**For Official Use** **DSTI/STP/BIO(2008)37**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

**10-Nov-2008**

**English - Or. English**

**DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INDUSTRY**
**COMMITTEE FOR SCIENTIFIC AND TECHNOLOGICAL POLICY**

## Working Party on Biotechnology

# ANALYTICAL PAPER: CLINICAL EVALUATION OF BIOMARKERS

**By Ron Zimmern and Carole Wright, PHG Foundation, United Kingdom**

*This analytical paper was submitted for discussion at the Workshop on Policy Issues in the Development and Use of Biomarkers in Health held on the 6-7 October 2008 in Hinxton, United Kingdom. It is submitted for information to the WPB.*

For further information, please contact: Ms. Marie-Ange Baucher, Tel: +33 1 45 24 94 22; email: marie-ange.baucher@oecd.org; or Ms. Benedicte Callan, Tel: +33 1 45 24 15 93; Fax: +33 1 44 30 63 36; email: benedicte.callan@oecd.org.

**JT03254954**

**Document complet disponible sur OLIS dans son format d'origine**
**Complete document available on OLIS in its original format**

**DSTI/STP/BIO(2008)37**
**For Official Use**

**English - Or. English**

## NOTE BY THE SECRETARIAT

This analytical paper was submitted as background material for discussion at the expert workshop organised by the Biotechnology Division on “Policy Issues in the Development and Use of Biomarkers in Health” held in Hinxton, United Kingdom on 6-7 October 2008. This workshop contributes to the fulfillment of Output Result 5 of the 2007-2008 PWB entitled “Analytical and policy reports on the impact of molecular markers and targeted therapies on Biomedicine”.

This analytical paper, written by the PHG Foundation, presents the challenges inherent in the clinical evaluation of biomarkers. It also gives some suggestions for discussion about the development of a methodology for the clinical evaluation of biomarkers which will be necessary to ensure the broad use of biomarkers in medical care.

This analytical paper, along with others developed for the Biomarker Workshop, will be used as input for the Policy Report entitled “Policy issues in the Development and Use of Biomarkers in Health” that will be submitted to WPB in early 2009.

Delegates to the Working Party on Biotechnology are invited to:

- **Note** the analytical paper.

## Table of Contents

## 1. Introduction

The purpose of this paper is to summarise some of the current thinking on the evaluation of *biomarkers*. It is presented not by way of a definitive proposal, but as a mechanism for stimulating debate and discussion. The definition of the term is set out on the title page. Many of the key concepts have arisen in the field of genetic testing, due to the rise in new and complex tests resulting from the Human Genome Project, but the concepts are applicable to all biomarkers. Specifically, the paper will cover the following key areas:

*i)* The turning of a biomarker assay into a clinical test;

*ii)* A description of the ACCE framework for evaluating analytical validity, clinical validity, clinical utility and ethical, legal and social issues;

*iii)* Specific frameworks for biomarker evaluation.

The paper will NOT discuss issues that pertain to the technical evaluation of a biomarker, the features of which will necessarily depend on the nature of the biomarker under evaluation.

*Biomarkers* have in recent years become the subject of much research activity, not only as a means for better understanding physiological and pathological processes in health and disease, but also specifically to use as tests for determining susceptibility to (or prognosis of) disease, and the response, dosage or risk of adverse events of pharmacological agents. These uses – disease susceptibility and pharmacogenetics – have been a potent driver for the recent interest in the evaluation of biomarkers.

## 2. Issues in Biomarker Evaluation

There are a number of challenges associated with the evaluation of biomarkers, many of which are addressed in this paper, whilst others should be borne in mind when considering the practicalities of test evaluation. Some of the key issues include:

- Increasing complexity of biomarkers, requiring knowledge beyond standard physician training for valid interpretation of the results;

- Increasing public accessibility to ‘direct-to-consumer’ biomarker tests;

- Frequent lack of data on clinical performance (validity and utility) due to the cost, size and length of studies required;

- Lack of agreement regarding who is responsible for funding studies, generating data and assessing evidence for clinical biomarker performance;

- Confusion on the part of biomarker developers and test manufacturers as to what level of clinical evidence is required, leading to different *ad hoc* standards;

- Lack of consensus standards or a quality assurance framework for biomarker evaluation;

- Publication bias and selective reporting of trials with positive outcomes;
- Difficulties associated with assessing predictive biomarkers indicating increased risk or susceptibility to a complex disease;
- Existence of many different types of biomarkers, requiring different expertise for proper use and evaluation;
- Enormity of the task of evaluating all biomarker tests.

## 3. Turning a Biomarker into a Test

### *3.1 Assay / test distinction*

For the purposes of evaluation and regulation, it is useful to distinguish between an *assay* and a *test*[1]. Simply put, an assay is a scientific measurement, whilst a test is its clinical interpretation. Specifically, an assay is a method to analyse or quantify a particular biomarker in a sample, such as a serum sodium level or the sequence of a gene; a test is the use of an assay within a specific context, *i.e.*

- For a particular disease
- In a particular population
- For a particular purpose

### *3.2 Defining purpose*

The definition of a test (above) requires that the purpose and target population for testing is clearly specified. A clinical test may have a variety of different purposes[2], including:

- Making or excluding a diagnosis
- Guiding further investigation
- Disease classification or stratification
- Evaluating prognosis
- Guiding and monitoring treatment
- Population screening and risk stratification

Some tests may have multiple clinical purposes. For example, the genetic test for cystic fibrosis maybe used to confirm a diagnosis of the disease, to guide treatment options, or to identify carrier status and aid reproductive decision-making. In each case, if the test is to achieve its purpose, it must be delivered in an appropriate way, in association with any relevant services or interventions.

Moreover, the extent to which a test (or intervention) meets the objectives or purposes for which it was designed is the formal definition of effectiveness. Therefore, defining the purpose of a test is a necessary requirement without which the effectiveness of a particular biomarker cannot be evaluated.

Policy makers therefore need to identify the potential purposes for which a test may be used in order to ensure appropriate priority setting, evaluation and logistical support. As biomarker assays become increasingly complex, both to perform and to interpret, consideration of test purpose will help health funders in setting priorities for coverage of services.

## 4. ACCE Framework

The evaluation of diagnostics has not featured highly on the priorities of policy makers or regulators. Tests were, for many years, believed to be a matter for clinical interpretation by physicians, based on their education and judicious reading of the medical literature. This model served the profession well, as tests were on the whole simple and unidimensional. For example, the interpretation of haemoglobin and white cell counts, urea and electrolytes and conventional liver function tests were all within the remit of practising physicians. But in recent years, with the increased understanding of cellular and molecular processes, and the growth in complex genetic and molecular tests, this paradigm can no longer be assumed. It is no longer reasonable to expect the general clinician to interpret a complex molecular biomarker, or to understand the subtleties of genetic variation in a DNA sequence, or the characteristics of a gene expression microarray used to determine the prognosis of a particular cancer; hence, there is a need to evaluate formally both the validation of the biomarker during its development and its utility in clinical practice.

The ACCE model, originally developed by the US Centre for Disease Control (CDC) for evaluating genetic tests, provides a theoretical framework for developing a robust evaluation process that we believe to be applicable for any diagnostic test[3]. One advantage is that it covers the issues of both clinical validity, which is so important during the development process, and clinical utility which should be (but often is not) demonstrated before use in a clinical or public health context. The key components are summarised below:

*Analytical validity* the accuracy and reliability with which the assay measures or detects the biomarker of interest, including laboratory quality assurance

*Clinical validity* the ability of the test to distinguish those who have, or will develop, a disorder from those who are, and will remain, healthy; divided into scientific validity and test performance

*Clinical utility* the risks and benefits of using the test, with particular reference to its purpose and feasibility of delivery

*'ELSI'* ethical, social and legal implications surrounding the test, including consideration of any safeguards and impediments

Although analytical validity is usually well covered by laboratory quality assurance procedures, clinical validity is only occasionally assessed in research publications, whilst clinical utility and ELSI are often not formally evaluated at all, despite being key to determining whether or not the test actually produces a benefit[4]. In part, this is due to the greater complexity, size and cost of studies to determine clinical validity and clinical utility, discussed further below. Therefore, full evaluation of a test will require the establishment of separate systems to generate the data needed for full evaluation of a test and associated services and undertake the analysis and assessment, in addition to providing a policy response and clinical guidance.

We take the view that there is nothing specific about the nature of genetic tests (which here is taken to mean tests based on the use of nucleic acids as the analyte) that require specific evaluative processes, and therefore the ACCE framework offers an excellent methodology for biomarker validation in general.

However, the distinction between diagnostic tests and predictive or susceptibility tests is relevant, and the evaluation of prediction or susceptibility to disease does require some further refinement of the ACCE framework, a matter that we deal with in 4.1.2 below.

### *4.1 Clinical validity*

The clinical validity of a test is its ability to successfully detect or predict a particular disorder. It can be usefully separated into two components: the biomarker-disease association and the clinical test performance.

#### *4.1.1 Scientific validity*

Scientific validity refers to the assessment of the evidence of a biomarker-disease association, which may be addressed through primary research, large well-controlled studies or systematic review and meta-analysis. Guidelines for the assessment of evidence of gene-disease associations have been produced[5], which are applicable to all biomarkers. It is important to note that evidence of association is necessary, but not sufficient, for a test to be clinically valid.

#### *4.1.2 Test performance*

The evaluation of test performance in a clinical context is essential for a physician to correctly interpret a test result. This is usually achieved through clinical trials to determine sensitivity, specificity and likelihood ratios using the standard '2 by 2' table to measure false positive and false negative test results[6]. The positive and negative predictive value of the test, which varies with disease and biomarker prevalence, can then be calculated for the target population and the test applied and interpreted accurately within the context of an individual patient.

Whilst this tried and tested model is satisfactory for *diagnostic* testing, it may be less so when applied to *predictive* testing. There is an implicit assumption that in diagnostic tests, the biomarker is an actual marker of disease; hence it is appropriate to dichotomise diagnostic tests as being either positive or negative and, within that paradigm, to establish parameters of test performance, such as sensitivity, specificity or predictive value.

However, in predictive tests, the biomarker is in effect a risk factor that indicates susceptibility to *future* disease. Therefore, it is perhaps less appropriate to use such a simple scheme for evaluation for several reasons. Firstly, it will prove extremely difficult to assess the clinical performance of such a test, as the required studies would be extremely lengthy, and likely to be confounded by any preventive measures taken in the interval between test and the development of disease. Secondly, the relative risk conferred by each individual susceptibility biomarker is often so small (usually less than 2) that the marker alone will be of little use for distinguishing between those who will, and those who will not, develop the disease in question[1]. Thirdly, it is unclear exactly what constitutes a false result in the case of a biomarker test that attempts to predict an individual's susceptibility to disease.

A new approach may therefore be needed to evaluate these types of susceptibility tests. We suggest that the fundamental basis for such an approach should lie in calculating the absolute risk of disease conferred by the presence of a biomarker, or combination of biomarkers. By considering each biomarker in combination with other clinically relevant information, the *relative* risk of disease associated with that

[1] Most experts now believe that relative risks of at least 20 to 30 are necessary for the test to have a practical discriminatory value on an individual level.

biomarker may be used to estimate the *absolute* risk of disease in a particular individual or subpopulation using, for example, the age-sex population risk as the base pre-test risk[2]. A susceptibility biomarker would thus form part of an overall strategy to categorise people into higher or lower risk groups, in order to target them for further testing or preventive interventions. Therefore, in addition to the relative risk conferred by the biomarker, the prevalence of both the disease and the biomarker itself in the population are crucial for evaluating the test performance and its utility. The absolute risk measurement would in effect determine whether or not a further intervention, such as a more invasive test or some form of treatment, should follow. This approach is consistent with the concept that all tests should have a purpose, and that their use and interpretation should result in a decision or action to carry out further tests, or to initiate some therapeutic intervention.

### *4.2 Clinical utility*

Prior to implementing a test, it is crucial to consider the associated risks and benefits. In addition to traditional measures of healthcare quality, the clinical utility of a test can be assessed by consideration of the different purposes proposed for a particular test, and the different ways in which a biomarker may be associated with a disease. Eight different dimensions to clinical utility are presented here, based on Donabedian's general work on quality assurance in health care[7,8]; the first four relate to test purpose, and the last four to feasibility of test delivery.

*i)* *Legitimacy* – the conformity of a test to social preferences expressed in ethical principle, value, norms, mores, laws and regulations. For example, a test to reduce morbidity and mortality (the primary goals of healthcare) has prior legitimacy, as long as it can be shown to achieve its purpose and is introduced within governing laws and regulations.

*ii)* *Efficacy* – the ability of a test (and any associated services) to bring about its intended purpose and deliver health benefits, when used under the most favourable circumstances. The efficacy of a test will vary with test performance and purpose, as well as with the nature of the target condition.

*iii)* *Effectiveness* – the degree to which the intended health benefits are actually attained under routine conditions. Factors that could result in differences between the research outcomes and the clinical outcomes are key to this evaluation, and the assessment must therefore consider logistical and other practical considerations.

*iv)* *Appropriateness* – the balance between anticipated benefits and adverse consequences of a test; the former should outweigh the latter by a sufficiently large margin that the service is considered to be worthwhile.

*v)* *Acceptability* – the conformity of a test to the wishes, desires and expectations of patients and their families. Evaluation of acceptability requires the inclusion of members of the target patient population and their families in the assessment.

*vi)* *Economic efficiency* – the ability of the test to lower the costs of care without diminishing the benefits. This quality is particularly important when a service is beneficial but costly. Differential distribution of care amongst different classes of patients can help to optimise the outcomes.

[2] In this context, the *relative risk* is a ratio of the risk of developing a particular disease given the presence or level of a particular biomarker relative to someone who does not have that biomarker (who is assumed to have a relative risk of 1). In contrast, the *absolute risk* is a measure of the overall probability of developing a particular disease over a specific time period, which is usually estimated from the age-specific incidence of the disease. Whilst relative risk can take a wide range of values, absolute risk only ranges from 0 to 1.

*vii)* *Economic optimality* – the balance of improvements in health against the cost of improvements. This cost-effectiveness assessment acknowledges the opportunity costs of medical innovation, as well as reimbursement and delivery cost of the test and associated services.

*viii)* *Equity* – ensuring a just and fair distribution of healthcare and its benefits among members of the population, which may require development of capacity before the test is introduced.

## 5. Other specific evaluation frameworks

Numerous attempts have been made to address the problem of how best to evaluate a particular specific biomarker[9,10]. A few examples are presented here, which are intended to be representative of the types of organisations involved in this area. They fall broadly into two categories: those that carry out test evaluations or offer an advisory service according to a specific evaluative framework, and those that develop data standards to facilitate evaluation.

1. Test evaluation

- The United Kingdom Genetic Testing Network (*www.ukgtn.nhs.uk*) has developed a 'gene dossier' process to evaluate genetic tests, based on the ACCE framework, in order to recommend which tests should be offered by the NHS[11]. The gene dossier was specifically developed to be completed and evaluated in a practicable and timely manner, by providing a standardised format for the presentation of key information about a particular test.
- The Human Genetics Quality Network (*www.hgqn.org*) was set up, in connection with The German Society of Clinical Genetics and the German Society of Human Genetics, to provide information on external quality assurance schemes. New entries to the database have to pass a quality review process prior to being accepted.
- The Evaluation of Genomic Applications in Practice and Prevention (*www.egappreviews.org*) initiative was launched in the US in 2004 to support a coordinated, systematic approach to the evaluation of genetic tests and other genomic applications that are in transition from research to clinical and public health practice. Its independent, multidisciplinary working group is developing a systematic process for assessing the available evidence regarding the validity and utility of rapidly emerging genetic tests for clinical practice. This panel prioritizes and selects tests, reviews CDC-commissioned evidence reports and other contextual factors, highlights critical knowledge gaps, and provides guidance on appropriate use of genetic tests in specific clinical scenarios.

2. *Development of standards*

- The Microarray Gene Expression Data Society (www.mged.org) develops standards for presenting and exchanging microarray data, in order to improve its quality and reproducibility. One major goal is the development of guidelines regarding the minimum information about a microarray experiment that should be reported so that others can unambiguously repeat and interpret microarray data. These efforts are complemented by the US FDA Microarray Quality Control Project, which seeks to publish standards for data from microarrays and associated technologies, to improve comparability between different platforms.

- The Standards for Reporting Diagnostic Accuracy (www.stard-statement.org), which evolved out of the Cochrane Collective, has produced a 25-item checklist to aid with the design and conduct of studies, execution of the tests and interpretation of the results.

## 6. Conclusion

The goal of test evaluation is simply to determine whether patients that undergo a diagnostic test fare better than similar untested patients[12]. Ultimately, the value of a medical test to society depends upon its diagnostic usefulness and cost-effectiveness. Test evaluation is therefore critical to providing excellence of care, ensuring patient safety and setting healthcare priorities. Central to evaluation is a clear understanding of the purpose(s) of testing, against which the clinical performance of a test can be judged. The ACCE framework has been developed, and successfully applied, to achieve a thorough, robust and evidence-based evaluation process, which is applicable to all medical tests.

In this paper, we suggest that:

*a)* The ACCE framework provides a good basis for evaluation of all biomarkers, but that it is not in itself sufficient;

*b)* To properly understand the nature of *clinical validity* it is necessary to subdivide this into two components, biomarker-disease association, which we call *scientific validity*, and *test performance*;

*c)* To take into account the complexities of future events, predictive or susceptibility testing requires a different approach, namely the determination of absolute risk and an understanding of the absolute risk thresholds at which preventive or therapeutic interventions should be used;

*d)* To link the evaluative process into healthcare pathways, greater emphasis should be given to the formal assessment of *clinical utility* and its various dimensions.

## REFERENCES

1. Zimmern RL, Kroese M. The evaluation of genetic tests. *Journal of Public Health,* 2007; 29:246-50.

2. Burke W, Zimmern RL, Kroese M. Defining purpose: a key step in genetic test evaluation. *Genetics in Medicine,* 2008; 9:675-81.

3. Haddow J, Palomaki G. ACCE: A Model Process for Evaluating Data on Emerging Genetic Tests. In Khoury MJ, Little J, Burke W, eds. *Human Genome Epidemiology*, Oxford University Press, 2004.

4. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genetics in Medicine,* 2007; 9:665-74.

5. Ioannidis JPA *et al.* Assessment of cumulative evidence on genetic associations: interim guidelines. *International Journal of Epidemiology*, 2008; 37:120-132

6. F letcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology The Essentials, 1996.

7. Burke W, Zimmern RL. Moving Beyond ACCE: an expanded framework for genetic test evaluation. *PHG Foundation,* 2007.

8. Donabedian A. An Introduction to Quality Assurance in Health Care. Oxford University Press, 2003.

9. Guidelines for Quality Assurance in Molecular Genetic Testing. *OECD*, 2007.

10. Cancer Biomarkers. Report from the *Institute of Medicine of the National Academies*, 2007.

11. Kroese M, Zimmern RL, Farndon P, Stewart F, Whittaker J. How can genetic tests be evaluated for clinical use? Experience of the UK Genetic Testing Network. *Eur J Hum Genet,* 2007; 15(9): 917-21

12. Sackett DL, Haynes RB. Evidence base of clinical diagnosis: The architecture of diagnostic research. *BMJ* 2002; 324:539-41.