring, however, showed that patients who self-monitor and -manage, using point-of-care INR data, can improve the quality and effectiveness of their oral anticoagulation therapy. The number of thromboembolic events and mortality were lower, without increase in harm from bleeding, and patients on the point-of-care group reported improved quality of life [10, 11]. One reason is probably that the slight reduction in analytical quality is compensated by more frequent measurements and more tailored self-management of warfarin dosing.

The importance of linking analytical performance and health outcomes of testing to optimize the development of tests has also been discussed by Werner, who pointed out that “the linkage between analytical performance goals and medical strategies is reciprocal; namely, that outcome can just as well be optimized by tailoring medical strategy to existing analytical performance, as by adapting analytical performance to medical strategy” [12]. If one decides to only accept the former and the profession’s role is simply to optimize medical management of conditions according to the existing analytical performance of tests (i.e., the state-of-the-art) then a number of clinically valuable biomarkers [i.e., higher sensitivity troponin (hs-Tn) or ultra-sensitive TSH tests, etc.] would probably not have been developed and put into practice. Outcome-based APS therefore not only facilitate research and development of newer and more advanced biomarkers but they also help laboratories in delivering tests that better support the needs of clinicians and patients and other stakeholders.
What outcomes are relevant for setting analytical performance specifications?

Outcomes of medical tests can be divided into clinical (health), operational and economic outcomes [13, 14]. These outcomes can be defined from various stakeholder perspectives: e.g., patient, population, healthcare staff, and policy makers. Clinical/health outcomes can be short- or long-term; and can include objective (i.e., mortality, morbidity and complication rates) and subjective measurements (i.e., the impact of medical care on patient well being, quality of life and satisfaction with care). Operational outcomes, also referred to as process outcomes, include time to test results, time to treatment, length of stay in hospital, and readmission rates. Economic outcomes include the cost of the test and all downstream consequences on health care including operational outcomes, such as the financial benefits of reductions in length of stay in hospital [13]. Comparison of health, operational and cost outcomes between alternative strategies or no testing (if feasible and ethical) allows assessment of the clinical and cost-effectiveness of a test.

When defining the health outcomes of testing, both the potential benefits and harms need to be considered. Early diagnosis of a condition which may not manifest itself, or manifests only much later in life and for which there is no effective treatment (e.g., predictive testing for certain inherited diseases) is one example where the harms can potentially outweigh the benefits of testing. Patient relevant harms of such testing in cases with a positive test result (i.e., increased anxiety, fear of making longer term life plans, impacts on family and social life and behavioral or psychological problems, etc.) may outweigh any potential benefits. These consequences of testing need to be considered when setting APS, as there is no need for perfecting the analytical performance of a test if testing leads to no effective treatment or incremental therapeutic yield.

Economic and organizational attributes of testing may also be relevant in achieving the best health outcomes for patients, sometimes even at the expense of analytical quality [e.g., provision of faster point-of-care troponin (Tn) results in acute chest pain in remote locations without hospital facilities, or more affordable rapid tests for HIV and tuberculosis in less developed countries]. Thus, when setting APS for tests that use different procedures and different costs, operational and economic outcomes are important but should remain only secondary considerations.

The below example, comparing the analytical and clinical performance of hs-Tn and former generation of conventional Tn assays, illustrates why the primary focus for setting APS should be on determining the impact of analytical reliability of tests on health outcomes before considering the additional impact of variations on test cost and process efficiency. High sensitivity Tn assays by picking up acute myocardial damage earlier than conventional Tn assays may lead to faster clinical decisions for initiating treatment to reduce myocardial damage (clinical outcome improved by better analytical test performance and diagnostic accuracy). Fast TAT of hs-Tn tests in an emergency setting may also result in shorter door to needle time, rapid thrombolytic treatment and thus saving as much myocardium as possible from necrosis (clinical outcome improved by a process measure, i.e., faster TAT). Rapid TAT can also result in patient satisfaction (patient relevant outcome improved by a process measure). The latter two scenarios of improved processes of care, whilst relevant for clinical- and patient-related outcomes, are less informative for developing APS. From the pathophysiological and clinical outcome point of view the first scenario, i.e., having hs-Tn versus conventional Tn assays is more relevant to setting APS as hs-Tn assays must meet the lowest possible analytical variation at the 99th percentile for being useful in making clinical decisions about subsequent treatment that ultimately will save patients’ lives and preserve their quality of life. A conventional Tn test compared to the hs-Tn test, even if faster, could still miss more acute myocardial infarction cases. Therefore, whilst operational outcomes in scenario 2 and 3 may be excellent with Tn testing, the health outcomes of patients will still be poorer than with a hs-Tn assay, even if the TAT of the latter had been poorer than that of the former.

What are the challenges and limitations of setting outcome-based analytical performance specifications?

Ideally APS should reflect real-life clinical scenarios, responding to the clinical needs of health professionals and patients. This approach, however, has several challenges and practical limitations:

- few diagnostic health outcome studies exist that can be used to set APS;
- the link between analytical performance and health outcomes is indirect and therefore the impact of testing and the quality of testing on health outcomes is difficult to measure;
it is only useful for situations where the links between the test results, clinical decisions and health outcomes are straightforward and well documented;
- no simple methodology and tools exist that could assist the profession in setting outcome-based APS.

The lack of high quality diagnostic RCTs or observational studies comparing different test-treatment strategies is partly due to the indirectness and complexity of the link between testing and health outcomes. The design of these studies, and their interpretation for setting APS is further complicated by the diversity of situations in which the test is used and the diversity of medical decisions and the choice of management strategies that the test would trigger in various clinical settings [12, 15].

The same laboratory test can be used for different purposes and in different roles in a given clinical pathway. These circumstances have a major influence on the clinical performance of the test and on the impact of the test on selecting downstream healthcare management options. Therefore, APS of the same measurand could be different for each testing scenario which prohibits the establishment of a universal quality specification in some cases.

In laboratory medicine, particularly with the emergence of new genetic biomarkers, the feasibility of assessing health outcomes can also be challenging even if the primary purpose of the test and link between test results and clinical decisions are well defined. Follow-up of patients for assessing changes in morbidity and mortality is not always practical, especially if health outcomes manifest many years after testing [16]. Moreover, measuring long-term outcomes may be confounded by multiple healthcare events and interventions, including clinician adherence to guidelines and patient adherence to treatment that may be unrelated to the information provided by the test. The need to include different types of outcomes and to take into account the perspectives of various stakeholders involved in the care of patients when assigning values to these outcomes may add to the complexity.

How can outcome-based analytical performance specifications be defined?

Considering these challenges, two models have been recommended for setting APS related to health outcomes at the 2014 conference in Milan: 1a) direct outcome studies, and 1b) indirect outcome studies.

**Model 1a: Analytical performance specifications based on direct outcomes data**

Defining APS on the basis of health outcomes data presumes that the outcomes of a test-treatment pathway are clearly defined and measurable. For defining APS, *direct* outcome studies should investigate the impact of the analytical performance of the test on health outcomes. This model requires a diagnostic RCT as the most appropriate study design. Ideally such a study would compare test-treatment strategies by randomizing patients who receive the test in question with a given analytical performance versus a strategy with the same test but an alternative analytical performance. For example, patients with symptoms of acute coronary syndrome could be randomized to receive a 4th generation Tn or a 5th generation hs-Tn assay. Positive and negative results would then be managed in the same way in both arms of the trial and patient outcomes recorded at discharge and after a 90-day follow-up for mortality, readmission rate, or quality of life.

In general, performing diagnostic RCTs is challenging as they require a relatively large sample size compared to treatment RCTs. The smaller the incremental benefits of testing, the larger the sample size required. Such trials may take a long time to complete, especially if a prolonged follow-up time is required to capture the full health consequences of treatment decisions made based on the test results. Once trial evidence has demonstrated the clinical value of a test, the need for further trials, to compare the marginal impact of differences in analytical performance of the test on its clinical effectiveness, may be seen as less urgent.

Alternative RCT designs have been described that can be used in some situations to reduce sample size requirements [17]. However, these alternatives are only suitable when the test results are proposed to inform a single well defined management strategy. As laboratory tests are used in many different ways in clinical practice, it is unfeasible to directly investigate all patient management options in RCTs. In addition to these challenges, biomarker technology can advance rapidly during the period of the trial which limits the future relevance of the results.

Given these limitations, diagnostic RCTs to assess the impact of analytical performance on health outcomes will be more feasible for tests used to inform well defined and standardized clinical decisions, with health outcomes that can be measured in a relatively short time frame. Alternatively, when diagnostic RCTs have already demonstrated the health outcomes of testing, further RCTs to assess the impact of differences in analytical performance may not
be necessary and indirect outcome data will suffice as outlined below for Model 1b.

**Model 1b: Analytical performance specifications based on indirect outcomes data**

A more pragmatic approach is to use indirect outcome data. This approach is commonly used to compare new versus existing tests and involves measuring differences in test performance to infer difference in health outcomes. When using an indirect approach, also referred to by some as a linked-evidence approach, evidence usually needs to be built-up from separate studies that investigate various elements of the testing – management – outcomes pathway [18].

The linked evidence approach recognizes that diagnostic RCTs are not necessary when:

- the clinical decisions associated with the test results are well defined;
- evidence about the diagnostic accuracy of the test (i.e., diagnostic sensitivity and specificity) to classify patients for these clinical decisions is available and is generalizable to the patient population and clinical setting of interest; and
- the consequences of correct/incorrect classification (i.e., true and false positives and negatives) have already been established.

Drawing this information together, one can conclude whether the diagnostic accuracy of a test versus no testing or an existing test may lead to improved outcomes with a favorable balance of benefits versus harms [19].

Klee has concluded that there is no single strategy for establishing outcome-related analytical performance goals, and multiple interrelated approaches may be necessary. Klee reviewed six indirect approaches for setting APS – all of which were related to the impact of tests on clinical decisions, but without providing a further clear link to outcomes [20]. The linked evidence approach offers the opportunity to go beyond the impact of tests on clinical decisions. Using this approach for the indirect assessment of analytical performance on health outcomes, conventional diagnostic accuracy studies can be designed to assess the impact of analytical performance on clinical classifications. Data from these studies can then be used to infer impact on health outcomes by considering the consequences of the test results on clinical decisions and the associated clinical benefits and harms.

A potential way forward is to extend the linked evidence approach to setting APS to recognize the link between analytical and clinical performance as outlined in Figure 1. This would involve defining the existing analytical performance of the tests used to generate diagnostic accuracy estimates. This state-of-the-art level can be considered as the benchmark if the test is already used in standard practice. If analytical performance varies between similar tests or laboratories, a study needs to be undertaken to investigate the impact on clinical performance. The aim of such a study would be to define the minimum acceptable threshold to achieve the clinical performance that supports current recommendations for testing.

If differences in the analytical performance of the test are shown not to alter the clinical classification of patients, then analytical performance goals may be set at the lower level studied. Conversely, if improved classification of patients is reported at the higher level of analytical performance, then further evidence is required to establish the health consequences of improved classification.

In some cases, assumptions about health benefits will be straightforward; e.g., when improved analytical performance leads to fewer false negative results where the benefits of treatment for cases are well established. In these situations, APS may shift up to improve the clinical effectiveness of the test. In other cases, these assumptions may be less straightforward; e.g., if the gain in test reliability leads to the detection of additional mild cases of disease, in which the net benefits of treatment are less well established. In these situations, a direct assessment of the impact of higher analytical performance levels on health outcomes may be justified before setting APS.

This approach can also be used when a new test is introduced that offers a higher level of analytical performance. Here, a study will be needed to evaluate the impact on clinical performance and a linked evidence approach can be used to estimate the health consequences. This evidence is required before raising the minimum acceptable threshold to match this new test. If the new test is very expensive we may demand robust evidence to be convinced it improves health outcomes (and by how much). Importantly, where the improved analytical performance leads to improved diagnostic accuracy for detecting a mild spectrum of disease (i.e., not detected when using the existing tests) treatment RCT may still be required to confirm that benefits of existing treatments outweigh the harms, or to investigate more appropriate treatment options for this subgroup of patients with “mild” disease.

This approach can also be used when a new test is proposed to have adequate analytical performance but
is cheaper or safer or more convenient than existing tests. We will need diagnostic accuracy studies to assess whether it provides adequate classification to meet the current benchmark and thereby does not compromise health outcomes. However, where there is a trade-off between the advantages of this test and analytical and clinical classification performance – diagnostic RCTs may still be required.

The approach outlined above is essentially qualitative and does not allow quantification of the health outcomes estimated for different analytical performance levels. Decision analysis provides a mathematical framework to assemble this evidence with the aim to assess trade-offs between benefits, risks, and costs of interventions [15]. Key input parameters in such simulation models are derived from studies on:
- disease prevalence in a particular setting;
- technical evaluation of the analytical performance of the test or various forms of the test;
- diagnostic or prognostic accuracy of the test in various patient populations;
- magnitude of benefits and harms of subsequent treatments using various measures of patient-centered outcomes.

For example, the effects of screening for type 2 diabetes mellitus on life expectancy have not been directly studied. Numerous studies exist on the prevalence, the analytical performance of glucose measurements by various methods, the diagnostic accuracy of screening by fasting glucose or oral glucose tolerance tests, and on the effectiveness of therapeutic interventions of type 2 diabetes.

Simulation modeling can account for uncertainty in key input quantities, and may be well suited for studying the effects of varying analytical performance on subsequent clinical decisions and outcomes. Karon et al. performed simulation modeling to investigate at what level of analytical performance glucose monitoring by point-of-care devices could cause harm to patients who were on a tight glycemic control protocol in intensive care units. They simulated the effects of 10%–15%–20% allowable total error of glucometers on insulin dosing decisions and concluded that 20% total error could lead to large insulin dosing errors resulting in hypoglycemia. At 15% total error, models predicted that such large insulin dosing errors would occur very infrequently and at 10% total error the glucometer can be used safely in intensive care units [21].

The advantages of decision analytic models are that, unlike RCTs, such studies are faster, easier and cheaper to perform, and various scenarios and clinical pathways can be modeled in a transparent manner using the best available clinical data about relevant model inputs. They provide the opportunity for examining the combined impact of differences in the analytical and clinical performance of a test with operational factors and costs to estimate clinical and cost-effectiveness. Varying the input parameters of the model allows for comparative analysis of test effectiveness across a range of plausible scenarios. A further advantage is that testing strategies that show promise in modeling studies can be prioritized for clinical trials [15]. Disadvantages are that these analyses usually oversimplify real clinical scenarios and may not capture important aspects of care or patient’s attitude to care plans. Another limitation is that these models are only as good as the input variables. When several sources are used for key input data, the uncertainties and biases of all those studies add up and may result in assumptions that cannot be reproduced or proven in clinical practice. Nevertheless if reliable key input parameters can be obtained, modeling is an accepted and useful tool for assessing the impact of testing on outcomes and could be the most pragmatic way of defining outcome-based APS.

Conclusions

The key driver for setting outcome-based APS is to enable the IVD industry, the laboratory profession and clinicians to develop and use laboratory tests that serve patients' needs best. When such specifications are defined, regardless of the approach used, the following key principles should be kept in mind.

Outcome-based APS:
- should reflect the clinical needs of patients and their carers and should be set at a level that achieves net health benefit for patients at reasonable costs;
- may not be possible to set for all measurands;
- are best suited for tests which are key determinants of subsequent patient management and where the link between testing and outcomes is obvious and strong;
- should be tailored to the purpose, role and significance of the test in a well defined clinical pathway;
- should be based on the contribution of the test to managing the health condition, to avoid setting APS that are too stringent and clinically inappropriate;
- The clinical actions triggered by test results should be well characterized and standardized to maintain a strong, valid and reproducible link between test performance and outcomes.

Defining outcome-based APS may be seen as a complex process but as the examples presented here illustrate, it
is not impossible. A guiding principle is that the effect of changes in APS should be evaluated in terms of the corresponding effects on patient outcome. Even though direct evaluations are difficult, all other forms of setting APS should be weighed against that standard, and regarded as approximations.

By working together to better define the relationship between the analytical performance of tests and health outcomes and determine appropriate specifications, laboratorians, researchers, epidemiologists, clinicians and the IVD industry can aim to improve this essential component of the clinical and cost-effectiveness of laboratory tests.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Financial support: European Federation of Clinical Chemistry and Laboratory Medicine, Test Evaluation Working Group.

Employment or leadership: None declared.

Honorarium: None declared.

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References