Translating Laboratory Tests into Clinical Practice: A Conceptual Framework

Michael Nagler[1]

[1] University Institute of Clinical Chemistry, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

Hämostaseologie 2020;40:420–429.

Address for correspondence Michael Nagler, MD, PhD, MSc, University Institute of Clinical Chemistry, Inselspital, Bern University Hospital, 3010 Bern, Switzerland (e-mail: michael.nagler@insel.ch).

Abstract

The use of laboratory biomarkers in clinical practice is rapidly increasing. Laboratory tests are, however, rarely evaluated adequately before implementation, and the utility of many tests is essentially unclear. An important reason for this knowledge gap is that a comprehensive and generally accepted methodological framework supporting evaluation studies is essentially lacking. Researchers, clinicians, and decision-makers are often not aware of the methodological tools available and face problems with the appraisal of a test's utility. With the present article, I aim to summarize current concepts and methodological tools and propose a framework for a phased approach that could be used in future evaluation projects. Future research will refine this suggested framework by identifying problems in current evaluation projects, specifying methodological criteria for all phases, as well as developing advanced methodological tools.

Keywords

- biomarkers/analysis
- predictive value of tests
- laboratory methods
- laboratory tests

Introduction: What Are Laboratory Tests for?

Laboratory tests are part of the clinical process rather than a distinct area of care.[1,2] Consequently, tests must be considered as a health care intervention.[3,4] This holds for laboratory methods in the field of thrombosis and hemostasis as for any other test in medicine. Consequently, the ultimate aim of laboratory testing is to improve health outcomes. Improved outcomes benefit patients as well as the caregiver team, the provider organization, the health care insurance, as well as society as a whole.[2,5] This is an enormous task that clinicians and laboratory staff accept to develop the future of laboratory medicine. If we neglect one or more of these areas of responsibility, our work will be questioned by patients, caregivers, health care organizations, and society.

Optimized laboratory tests can have a significant beneficial effect on clinical decision making as well as on health care processes. Asymptomatic patients would be screened effectively, diagnoses made solidly, reliable predictions about future clinical events could be made, and the effects of treatments are monitored sensitively. How can we achieve this? As Christopher P. Price and Robert H. Christensen put it in the preface of their excellent book *Evidence-based laboratory medicine*: "We have to ask (1) the *right* clinical question, (2) using the *right* test, (3) in the *right* patient, (4) at the *right* time, (5) with the analysis using the *right* result, (6) which yields the *right* interpretation, (7) with the *right* decision being made, and (8) the *right* action being taken, (9) at the *right* cost."[2] The task of *academic laboratory medicine* is to generate knowledge to decide what *right* is.

To date, we have done this job insufficiently. It appears that most laboratory tests implemented in clinical practice have not been adequately evaluated and that current biomarker research does often not address unmet clinical needs.[5,6] International standards are scarce, and little research has been done on developing methodology.[7] Hence, researchers are often not aware of the various methodological tools available.

With the present article, I summarize current problems as well as concepts and methodological tools for evaluating laboratory tests and propose a framework to be used in future evaluation projects.

received
March 14, 2020
accepted after revision
August 25, 2020

DOI https://doi.org/10.1055/a-1227-8008.
ISSN 0720-9355.

Current Problems

The use of biomarkers in clinical practice and scientific inquiry is rapidly increasing due to new technologies and the potential associated with precision medicine. The implementation of these biomarkers, defined as the process of putting to use in routine clinical practice,[8] is often done without appropriate evaluation studies.[7,9–12] Implementing and applying medical tests before adequate evaluation might lead not only to high health care costs but may also harm patients and healthy individuals through unnecessary labeling, incorrect diagnoses, delays in starting appropriate treatment, or hazardous therapy.[13,14] An overview of potential risks of applying laboratory tests to patients is given in ►Table 1. As a consequence, overdiagnosis and overtreatment are regarded as a significant threat to human health and represent a significant contributor to health care costs.[13] Data from a large number of studies using a variety of study designs suggest that the incidence of diagnostic error is unacceptably high.[15–17] Recognition of this problem has led to international initiatives such as the "Choosing Wisely" campaign[18] or the IFCC Task Force on the Impact of Laboratory Medicine on Clinical Management and Outcomes.[19]

The following examples illustrate some of the problems. Sensitivity and specificity are often used as fixed measures of test performance rather than a description of test behavior under specific circumstances (such as prevalence), and little attention is paid to issues of validity, variability, applicability, and precision.[7,20] Sample size calculations are rarely employed in diagnostic accuracy studies,[21] and a range of methodological shortcomings can result in biased estimates.[22–25] Reproducibility and consistency of measurements have not been evaluated adequately in a relevant proportion of laboratory tests.[9,10,12,26] The application of sensitivity, specificity, likelihood ratios, and Bayes' theorem to diagnostic reasoning has limitations because these measures vary between subgroups of patients.[27–29] Also, studies examining the various outcomes associated with the use of laboratory tests are lacking in most cases.[30] In general, the majority of factors known to affect the utility of laboratory measurements have not been studied adequately, and this is also the case of laboratory tests used in the area of thrombosis and hemostasis.[10] This is a major challenge for scientists, physicians, policymakers, and funding agencies.

What are the reasons for this evaluation gap? This knowledge void has most probably arisen from the absence of a *comprehensive conceptual framework* and a generally accepted, standardized approach to research. Researchers and clinicians are often not aware of the methodological tools available. Summarizing the evidence from existing studies is also difficult because generally accepted criteria for completeness and methodological quality are only available regarding diagnostic accuracy studies. Also, current methodological tools have limited applicability to diagnostic problems where an appropriate reference standard test is absent or to very rare diseases.

Concepts of Evidence-Based Laboratory Medicine

What Does Evidence-Based Laboratory Medicine Mean?

The concept of evidence-based medicine (EBM) was introduced in the 1990s to promote evidence from clinical research as the primary approach to clinical decision making.[31] EBM was a significant advance because it has provided methodological tools to allow clinicians to identify relevant studies, critically appraise the literature, and apply the findings to their

Table 1 Potential risks of applying laboratory tests to patients

Risk	Description
Direct adverse event	Blood drawing might lead to hematoma, syncope, arterial puncture, and/or thrombus formation
False diagnosis	False-positive results may result in incorrect diagnoses and application of unnecessary and/or risky treatments
Rejection of correct diagnosis	False-negative results may lead to rejection of correct diagnoses and unnecessary delays in starting appropriate treatment
Initiation of additional investigations	Uncertainty in interpretation of positive or negative test results may result in further investigations
Withdrawal of treatments	False-positive results may lead to suspicion of certain diseases, which constitutes a contraindication for treatment of other diseases
Increased costs	False-positive results may lead to additional investigations and/or treatments with a relevant increase of costs
Adverse emotional effects	Receiving a test result may have a lasting impact on mental health, increase anxiety, stress, and/or lead to depression
Adverse social effects	Results of medical tests may affect relationships and social interactions
Adverse cognitive effects	Receiving a test result influences patients' beliefs, perceptions and understanding of their condition that may affect patients' adaptive behavior
Adverse behavioral effects	Test results alter risk perceptions and anxiety, which may influence patients' behavior, for example, with regard to adherence with follow-ups, investigations, and treatments as well as preventive lifestyle

Source: Adapted from Nagler.[10]

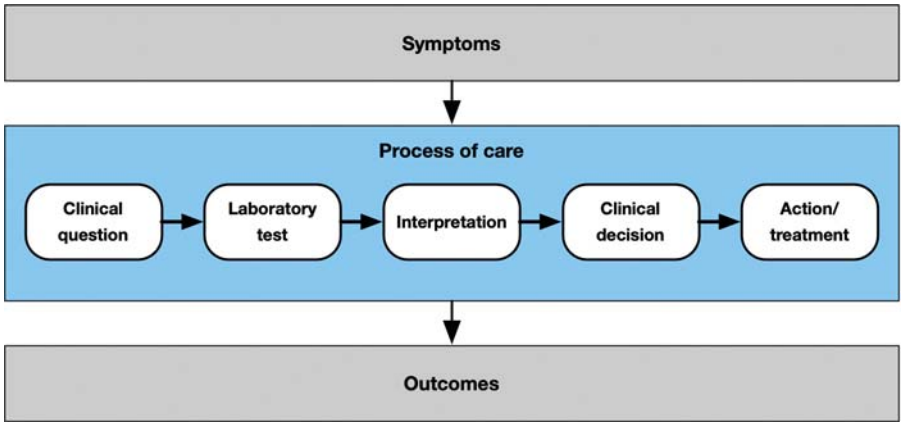

Fig. 1 Process of decision making supported by laboratory tests embedded in the process of care.

clinical context.[32] Yet, EBM emphasizes patient values and preferences as essential contributors to decision making.[33] Applying EBM to diagnostic processes means asking clinical questions as well as probabilistic reasoning. Both issues are discussed later, and a comprehensive overview of evidence-based laboratory medicine is given elsewhere.[3]

Applying Clinical Questions

If laboratory medicine is to support clinical decision making, and if clinical decision making is to be based on rational arguments rather than intuition, then a precise clinical question must be raised.[32] The question should capture the clinical problem that arises in the process of care, and there must also be evidence that the laboratory test can answer the question posed.[3] Any given result from a laboratory test can only have an impact on clinical decision making if the clinical question is clearly defined upfront. This holds not only for individual patients and particular clinical problems but also for the evaluation and implementation of new laboratory tests into clinical practice. Of course, a precise question is essential, but not a sufficient requirement for a laboratory test to be useful. ►Fig. 1 illustrates the process of decision making using laboratory tests embedded in the process of care. Broad categories of clinical questions that can be addressed using laboratory tests are displayed in ►Table 2.

How should we formulate clinical questions appropriately? Drawing from the principles of *EBM* in general and with a view on diagnostic accuracy studies, a clinical question should include (1) the patient population to which the respective laboratory test is applied (*population*), (2) the diagnostic test under investigation (*intervention*), (3) the reference (gold) standard test (*comparator*), and (4) a measure of diagnostic performance (*outcome*). This approach is generally described as the *PICO* method.[34] Of course, this principle of questioning can be adapted to fit studies in other areas of research as well as decision making for individual patients.

To give an example, I would like to sketch a scenario relating to prognosis. *The head of vascular medicine calls the laboratory manager and asks for the implementation of a new "ThrombClot" device. He heard at a scientific meeting that this laboratory assay could predict recurrent venous thromboembolism better than the current standard.* Based on this summary, the following questions might be raised: In adult patients with a recent deep vein thrombosis or pulmonary embolism (*population*), does determination of the "ThrombClot" assay in addition to the established "unprovoked index event" criterion (*intervention*) result in a more accurate prediction of recurrent venous thromboembolic events per 100 patient-years (*outcome*) than using the "unprovoked" criterion alone (*comparator*)?

The Diagnostic Process

The diagnostic process will be illustrated as an example of probabilistic reasoning. Following history taking and physical examination, clinicians develop a list of potential diagnoses.[35] Probabilistic reasoning then seeks to estimate the probability of each diagnosis and to adjust the probability as new information such as laboratory test results becomes available.[36] The probability of the presence of a specific disease is called the

Table 2 Main clinical questions to be addressed by laboratory tests in the process of care

Clinical problem	Associated question	Possible actions	Clinical outcomes
Diagnostic	Does my patient have the disease?	Treat, test further, wait	Improvement or deterioration
Prognostic	Will my patient suffer an adverse event?	Start or stop treatment and decide on treatment details	Improvement or deterioration
Monitoring	Is the treatment effective and safe?	Intensify or reduce treatment	Improvement or deterioration

Notes: Diagnostic problems also apply to screening programs. Prognostic information is used in risk stratification. A combination of a diagnostic and a monitoring problem appears in critically ill patients: the extend of physiologic derangement is looked for (*diagnostic problem*) and the treatment is adapted closely (*monitoring*).

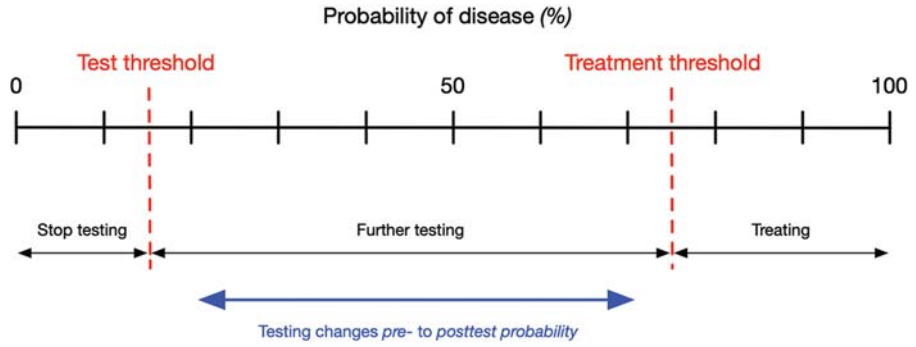

Fig. 2 Changing probabilities as well as test and treatment thresholds in the diagnostic process.

pre-test probability, and it corresponds to the proportion of patients who suffer from the disease among all patients with the same presentation.[32] As estimation of the *pre-test probability* is generally done intuitively, it is prone to error. Intuition is strongly influenced by recent or previous dramatic events, and clinicians may have limited experience with certain diseases. The probability of the disease after new information is incorporated into the assessment is called the *post-test probability*. The extent to which a laboratory test result changes the probability can be expressed by *likelihood ratios*, which are estimated from diagnostic accuracy studies. ►Fig. 2 illustrates the diagnostic process from *pre-* to *post-test probabilities* and illustrates *test and treatment thresholds*. If the *post-test probability* is above a certain threshold, the disease is regarded as sufficiently likely, and treatment is started. If the *post-test probability* is below another threshold, the disease is unlikely, and any pre-test interventions that may have been initiated are stopped. Additional testing will be performed if the *post-test probability* is between the test and treatment threshold. Informing clinicians clearly as to how test results change the probability of a disease would be anticipated to improve health care processes considerably.

A Conceptual Framework for the Implementation of Laboratory Tests

A Phased Approach

Building on the work of several others[3,5,19,30,37] and existing guidance for diagnostic accuracy studies (STARD,[38] QUADAS[39]) and drawing from our previous study publications, I propose a conceptual framework to be used in future projects evaluating laboratory tests in thrombosis and hemostasis as well as in other disciplines (►Fig. 3). Our proposed phased approach has several essential advantages. In particular, the completeness and methodological quality of evaluation studies can be defined and monitored more easily. A full list of

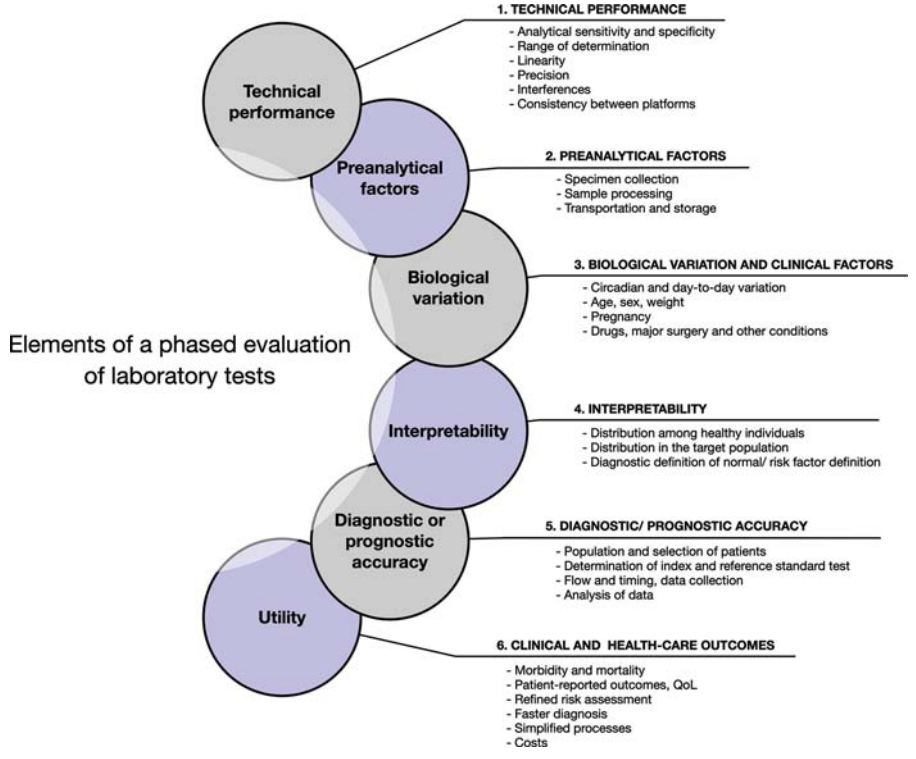

Fig. 3 A conceptual framework for the implementation of laboratory test.

Table 3 Arguments for a phased approach to laboratory test evaluation

Arguments	Rational
Completeness	All test characteristics are assigned to a particular study phase and incomplete evaluation can be identified easily
Structure	Essential characteristics are determined first and studies of a particular phase take results of the previous phase into account
Quality	Aims and methodological requirements can clearly be defined for every study phase
Synthesis	Review of the literature for assessing the value of a test will be simplified
Standards	Scientific societies and authorities can define minimal methodological requirements for every phase of the evaluation process
Costs	Evaluation can be stopped early in case of inadequate study results and cost associated with more expensive later stages will be saved

arguments for a phased approach is given in ►**Table 3**. The following sections describe the components of the phased conceptual framework.

Phase 1: Technical Performance

There is consensus that studies of *analytical accuracy* and *technical performance* should precede more elaborate studies.[40] Once a biomarker is identified, an analytical technique shall be selected that can determine the marker accurately while meeting the application needs.[5,10] If this selection and evaluation procedure is done improperly, subsequent diagnostic accuracy studies might have inaccurate or even biased results. Besides, imprecise and inaccurate results, as well as inconvenient application characteristics, hamper the implementation in routine practice.[40] *Analytical sensitivity* and *analytical specificity* are essential characteristics to determine. (Does the test measure what it aims to measure?) The *detection method* will be defined as well as methods of *standardization* and *calibration*. Important performance characteristics to study in this phase are the *reportable range* (including limits of detection and quantification), *linearity*, *precision* (agreement within and between series), *reproducibility* (agreement between observers, devices, as well as laboratories), as well as *interferences* (including intravascular hemolysis, lipemia, and icterus). A different problem is that analytes might not be consistent if measured with different analytical platforms. A particular issue of test performance, which is often neglected, is the *interobserver reproducibility*. Many tests require some degree of interpretation. One example is flow cytometry, which is used in the hemostasis laboratory to observe platelet function. Poor interobserver reproducibility severely reduces the value of a diagnostic test, also with regard to thrombosis and hemostasis.[41]

How shall we design evaluation studies adequately addressing the analytical accuracy and technical performance of a laboratory test, and how do we select suitable performance criteria? I argue that two critical criteria must be taken into account to generate informative results: (1) what is the clinical question to be answered and (2) what is the adequate experimental design for the specific analytical method? Uniform checklists considering only the number of samples might grossly overestimate technical performance obtained in routine laboratory practice. As an example, we aimed to study the consistency of thromboelastometry measurements, a point-of-care device that is often used in the perioperative and acute trauma setting. It was unclear whether the measurements are reproducible among different devices, among different channels of the same device, and within individuals. Faced with the problem of labile sample material, we used a distinct study design and data analysis method, thus identifying critical inconsistencies.[14]

Phase 2: Preanalytical Factors

Preanalytical issues are regarded as the most frequent source of error in laboratory medicine.[1,42] The preanalytical phase is consequently the second issue to be studied in evaluation projects. It covers *specimen collection*, *sample processing* (e.g., centrifugation), *transportation*, as well as *storage*. Tests conducted in the hemostasis laboratory are particularly susceptible to preanalytical artifacts due to the activation of platelets, activation factors, as well as inhibitors.[1] A large number of such factors were found to affect hemostasis tests which included misidentification of samples, traumatic blood drawing, drawing from vascular access devices, incomplete distribution of anticoagulant within the tube, underfilling of tubes, vigorous shaking, use of very small and large needles, use of activating collection containers, use of anticoagulants other than citrate, tourniquet use, pneumatic tube transport, long processing times, and incorrect centrifugation schemes. Thus, evaluation projects should adequately address all aspects of sample handling and processing. Again, the study design and the performance criteria are defined by the analytical method and the clinical question to be answered. As an example, we aimed to evaluate a rapid, high-speed centrifugation scheme to be implemented in a routine hemostasis laboratory, thus promoting efficiency in laboratory automation. An experimental design in consecutive patients with suspected abnormal hemostasis tests was chosen in order to study the full range of values obtained in clinical practice.[43]

Phase 3: Biological Variation and Clinical Factors

The results of laboratory tests vary between individuals, and this must be taken into account for interpretation.[44] The most obvious examples are differences between males and females, newborns and adults, pregnancy, and over- and under-weight. Some analytes vary according to circadian rhythms or longer cycles (e.g., menstrual cycle). Also, laboratory test results might differ in patients taking particular drugs or undergoing major surgery.[45,46] If recognized and taken into account, these effects can even improve the accuracy of laboratory tests. As an example, the age-dependent increase of D-dimer levels has

been implemented in diagnostic algorithms to rule out venous thromboembolism, thus improving their diagnostic accuracy.[47,48]

Phase 4: Interpretability

Interpretation of test results is a critical step in the process of care because it determines the course of action to be taken (►Fig. 1). It can be routine in the case of dichotomous outcomes: i.e., the outcome is either positive (*factor is present*) or negative (*factor is absent*). Typical examples are genetic polymorphisms or the presence of viruses and antibodies. However, most test results are provided on a continuous scale that makes decision making much more difficult (even dichotomous test results reflect quantitative values using certain thresholds). At what level should the test result be regarded as *normal*, implying that the clinical question is to be answered with a *no* (e.g., the diagnosis is rejected)? And, at which point should the test result be regarded as *abnormal*, implying that the clinical question is to be answered with a *yes* (e.g., the diagnosis is accepted)?

Several different approaches have been used to solve this problem: (1) the *reference range* approach, (2) the *target cohort* approach, (3) the *diagnostic definition of normal* or *risk factor definition*, and (4) the *therapeutic definition of normal*.[32] Each of these options are all associated with certain drawbacks.[32] The *reference range* approach is used the most often, and clinicians and laboratory specialists are generally familiar with it. Here, the distribution of test results is determined in a cohort of healthy individuals (e.g., blood donors) and a statistical cutoff on both sides of the mean or median is defined (2 standard deviations or the interval between the 2.5th and the 97.5th percentile). The main problem with this method is that an abnormal test result does not automatically mean that a disease is present, and a normal test result does not always exclude a disease state. As an example, the prothrombin time (PT) is not only used for the monitoring of vitamin K antagonists but as a screening tool in patients with suspected bleeding disorders. The results are usually reported in seconds, percentages (PT ratio; quick percent), or as international normalized ratio against a reference range established for the respective reagent and coagulometer. An abnormal PT will not be associated with a bleeding disorder in the majority of cases, and a bleeding disorder might be present in some patients with normal PT.

Through the use of the *target cohort approach*, clinicians can differentiate patients with the disease from patients without the disease in a cohort of patients with similar signs and symptoms. The advantage of this approach is that it reflects the clinical question. However, studies must be conducted with adequately powered cohorts of patients with signs and symptoms of the target indication, which are tested against a reference standard test. Another drawback of this method is that some tests are used to answer several different clinical questions, which makes reporting of the test results challenging. A typical example is the platelet function analyzer (PFA). The diagnostic accuracy for von Willebrand disease was established in patients with suspected bleeding disorders, and respective cutoffs were established.[49] Among other reasons, the interpretation of test results is difficult because the PFA does not capture other common bleeding disorders (platelet function disorders).

In the *diagnostic definition of normal/risk factor definition*, the interpretation of test results is achieved according to the diagnostic or predictive value at certain thresholds. For example, D-dimer tests to rule out pulmonary embolism utilizes a cutoff level of 500 µg/L because this corresponds to a high predictive value of not having the disease (the likelihood ratio is well below 1).[47] D-dimer tests might also be used in the risk assessment for recurrent venous thromboembolism. The higher the level of D-dimer, the higher the risk for recurrent venous thromboembolism. This information can even be quantitatively implemented in clinical prediction models.[50] Another example is immunoassays for the diagnosis of heparin-induced thrombocytopenia. Higher cutoffs are associated with higher (positive) likelihood ratios, which facilitate clear clinical decisions.[51] Even though the *diagnostic definition of normal/risk factor definition* approach ensures that clinical decisions are made taking the actual risk of the patients into account, it does not automatically mean that this is associated with an improvement in clinical outcomes. The drawback of this approach is that large and well-designed clinical studies are necessary to obtain the estimates needed. In addition, the definition may change regularly as new studies come up.

The *therapeutic definition of normal* is the most intuitive definition of *normal*. Laboratory values consistent with a patient population that benefits from a certain treatment are used as a cutoff. The recent example is treatment with intravenous iron in patients with heart failure. Two randomized controlled trials demonstrated that intravenous iron is beneficial with regard to clinical outcomes in patients with heart failure and iron deficiency. To define iron deficiency, a ferritin cutoff level of 100 ng/mL was chosen.[52] The drawbacks of this approach are, however, that the *abnormal* definition is applicable only to a certain patient population, and it is very costly to conduct the underlying studies.

Phase 5: Diagnostic or Prognostic Accuracy

Laboratory test results are typically used to substantiate a suspected diagnosis (*diagnostic problem*) or to inform a risk assessment to decide on treatment characteristics (*prognostic problem*). Thus, the diagnostic (or prognostic) accuracy is a crucial characteristic of a laboratory test. Unfortunately, sensitivity and specificity have been regarded as fixed properties of a test and too little attention is given to how these measures are generated, the settings and circumstances to which these values apply, and what these parameters mean for clinical decision making. In this paragraph, I will discuss the major issues that apply to diagnostic accuracy studies. As prognostic studies correspond to classic epidemiological studies, the reader is referred to major textbooks of epidemiology.

Evidence from a large number of studies has made it clear that sensitivity, specificity, and related measures only describe the behavior of a test *under specific circumstances* (the circumstances of the evaluation study) and that these circumstances

often do not resemble the situation in clinical practice.[7] Besides, it has become clear that diagnostic accuracy studies with a suboptimal design can result in biased results.[22–25]

The *selection of patients* is an essential characteristic of the study design because this selected population defines the target population to which the test can be applied. Ideally, the study population resembles the target population perfectly in terms of patient characteristics as well as signs and symptoms. Thus, the diagnostic accuracy measures are determined from a representative range of patients including those that are "slightly" ill (usually with lower levels of the index test than "seriously" ill patients) as well as patients with other disorders that exhibit similar signs and symptoms (often with higher levels of the index test than healthy volunteers). In contrast, previous studies have generally selected a group of seriously ill patients and a control group of healthy volunteers, which has resulted in impressive (biased) diagnostic accuracy measures. This can result in harm to many patients if the test is implemented prematurely. A well-known example is the screening for prostate-specific antigen as an indication of prostate cancer.[53] In this instance, including selected rather than consecutive patients referred for the workup of the suspected disorder leads to a *spectrum* or *selection bias*. Verifying the workup modalities of patients can help identify a respective risk of bias in study populations.[7]

The determination of the *index test* should be done in exactly the same way as done in clinical practice. Overestimation of the diagnostic accuracy can occur if the conditions are better in the study situation. Typical examples are duplicate measurements, test modifications, and interpretation by specially trained investigators. The operators performing the index test must not be aware of the results of the reference standard test.

The diagnostic accuracy is estimated against a *reference standard test*. Suboptimal reference standard tests will result in biased estimates of diagnostic accuracy. The *best available method* should instead be selected. Less stringent reference standards might lead to misclassification and *reference standard bias*.[7] *Partial verification bias* can occur if only a population subgroup is tested against the reference standard[54] and *differential verification bias* if several reference standard tests are used. Sometimes, a panel of experts reviews patient charts, clinical data, index test results, and treatment course of the patient and this represents the reference standard. In this situation, the diagnostic accuracy is often overestimated due to *incorporation bias* and that the experts consider cases with a typical presentation (whether or not the diagnosis is true) and atypical cases tend to be neglected.[7] Again, knowledge of the index test result while interpreting the reference standard test can lead to biased estimates.

Characteristics of *flow and timing* of the study procedures might additionally introduce biased estimates. Inappropriately designed studies may result in awareness of the index test results while interpreting the reference standard (and vice versa), and the detectable presence of the disease may vary with time. Besides, the natural course of the disease might be affected by each intervention. This might change the detectable presence of the disease in the period between the performance of the index test and the performance of the reference standard. And may cause bias if both tests are interpreted at different time points during the disease course. Ideally, the index test is compared with the reference index test at the same time point.

How shall we analyze data from *diagnostic accuracy studies*? Traditionally, data are arranged in 2×2 tables, and sensitivity and specificity are calculated. This approach is associated with a number of pitfalls and drawbacks, however. First, 2×2 tables neglect inconclusive results, both with the index as well as the reference standard test. Excluding inconclusive results from the analysis is an important source of bias. Thus, data should be analyzed according to the *intention-to-diagnose approach*.[55] Inconclusive results of the index test are rated as negative if the reference standard is positive and classified as positive if the reference standard is negative. Observations are excluded if the reference standard test is inconclusive. Second, the post hoc definition of the index test cutoff often leads to overestimation of the diagnostic accuracy.[7] Thus, the threshold should already be defined in the study protocol based on the preliminary studies (or alternatively via a separate training set of observations). Third, point estimates of diagnostic accuracy may be imprecise and spurious in smaller studies. A power calculation is, however, rarely done in diagnostic accuracy studies.[21] A power calculation based on realistic assumptions and reporting of confidence intervals are essential aspects of diagnostic accuracy studies.[21,56,57] Fourth, sensitivity and specificity as well as (positive and negative) predictive values are often used as measures of diagnostic accuracy, but it is known that they are generally not applicable to other patient cohorts because of changes in prevalence and various patient characteristics.[7,20,28] Fifth, the use of likelihood ratios and Bayes' theorem to aid diagnostic reasoning is questionable because these measures vary between subgroups of patients.[27–29] In contrast, multivariate prediction models not only take covariables into account (representing subgroups of patients), but they can also determine the added value of a new laboratory test to existing diagnostic pathways.[58]

Phase 6: Utility (Clinical and Health Care Outcomes)

The ultimate means of assessing the utility of a laboratory test is to study its effect on *health outcomes*.[3,30] What outcomes should we focus on? Most interestingly and obvious for clinicians are *clinical outcomes*. Similar to outcome assessment of randomized controlled trials of interventions, clinical characteristics are *mortality* and *morbidity* crucial. Randomized controlled trials assessing a *testing-and-treatment strategy* against the absence of such a strategy represent the most rigorous of studies. The particular morbidity measure depends on the individual disease involved. In the case of deep vein thrombosis, this might be *recurrent venous thromboembolism* or the presence of severe *postthrombotic syndrome according to the Villalta score* and *major bleeding events*. However, a testing-and-treatment strategy may lead to a variety of adverse events, and testing for all possible events can be difficult.

The assessment of unmet clinical needs is suggested as the first step in the development and evaluation of new laboratory tests and a respective checklist is available[5] (►Fig. 4). This

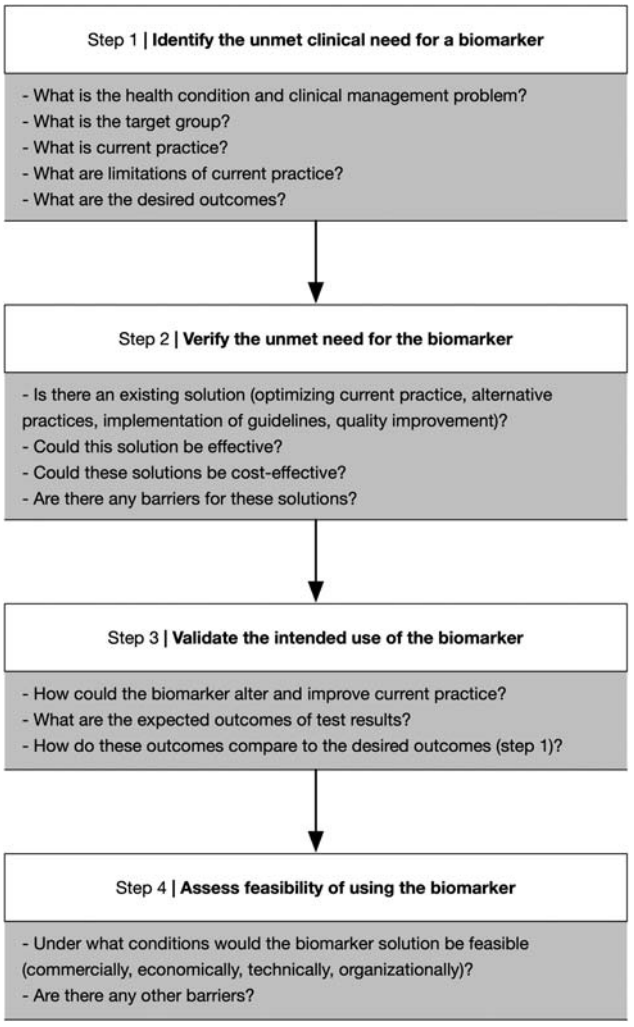

Fig. 4 Proposed checklist for the identification of unmet clinical needs (adapted from Monaghan et al[5] with modifications). The list can be applied by researchers to identify unmet clinical needs before developing, evaluating, and implementing new laboratory tests.

approach has the potential to increase value and reduce waste in biomedical research. As long as development of biomarkers usually evolves from new analytical technologies and knowledge from basic science, this checklist might be difficult to implement, however.

Patient-reported outcomes such as pain, anxiety, and functioning add valuable additional measures to the assessment of clinical outcomes. Also, generic and disease-specific questionnaires measuring the quality of life are available.

Process outcomes measure how the use of the laboratory test affects health care processes. Is the risk assessment refined? Is the diagnosis obtained more quickly? Are the processes simplified? Studies investigating these issues are usually performed in clinical practice and the study design must be highly adapted to the individual research question.

It is a matter of fact that laboratory medicine must be *cost-effective* to be relevant for patients, caregivers, provider organizations, health care insurances, and society as a whole. Thus, *costs* are an important outcome to study in order to assess the utility of a laboratory test. Conducting a cost-effectiveness study is, however, difficult in the diagnostic area because a large number of variables must be taken into account.[59] Thus, such an evaluation is rarely performed in laboratory medicine, though it would be beneficial given the rising costs of health care.

Perspectives

Laboratory testing aims to improve outcomes—not only for patients but also for caregivers, provider organizations, health care insurances, and society generally. Clinical decision making must effectively and efficiently be promoted by laboratory testing to achieve this goal. A broad range of test characteristics must be assessed, and a number of methodological tools used to demonstrate the utility of a laboratory test in this process.

To date, laboratory tests are rarely assessed adequately prior to implementation. Consequently, overdiagnosis and overtreatment is regarded as a major threat to human health and health care systems. To address this issue, a comprehensive methodological framework to be used by researchers, clinicians, and decision makers in authorities, health care insurances, and provider organizations should be developed and provided. A phased approach—similar to that used in the assessment of new treatment innovations—has many important advantages, and an outline has been proposed in this article.

An important question is who should be responsible to do the evaluation projects for individual laboratory tests? One might argue that authorization processes similar to the approval of new drugs shall be established. This approach needs a clear and detailed catalogue of requirements, however, which is not available so far. Future research can be anticipated to identify methodological shortcomings in current evaluation projects and to develop new methodological tools to address these issues. Specific issues, such as the absence of a reference standard or appropriate testing of rare diseases, should also be addressed. We look forward to scientific societies and authorities supporting this process through the development of definitive requirements and acceptance criteria for every phase of the evaluation process of various laboratory medicine tests.

Funding

M. Nagler is supported by a research grant of the Swiss National Science Foundation (SNSF; #179334).

Conflict of Interest

The author declares that there is no conflict of interest.

References

1 Funk D. Sample integrity and preanalytical variables. In: Kitchen S, Olson JD, Preston FE, eds. . Quality in Laboratory Hemostasis and Thrombosis. John Wiley & Sons; 2013
2 Price CP, Christenson RH. Preface to the second edition. In: Price CP, Christenson RH, eds. . Evidence-Based Laboratory Medicine. 2nd ed. AACC Press; 2007:9–11

3 Price CP. Evidence-based laboratory medicine: supporting decision-making. Clin Chem 2000;46(8, Pt 1):1041–1050
4 World Health Organization. International Classification of Health Interventions (ICHI). 2020. https://www.who.int/classifications/ichi/en/
5 Monaghan PJ, Lord SJ, St John A, et al; Test Evaluation Working Group of the European Federation of Clinical Chemistry and Laboratory Medicine. Biomarker development targeting unmet clinical needs. Clin Chim Acta 2016;460:211–219
6 Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. BMJ 2002;324(7335):477–480
7 Bossuyt PMM. Studies for evaluating diagnostic and prognostic accuracy. In: Price CP, Christenson RH, eds. . Evidence-Based Laboratory Medicine. 2nd ed. AACC Press; 2007:67–81
8 Rabin BA, Brownson RC, Haire-Joshu D, Kreuter MW, Weaver NL. A glossary for dissemination and implementation research in health. J Public Health Manag Pract 2008;14(02):117–123
9 Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol 2015;13(06):e1002165
10 Nagler M. Validity and Diagnostic Value of Tests Used in the Diagnostic Work-up of Haemostatic Disorders. Maastricht; 2014
11 Rubin EH, Gilliland DG. Drug development and clinical trials–the path to an approved cancer drug. Nat Rev Clin Oncol 2012;9(04):215–222
12 Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 2011;10(09):712
13 Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. BMJ 2012;344:e3502
14 Nagler M, ten Cate H, Kathriner S, Casutt M, Bachmann LM, Wuillemin WA. Consistency of thromboelastometry analysis under scrutiny: results of a systematic evaluation within and between analysers. Thromb Haemost 2014;111(06):1161–1166
15 Graber ML. The incidence of diagnostic error in medicine. BMJ Qual Saf 2013;22(Suppl 2):ii21–ii27
16 Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf 2014;23(09):727–731
17 Mangalmurti SS, Harold JG, Parikh PD, Flannery FT, Oetgen WJ. Characteristics of medical professional liability claims against internists. JAMA Intern Med 2014;174(06):993–995
18 Levinson W, Kallewaard M, Bhatia RS, Wolfson D, Shortt S, Kerr EAChoosing Wisely International Working Group. 'Choosing Wisely': a growing international campaign. BMJ Qual Saf 2015;24(02):167–174
19 Hallworth MJ, Epner PL, Ebert C, et al; IFCC Task Force on the Impact of Laboratory Medicine on Clinical Management and Outcomes. Current evidence and future perspectives on the effective practice of patient-centered laboratory medicine. Clin Chem 2015;61(04):589–599
20 Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013;185(11):E537–E544
21 Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006;332(7550):1127–1129
22 Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282(11):1061–1066
23 Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174(04):469–476
24 Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med 2013;137(04):558–565
25 Whiting PF, Rutjes AWS, Westwood ME, Mallett SQUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013;66(10):1093–1104
26 Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature 2012;483(7391):531–533
27 Miettinen OS, Henschke CI, Yankelevitz DF. Evaluation of diagnostic imaging tests: diagnostic probability estimation. J Clin Epidemiol 1998;51(12):1293–1298
28 Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8(01):12–17
29 Moons KGM, Grobbee DE. Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 2002;56(05):337–338
30 Moons KGM. Criteria for scientific evaluation of novel markers: a perspective. Clin Chem 2010;56(04):537–541
31 Sackett DL, Rosenberg WMC. The need for evidence-based medicine. J R Soc Med 1995;88(11):620–624
32 Karanicolas PJ, Guyatt GH. Evidence-based medicine and the diagnostic process. In: Price CP, Christenson RH, eds. . Evidence-Based Laboratory Medicine. AACC Press; 2007
33 Haynes RB, Sackett DL, Gray JM, Cook DJ, Guyatt GH. Transferring evidence from research into practice: 1. The role of clinical care research evidence in clinical decisions. ACP J Club 1996;125(03):A14–A16
34 Price CP, Christenson RH. The clinical question: a system for formulating answerable questions in laboratory medicine. In: Price CP, Christenson RH, eds. . Evidence-Based Laboratory Medicine. AACC Press; 2007:25–52
35 Bickley LS. Bates Guide to Physical Examination and History Taking. 12th ed. Wolters Kluwer; 2017
36 Guyatt G, Rennie D, Meade MO, Cook DJ. Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice. McGraw-Hill Education; 2015
37 Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test. Lessons from the rise and fall of dexamethasone suppression test. JAMA 1988;259(11):1699–1702
38 Bossuyt PM, Reitsma JB, Bruns DE, et al; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527
39 Whiting PF, Rutjes AWS, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(08):529–536
40 Westgard JO. Basic Method Validation. 3rd ed. Westgard QC, Inc.; 2008
41 Nagler M, Fabbro T, Wuillemin WA. Prospective evaluation of the interobserver reliability of the 4Ts score in patients with suspected heparin-induced thrombocytopenia. J Thromb Haemost 2012;10(01):151–152
42 Bonini P, Plebani M, Ceriotti F, Rubboli F. Errors in laboratory medicine. Clin Chem 2002;48(05):691–698
43 Wolfensberger N, Georgiou G, Giabbani E, et al. Rapid centrifugation in the routine hemostasis laboratory. Thromb Haemost 2019;119(12):2025–2033
44 Fraser CG. Biological Variation: From Principle to Practice. AACC Press; 2001
45 Banfi G, Del Fabbro M. Biological variation in tests of hemostasis. Semin Thromb Hemost 2009;35(01):119–126
46 Erdoes G, Dietrich W, Stucki MP, et al. Short-term recovery pattern of plasma fibrinogen after cardiac surgery: a prospective observational study. PLoS One 2018;13(08):e0201647
47 Righini M, Van Es J, Den Exter PL, et al. Age-adjusted D-dimer cutoff levels to rule out pulmonary embolism: the ADJUST-PE study. JAMA 2014;311(11):1117–1124
48 Parpia S, Takach Lapner S, Schutgens R, Elf J, Geersing GJ, Kearon C. Clinical pre-test probability adjusted versus age-adjusted D-dimer interpretation strategy for DVT diagnosis: a diagnostic individual patient data meta-analysis. J Thromb Haemost 2020;18(03):669–675

49 Fressinaud E, Veyradier A, Truchaud F, et al. Screening for von Willebrand disease with a new analyzer using high shear stress: a study of 60 cases. Blood 1998;91(04):1325–1331
50 Eichinger S, Heinze G, Jandeck LM, Kyrle PA. Risk assessment of recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism: the Vienna prediction model. Circulation 2010;121(14):1630–1636
51 Nagler M, Bachmann LM, ten Cate H, ten Cate-Hoek A. Diagnostic value of immunoassays for heparin-induced thrombocytopenia: a systematic review and meta-analysis. Blood 2016;127(05):546–557
52 Ponikowski P, van Veldhuisen DJ, Comin-Colet J, et al; CONFIRM-HF Investigators. Beneficial effects of long-term intravenous iron therapy with ferric carboxymaltose in patients with symptomatic heart failure and iron deficiency. Eur Heart J 2015;36(11):657–668
53 Andriole GL, Crawford ED, Grubb RL III, et al; PLCO Project Team. Mortality results from a randomized prostate-cancer screening trial. N Engl J Med 2009;360(13):1310–1319
54 Baker SG. Evaluating multiple diagnostic tests with partial verification. Biometrics 1995;51(01):330–337
55 Schuetz GM, Schlattmann P, Dewey M. Use of 3×2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. BMJ 2012;345:e6717
56 Bujang MA, Adnan TH. Requirements for minimum sample size for sensitivity and specificity analysis. J Clin Diagn Res 2016;10(10): YE01–YE06
57 Alonzo TA, Pepe MS, Moskowitz CS. Sample size calculations for comparative studies of medical tests for detecting presence of disease. Stat Med 2002;21(06):835–852
58 Moons KGM, de Groot JAH, Linnet K, Reitsma JBR, Bossuyt PMM. Quantifying the added value of a diagnostic test or marker. Clin Chem 2012;58(10):1408–1417
59 Hernandez JS. Cost-effectiveness of laboratory testing. Arch Pathol Lab Med 2003;127(04):440–445