Clinica Chimica Acta 427 (2014) 49–57

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/clinchim

Special report

From biomarkers to medical tests: The changing landscape of test evaluation

Andrea R. Horvath [a,b,*], Sarah J. Lord [b,c,l], Andrew StJohn [d], Sverre Sandberg [e], Christa M. Cobbaert [f], Stefan Lorenz [g], Phillip J. Monaghan [h], Wilma D.J. Verhagen-Kamerbeek [i], Christoph Ebert [j], Patrick M.M. Bossuyt [k],

For the Test Evaluation Working Group of the European Federation of Clinical Chemistry Laboratory Medicine

[a] SEALS Department of Clinical Chemistry, Prince of Wales Hospital and School of Medical Sciences, University of New South Wales, Australia
[b] Screening and Test Evaluation Program, School of Public Health, University of Sydney, Australia
[c] National Health and Medical Research Council (NHMRC) Clinical Trials Centre, University of Sydney, Australia
[d] ARC Consulting, Perth, Australia
[e] Laboratory of Clinical Biochemistry, Haukeland University Hospital, and The Norwegian Quality Improvement of Primary Care Laboratories (NOKLUS), Department of Public Health and Primary Health Care, University of Bergen, Norway
[f] Department of Clinical Chemistry and Laboratory Medicine, Leiden University Medical Center, The Netherlands
[g] Abbott Diagnostics, Wiesbaden, Germany
[h] Department of Clinical Biochemistry, The Christie Hospital NHS Foundation Trust, Manchester, UK
[i] Medical and Scientific Affairs, Roche Diagnostics International Ltd., Rotkreuz, Switzerland
[j] Medical and Scientific Affairs, Roche Diagnostics GmbH, Penzberg, Germany
[k] Department of Clinical Epidemiology, Biostatistics & Bioinformatics, Academic Medical Center, University of Amsterdam, The Netherlands
[l] School of Medicine, Sydney, The University of Notre Dame Australia, Australia

ARTICLE INFO

Article history:
Received 22 April 2013
Received in revised form 15 September 2013
Accepted 15 September 2013
Available online 27 September 2013

Keywords:
Biomarker
Clinical effectiveness
Evidence-based laboratory medicine
Medical test

ABSTRACT

Regulators and healthcare payers are increasingly demanding evidence that biomarkers deliver patient benefits to justify their use in clinical practice. Laboratory professionals need to be familiar with these evidence requirements to better engage in biomarker research and decisions about their appropriate use.

This paper by a multidisciplinary group of the European Federation of Clinical Chemistry and Laboratory Medicine describes the pathway of a laboratory assay measuring a biomarker to becoming a medically useful test. We define the key terms, principles and components of the test evaluation process. Unlike previously described linearly staged models, we illustrate how the essential components of analytical and clinical performances, clinical and cost-effectiveness and the broader impact of testing assemble in a dynamic cycle. We highlight the importance of defining clinical goals and how the intended application of the biomarker in the clinical pathway should drive each component of test evaluation. This approach emphasizes the interaction of the different components, and that clinical effectiveness data should be fed back to refine analytical and clinical performances to achieve improved outcomes.

The framework aims to support the understanding of key stakeholders. The laboratory profession needs to strengthen collaboration with industry and experts in evidence-based medicine, regulatory bodies and policy makers for better decisions about the use of new and existing medical tests.

Abbreviations: ACR, albumin:creatinine ratio; ACS, acute coronary syndrome; AMI, acute myocardial infarction; CRP, C-reactive protein; BNP, B-type natriuretic peptide; CE, Conformité Européenne; CK-MB, creatine kinase MB isoform; cTn, cardiac Troponin; EFLM, European Federation of Clinical Chemistry and Laboratory Medicine (formerly abbreviated as EFCC); EU, European Union; FDA, Food and Drug Administration; HbA_{1c}, Hemoglobin A_{1c}; hs-cTn, high sensitivity cardiac Troponin; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; INR, international normalized ratio; IVD, in vitro diagnostics; NGSP, National Glycohemoglobin Standardization Program; NHS, National Health Service in the United Kingdom; NICE, National Institute for Health and Care Excellence; PoCT, point of care testing; RCT, randomized controlled trial.

* Corresponding author at: SEALS North, Department of Clinical Chemistry Level 4, Campus Centre, Prince of Wales Hospital Barker Street, Randwick NSW 2031, Sydney, Australia. Tel.: +61 404 027 843.
E-mail address: rita.horvath@sesiahs.health.nsw.gov.au (A.R. Horvath).

0009-8981/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cca.2013.09.018

1. Introduction

There is an increasing awareness that the introduction of costly new medical interventions, including medical tests, can only be justified if they deliver proportionate benefits to patients. Increased public, media and political awareness of the harms from medical tests has come from debates about the potential for over-diagnosis in asymptomatic patients [1] and concerns about the harms of direct-to-consumer testing [2]. Weaknesses of the current systems to assure the quality and clinical utility of in vitro medical devices (IVDs) have been pointed out [3]. Amidst all this, the regulatory environment for therapeutic and diagnostic technologies is changing rapidly. Revisions of the European directives on medical [4] and in vitro diagnostic devices [5] are being prepared in parallel. The ability of novel medical tests to improve health outcomes is also becoming more central in discussions about their market entry and reimbursement. The increasing requirements for clinical benefits and patient safety mirror public and political pressures for more transparency. These processes are affecting the way novel medical tests and biomarkers are being developed, and are likely to reshape the landscape of medical test evaluation. Laboratory professionals need to be familiar with these evidence requirements to better engage in biomarker research and in clinical and policy decisions about the appropriate use of laboratory tests.

Over the past decade landmark advances have been made to define the types of evidence required to evaluate medical tests and distinguish between the different phases of test evaluation from discovery to assessment of cost-effectiveness [6–10]. There is less guidance, however, about the most efficient approaches to produce this evidence and judging whether it is adequate for proving the clinical effectiveness of biomarkers.

The evaluation of medical tests differs from comparable processes for therapeutic interventions. One of the most important differences is that medical testing rarely improves health outcomes directly. Testing is usually part of a more complex clinical pathway where test results guide treatment decisions, which include a variety of medical actions and processes. All of these shape the final health outcomes for the patients tested. Test evaluation therefore requires the consideration of all the consequences of clinical management decisions that are guided by the test results. An understanding of these more complex concepts for test evaluation is becoming essential for informed decision making by all potential stakeholders.

To help address these issues, the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) has formed a Working Group on Test Evaluation which consists of laboratory professionals, clinical epidemiologists, health technology assessment experts and representatives of the in vitro diagnostics industry. The primary purpose of this working group is to provide key stakeholders, i.e. laboratory professionals, clinicians, researchers, manufacturers, policy makers and purchasers, with guidance and practical tools for assessing the clinical benefits of in vitro medical tests. In this introductory paper, the working group outlines the key principles and defines some of the key components of contemporary approaches to test evaluation, such as analytical performance, clinical performance, clinical effectiveness, cost-effectiveness, and the broader impact of testing on social, psychological, legal, ethical, societal, organizational and other consequences. We additionally present a framework for the evaluation of medical tests that integrates these components into a dynamic process. We illustrate the key principles and components with examples from the literature on Hemoglobin A_{1c} (HbA_{1c}) and cardiac markers, including cardiac Troponins (cTn) and B-type natriuretic peptides (BNP).

2. Key definitions and principles

There is no international consensus on the terminology related to test evaluation and numerous definitions exist in the literature. Table 1 lists a number of alternative terms and illustrates the proposed definitions with examples.

Under the general umbrella term of *medical tests*, which encompasses tests from all clinical disciplines, specialties, or types (laboratory, histopathology, imaging, and others), we define and focus this paper on in vitro medical tests only; yet the key messages outlined here can also be adapted to any other forms of medical tests. We distinguish in vitro medical or laboratory assays and measurement procedures [11] from biomarkers that are measured by these assays [12].

From the regulatory and medical laboratory perspective, *test evaluation* refers to a set of processes which start when a laboratory assay capable of measuring a biomarker with potential application in clinical care becomes available. Ideally, and before the test evaluation process starts, the potential purpose of the new marker is defined to address an unmet clinical need. Laboratory assays can then be developed to measure the marker with this purpose in mind. For example, in the field of cardiac biomarkers, CK-MB and Troponins are considered as tests of myocardial damage, but clinicians have long been waiting for non-invasive markers that can predict myocardial infarction before cell damage happens. An early research finding revealed that endothelial cells are shed from coronary arteries several days to weeks before heart attack. This finding led to the development of a method to measure circulating endothelial cells [13]. Translation of such primary findings usually starts with proof-of-concept studies which explore the association of the disease or condition with the new potential biomarker, usually in diseased versus control patients. Such 'case–control' designs tend to overestimate the clinical performance of a diagnostic assay, as they are designed to test proof of concept. Additional study designs are required for other phases in the medical test evaluation process.

The key principle of medical test evaluation is the fundamental premise that the introduction of any new test should eventually improve health outcomes, or provide other benefits, e.g. reduce costs, or simplify health care delivery without compromising the well-being of patients. Therefore evaluation begins with defining the potential health outcomes (benefits and harms) of introducing the test. Health outcomes should include consequences most relevant to patients. As discussed by Porter, these include survival, sustaining health, achieving recovery, improving functioning and reducing complications [14]. It also includes process outcomes such as reducing delays in time to diagnosis which also have direct patient benefits for reducing anxiety and improving treatment outcomes [14].

New tests should provide added benefit for patients or society over currently existing clinical pathways. *Clinical pathways* (also termed as clinical care pathways or test-treatment pathways) describe the typical processes of care for managing a specific condition in a specific group of patients [15], and provide a map that links testing to health and other outcomes (Table 1). The clinical pathway therefore plays a central role in the test evaluation process. Its description can be supported by information in well accepted best practice guidelines. For example, the National Institute for Health and Care Excellence (NICE) provides interactive clinical pathways supported by existing evidence-based guideline recommendations and tools for implementation (http://pathways.nice.org.uk/). For the assessment and management of suspected acute coronary syndrome the pathway shows the use of ECG and describes the timing and role of Troponin T or Troponin I tests and how they should inform subsequent clinical decisions for management based on test results (Table 1).

Since tests usually do not affect health outcomes directly, one has to define, right at the beginning of the evaluation process, the *purpose* and *role* of the medical test in the clinical pathway and the relevant patient population for each testing application. *Test purpose* describes the intended clinical application of the test and how the test information will be used to improve clinical management in practice (Table 1). Medical tests can be used for diagnosis and prognosis, but also for monitoring, early detection, screening, risk classification, treatment selection, surveillance after treatment, and many more. Within these applications,

the purpose may vary: in diagnosis, the test may be used to 'rule in' or 'rule out' disease, for example.

Sometimes the same laboratory test can be used for different purposes at different thresholds and in different patient populations. HbA_{1c} at 7% (NGSP unit) or 53 mmol/mol (IFCC unit), for example, is used as a monitoring and treatment target in patients with established diagnosis of and treatment for diabetes mellitus; while in patients suspected of having diabetes HbA_{1c} values of ≥6.5% (NGSP unit) or ≥48 mmol/mol (IFCC unit) are now recommended as diagnostic on the basis of the observed relationship with retinopathy [16]. Another example is the high sensitivity cardiac Troponin assay (hs-cTn), which can be used both for diagnosis of acute myocardial infarction (AMI) and for prognosis in non-AMI patients, with different cut-points defined for each purpose [17,18].

The role of a test describes how the test, used for a specific clinical purpose, will be positioned to alter the current clinical pathway (Table 1). A test can be proposed as a *replacement* for an existing test (e.g. cTn replaced CK-MB in the diagnosis of AMI). The new test can also act as a *triage* instrument before the existing test (e.g. BNP before echocardiography for diagnosing congestive heart failure), or as an *add-on* concomitantly with or after the existing test. In pathology, add-on tests include reflex or reflexive tests that are automatically initiated follow-up tests using a predefined algorithm when certain primary test results identify some changes that need to be further clarified or elaborated to achieve a more accurate diagnosis [19]. For example, free T4 is a common automatically generated reflex test when TSH is abnormal in certain thyroid conditions. Another term is "reflective testing" which refers to a discretionary process by which laboratory professionals might add on further tests using their clinical judgment based on a combination of available clinical information and test results [19]. For example, addition of iron studies when the possibility of hemochromatosis is suspected based on clinical information, raised transaminases, ferritin, or a combination of these [19].

Defining the role of the test is crucial to determine the most appropriate strategy and comparator for test evaluation studies. For example, when the new diagnostic test will be used as a 'replacement' test, we are interested in comparing the diagnostic accuracy of the new test versus the existing test (against a common reference standard); while if it will be used as an 'add-on' test we need information to compare the diagnostic accuracy of the combination of tests versus the existing test alone [6].

3. Key components of test evaluation

Within the evaluation process of medical tests, we can distinguish between five essential components. Laboratory professionals are familiar with the assessment of analytical performance (Table 1). This refers to the ability of a laboratory assay to conform to predefined technical specifications [7,20]. Specific elements of analytical performance include analytical sensitivity and specificity, limit of detection and quantitation, measurement range, linearity, metrological traceability, measurement accuracy (imprecision and trueness) and consideration of pre-analytical variables including interferences and cross-reactions [21]. Overall, verification and validation of the analytical performance of measurement procedures are well embedded in the routine quality management and accreditation activities of medical laboratories and in the FDA and CE marking approval processes of in vitro diagnostic medical devices.

A second component is clinical performance, alternatively referred to as clinical validity [8,9]. This refers to the ability of a laboratory assay to detect patients with a particular clinical condition or in a physiological state (Table 1). The Global Harmonization Task Force recommends that clinical performance should be evaluated by demonstrating that the results of the biomarker measurement procedure are well associated with the clinical condition or physiological state in the relevant target population [21]. For a biomarker that is intended to be used for diagnosing a condition, clinical performance can be evaluated in a diagnostic accuracy study. In such studies results of the biomarker are compared to those of the clinical reference standard in a sample of the patient population suspected of having the condition of interest. For example, comparison of hs-cTn concentrations in acute chest pain patients versus the final clinical diagnosis of acute coronary syndrome established by cardiac computed tomography ([22], Table 1). For a prognostic marker, clinical performance is demonstrated by establishing meaningful associations between the marker and the event of interest in observational studies. For example, increased concentrations of hs-cTn and markers of heart failure are associated with 10-year cardiac mortality ([23], Table 1). Our working group holds the view that, similarly to analytical performance specifications, desirable clinical performance criteria can and should be predefined for each clinical purpose proposed for a biomarker. These criteria should be based on actual clinical needs and the expected health outcomes from testing, and should be used to select the appropriate study design for evaluation. For example, desirable clinical performance criteria for a diagnostic cancer biomarker would be that the test is only positive in a specific cancer type in patients and negative in all other cases suspected of having cancer. For a staging biomarker, the biomarker test should correlate with disease stage in those diagnosed with cancer [24]. Desirable clinical performance criteria for a monitoring test would be that the test is able to predict clinically significant events, responds to changes in the condition or treatment, and has a high signal-to-noise ratio (i.e. a single test result is likely to detect a clinically significant change and to differentiate it from background biological and analytical variation) [25]. This approach has important advantages in improving the efficiency of the test evaluation cycle. If evaluation shows the test does not achieve the pre-specified performance criteria, further evaluation should not proceed and other purposes or roles for the test will need to be considered.

A third component is the clinical effectiveness of the test. This refers to the ability of a test to improve health outcomes that are relevant to the individuals being tested (Table 1). Sometimes it is also referred to as the 'clinical utility' of a test. It is widely accepted that individual randomized controlled trials (RCTs) and their incorporation into systematic reviews represent the highest levels of evidence for estimating the clinical effectiveness of interventions, including medical tests. Whilst RCTs are the cornerstone of pharmaceutical trials the difficulties of designing RCTs for medical tests are well documented [26]. Lord et al. suggested that when the exact purpose and role of a test are clearly determined, more practical study designs and alternatives to RCTs may be sufficient to evaluate clinical effectiveness; for example where diagnostic accuracy studies can demonstrate the new test is more sensitive than the old test, and trial evidence about the effectiveness of treatment for the additional cases detected already exists. [27].

The fourth component is cost-effectiveness. The introduction of medical tests could increase or reduce the use of resources. In economic evaluations, changes in costs are evaluated as well as changes in health outcomes. Similar to estimates of effectiveness, the costs of the test and all consequences of testing are compared to the costs of the existing clinical pathway. The benefits are expressed in non-monetary terms related to health, such as symptom-free days, morbidity and mortality avoided, life years gained or quality adjusted life years gained [28]. Costs can be calculated from several perspectives: from that of society at large, from the third-party payer, or from the hospital. The cost-effectiveness ratio is an expression of the change in costs from testing relative to the change in health outcomes after testing.

The final component addresses the broader impact of using the medical test, which refers to all other consequences of testing beyond clinical effectiveness and cost-effectiveness (Table 1). The impact of medical testing covers social, psychological, legal, ethical, societal and organizational consequences of testing. For example, genetic testing for hereditary conditions can have psychological impacts [29] and important ethical implications for sharing of information with family members [30]; the availability of point of care tests or rapid tests can lead to organizational changes in the provision of health care, with shifts

Table 1
Key definitions for in vitro medical test evaluations.

Proposed term	Related terms	Explanation (reference)	Examples[a]
In vitro medical assay	In vitro diagnostic medical device Laboratory assay Measurement procedure	A measurement procedure undertaken on a biological specimen which measures the quantity of the biomarker (see below) intended to be measured; i.e. the measurand [11].	– Two-site immunoenzymatic ("sandwich") assay using electrochemiluminescence detection for cardiac Troponin (cTn) measurement – HbA_{1c} measured by cation exchange chromatography or boronate affinity chromatography or by a latex agglutination immunoassay
Biomarker	Biological marker	A characteristic that is an indicator of normal biological or pathogenic processes, or pharmacologic responses to a therapeutic intervention [12].	– cTn-s are biomarkers of cardiac diseases associated with myocardial ischemia and necrosis – HbA_{1c} is a biomarker of altered glycosylation in hyperglycemic states, such as diabetes mellitus
In vitro medical test	Medical test or testing strategy	In vitro medical tests or testing strategies utilize laboratory assays of biomarkers in a specific clinical context and for a specific clinical purpose (see below), in a specific patient population.	– Serial cTn testing for diagnosing acute coronary syndrome (ACS) in patients with symptoms of acute chest pain – HbA_{1c} as a monitoring test to assess treatment effect in type 1 or type 2 diabetic patients
Clinical pathway	Test-treatment pathway Clinical algorithm Care pathway Critical pathway Care map Guideline	A description of typical processes of care in managing a specific condition in a specific group of patients [15].	– Clinical pathways by NICE in the UK: http://pathways.nice.org.uk/ – NICE clinical pathway for the assessment, diagnosis and immediate management of acute coronary syndrome where patients with chest pain receive ECG and biochemical tests, and further management is determined by test results (http://pathways.nice.org.uk/pathways/acute-coronary-syndrome#path=view%3A/pathways/acute-coronary-syndrome/tests-in-hospital-for-people-with-a-suspected-acute-coronary-syndrome.xml&content=view-node%3Anodes-initial-assessment)
Test purpose	Intended use of test Indication for testing Claim (in the context of manufacturer's claim for the intended use of the test)	Test purpose describes the intended clinical application of the test and how the test information will be used to improve clinical management in practice. Test purpose includes: diagnosis. prognosis, monitoring, early detection, screening, risk classification, treatment selection and surveillance after treatment, etc.	– HbA_{1c} as a diagnostic marker of diabetes mellitus [16] – HbA_{1c} as a monitoring test to assess diabetes control [16] – hs-cTn for diagnosing ACS [17] – hs-cTn as a prognostic marker for cardiovascular morbidity and mortality [18]
Test role	Replacement test Triage test Add-on test (includes Reflex testing and Reflective testing)	Test role describes how the test, used for a specific clinical purpose, will be positioned to alter the existing clinical care pathway in a specific condition or target population [6,19]: – Replacement: When a new test replaces an existing test in the testing pathway. – Triage: When the new test is used before the existing test or testing pathway, and only patients with a particular result on the triage test continue on the testing pathway. – Add-on: When a test is added to an existing test or testing pathway. Includes automatically (reflex testing) or by a laboratory professional (reflective testing) to the existing test or testing pathway, either to help interpreting results of analyses when establishing a diagnosis or to assist patient management [19].	Replacement: – cTn-s replacing CK-MB as a biomarker of myocardial damage; – CRP replacing erythrocyte sedimentation rate as marker of acute inflammation Triage: – Natriuretic peptides before echocardiography for congestive heart failure Add-on: – HbA_{1c} monitoring together with self-monitoring of blood glucose in managing Type 1 diabetes patients – free T4 added on to abnormal TSH results (reflex testing) – Iron studies added on by a laboratory professional when biochemical results, available clinical information, demographic data, and clinical experience or their combinations suggest the possibility of hemochromatosis (reflective testing) [19]

Analytical performance	Analytical validity Technical efficacy	Ability of an in vitro medical assay to conform to predefined quality specifications (e.g. as defined by the Stockholm Conference hierarchy or in clinical guidelines) [7,20]	Universal definition of myocardial infarction recommends that high sensitivity Tn assays must have acceptable imprecision, i.e. ≤10% CV, at the 99th percentile of normal [16]. For example, in an analytical performance study, a hs-cTnT assay had a CV of 9% at 13.5 ng/L [36].
Clinical performance	Clinical validity Test performance Performance or Operating characteristics Test accuracy or diagnostic accuracy Diagnostic accuracy efficacy Test efficacy Prognostic accuracy Net reclassification improvement index	Ability of a biomarker to conform to predefined clinical specifications in detecting patients with a particular clinical condition or in a physiological state (adapted from Ref. [21]).	Diagnostic test: In patients presented to emergency with chest pain and low to intermediate likelihood for ACS, the hs-cTnT assay was compared with a conventional cTnT method and CT angiography as the gold standard for diagnosing ACS. At the optimal hs-cTnT cut point of 8.62 ng/L, sensitivity for ACS was 76% and specificity was 78%, and hs-cTnT above the 99th percentile strongly predicted ACS. Compared with the conventional cTnT method, hsTnT detected 27% more ACS cases [22]. Prognostic test: In elderly patients presenting to primary care with symptoms of heart failure the risk for cardiovascular mortality (adjusted for age, sex, impaired estimated glomerular filtration rate, and anemia) increased 2.5-fold with a plasma NT-proBNP concentration >507 ng/L; 2-fold with hs-cTnT >99th percentile; 3-fold when both biomarkers were elevated [23].
Clinical effectiveness	Clinical utility Clinical usefulness	Ability of a test to improve health outcomes that are relevant to the individual patient (adapted from Ref. [7,8,21])	The clinical effectiveness of BNP testing for diagnosis of heart failure in patients presenting to emergency with acute dyspnea were investigated in RCTs that compared the addition of BNP testing with standard investigations alone followed by routine care. A meta-analysis of these RCTs reported that addition of BNP testing decreased length of hospital stay by ~1 day; possibly reduced admission rates, but did not affect 30-day mortality rates [39].
Cost effectiveness	Efficiency	A cost-effectiveness analysis compares the changes in costs and in health effects of introducing a test, to assess the extent to which the test can be regarded as providing value for money.	Point of care testing (PoCT) in general practice: A cost-effectiveness analysis based on an RCT of nearly 5,000 patients followed up for 18 months in Australian general practices compared the incremental costs and health outcomes associated with a clinical strategy of PoCT for INR, HbA1c, lipids, and albumin:creatinine ratio (ACR) to those of pathology laboratory testing. Under base-case assumptions, PoCT was more cost-effective and effective for ACR than standard pathology. For HbA1c, POCT was more expensive but also more effective than standard pathology with an incremental cost-effectiveness ratio of $40 per patient maintained in the therapeutic range, while INR was more costly but less effective and therefore not cost-effective [31].
Broader impact		Consequences of testing beyond clinical effectiveness and cost-effectiveness (e.g. acceptability, social, psychological, legal, ethical, societal, organizational consequences and other aspects).	Psychological impact of genetic testing for familial hypercholesterolemia: An RCT in a population previously aware of their hypercholesterolemia found that finding a mutation by genetic testing did not reduce patients' perceptions of control over the disease and adherence to risk-reducing behaviors, but did affect their perceptions of how control is most effectively achieved [29].

[a] The examples given in the table are for illustration only.

to local health care units or family practitioners that go beyond testing itself [31].

4. The test evaluation framework

The sequence of activities that describe the pathway of newly discovered biomarkers from the "bench to the bedside" has been described in the literature [8–10,21,24]. A systematic review identified at least 19 different test evaluation frameworks [10]. The evaluation process for medical tests has been described in these frameworks in a linear fashion, similarly to the staged evaluation of drugs.

Only certain elements of test evaluation are addressed before in vitro medical tests are released to the market. Regulatory bodies currently require that manufacturers demonstrate the analytical performance and limited information related to the clinical performance for all in vitro diagnostic medical devices [32]. While well elaborated approaches to the assessment of the clinical performance of diagnostic tests have been published [33], these are not directly applicable to monitoring and to predictive or prognostic markers, which probably represent the majority of laboratory assay applications in practice. In the absence of international standards and the paucity of methodological criteria for the assessment of tests used for a variety of purposes and roles, manufacturers, laboratory professionals, researchers and regulators are equally confused on what studies to do or accept as evidence for the clinical performance and effectiveness of medical tests. There is no consensus on the terminology or on any single approach for evaluation. In particular, there is a continuing debate about the relative importance and appropriate sequence of the five key elements in the test assessment process, how they should relate and link to one another and what hierarchy such a multiphase model should follow.

To address these issues we describe the test evaluation process in a way that incorporates the definitions described earlier and emphasizes the importance of the relationship of the five key evaluation components. We discuss departures from the traditional linear process through illustrative examples and, as a consequence, the need for a consensus on a more cyclical approach in order to improve the selection of the most appropriate study designs that provide best evidence for medical and policy decisions.

Once early research studies identify a link between the new potential biomarker and the disease or condition, a laboratory assay is developed that is capable of measuring the biomarker. Before further evaluation of the discriminatory power of the new laboratory assay on a large number of patient samples, it is recommended that the context in which the new test is to be used in practice is addressed and that the potential purpose and role of the new biomarker are defined. There is little point in investing a lot of effort into medical test development and evaluation unless the clinical pathway from testing to treatment decisions and to health or other outcomes is clearly mapped and well understood, and there is an indication that the new biomarker will offer incremental benefits over existing clinical pathways. This is essential for formulating appropriate research questions that can be addressed by well-designed studies in the subsequent process. This can be done by drawing a clinical algorithm or decision tree modeling the typical scenario for the potential application of the biomarker [34]. As the test evaluation process described below is cyclical, the originally intended purpose and role of the test can be redefined if an alternative use for the test becomes apparent during the investigation process.

By mapping the intended application and outcomes of the new test, clinical needs can be further refined. This should ideally dictate the *quality requirements for analytical performance* of the assay(s) developed for the intended use of the biomarker. Analytical performance goals of the same assay can – and probably should – be different for diagnostic or monitoring applications. Analytical performance criteria can be defined by using the Stockholm hierarchy of preferred methods [20]. In this consensus-based hierarchy, analytical performance goals based on health outcomes and clinical decision making represent the highest levels, and criteria based on biological variation, or clinicians' or experts' opinion are positioned lower. For example, the health outcome- and clinical decision-driven analytical performance goals for cardiac Troponins for diagnosing AMI are agreed on the basis of how many diagnostic misclassifications of AMI are acceptable or tolerated by clinicians. A cTn assay, assuming to have zero bias, and a CV of $<10\%$ or $<6\%$ at the 99th percentile decision limit can result in a misclassification rate of 1% and 0.5%, respectively [17,35]. Systematic errors in analysis can affect diagnostic accuracy even more. The diabetes mellitus guideline defines analytical performance goals for glucose measurements on the basis of biological variation. To avoid misdiagnosis of patients, the goal for glucose analysis is to minimize total analytical error, and methods should be without measurable bias. This translates to goals for analytical imprecision of $\leq 2.9\%$, bias $\leq 2.2\%$, and a total error $\leq 6.9\%$ [16].

Analytical performance studies, as described earlier, verify how the laboratory assay developed for the biomarker meets these predefined criteria (for example see Table 1 and [36]). It is worth emphasizing that the originally defined purpose or the role of testing might change as the analytical performance of an assay is improved during the manufacturer's assay development process. For example, HbA_{1c} was originally developed as a monitoring test for diabetes and now, with improved analytical performance after standardization, it is recommended as a diagnostic test and is now also used as part of the definition of the disease (Table 1 and [16]). Likewise, Troponin was primarily developed as a biomarker of AMI. Earlier generations of this biomarker had low analytical sensitivity; they detected only more extensive myocardial injury and did not allow the measurement of this marker in asymptomatic and healthy individuals. With the development of assays with higher analytical sensitivity, their analytical and clinical performance has been re-evaluated [36], and the universal definition of AMI has been revised. The rise and/or fall of hs-cTn with at least one value above the 99th percentile of the reference population, in association with signs and symptoms of myocardial ischemia, now define AMI [17]. The new hs-cTn test is now also utilized as a potential prognostic and risk stratification marker of cardiovascular events as non-AMI patients who have hs-cTn between the limit of detection and the 99th percentile at presentation have been shown to have increased risk of cardiac death or AMI over the subsequent 2 years (Table 1) [18,37]. This highlights how improvements in analytical performance are driven by clinical needs and the clinical pathway, and reciprocally, how improved analytical performance can lead to new definitions of disease and clinical pathways.

The intended use of the test and the analytical performance of the assay determine the investigation of the clinical performance of the new biomarker. If the assay is intended to be used for diagnosis, clinical performance is best investigated in a diagnostic accuracy study, and expressed as (changes in) diagnostic sensitivity and specificity or other accuracy statistics. If the assay is proposed as a prognostic marker, its performance for discriminating between high and low risk patients is evaluated in an observational study, expressed as prognostic accuracy, or by reporting risk reclassification, for example using the net reclassification improvement index [38]. The actual clinical performance of the assay may modify existing clinical pathways (e.g. by limiting the use of the assay to certain subgroups of patients; or only enabling the use as a rule out but not as a rule in diagnostic test; or only as a monitoring but not as a diagnostic test). Insufficient clinical performance may also call for further improvement of analytical performance before the assay becomes suitable for implementation into clinical practice.

Once the clinical performance of the biomarker is demonstrated for the well-defined clinical purpose, the actual clinical effectiveness of the test can be further explored. This includes demonstration of the benefits and harms related to testing to the individual patient, relative to current best practice. Examples include trials that have randomized patients with suspected heart failure to BNP testing plus routine care versus routine care alone ([39], Table 1). As pointed out earlier, these RCT designs

for medical testing are not always efficient, and RCTs of the complete clinical pathway from testing to treatment are not always practical or even necessary [26,40]. Our working group takes the view that considerations of clinical effectiveness should be part of the pre-market test evaluation process. Necessary conditions for the intended application in the clinical pathway (i.e., purpose and role of testing) in the targeted patient population, in terms of required levels of analytical and clinical performance, should definitely be demonstrated before a new biomarker can be safely released. The possibility of harms from testing should also be considered. An example is the use of PSA as a screening test, where the harms from screening low risk symptom free men may outweigh the previously perceived benefits [41].

The benefit of a cyclical test evaluation approach is that manufacturers' claimed purpose for the assay and its position in the clinical pathway can be changed as more evidence is published about clinical effectiveness. Clinical effectiveness data on a given test application in a certain patient population may not only modify the clinical pathway or alter the role of the test in the clinical pathway but also may inform the design of further clinical performance studies. For example, D-dimer may not be clinically effective as a stand-alone diagnostic marker for pulmonary embolism. However, as an add-on test in a diagnostic protocol to rule out pulmonary embolism in low risk patients could improve the clinical and cost-effectiveness of the existing clinical pathway by limiting the number of further more invasive, or harmful, or costly diagnostic procedures [42].

If the test or testing strategy is found to be effective in a given clinical indication, estimation of cost-effectiveness will be key for reimbursement decisions. This may drive the need for comparative clinical studies to better estimate the size of effects, and for economic models to capture long term consequences of testing and potential uncertainty for both costs and effects. An example of a cost-effectiveness study is illustrated in Table 1 [31]. Cost-effectiveness data may again alter the clinical pathway and limit the use of the test to higher risk patient groups in whom the benefits from testing offer more value.

Test evaluation should not finish with the introduction of the new biomarker into clinical practice. The broader impact and consequences of testing, beyond clinical and cost-effectiveness, can be investigated when the medical test is more widely and routinely utilized in various clinical settings. In some cases one can build on prior experience with the introduction of similar tests for other clinical purposes. In other cases, the actual use of the test may change after its introduction to the market. Post-market surveillance of the performance of medical tests is an often neglected, but very important activity which is usually based on data from quality improvement and clinical audit projects. This is especially relevant when tests eventually become utilized for additional clinical indications that have not been formally evaluated; or regular use of the test reveals so far unknown analytical variation, or interferences, or when lot-to-lot variations impact both analytical and subsequent clinical test performances. Such feedback from laboratories or clinicians could be invaluable to manufacturers and researchers on how to improve the analytical and clinical performance of their assays in order to make them a clinically more effective biomarker and a medically more useful test in their intended or newly introduced application.

These examples demonstrate that test evaluation is a dynamic process with many inter-related elements as described above rather than proceeding in rigid sequential phases. Our Working Group believes, based on the observation also made by others [8,10], that a rigidly phased, linear sequential model is overly simplistic and that biomarker evaluation should rather operate as a cycle. We propose that the five key elements should be tightly integrated with each other along the key principles defined earlier, and linked to the test's purpose and role within a specified clinical pathway in order to develop a systematic approach for test evaluation. We believe that a more unifying framework is needed not only to reduce the confusion of stakeholders, but also to allow for an integrated system approach, whereby larger and more costly studies are only initiated if there is prior evidence to support the test's worth; a principle that also provides the ethical justification for inviting patients to participate in clinical studies. We therefore illustrate the journey of a laboratory assay measuring a biomarker in becoming a medical test or testing strategy as a dynamic cycle driven by the clinical pathway (Fig. 1). In this cycle, we visualize each component as being interconnected with and influencing one another. In this model the clinical pathway, describing the purpose and role of testing in a specific group of patients and linking testing to health or other outcomes, is at the core of the test evaluation process and influences all other components in a reciprocal fashion.

5. Conclusions

There is an increasing awareness that the introduction of new technology can lead to increased costs which are often not in direct proportion to the benefits for patients. This is particularly the case where new technologies offer only marginally better alternatives to existing processes. Healthcare funders and regulators are indicating that such marginal improvements are less likely to be funded or reimbursed in the future. Professional, public and government organizations have criticized the current FDA approval and CE marking processes for failing to provide patients with safer, higher quality and more effective devices [3,32,43]. While it is clear that a more integrated regulatory framework is needed, it must be one that does not stifle innovation in the medical and in vitro diagnostic medical device industry. Revised policies and procedures of IVD regulatory bodies in the US and the European Commission discuss the need of pre-market presentation of high quality evidence and the post-market surveillance of the clinical safety and performance of in vitro diagnostic medical devices [32].

Researchers, manufacturers of IVDs and laboratory professionals face a major challenge to provide evidence that testing improves actual

Fig. 1. Cyclical framework for the evaluation of in vitro medical tests. This framework illustrates that the key components of the test evaluation process are driven by the purpose and role of using a test in the clinical pathway. Reciprocally, the key test evaluation elements may influence or modify existing clinical pathways. The outer circle linking the various elements of the test evaluation cycle highlights the interplay between the various components and how e.g. analytical performance may impact clinical performance and vice versa; how clinical performance or effectiveness of a test may call for improved analytical performance and sets new analytical goals for improving the clinical and cost-effectiveness of the test-treatment pathway. The dynamic relationship of the various key components is further explained in the text and is illustrated with examples.

health outcomes, in order to meet regulatory approval and achieve reimbursement requirements. Statutory regulators and consumers are also starting to require that evidence relating to test performance (or the lack thereof) is now also placed in the public domain. Therefore clear guidance and a multidisciplinary, more responsive and proportionate risk assessment during pre-market approval and post-market surveillance of tests are needed to ensure patient safety. After initial discovery of potential biomarkers, careful consideration should be given to their intended use and the consequences and health outcomes of testing in clinical practice. No new test should be subjected to tedious evaluation and released to the market if it is unlikely, or if it cannot convincingly be substantiated, that using the test will result in changes in clinical actions and health outcomes.

Clear guidelines on pre-market test evaluation methods and standardized performance measures for post-market assessments are needed for improving the effectiveness and cost-effectiveness of laboratory practice. Feedback on the clinical effectiveness of new biomarkers needs to be communicated to scientists, industry, clinicians and guideline teams to enable them to refine analytical and clinical performance, ask new research questions, design new biomarker evaluation studies, and formulate better evidence-based recommendations that are more responsive to real clinical and patients needs. Collaboration between experts in laboratory medicine, epidemiology, evidence-based medicine and industry needs to continue. This collaboration could be extended to regulatory bodies and policy makers. This would allow collective and informed decisions about the appropriate adoption of new or existing medical tests and testing practices.

Acknowledgments

The Working Group wishes to thank the European Federation of Clinical Chemistry and Laboratory Medicine for supporting its work. Roche Diagnostics is thanked for supporting the Working Group with an independent educational grant to EFLM.

Role of sponsor

Neither the funding organization (EFLM) nor the independent educational support by Roche Diagnostics influenced the content of this paper. This publication reflects the collective view of the working group. The authors include representatives from Roche Diagnostics and Abbott Diagnostics as they provided important insights from the IVD industry. The IVD industry is a key stakeholder in the research and development of new biomarker assays and therefore their input was essential to providing a balanced view of all professions involved in the field.

References

[1] Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. BMJ 2012;344:e3502.
[2] Lovett KM, Liang BA. Direct-to-consumer disease screening with finger-stick testing: online patient safety risks. Clin Chem 2012;58(7):1091–3.
[3] Wilcox M. Assuring the quality of diagnostic tests. How do we know that they do what we think they do? BMJ 2013;346:f836.
[4] Directive 2007/47/EC of the European Parliament and of the Council of 5 September 2007 amending Council Directive 90/385/EEC on the approximation of the laws of the Member States relating to active implantable medical devices, Council Directive 93/42/EEC concerning medical devices and Directive 98/8/EC concerning the placing of biocidal products on the market. Off J Eur Union 21.9.2007;L247:21–55 [Available at: http://ec.europa.eu/consumers/sectors/medical-devices/files/revision_docs/2007-47-en_en.pdf (Last accessed on 30 March 2013)].
[5] Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostic medical devices. Off J Eur Communities December 12, 1998;L331:1–37 [Available at: http://eurotechcertification.com/docs/ivd.pdf (Last accessed on 30 March 2013)].
[6] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089–92.
[7] Methods guide for medical test reviews. AHRQ publication no. 12–EC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012 188 [Available at: http://www.effectivehealthcare.ahrq.gov/ehc/products/246/558/Methods-Guide-for-Medical-Test-Reviews_Full-Guide_20120530.pdf (Last accessed on 30 March 2013) Also published as a special supplement to the J Gen Int Med 2012;27(Suppl1):1–101].
[8] Haddow JE, Palomaki GE. ACCE: a model process for evaluating data on emerging genetic tests. In: Khoury M, Little J, Burke W, editors. Human genome epidemiology: a scientific foundation for using genetic information to improve health and prevent disease. Oxford: Oxford University Press; 2003. p. 217–33.
[9] Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. Genet Med 2009;11(1):3–14.
[10] Lijmer JG, Leeflang M, Bossuyt PMM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29:E13–21.
[11] International vocabulary of metrology — basic and general concepts and associated terms (VIM) 3rd edition; 2008 version with minor corrections. JCGM 2012;200 [(E/F) Available at: http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf (Last accessed on 30 March 2013)].
[12] Atkinson AJ, Colburn WA, DeGruttola VG, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001;69:89–95.
[13] Damani S, Bacconi A, Libiger O, et al. Characterization of circulating endothelial cells in acute myocardial infarction. Sci Transl Med 2012;4:126ra33.
[14] Porter ME. What is value in health care? N Engl J Med 2010;363(26):2477–81.
[15] Kinsman L, Rotter T, James E, Snow P, Willis J. What is a clinical pathway? Development of a definition to inform the debate. BMC Med 2010;8:31 [Available at: http://www.biomedcentral.com/1741-7015/8/31 (Last accessed on 30 March 2013)].
[16] Sacks DB, Arnold M, Bakris GL, et al. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. Clin Chem 2011;57(6):e1-47.
[17] Thygesen K, Alpert JS, Jaffe AS, Simoons ML, Chaitman BR, White HD: the writing group on behalf of the Joint ESC/ACCF/AHA/WHF task force for the Universal Definition of Myocardial Infarction. Third universal definition of myocardial infarction. Eur Heart J 2012;33:2551–67.
[18] Aldous SJ, Florkowski CM, Crozier IG, George P, Mackay R, Than M. High sensitivity troponin outperforms contemporary assays in predicting major adverse cardiac events up to two years in patients with chest pain. Ann Clin Biochem 2011;48:249–55.
[19] Paterson JR, Paterson R. Reflective testing: how useful is the practice of adding on tests by laboratory clinicians? J Clin Pathol 2004;57(3):273–5.
[20] Fraser CG, Kallner A, Kenny D, Petersen PH. Introduction: strategies to set global quality specifications in laboratory medicine. Scand J Clin Lab Invest 1999;59(7):477–8.
[21] Study Group 5 of the Global Harmonization Task Force. Clinical evidence for IVD medical devices — key definitions and concepts. GHTF/SG5/N6; 2012 1–11 [Available at: http://www.imdrf.org/docs/ghtf/final/sg5/technical-docs/ghtf-sg5-n6-2012-clinical-evidence-ivd-medical-devices-121102.pdf (Last accessed on 30 March 2013)].
[22] Januzzi JL, Bamberg F, Lee H, et al. High-sensitivity Troponin T concentrations in acute chest pain patients evaluated with cardiac computed tomography. Circulation 2010;121:1227–34.
[23] Alehagen U, Dahlström U, Rehfeld JF, Goetze JP. Prognostic assessment of elderly patients with symptoms of heart failure by combining high-sensitivity Troponin T and N-Terminal Pro-B-type Natriuretic Peptide measurements. Clin Chem 2010;56(11):1718–24.
[24] Pavlou MP, Diamandis EP, Blasutig IM. The long journey of cancer biomarkers from the bench to the clinic. Clin Chem 2013;59(1):147–57.
[25] Mant D. A framework for developing and evaluating a monitoring strategy. In: Glasziou P, Aronson J, Irwig L, editors. Evidence-based medical monitoring: from principles to practice. Oxford: Blackwell Publishing; 2007. p. 15–30.
[26] Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844–7.
[27] Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009;29(5):E1-12.
[28] NICE glossary. NHS National Institute for Health and Care Excellence. Available at: http://www.nice.org.uk/website/glossary/glossary.jsp?alpha=C . [Last accessed on 30 March 2013].
[29] Marteau T, Senior V, Humphries SE, et al. Psychological impact of genetic testing for familial hypercholesterolemia within a previously aware population: a randomized controlled trial. Am J Med Genet A 2004;128A:285–93.
[30] Offit K, Groeger E, Turner S, Wadsworth EA, Weiser MA. The "duty to warn" a patient's family members about hereditary disease risks. JAMA 2004;292(12):1469–73.
[31] Laurence CO, Moss JR, Briggs NE, Beilby JJ, for PoCT Trial Management Group. The cost-effectiveness of point of care testing in a general practice setting: results from a randomised controlled trial. BMC Health Serv Res 2010;10:165 [Available at: http://www.biomedcentral.com/1472-6963/10/165 (Last accessed on 30 March 2013)].
[32] National Research Council. Medical devices and the public's health: the FDA 510(k) clearance process at 35 years. Washington DC: The National Academies Press; 2011 1–280 [Available at: http://www.nap.edu/catalog.php?record_id=13150 (Last accessed on 30 March 2013)].
[33] Bossuyt PM, Reitsma JB, Bruns DE, et al. Standards for reporting of diagnostic accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49(1):7–18 [Available at: www.stard-statement.org (Last accessed 30 March 2013)].
[34] Samson D, Schoelles KM. Chapter 2: medical tests guidance (2) developing the topic and structuring systematic reviews of medical tests: utility of PICOTS, analytic frameworks, decision trees, and other frameworks. J Gen Intern Med 2012;27(Suppl. 1):

S11– [Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3364354/ (Last accessed on 30 March 2013)].
[35] Sheehan P, Blennerhassett J, Vasikaran SD. Decision limit for Troponin I and assay performance. Ann Clin Biochem 2002;39:231–6.
[36] Giannitsis E, Kurz K, Hallermayer K, Jarausch J, Jaffe AS, Katus HA. Analytical validation of a high-sensitivity cardiac Troponin T assay. Clin Chem 2010;56(2): 254–61.
[37] Vasikaran SD, Macdonald SPJ, Sikaris KA. Editorial: high-sensitivity cardiac troponin assays for risk stratification and for the diagnosis of acute myocardial infarction. Ann Clin Biochem 2012;49:209–10.
[38] Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128–38.
[39] Lam LL, Cameron PA, Schneider HG, Abramson MJ, Muller C, Krum H. Meta-analysis: effect of B-type natriuretic peptide testing on clinical outcomes in patients with acute dyspnea in the emergency setting. Ann Intern Med 2010;153:728–35.
[40] Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850–5.
[41] Chou R, Croswell JM, Dana T, et al. Screening for prostate cancer: a review of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2011;155:762–71.
[42] Geersing G, Erkens PMG, Lucassen WAM, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study. BMJ 2012;345:e6564.
[43] Horton R. Offline: the scandal of device regulation in the UK. Lancet 2012;379:204.