Procedural guidance for the systematic evaluation of biomarker tests

Version 1.2, November 2014

Ludwig Boltzmann Institut
Health Technology Assessment
Decision Support Document Nr. 77
ISSN-online: 1998-0469

Procedural guidance for the systematic evaluation of biomarker tests

Version 1.2, November 2014

Ludwig Boltzmann Institut
Health Technology Assessment

Vienna, November 2014

Project team

Project head:	Dr. Ingrid Zechmeister-Koss
Project team:	Dr. Agnes Kisser

Project support

External review:	Tracy Merlin, AHTA, Australia Stefan Lange, IQWIG, Germany Petra Schnell-Inderst, UMIT, Austria
Internal review:	PD Dr. Claudia Wild

Correspondence: Dr. Agnes Kisser, agnes.kisser@hta.lbg.ac.at

Dieser Bericht soll folgendermaßen zitiert werden/This report should be referenced as follows:

Kisser A & Zechmeister-Koss I. Procedural guidance for the systematic evaluation of biomarker tests.
Decision Support Document Nr. 77; 2014. Vienna: Ludwig Boltzmann Institute for Health Technology Assessment.

Conflicts of interest

The authors declare no conflicts of interest according to the Uniform Requirements of Manuscripts Statement of Medical Journal Editors (www.icmje.org).

IMPRESSUM

Medieninhaber und Herausgeber:

Ludwig Boltzmann Gesellschaft GmbH
Nußdorferstr. 64, 6 Stock, A-1090 Wien
http://www.lbg.ac.at/de/themen/impressum

Für den Inhalt verantwortlich:

Ludwig Boltzmann Institut für Health Technology Assessment (LBI-HTA)
Garnisongasse 7/20, A-1090 Wien
http://hta.lbg.ac.at/

Die HTA-Projektberichte erscheinen unregelmäßig und dienen der Veröffentlichung der Forschungsergebnisse des Ludwig Boltzmann Instituts für Health Technology Assessment.

Die HTA-Projektberichte erscheinen in geringer Auflage im Druck und werden über den Dokumentenserver „http://eprints.hta.lbg.ac.at" der Öffentlichkeit zur Verfügung gestellt:

Decision Support Document Nr.: 77

ISSN: 1992-0488
ISSN-online: 1992-0496

© 2014 LBI-HTA – Alle Rechte vorbehalten

Contents

Figures

Tables

1 Introduction

Personalised Medicine seeks to improve stratification and timing of health care by utilising biological information and biomarkers on the level of molecular disease pathways, genetics, proteomics as well as metabolomics [1]. These biomarker tests only deploy their value through influencing subsequent clinical decisions and in conjunction with subsequent treatments. The increasing numbers of biomarker tests in clinical routine require an adaptation of current assessment of diagnostics and medical tests to precisely determine the value of each test in a given setting.

increasing numbers of biomarkers in clinical routine:
adaptation of current assessment of diagnostics and medical tests needed

With the sequencing of the human genome in 2001 and the fast development of the –omics technologies, expectations are high, that biological processes including disease progression and treatment reactions can be measured accurately and even predicted on a molecular level. Association studies yield vast numbers of potential biomarker candidates, but so far there is no definitive consensus on the evidentiary requirements for the evaluation of biomarker tests for clinical routine. Diagnostic accuracy represents an important but not by itself sufficient characteristic of biomarker tests. Common challenges in the evaluation of biomarkers and medical tests are the assessment of multiple steps in a clinical path (test and treatment), methodological challenges in systematic reviews of diagnostic accuracy and prognostic tests, the lack of direct evidence from (double) randomised clinical trials and the complex assessment of the applicability of the test and treatment strategy.

challenges:
multiple steps in clinical path, diagnostic accuracy insufficient characteristic, lack of direct evidence on clinical utility

It was our aim to analyse the approaches proposed in methodological guidelines by leading HTA institutes and to extract a procedural guidance for the evaluation of research questions beyond the classical therapeutic intervention, triggered by the increasing importance of biomarkers and medical tests in medical routine. The document is meant to complement the LBI internal manual for qualitative evidence synthesis on effectiveness and safety of medical technologies [2]. Only steps in the assessment which deviate or need special attention are covered, and methods of quantitative evidence synthesis and cost-effectiveness are excluded. This report is a living document and is meant to be subject for development over time.

aim:
complement the LBI internal manual for qualitative evidence synthesis on effectiveness and safety of biomarker tests

2 Method

The report is not intended to provide a systematic review of the topic. Findings are based on a limited literature search that was conducted using the Pubmed and Medline bibliographic databases and inclusion of references was restricted to English and German language documents. Grey literature was identified by searching a variety of websites from HTA institutes or EBM research institutes for reports and guidelines addressing the specific challenges in the assessment of diagnostics, biomarkers and medical tests and providing methodological guidance. Further references were gathered by a snowball system in the references of the reports identified and conference abstracts from the 2013 Symposium "Methods for Evaluating Medical Tests and Biomarkers" at the University of Birmingham or were suggested by the reviewers. We compared the terminology used in the methodology papers and mapped them to the steps in the evaluation process. We identified areas with missing methodological consensus and compared the solutions proposed. The method guidelines and reports included in this report are summarised in Table 2-1. In line with the scope of the report, we did not include publications covering meta-analysis, decision modelling or cost-effectiveness analysis.

unsystematic review: literature search in Pubmed, Medline and hand search for grey literature

The structure of the report corresponds to the sequence of steps involved in the assessment of a medical test from formulating the research question to the synthesis of the evidence as described in the LBI-HTA's internal manual [2]. In a first section, we provide definitions and clarifications of the often varying terminology used in the field. This is followed by guidance for the development of the analytical framework and the formulation of the PICO question. New study designs likely to be encountered in reviews of biomarker tests are presented together with alternative evidence hierarchies. Finally guidance is provided on how to assess bias and applicability in alternative study designs and how to grade the available evidence. The report is not a comprehensive catalogue of methods, but aims to emphasise the issues specific to the evaluation of biomarkers and medical tests and provide references to detailed method guidelines, where necessary.

structure of the report corresponds to steps in assessment

highlight specific issues

Table 2-1: Method guidelines and reports included

Institution	Title	Reference
Agency for Healthcare Research and Quality's (AHRQ)	Methods Guide for Medical Test Reviews	AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. Available from: www.effectivehealthcare.ahrq.gov/reports/final.cfm Also published as a special supplement to the Journal of General Internal Medicine, July 2012.
The Cochrane Collaboration	Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9	Deeks JJ, Bossuyt PM, Gatsonis C (editors), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9.0.The Cochrane Collaboration, 2010 Available from: http://srdta.cochrane.org/
National Health Care Institute (Zorginstituut Nederland, formerly CVZ)	Medical tests (assessment of established medical science and medical practice)	Derksen J. 2011; CVZ Report 293. Available from http://www.zorginstituutnederland.nl/publicaties
GRADE Working Group.	Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies	Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, Helfand M, Ueffing E, Alonso-Coello P, Meerpohl J, Phillips B, Horvath AR, Bousquet J, Guyatt GH, Schünemann HJ; GRADE Working Group. Allergy. 2009 Aug;64(8):1109-16 Available from http://dx.doi.org/10.1111/j.1398-9995.2009.02083.x
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWIG)	Allgemeine Methoden Version 4.1. 2013	Available from http://www.iqwig.de
Medical Services Advisory Committee (MSAC)	Guidelines for the Assessment of Diagnostic Technologies	Canberra, Australia, Commonwealth of Australia, 2005. Available from http://www.msac.gov.au/
Adelaide Health Technology Assessment (AHTA)	Assessing Personalized Medicines in Australia: A National Framework for Reviewing Codependent Technologies	Tracy Merlin, Claude Farah, Camille Schubert, Andrew Mitchell, Janet E. Hiller and Philip Ryan. Med Decis Making. Apr 2013; 33(3): 333–342. Available from http://dx.doi.org/10.1177/0272989X12452341
National Institute of Health and Clinical Excellence (NICE)	Diagnostics Assessment Programme manual	December 2011. Available from http://www.nice.org.uk/
Institute of Medicine (IOM)	Evaluation of biomarkers and surrogate endpoints in chronic disease	Washington; DC: The National Academies Press, 2010. Available from http://www.iom.edu/Reports/2010/Evaluation-of-Biomarkers-and-Surrogate-Endpoints-in-Chronic-Disease.aspx
Ludwig Boltzmann Institute for Health Technology Assessment (LBI-HTA)	Evaluation von Diagnostika – Hintergrund, Probleme, Methoden	Nachtnebel A. HTA Projektbericht Nr 36. 2010. Available from http://eprints.hta.lbg.ac.at/898/
Centre for Reviews and Dissemination, University of York (CRD)	Chapter 2 Systematic reviews of clinical tests in: Systematic Reviews. The CRD's guidance for undertaking reviews in health care., 2009	Centre for Reviews and Dissemination, University of York, 2009. P.109ff Available from http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf

3 Terminology and classifications

3.1 Definition: Biomarker

NIH definition

According to the definition of the National Institute for Health (NIH) a biomarker is „a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention“ [13]. This broad definition allows for a large variety of biomarkers, from the analysis of small molecules up to the examination of physical parameters such as blood pressure. Table 3.1-1 presents an overview of the analytical technologies used in biomarker research and the biological processes, entities and examples of applications associated with each technology.

Table 3.1-1: Analytical technologies in biomarker research: overview of research areas, biological entities and processes studied; present and future applications

Research area	Biological entity	Biological Process	Examples of applications
Genomics	Gene characteristics (Variations in sequence, copy number, epigenetic modification)	Gene expression Gene function Gene regulation	Pharmacogenomics/-genetics: response to medication Nutrigenomics: effects of food and food constituents Epigenomics: effects of epigenetic modifications Metagenomics: study of communities of microbial organism (*e.g.* gut flora) Immunomics: study of immune system regulation and responses to pathogens
Transcriptomics	mRNA characteristics (Variations in sequence, expression levels, processing, splicing, editing)	Gene expression	Metatranscriptomics: study of communities of microbial organism (*e.g.* gut flora)
Proteomics	Proteins (Structure, posttranslational modifications)	Protein expression Protein-protein-interactions Protein function and activity Protein secretion	Immunoproteomics: study of proteomes in immune response Secretomics: analysis of the secreted proteins in a cell
Metabolomics	Metabolites (small molecules) Metabolic profile	Biochemical pathways	Toxicology: metabolic profiling of the response to toxic insult of a chemical or drug Nutrigenomics (see above) Lipidomics: analysis of lipid species within a cell or tissue
Imaging	Cell Tissue Organ	Growth rate, metabolic rate, Plaque formation Inflammation	Pharmacokinetics: analysis of time course of drug absorption, distribution, metabolism, and excretion. Pharmacodynamics: biochemical and physical effects of a drug on the body
Physical or physiological measurements	Whole body	various	Biomarkers of ageing (*e.g.* muscle mass, muscle strength) Multiple sclerosis biomarkers (*e.g.* walking capacity)

(Source: own presentation)

These technological categories however do neither align with functional differentiation of biomarkers nor with specific diseases (despite certain associations such as *e.g.* cancer – genomics or diabetes – metabolomics). The applications widely vary with regards to their stage of development: while a number of genomic, imaging and whole body biomarkers are already in use in clinical practice, many others areas are still in the stage of early basic research and hypothesis generation in association studies.

3.2 Differentiation of functions of biomarkers

differentiation by type of information

The methodology of HTA evaluation of a biomarker test is independent of the technological category of the biomarker with one exception: genetic/genomic biomarkers require a particular consideration as health and non-health related (*e.g.* ethical, social and legal) effects may affect more than one generation. To identify the outcomes relevant for the evaluation in a specific context of use, it is useful to differentiate biomarkers by the type of information they provide and the context of their use in clinical practice, detailed below [13-15]. As a principle, any biomarker is only useful if the test results are associated with appropriate differential treatment options.

Differentiation by type of information

- **Diagnostic** biomarkers are used to identify patients with a particular health condition, and to differentiate it from other conditions with similar symptoms, requiring differential treatment. Tests for diagnostic biomarkers may be used as replacement for time-consuming, expensive or invasive diagnostic procedures (e.g. replacement of echocardiogramm by testing levels of brain natriuretic peptides to rule out heart disease), or as an Add-on to existing methods to further refine the diagnosis. Diagnostic biomarkers can be used as screening markers to identify persons with an underlying disease in a screening population. (Example: elevated blood glucose concentration for the diagnosis of diabetes mellitus).
- **Prognostic** biomarkers predict the likely course of a disease in patients regardless of the treatment given. They can be used for risk stratification (triage) of patients based on their risk of disease progression to avoid expensive or invasive treatments and to ensure an optimal distribution of resources. Prognostic biomarkers may also be used in a screening population to identify people at risk to develop a specific health condition: these so-called risk biomarkers are indicative of a changed physiological state that is associated with a risk of disease. Several prognostic markers can be combined in a prediction model. (Example: fibrinogen for prognosis of primary stroke [16]).
- **Predictive** biomarkers (theranostic biomarker) predict the response of the patients to a specific treatment in comparison to the standard treatment, placebo or observation only (the biomarker (information) x treatment (effect) interaction). They are used for treatment stratification and guide choice of treatment [17]. (In vitro) Companion diagnostics are defined as "*in vitro diagnostics that provide information that is essential for the safe and effective use of a corresponding therapeutic*" (FDA Definition, [18]): here the use of a specific treatment is obligatorily preceded by a test for this predictive biomarker (Example: Trastuzumab/HER2 testing for breast cancer).

Differentiation by the context of use

differentiation by context of use

- **Screening** markers are used to detect disease in asymptomatic or presymptomatic persons and are used for prevention. They may provide diagnostic or prognostic information.
- **Aetiologic** markers are similar to prognostic markers in that both are risk factors for a specific outcome, but should be differentiated by the population they relate to: in prognosis all the population has the same disease/condition, in aetiology the marker serves to differentiate between individuals with/without the condition and to identify causal risk factors for a specific condition.
- **Monitoring** biomarkers are used for surveillance of the response to treatment (Example CEA and PSA for monitoring of tumor status during therapy and between image evaluations [19]).
- **Surrogate** endpoints are laboratory measurements or physical signs used in therapeutic trials as a substitute for a clinically meaningful endpoint. “A surrogate endpoint is expected to predict clinical benefit (or harm or lack of benefit or harm) based on epidemiologic, therapeutic, pathophysiologic, or other scientific evidence”[13].

> ⓘ For a more in-depth overview and discussion of definitions refer to:
> Surrogate biomarkers: [11], Chapter 1, p.17ff
>
> Diagnostic/Prognostic/Predictive biomarkers: [13, 14]

3.3 Medical tests

biomarkers must be objectively measurable

By definition a biomarker must not only represent a specific biological process, but it must also be objectively measurable. Thus, biomarkers are always associated with a corresponding medical test. The test must be applicable in clinical routine, reproducible and accurately translate the biomarker into a measurement parameter. Often several tests are available to measure the same biomarker – in this case evaluation not only takes into account the specific test performance characteristics but also costs and handling of the test might become decisive factors.

Medical tests can be differentiated by test methodology:

types of tests can be differentiated by test methodology

1. **Imaging**
   This category comprises Radiology (classic X-Ray), Sonography (Ultrasound), X ray, Computed tomography, magnetic resonance imaging, angiography, positron emission tomography, and other methods of nuclear medicine.
2. **Analysis of body fluids or smears**
   Most frequently analysed body fluids are blood, urine and cerebrospinal fluid; less common are synovial fluid, sweat, saliva and gastric juices. The analysis involves chemical or molecular biologic assays and cytologic examinations of cell smears or suspensions.
3. **Endoscopy**
   This consists in the investigation of interiors of organs or body cavities with an endoscope (viewing tube), that is introduced through a small incision or an existing body orifice (nose, mouth, anus, urethra, vagina).

4. **Measurement of body functions**
   Examples for this category are the measurement of blood pressure, the measurement of the electrical activity of the heart (Electrocardiography, ECG) or of the brain (Electroencephalography, EEG).
5. **Examination of biopsies**
   This consists in the removal of a tissue sample followed by the histological examination of the tissue. The analysis of the biopsy may also involve chemical or molecular biologic tests.

3.4 Evaluation framework

clinical outcomes of medical tests are induced indirectly

The purpose of a systematic review of any medical test is to identify and present evidence of its clinical utility: the health outcomes associated with its use [3, 5, 10, 14, 20]. Unlike the outcomes of therapeutic interventions, the clinical outcomes of medical tests are only in part directly induced by the test procedure but most of them indirectly by patient management decisions and treatments initiated according to the test results. The majority of the studies evaluating medical tests cover only segments of this path.

evaluation framework to identify and categorise key review questions

A number of frameworks have already been proposed for the evaluation of medical tests [20, 21]. Organising frameworks can and should be "used to categorise key questions and suggest which types of studies would be most useful for the review" [3, p.2-6]. These parameters can be mapped to the three-step evaluation process (Figure 3.4-1) proposed by IOM, USA, in 2010 [11]. The IOM framework shows the interdependency of the various steps, *i.e.* a change in technical parameters will require a re-evaluation of patient-relevant outcomes, and it specifically includes the contextual analysis and the assessment with regards to the specific context of use. Especially for diagnostics clinical practice of testing algorithms may strongly vary between countries. This process comprises three interrelated tiers, that can be assessed individually but need to be correlated to come to a final evaluation:

analytical validation

- Analytical validation describes the ability of a test to reliably and accurately measure a biomarker of interest, including limits of detection, limits of quantitation, reference value cut-off concentration, reliability and reproducibility.

qualification

- The qualification step comprises the actual evidentiary assessment of the association between biomarker and disease states. In the case of surrogate biomarkers the qualification step includes in addition assessment of evidence that interventions targeting the biomarker have an impact on health outcomes. Evidence on the impact on health outcomes can be provided by direct or by indirect evidence.

utilisation

- The third step of biomarker evaluation – utilisation – consists in the analysis of the evidence "with regard to the proposed use of the biomarker". In this step, evaluators should take into consideration the specific context of use of the biomarker with regards to target population, setting and purpose of the biomarker.

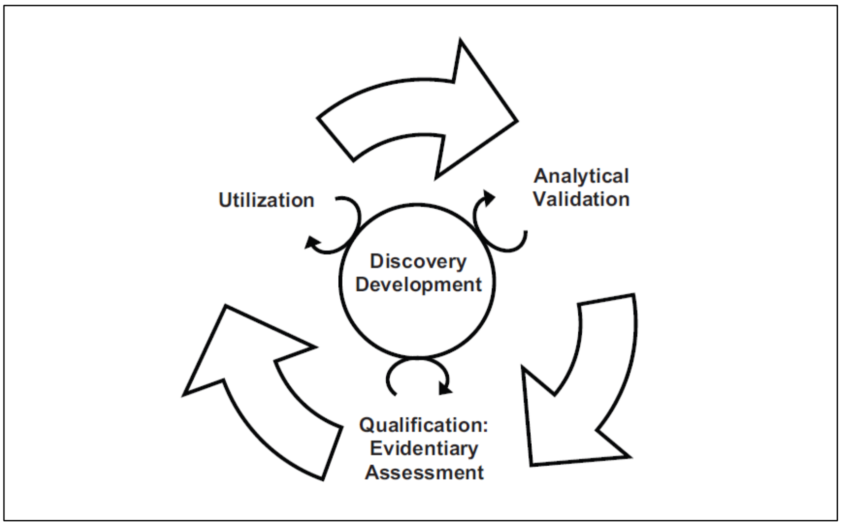

Figure 3.4-1: The three interdependent tiers of biomarker evaluation
Source: [11]

Steps in biomarker evaluation

Table 3.4-1 gives an overview of the characteristics evaluated in each step of the biomarker evaluation, with example parameters. Where varying terms were used in the literature, they are listed in brackets below each of the evaluation steps.

Table 3.4-1: Steps in the evaluation of biomarkers

Steps in Biomarker evaluation	Biomarker characteristics evaluated	Parameters (Examples)
Analytical validation (Technical efficacy, analytical validity)	Ability of a test/chemical assay to quantitate a biomarker of interest	Technical quality of a radiological image, reproducibility, repeatability
Qualification 1 (Test accuracy –diagnostic/prognostic accuracy, clinical validity)	Ability of a test to classify a patient into a disease, phenotype or prognosis category	Sensitivity, specificity, SROC curve
Qualification 2 (Clinical utility, therapeutic efficacy, patient outcome efficacy)	Ability of a test to improve patient outcomes	Changes in patient management, mortality, morbidity.
Utilization (Societal aspects, economic aspects)	Contextual analysis and risk-benefit assessment with regard to the proposed use	Opportunity costs

SROC – Summary Receiver Operating Characteristic

ⓘ For a more in-depth overview refer to:
Three steps in the biomarker evaluation process: [11] p.5ff
Review of different types of analytical frameworks: [21]

4 Context for analysis

4.1 Developing an analytical framework

more time needed to develop review question

A systematic review is the method of choice for the evaluation of medical tests. The same quality criteria as for systematic reviews of therapeutic interventions apply [2, 3, 12]. A specific challenge in the evaluation of medical tests is that more time needs to be dedicated to a careful definition of the review question to avoid ambiguity during literature search and literature selection and the evaluation of the selected studies.

clarify embedment of test in clinical routine and implications

Due to the indirect influence on patient outcomes, a systematic review of a medical test starts with clarifying the embedment of the test in clinical routine and the implications resulting thereof. A useful tool is to create an analytical framework (Figure 4.1-1 for a schematic representation), including decision making, further tests and treatments, patient outcomes and their surrogates [22]. This might include liaising with relevant experts and also by taking into account the indications of the manufacturer. Often there are several scenarios possible with varying positions of the test in the diagnostic chain and in a first step the scenario(s) relevant to the commissioner of the study need(s) to be clarified.

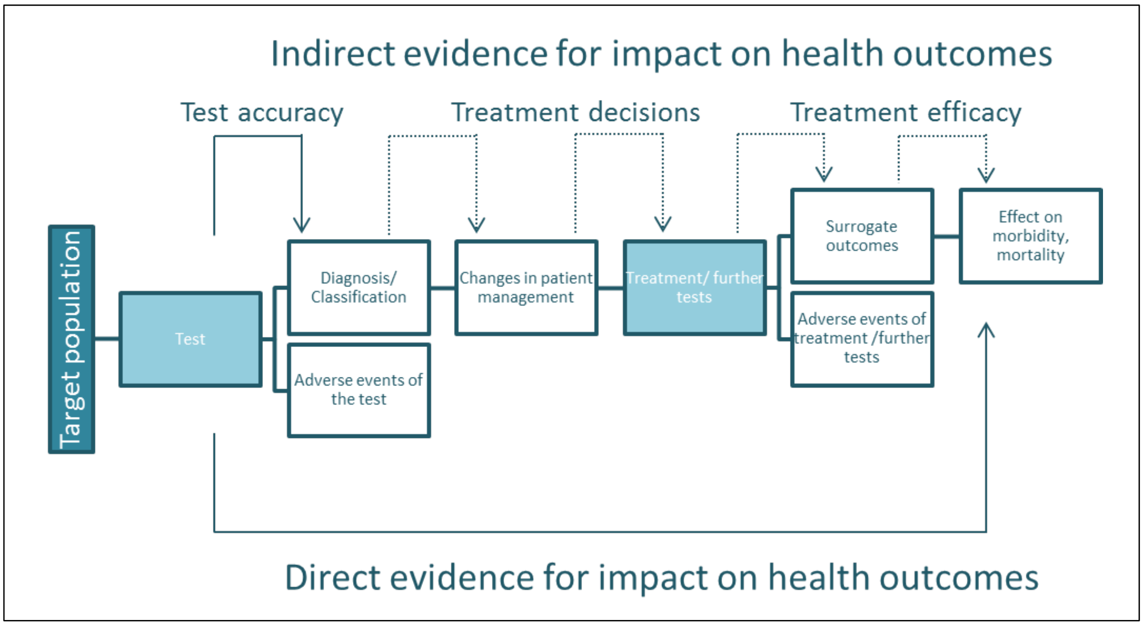
Indirect evidence for impact on health outcomes
Test accuracy
Treatment decisions
Treatment efficacy
Target population
Test
Diagnosis/ Classification
Adverse events of the test
Changes in patient management
Treatment/ further tests
Surrogate outcomes
Adverse events of treatment /further tests
Effect on morbidity, mortality
Direct evidence for impact on health outcomes

Figure 4.1-1: Analytical framework for the evaluation of medical tests
Source: (Adapted from [8] and [22])

take into account specific context

The evaluation of a medical test in a defined scenario has to take into account the specific context in which the test will be applied, as the context might influence test performance and vice-versa [3]. A widely accepted tool for the description of the context is the formulation of a PICO question, where P stands for Population, I for Intervention, C for Control intervention and O for Outcomes [2].

4.2 Formulation of review question with PICO

4.2.1 Population

specify target population

The review should clarify the target population of the test, *i.e.* to which patients the test is planned to be applied to. This includes information on

- demographic characteristics (age, sex, ...),
- medical history (prior diseases and treatments, co-morbidity, ...)
- the clinical setting in which the test will be used (inpatient, ambulant/outpatient, doctor's office, self-administration etc.) and
- the prevalence of the target disease in the target population.
- For prognostic markers: the observed probability (i.e. the observed proportion of an event in a given time period, [23, 24]) of the outcome being predicted.

From the prevalence of the disease, the reviewer may deduce the pre-test probability of the target health condition in the population. In combination with the diagnostic accuracy parameters, the pre-test probability allows to calculate the post-test probabilities.

The detailed description of the target population is further required for the assessment of the available studies with regards to the applicability of their results to the review question.

4.2.2 Intervention: new test- and treatment strategy

identify proposed test and treatment strategy

An essential characteristic of medical tests is that they are generally not standalone interventions, but embedded in a testing and treatment strategy, often with several sequential or parallel diagnostic tests and various treatment options according to specific combinations of test results, which will be influenced by the embedding of the new test. By formulating the review question therefore the entire sequence of tests including the new test and treatments should be considered as the intervention under review.

options to integrate a test in an existing strategy

Bossuyt *et al.* identified three options how to integrate a medical test in an existing testing and treatment strategy depicted in Figure 4.2-1: Replacement, Triage or Add-on [25]. The purpose of a replacement test is usually to maintain the same test performance (sensitivity and specificity) as the existing test, while increasing cost-effectiveness or reducing adverse events due to invasive procedures. Triage tests serve to avoid invasive or expensive procedures by decreasing unnecessary referrals. They should maintain the same sensitivity, since test-negatives of the new test will not be tested by the existing test, but may have lower specificity. Add-on tests finally serve to refine a diagnosis with the goal to improve treatment decisions and, thus, outcomes [26].

clarify preceding tests and patient management decisions with experts

Initial tests and patient management and treatment decisions need to be clarified with relevant experts and guidelines to adjust to the national context.

Description of the intervention should further include variants of the test, a definition of the cut-off point(s) and the timing of the application of the test (follow-up). Moreover the description of the test should include an assessment of infrastructure and processes necessary for implementation of the test in practice.

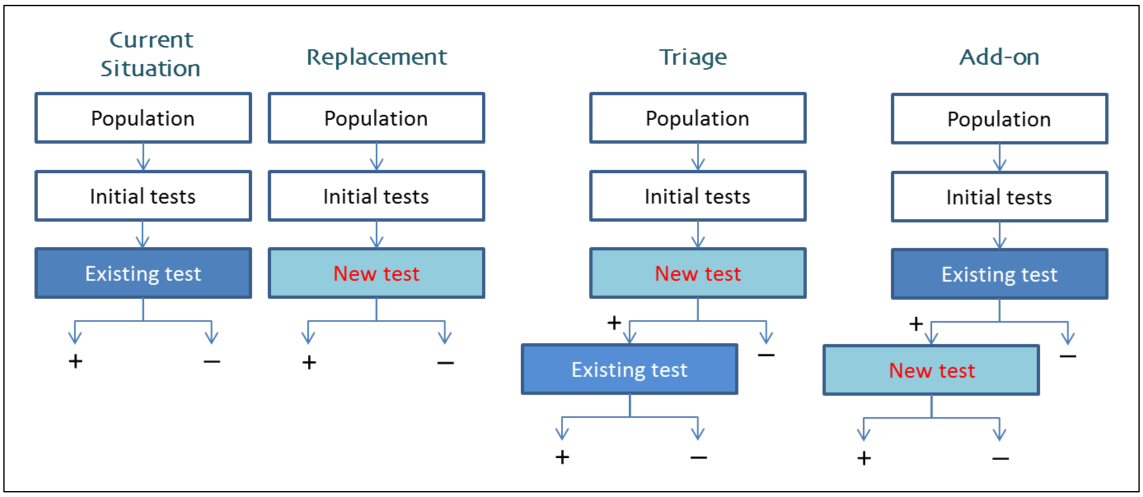

Figure 4.2-1: Options for integration of a medical test in an existing testing strategy (+ and – indicate positive or negative test results).
Source: [25]

In the assessment of biomarkers, the "intervention", however, does not necessarily include subsequent treatments: prognostic or etiologic marker may have as primary purpose to identify risk factors for a specific outcome.

In diagnostic accuracy studies, the test, whose performance is evaluated is called "index test".

4.2.3 Comparator: existing test and treatment strategy

As a principle, the comparator in reviews of medical tests is the current test and treatment strategy, *i.e.* the sequence of tests, patient management decisions and treatments.

comparator for the review: existing sequence of tests and treatment decisions

In diagnostic accuracy studies, the comparison of a new and an existing test may be direct: in fully paired direct comparisons, where all participants receive the index test and one or more comparator tests. As an alternative participants may be randomly allocated to receive the index or the comparator test (randomized direct comparison). The comparison is indirect, if the estimates of diagnostic accuracy from different study populations are compared [27].

diagnostic accuracy studies: direct or indirect comparison with other tests

The reference standard is used to define the target condition [28], *i.e.* to provide a classification of the disease for diagnosis or treatment decisions based on the best available evidence and to identify clinically relevant subgroups. The reference standard is the best available method to detect the target condition. Downstream management decisions and the impact of treatments may vary in dependence of the chosen target condition and thus, the choice of the most appropriate reference standard for the review question should be clarified by involving clinical guidelines and possibly experts. [28].

reference standard: used to define target condition in diagnostic accuracy studies same test cannot be used as reference standard and comparator test in diagnostic accuracy studies

It is not possible to use the same test as reference standard and comparator test in a diagnostic accuracy study: diagnostic accuracy calculations are based on the assumption that the reference standard is "perfect", *i.e.* capable of identifying "true" cases of disease and non-disease with 100% accuracy. This is problematic in circumstances where, in fact, the reference standard is poor and in cases where no reference standard is available. In these cases naïve estimates[1] of diagnostic accuracy are likely to be biased and are unsuitable to substitute valid measurements of clinical outcomes [29]. In the absence of a perfect reference standard, the best way to determine diagnostic effectiveness would be a trial. A trial would also be required, if the index test is expected to have superior diagnostic accuracy to the current reference standard.

4.2.4 Outcomes

selection and categorisation of relevant outcomes for the review

Accurate diagnosis is a prerequisite for a successful therapy, but it should not be seen in isolation. Instead, the benefit to patients resulting from diagnosis should be measured in patient-relevant outcomes [30], such as survival (mortality), clinical events, adverse events, patient-reported outcomes (health related quality of life), activity and function.

Based on the analytical framework developed, the reviewer should first explore all outcomes resulting from embedment of the test in the testing and treatment strategy in comparison to clinical practice without the test (as described in 'Intervention' and 'Comparator'). Reviewers should then make a careful selection of the relevant outcomes both to the process of testing and to the results of the test [31] by mapping them according to the following categories [31, 32][2]:

1. Clinical management effects due to testing
2. Direct health effects of testing
3. Emotional, social, cognitive, behavioural responses to testing
4. Legal and ethical effects of testing

dependent of type of test and purpose of the review

A decision which outcomes are relevant for a review depends on the type of test under review and on the needs of the stakeholders of the study [31]. A review assessing the inclusion of a test in the benefits catalogue of health insurances or hospital interventions might be more restricted on outcomes directly affecting the patient, while a review serving medical guideline development or the choice of a screening algorithm needs to take into account the outcomes on a societal level. As a consequence the prioritization of the relevant outcomes should involve the commissioner of the study. The outcomes are decisive only if they differ between current and new testing and treatment strategy.

The outcomes should explicitly be rated by importance *a priori [30]*.

[1] Naïve estimates of diagnostic accuracy: parameters like sensitivity, specificity or positive and negative predictive value are calculated using 2x2 tables in which the positive and negative results of the reference standard are assumed to be the true numbers of sick and healthy, respectively.

[2] Only outcomes within the scope of this guidance were included (notably, costs were excluded).

Clinical management effects due to testing

expected consequences in TN, FN, TP, FP groups

This category describes the clinical consequences (in patient relevant outcomes: mortality, morbidity and quality of life) that the use of a test will induce. This includes expected consequences based on a negative (true negative, TN and false negative, FN) or positive (true positive, TP and false positive, FP) test result, as supported by primary literature on therapy decisions and outcomes in the particular setting under review. Further consequences might be induced by unexpected findings: they are particularly prominent in imaging methods.

Clinical management effects might be of particular importance in the evaluation of diagnostic and prognostic tests [31].

The desired management effects define which test performance parameters are of greatest importance: a test used to rule out the presence of a disease or a high risk for a disease should have high sensitivity and negative predictive value, NPV; a test used as "Add-on" to refine diagnosis should have high specificity or positive predictive value, PPV.

Direct health effects of testing

relevant for invasive procedures

This category relates to the health effects that are directly induced by the test procedure and might be particularly relevant for invasive procedures or procedures that involve radiation, while other forms of testing (*e.g.* a vaginal swab) most likely will not have any health consequences.

Emotional, social, cognitive, behavioral responses to testing

e.g. change of behaviour

Emotional responses might include relief or anxiety as a consequence of a test result – outcomes of this type might be particularly relevant in screening or prognostic tests. Test results might induce a change in behaviour in the testees – for example opting for a healthier lifestyle to compensate a high-risk prognosis or sustaining an unhealthy lifestyle as a consequence of a low-risk prognosis. Emotional responses might also be related to the test procedure itself – for example psychological symptoms following colonoscopy [33]. The effects might also extend to family members. Social issues such as stigmatization, discrimination and privacy/confidentiality should also be considered. Genetic tests might have complex impact on behaviour, *e.g.* regarding family planning.

Legal and ethical effects

e.g. test of reportable diseases

Legal consequences might arise in the case of reportable diseases (to be reported by the health care provider) or diseases representing a safety threat in certain professions (pilots, surgeons, gastronomy ...), which warrant disclosure to the employer. Besides disclosure or reporting requirements, legal issues might involve consent, ownership of data and/or samples, patents, licensing, proprietary testing.

Genetic tests have a special status with regards to the ethical and legal effects because of the possible impact on family members. This depends on whether the testing is for inherited or acquired genetic mutations and the inheritance pattern of the trait.

select relevant outcomes

Based on the inventory of outcomes the reviewer will need to choose the relevant outcomes for the review, depending on time and resources available and the intended purpose of the review.

> ⓘ A checklist for assessing the context of submission and the proposed impact of a biomarker test/technology on current clinical practice is proposed in [9](see also Annex).
>
> The ACCE model proposes a checklist of targeted questions specifically for genetic testing [34].

5 Identifying the evidence base

systematic literature search

To identify the evidence base relevant to a particular research question, a systematic literature search is conducted to identify all studies relevant to the research question and the evidence base is established based on the quantity and quality of the studies included [35].

5.1 Literature search and selection

studies on diagnostic tests poorly indexed: other search strategies need to be applied

Several common challenges are associated with literature search for medical tests with the following key points [36]:

1. Due to still underdeveloped indexing and reporting of studies of diagnostic tests, literature search should not rely (exclusively) on diagnostic search filters, in particular these filters are inappropriate for systematic review of clinical effectiveness.
2. If the name(s) of the diagnostic test(s) relevant for the research question is not known, search strategies should capture the "concept of diagnostic tests"[3].
3. To identify all studies for a systematic review, searches should include text words (not subject headings alone) and be combined with hand search including additional sources of information: specialised databases, citation tracking and regulatory documents ([37]).

ⓘ A thorough description of "Effective search strategies for systematic reviews of medical tests" is given by Relevo *et al.* in the AHRQ Methods Guide for Medical Test Reviews [36], p.4-1ff

5.2 Hierarchies of evidence/Study designs

classical intervention: established evidence hierarchy

For classical intervention research questions (therapeutic effectiveness), a hierarchy of evidence has been established and is widely accepted based on the degree of bias associated with observational and non-randomised studies in comparison to randomised controlled trials [38-41]: this hierarchy attributes the highest level of evidence (Level I) to systematic reviews and meta-analyses of RCT and Level II to evidence obtained from at least one (properly designed) RCT. Level III and IV subsequently refer to non-randomised comparative studies and case series, respectively. Similarly, according to GRADE (Grading of Recommendations Assessment, Development and Evaluation), only RCTs are *a priori* considered to provide high quality evidence about treatment effects [42].

[3] *E.g.* diagnosis OR diagnose OR diagnostic OR di[sh] OR "gold standard" OR "ROC" OR "receiver operating characteristic" OR sensitivity and specificity[mh] OR likelihood OR "false positive" OR "false negative" OR "true positive" OR "true negative" OR "predictive value" OR accuracy OR precision.

Study designs of biomarker studies may vary from classical intervention studies. This chapter is therefore meant to help reviewers to categorise the studies identified during literature search.

As described in Chapter 3.4 – Evaluation framework, in order to fulfil the qualification step, evaluators first perform an evidentiary assessment on the causal relationship between biomarker and disease pathogenesis and second an assessment of the evidence that interventions targeting the (surrogate) biomarker impact the health outcome of interest [11].

assessment may include other research questions: prognosis, aetiology, diagnostic accuracy

In principle, the most appropriate study design to evaluate the impact of a biomarker test on clinical management effects and health outcomes is a randomised controlled trial (RCT) [43, 44]. However the assessment of biomarkers does not necessarily include a classical intervention research question (therapeutic effectiveness), but instead may include questions on prognosis, aetiology, or diagnostic accuracy, depending on the context of use. For some of these questions the only evidence feasible and/or ethical will be from observational studies and different evidence hierarchies may apply [11, 43, 44]. To reflect this, revised evidence hierarchies have been elaborated by the NHMRC [43] and the Oxford Centre for Evidence-Based Medicine [44], which should be considered when prioritising available evidence. Table 5.2-1 presents an amalgamation of both evidence hierarchies. In all instances the highest level of evidence is provided by a systematic review of level II studies.

5.2.1 Diagnostic accuracy

Diagnostic accuracy studies are cross-sectional by nature. Study designs are differentiated in „Single-gate“ (diagnostic cohort study[4]) and „Two-gates“ (diagnostic case-control studies[4]) studies.

"single-gate" vs. "double-gate" study designs

In „Single-gate“ studies all study participants are first tested with an index test and then with the reference standard. Provided, all participants undergo both tests, the sequence of testing may be reversed (Reversed Flow Design), without influencing estimates of diagnostic accuracy [45]. "Two-gates" studies make comparisons between participants with confirmed disease/condition and healthy participants. The "Two-gate" design is intrinsically prone to spectrum bias, potentially leading to inflated estimates of the diagnostic accuracy [45, 46]. Quality assessment should identify if a "Two-gate" study represents only a limited spectrum of disease and non-disease and if so, omit the study from the meta-analysis (if done) [47].

one or more comparator tests

Diagnostic accuracy studies may include one or more comparator tests to which the index test is compared. In fully paired direct comparisons, all participants receive index test, comparator test(s) and reference test. As an alternative, participants may be randomly allocated to receive the index or the comparator test (randomized direct comparison). The comparison is indirect, if the estimates of diagnostic accuracy from different study populations are compared[27].

The identification of sources of bias in diagnostic accuracy studies requires a detailed assessment of the study design, including criteria on reference standard, study population and blinding.

[4] For Diagnostic accuracy studies, the terms cohort and case-control study relate to the inclusion modes only – in contrast to conventional case-control and cohort studies, they are always cross-sectional with the aim to determine a status quo at one timepoint.

ⓘ **When can diagnostic accuracy be used as surrogate for health outcomes?** [38, 48]

- The index test has similar sensitivity as the comparator test but other positive attributes such as higher specificity, lower costs, fewer adverse events or being less invasive: the value of the test corresponds to the benefits of avoiding adverse events or costs associated with the comparator test.
- The index test has higher sensitivity and similar specificity than the comparator test and the extra cases detected by the new, more sensitive index test represent the same spectrum of disease (size, grade, severity) or a same definition of disease for which treatment response is known. This condition may e.g. be fulfilled if in clinical diagnostic routine, test cases are subsequently confirmed by the reference standard (linked-evidence approach[5]).
- The index test has higher sensitivity and similar specificity than the comparator test and treatment response of the extra cases detected by the index test has been shown in trials (linked-evidence approach[5]).

If the index test is less sensitive or less specific than the comparator test but has other positive attributes, assessing the trade-off of benefits and harms of using the index test will require direct evidence from an RCT [48].

RCTs are also needed if no or only a poor reference standard is available to determine diagnostic accuracy or if the index test is expected to perform better than the current reference standard, which by definition cannot be demonstrated in diagnostic accuracy studies.

5.2.2 Prognosis and Aetiology

prognosis:
likelihood to experience
a medical event

A prognostic test is used to predict a patient's likelihood to experience a medical event (disease development or progression), within a defined time interval and using the observed proportion of the population experiencing this event as reference.

Biomarkers are indicative of a physiological state and therefore not necessarily causal. Cross-sectional studies do not allow for causal inferences to be made since in these studies biomarker-disease measurements occur simultaneously. In order to show causality and hence, prognostic value, prospective cohort studies are required that follow health outcomes over time in a population characterised by the levels of the biomarker [11].

adaptation of the
GRADE framework

Huguet *et al.* recently proposed an adaptation of the GRADE framework to research on prognostic factors in which they suggest to consider the phase of investigation in the ranking of evidence: a high level of evidence for prognosis would be provided by prospective or retrospective cohort studies that test a fully developed hypothesis and conceptual framework on the underlying processes for the prognosis of a health condition[49]. Studies in an early stage of investigation, to generate hypotheses, would be attributed a moderate level of evidence.

[5] See Section 5.2.2

Table 5.2-1: Evidence hierarchies by research questions

Evidence Level		Intervention	Diagnostic Accuracy[6]	Prognosis	Screening	Aetiology[7]
Highest[8]	I	Systematic review of Level II studies	Systematic review of Level II studies	Systematic review of Level II studies	Systematic review of Level II studies	Systematic review of Level II studies
High	II	RCT	Diagnostic accuracy study[9] Independent blinded comparison Valid reference standard Consecutive patient sample Defined clinical presentation	Prospective (inception[10]) cohort studies (Phase 2 or Phase 3 explanatory studies)	RCT	Prospective cohort study
Moderate	III	Non randomised controlled trial/cohort/ follow-up study	Diagnostic Accuracy Study not meeting the criteria for level II; Diagnostic case-control study	Cohort study (Phase 1 explanatory study) or control arm of randomised trial	Non randomised controlled trial/cohort/ follow up study	Retrospectiv e cohort study or case-control study
Low	IV	Case series, case-control, or historically controlled studies	Diagnostic Accuracy Study with poor reference standard; study of diagnostic yield	Case series or case control studies	Case series, case-control, or historically controlled studies	Cross-sectional study or case series

Source: (adapted from [43, 44, 49])

[6] This column only applies to reviews assessing diagnostic accuracy. For the evaluation of the impact of a diagnostic test on health outcomes, the "intervention" column should be used.

[7] This column should only be used if a causal relationship cannot be determined using RCT. Otherwise the "intervention" column should be used.

[8] The highest evidence level applies only to SR of Level II studies. SR of studies from other evidence levels should not be ranked higher than the levels of the studies included.

[9] The fulfilment of these criteria can be determined through quality appraisal of the diagnostic accuracy study using the QUADAS-2 tool [54]

[10] All persons in an inception cohort are non-diseased or in the same status of the disease.

5.2.3 Clinical trial designs

The evaluation of biomarkers used for treatment selection necessarily requires a randomised design to isolate the effect of the marker on therapeutic efficacy from all the other factors influencing a treatment choice [50]. To this end, new trial designs have been proposed, with substantial variability in the labelling of the trial designs [51]. Trial designs can be classified according to the patient flow in the studies; each category allowing assessing different effects (Figure 5.2-1).

isolate the effect of the marker on therapeutic efficacy from all the other factors influencing a treatment choice

Targeted or Enrichment designs [51, 52]

Patients are screened for the presence or absence of the marker and only the subgroup of patients defined by a specific marker status is studied (Example: HER2/Trastuzumab trial, [53]). The study is powered to detect a clinically meaningful effect in the marker-positive subgroup, but does not provide information on the treatment benefit in the marker-negative group. It therefore does not allow to answer the question, whether costs and inconvenience associated with biomarker-based treatment allocation is worthwhile [52]. Furthermore causality of any association of treatment effect with biomarker status may not be established, as there is no comparison with a biomarker negative group (Table 5.2-2).

only the subgroup of patients defined by a specific marker status is studied

Alternatively, marker-negative patients may be assigned a control treatment – this form of hybrid design was used for example in the TAILORx trial for the evaluation of Oncotype DX [54].

Marker by treatment interaction design or tests at baseline in RCT [17, 55, 56]

All patients are randomly assigned to treatments based on a randomisation that may or may not be stratified based on biomarker status. This is the most efficient trials design in situations in which there are two or more existing treatment options with no definitive evidence for one being preferred in a given population [57].

all patients randomly assigned to treatments

randomisation may or may not be stratified based on biomarker status

This design allows the embedment of the evaluation of various diagnostic strategies in classical intervention studies. To this end, either before or after randomisation, all study participants are tested with one or more diagnostic test strategies (*e.g.* biomarker-based and non-marker based/standard). Then all participants are randomised to either of two treatments: this randomisation should ideally be blinded and independent of test results. In principle this study design may also be performed retrospectively in archived samples of a classical intervention study of the treatment under consideration, provided that the standard diagnostic treatment decision can be determined retrospectively [56]. Retrospective stratification, however, involves a higher risk of bias by confounding. A stratified randomisation by biomarker will ensure appropriate power of the study and minimise selection bias. Because all patients are randomised to both treatment and control, this study design is also designated "randomize-all".

"Randomise-all"

With randomisation stratified according to biomarker status, this study design allows to assess the relationship between the test (biomarker) and the treatment, *i.e.* whether the biomarker is predictive (modifier of treatment effect) or prognostic (favourable outcome in marker-positive patients regardless of treatment) (Table 5.2-2).

Biomarker strategy designs [17, 58]

patients are randomised to either a biomarker-based or a biomarker-independent strategy for allocation of treatment

In this design, marker status is first determined in all patients, and then patients are randomised to either a biomarker-based or a biomarker-independent strategy for allocation of treatment. The biomarker-independent strategy may be a standard diagnostic pathway or a randomised treatment allocation. This type of trial design allows to directly comparing the outcome of all patients in the marker-based arm to the outcome of the patients in the non-marker based arm. The efficiency of this trial design is limited by the fact that in many cases treatment choices by either diagnostic strategy would be the same for a large number of the patients, reducing the potential observable differences between the groups and thus increasing the required participants number [50]. This study design allows to directly comparing the potential of two diagnostic strategies to differentiate between patients likely to profit and patients unlikely to profit of a specific treatment. This study design does not allow to assess the relationship between the test (biomarker) and the treatment: *i.e.* whether the biomarker is predictive (modifier of the treatment effect) or prognostic (favourable outcome in marker-positive patients regardless of treatment) [59] (Table 5.2-2).

RCT of discordant test results[55, 56]

all participants are tested with both strategies, only those with discordant test results are randomised to the treatment options

This is an adaptation of the marker-based strategy design, enriching on the fraction of participants that receive discordant results using the two diagnostic strategies, sparing a treatment randomisation to those in which both diagnostic strategies come to the same treatment decision.

All participants are tested with both strategies, only those with discordant test results are randomised to the treatment options under consideration. In both groups (new +/standard –) and (new –/standard +) superiority (or: non-inferiority) of the treatment chosen by the biomarker based strategy over the treatment option chosen by standard diagnosis needs then to be demonstrated [56] (Table 5.2-2).

Double-randomised controlled trial

two randomisation steps

This design combines a randomisation to the testing strategy, similar to the biomarker-strategy design, with a randomisation to the treatment in both biomarker positive and negative groups to allow to explain the biomarker-drug relationship (*i.e.* predictive vs. prognostic factor) [9] (Table 5.2-2). This design poses practical limitations, especially if the biomarker is uncommon. Ethical challenges might arise if a new treatment is tested as a replacement for an old therapy and an effect is only plausible in biomarker-positive patients: in this case it might be considered unethical to randomise biomarker-negative patients to a treatment, where no effect is to expected, while forgoing an effective treatment [9], which is also of concern for the marker by treatment interaction design.

Table 5.2-2: List of effects that can be assessed and questions that can be answered by the trials of each design category

		Marker by treatment	Biomarker-strategy		
Questions trial can answer	Enrichment	With randomisation stratified by biomarker status	With treatment randomization in the control arm	RCT of discordant results	Double RCT
Treatment effects					
Q1. How does the experimental treatment compare with the control treatment in biomarker-positives?	✓	✓	✓	–	✓
Q2. How does the experimental treatment compare with the control treatment in biomarker-negatives?	–	✓	✓	–	✓
Q3. How does the experimental treatment compare with the control treatment in overall study population?	–	✓	✓	–	✓
Biomarker effects					
Q4. Is the biomarker status associated with the outcome in the standard of care group? (Is the biomarker prognostic?)	–	✓	✓	–	✓
Q5. Is the biomarker status associated with the outcome in the experimental treatment group?	–	✓	✓	–	✓
Biomarker by treatment effect					
Q6. Is the biomarker status associated with a benefit of experimental treatment? (Is the biomarker is predictive?)	–	✓	✓	–	✓
Strategy effects			✓		✓
Q7. How does the biomarker-based treatment strategy compare with the control treatment in the overall study population?	–	✓ Indirect	✓	✓	✓
Q8. How does the biomarker-based treatment strategy compare with the experimental treatment in the overall study population?	–	✓ Indirect	✓	✓	✓

Source: (adapted from [51])

ⓘ For a classification of varying labels used by study authors, see [51].

For a comparison of which research questions can be answered by which trial design, see [57] or [51].

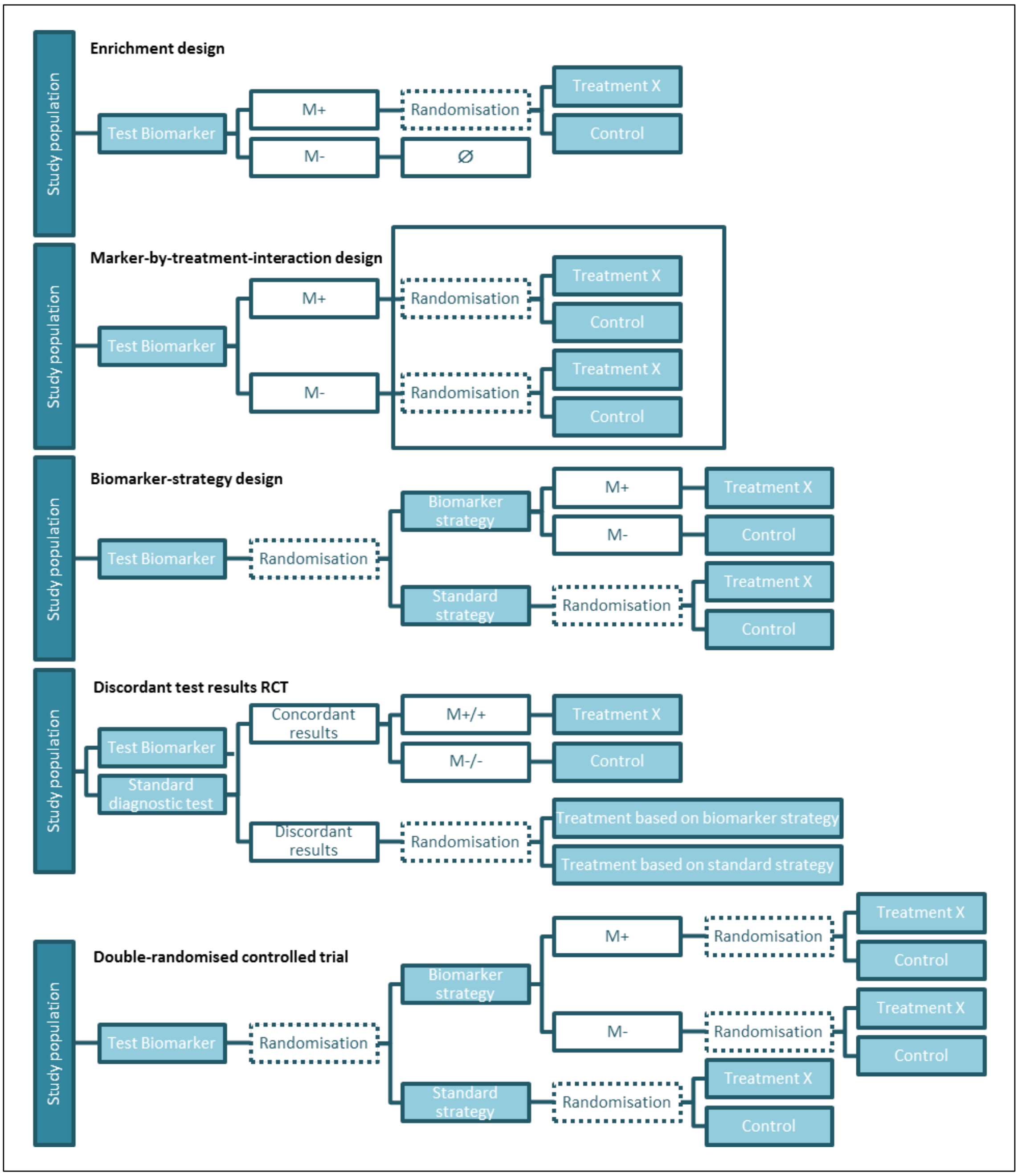

Figure 5.2-1: Randomised clinical trial designs for biomarker evaluation

Source: [17, 51, 56, 59]

6 Critical appraisal

6.1 Internal validity

how well the design and conduct of the study minimises potential biases

Internal validity describes how well the design and conduct of the study minimises potential biases, *i.e.* the risk of bias resulting from "study limitations" [60] or "study quality". Other possible sources of bias, such as inadequate reporting or publication bias are discussed later. Different checklists apply to the assessment of internal validity of different study designs. Reviewers should define *a priori* the relevant quality criteria. Explicit rules for rating each quality criterion adapted to the specific review question and definitions for ranking the overall risk of bias as "High", "Low" or "Unclear" should be provided. The appraisal should not follow an automatised scoring process using predefined values for each criterion or a point system, rather risk assessment will require a thoughtful approach about the methodological issues and the direction of the potential bias [2].

Internal validity criteria may come into play at several time points during the assessment: either as inclusion/exclusion ("fatal flaw") criteria in the initial or a secondary screening of eligible articles, or as exclusion criteria before proceeding to the evidence synthesis, or quality criteria may be used to test the association between quality and outcomes. Adequate reporting is conditional for quality appraisal, but it is in itself not yet a quality criterion [60].

Standard internal validity critical appraisal checklists for randomised controlled trials and cohort studies for interventions may be used [2]. In the following we will present quality checklists for the special cases of diagnostic accuracy and prognosis studies.

6.1.1 Diagnostic accuracy studies

quality checklist for diagnostic accuracy: QUADAS-2 to assess bias introduced by …

Diagnostic accuracy relates to the strength of association between the results of the index test and a reference standard (or „gold standard", meaning the best available diagnostic approach).

All sources of bias particularly common to diagnostic accuracy studies have been comprehensively reviewed by Whiting *et al.* [61] and, based on these results, the QUADAS-2 checklist has been developed to assess the main empirically validated sources of bias in Diagnostic accuracy studies [62] (Table 6.1-1).

Table 6.1-1: QUADAS-2 Checklist for assessment of Risk of Bias in Diagnostic Accuracy Studies

Domain 1: Patient selection	
Was a consecutive or random sample of patients enrolled?	Yes □ No □ Unclear □☐
Was a case-control design avoided?	Yes □ No □ Unclear □
Did the study avoid inappropriate exclusions?	Yes □ No □ Unclear □
Could the selection of patients have introduced bias?	**Risk: Low □ High □ Unclear □**
Domain 2: Index test(s) (complete for each index test used)	
Were the index test results interpreted without knowledge of the reference standard?	Yes □ No □ Unclear □
If a threshold was used, was it pre-specified?	Yes □ No □ Unclear □
Could the conduct or interpretation of the index test have introduced bias?	**Risk: Low □ High □ Unclear □**
Domain 3: Reference Standard	
Is the reference standard likely to correctly classify the target condition?	Yes □ No □ Unclear □
Were the reference results interpreted without knowledge of the results of the index test?	Yes □ No □ Unclear □
Could the reference standard, its conduct, or its interpretation have introduced bias?	**Risk: Low □ High □ Unclear □**
Domain 4: Flow and Timing	
Was there an appropriate interval between index test(s) and reference standard?	Yes □ No □ Unclear □
Did all patients receive a reference standard?	Yes □ No □ Unclear □
Did all patients receive the same reference standard?	Yes □ No □ Unclear □
Were all patients included in the analysis?	Yes □ No □ Unclear □
Could the patient flow have introduced bias?	**Risk: Low □ High □ Unclear □**

Source: [62]

... patient selection

We have previously explained that a “Two-gate” study design, or “diagnostic case-control” study design is prone to spectrum bias, which is reflected in the questions regarding “Patient selection”, in this case reviewers would need to estimate if despite the case control design the full spectrum of disease and non-disease is represented.

... flow and timing

The appropriate interval between index test(s) and reference standard needs to be defined in context with the review question. There is risk of bias, if:

- Interventions have been initiated between index test and reference standard and it cannot be excluded that they have (already) influenced the condition of the participant at the time of conduct of the reference standard.
- It cannot be excluded that the condition of the participant has significantly improved (*e.g.* clearance of an infection) or worsened (*e.g.* tumour progression) in the time interval chosen
- It cannot be ascertained that the reference standard can detect the condition after the time interval chosen (*e.g.* motor symptoms of multiple sclerosis)

… reference standard and index test

Per definition diagnostic accuracy is inferred from the comparison of the index test with a reference test, which for this purpose is considered „ideal“, *i.e.* able to detect "true" presence or absence of disease. As a consequence of this assumption, estimates of diagnostic accuracy are not possible or potentially biased if

- No reference standard is available
- Available reference standards are known to be imperfect
- Not all participants receive the same reference standard (Differential verification bias)
- Not all participants receive a reference standard (Partial verification bias)
- Interpretation of the results of index test or reference test is not blinded to the results of the other test respectively (Review bias).

indeterminate results need to be classified

There is no consensus how to process indeterminate results. Often these results are either excluded or classified as positive – both procedures may lead to overestimation of test performance. It is recommended to classify indeterminate results by "intention to diagnose": *i.e.* false negative if the reference standard result is positive and false positive if the reference standard result is negative [63, 64].

Table 6.1-2: Classification of indeterminate results by "Intention to Diagnose"

		Reference standard	
		+	-
Index test	+	True Positive A	False Positive B
	?	Indeterminate (RS Positive) ↓	↑ Indeterminate (RS Negative)
	-	False Negative C	True Negative D

Source: [63, 64]

6.1.2 Prognosis studies

quality checklist for prognosis studies: QUIPS

In a systematic review Hayden *et al.* have identified quality items used in assessing 6 sources of bias in prognosis studies: study participation, study attrition, measurement of prognostic factors, measurement of and controlling for confounding variables, measurement of outcomes and analysis approaches [65].

Based on this work, a team of epidemiologists, statisticians and clinicians developed the QUIPS (Quality in Prognosis Studies) tool for assessing bias in studies of prognostic factors which demonstrated acceptable reliability [66] and so far is the only tool available for this purpose (Table 6.1-3).

The authors reported reasonable interrater agreement scores for the bias domains 1-4 and 6, but a κ statistic of only 0.4 for the domain 5 (Study confounding). As a consequence they recommend operationalising the tool *a priori* by including specifying key characteristics and omitting items irrelevant for the review question (marked with LIST in Table 6.1-3).

Table 6.1-3: QUIPS tool for the assessment of bias in studies of prognostic factors

Biases	Issues to consider for judging overall rating of "Risk of bias"	
Domain 1: Study Participation		
Goal: To judge the risk of selection bias (likelihood that relationship between prognostic factor and outcome is different for participants and eligible non-participants).		
Source of target population	The source population or population of interest is adequately described for key characteristics (LIST).	Yes □ Partial □ No □ Unclear □
Method used to identify population	The sampling frame and recruitment are adequately described, including methods to identify the sample sufficient to limit potential bias (number and type used, *e.g.*, referral patterns in health care)	Yes □ Partial □ No □ Unclear □
Recruitment period	Period of recruitment is adequately described	Yes □ Partial □ No □ Unclear □
Place of recruitment	Place of recruitment (setting and geographic location) are adequately described	Yes □ Partial □ No □ Unclear □
Inclusion and exclusion criteria	Inclusion and exclusion criteria are adequately described (*e.g.*, including explicit diagnostic criteria or "zero time" description).	Yes □ Partial □ No □ Unclear □
Adequate study participation	There is adequate participation in the study by eligible individuals	Yes □ Partial □ No □ Unclear □
Baseline characteristics	The baseline study sample (*i.e.*, individuals entering the study) is adequately described for key characteristics (LIST).	Yes □ Partial □ No □ Unclear □
Summary Study participation	**The study sample represents the population of interest on key characteristics, sufficient to limit potential bias of the observed relationship between PF and outcome.**	**Risk: High □ Moderate □ Low □**
Domain 2: Study Attrition		
Goal: To judge the risk of attrition bias (likelihood that relationship between PF and outcome are different for completing and non-completing participants).		
Proportion of baseline sample available for analysis	Response rate (*i.e.*, proportion of study sample completing the study and providing outcome data) is adequate.	Yes □ Partial □ No □ Unclear □
Attempts to collect information on participants who dropped out	Attempts to collect information on participants who dropped out of the study are described.	Yes □ Partial □ No □ Unclear □
Reasons and potential impact of subjects lost to follow-up	Reasons for loss to follow-up are provided.	Yes □ Partial □ No □ Unclear □
Outcome and prognostic factor information on those lost to follow-up	Participants lost to follow-up are adequately described for key characteristics (LIST).	Yes □ Partial □ No □ Unclear □
	There are no important differences between key characteristics (LIST) and outcomes in participants who completed the study and those who did not.	Yes □ Partial □ No □ Unclear □
Study Attrition Summary	**Loss to follow-up (from baseline sample to study population analysed) is not associated with key characteristics (*i.e.*, the study data adequately represent the sample) sufficient to limit potential bias to the observed relationship between PF and outcome.**	**Risk: High □ Moderate □ Low □**
Domain 3: Prognostic Factor Measurement		
Goal: To judge the risk of measurement bias related to how PF was measured (differential measurement of PF related to the level of outcome).		
Definition of the PF	A clear definition or description of 'PF' is provided (*e.g.*, including dose, level, duration of exposure, and clear specification of the method of measurement).	Yes □ Partial □ No □ Unclear □
Valid and Reliable Measurement of PF	Method of PF measurement is adequately valid and reliable to limit misclassification bias (*e.g.*, may include relevant outside sources of information on measurement properties, also characteristics, such as blind measurement and limited reliance on recall).	Yes □ Partial □ No □ Unclear □
	Continuous variables are reported or appropriate cut-points (*i.e.*, not data-dependent) are used.	Yes □ Partial □ No □ Unclear □

Biases	Issues to consider for judging overall rating of "Risk of bias"	
Method and Setting of PF Measurement	The method and setting of measurement of PF is the same for all study participants.	Yes □ Partial □ No □ Unclear □
Proportion of data on PF available for analysis	Adequate proportion of the study sample has complete data for PF variable.	Yes □ Partial □ No □ Unclear □
Method used for missing data	Appropriate methods of imputation are used for missing 'PF' data.	Yes □ Partial □ No □ Unclear □
PF Measurement Summary	**PF is adequately measured in study participants to sufficiently limit potential bias.**	**Risk: High □ Moderate □ Low □**
Domain 4: Outcome Measurement		
Goal: To judge the risk of bias related to the measurement of outcome (differential measurement of outcome related to the baseline level of PF).		
Definition of the Outcome	A clear definition of outcome is provided, including duration of follow-up and level and extent of the outcome construct.	Yes □ Partial □ No □ Unclear □
Valid and Reliable Measurement of Outcome	The method of outcome measurement used is adequately valid and reliable to limit misclassification bias (*e.g.*, may include relevant outside sources of information on measurement properties, also characteristics, such as blind measurement and confirmation of outcome with valid and reliable test).	Yes □ Partial □ No □ Unclear □
Method and Setting of Outcome Measurement	The method and setting of outcome measurement is the same for all study participants.	Yes □ Partial □ No □ Unclear □
Outcome Measurement Summary	**Outcome of interest is adequately measured in study participants to sufficiently limit potential bias.**	**Risk: High □ Moderate □ Low □**
Domain 5: Study Confounding		
Goal: To judge the risk of bias due to confounding (*i.e.* the effect of PF is distorted by another factor that is related to PF and outcome).		
Important Confounders Measured	All important confounders, including treatments (key variables in conceptual model: LIST), are measured.	Yes □ Partial □ No □ Unclear □
Definition of the confounding factor	Clear definitions of the important confounders measured are provided (*e.g.*, including dose, level, and duration of exposures).	Yes □ Partial □ No □ Unclear □
Valid and Reliable Measurement of Confounders	Measurement of all important confounders is adequately valid and reliable (*e.g.*, may include relevant outside sources of information on measurement properties, also characteristics, such as blind measurement and limited reliance on recall).	Yes □ Partial □ No □ Unclear □
Method and Setting of Confounding Measurement	The method and setting of confounding measurement are the same for all study participants.	Yes □ Partial □ No □ Unclear □
Method used for missing data	Appropriate methods are used if imputation is used for missing confounder data.	Yes □ Partial □ No □ Unclear □
Appropriate Accounting for Confounding	Important potential confounders are accounted for in the study design (*e.g.*, matching for key variables, stratification, or initial assembly of comparable groups).	Yes □ Partial □ No □ Unclear □
	Important potential confounders are accounted for in the analysis (*i.e.*, appropriate adjustment).	Yes □ Partial □ No □ Unclear □
Study Confounding Summary	**Important potential confounders are appropriately accounted for, limiting potential bias with respect to the relationship between PF and outcome.**	**Risk: High □ Moderate □ Low □**
Domain 6: Statistical Analysis and Reporting		
Goal: To judge the risk of bias related to the statistical analysis and presentation of results.		
Presentation of analytical strategy	There is sufficient presentation of data to assess the adequacy of the analysis.	Yes □ Partial □ No □ Unclear □

Biases	Issues to consider for judging overall rating of "Risk of bias"	
Model development strategy	The strategy for model building (*i.e.*, inclusion of variables in the statistical model) is appropriate and is based on a conceptual framework or model.	Yes □ Partial □ No □ Unclear □
	The selected statistical model is adequate for the design of the study.	Yes □ Partial □ No □ Unclear □
Reporting of results	There is no selective reporting of results.	Yes □ Partial □ No □ Unclear □
Statistical Analysis and Presentation Summary	**The statistical analysis is appropriate for the design of the study, limiting potential for presentation of invalid or spurious results.**	**Risk: High □ Moderate □ Low □**

Source: [66]

6.2 Strength of the body of evidence

strength of whole body of evidence according to GRADE

Following the assessment of the single studies, the whole body of evidence relevant to the research question, is assessed in the context of the setting in which it is to be applied, now per outcome and across studies [35].

For the evaluation of the strength of evidence using GRADE, a starting level of evidence is attributed based on the study design, which is then modified by five potentially downgrading factors (study limitations, consistency, directness, precision, publication bias) and three potentially upgrading factors (strength of association, exposure-response gradient, plausible confounding) [67].

appropriate evidence hierarchy dependent of review question

For research questions on interventions, GRADE attributes a starting level of evidence based on the research design – ranking RCT as “High” and observational studies as “Low”. As outlined in Section 4.2 of this report, under specific circumstances alternative evidence hierarchies should be considered (Table 5.2-1) [49]. In the following we summarize proposed adaptations of the GRADE framework to the special cases of diagnostic accuracy studies and prognosis studies.

6.2.1 Adaptations of the GRADE framework to diagnostic accuracy and prognosis studies

Diagnostic accuracy studies

study limitations

Study limitations are identified based on the quality assessment of the single studies; the same rules as for grading of interventions apply to the overall grading of study limitations in a body of evidence [60].

directness: direct or indirect evidence needed for review question

Directness. An important challenge is to decide on whether diagnostic accuracy alone, a linked evidence approach or direct evidence from RCT is needed to answer the specific review question [68]. It being in most cases the only evidence available, it is tempting to use diagnostic accuracy as a surrogate outcome for patient outcomes. The linkage between diagnostic accuracy and clinical outcomes, however, is in most cases indirect and is challenging to establish.

scenario 1: conditions for use of diagnostic accuracy alone

Scenario 1. In Section 5.2 we have described the conditions under which diagnostic accuracy alone can be used as surrogate outcome for clinical outcomes. In this case the evidence hierarchy (Table 5.2-1) and the risk of bias assessment tool for diagnostic accuracy studies (Table 6.1-1) should be applied to determine the starting level of evidence and potential study limitations.

scenario 2: conditions for use of diagnostic accuracy in linked-evidence approach

Scenario 2. Section 5.2 further describes the conditions in which diagnostic accuracy may be used as intermediate outcome in a "linked evidence" approach. This approach would require separate grading of the body of evidence for each link in the chain (based on the analytical framework established, see Section 3.1).

If the conditions for Scenario 1 and 2 are fulfilled, grading of the body of evidence from diagnostic accuracy studies would NOT require a downgrading for **indirectness**.

If however, the research question necessitates direct evidence from RCT, evidence from studies on diagnostic accuracy would need to be downgraded for indirectness.

According to GRADE, indirectness also includes an assessment of applicability (of the results). For diagnostic accuracy studies, this is evaluated using the QUADAS-2 tool [62]. Investigators need information on inclusion and exclusion criteria, settings and locations of data collections, methods of patient recruitment and sampling to decide whether evidence about the test is valid, clinically relevant and applicable to specific patient groups or individuals [45].

Table 6.2-1: QUADAS-2 checklist for the assessment of applicability in diagnostic accuracy studies

Domain 1: Patient selection	
Describe included patients (prior testing, presentation, intended use of index test and setting)	
Is there concern that the included patients do not match the review question?	Low □ High □ Unclear □
Domain 2: Index test(s) (complete for each index test used)	
Describe the index test and how it was conducted and interpreted.	
Is there concern that the index test, its conduct, or interpretation differ from the review question?	Low □ High □ Unclear □
Domain 3: Reference Standard	
Describe the reference standard and how it was conducted and interpreted.	
Is there concern that the target condition as defined by the reference standard does not match the review question?	Low □ High □ Unclear □

Source: [62]

ROC plot to assess inconsistency in diagnostic accuracy studies

Guyatt *et al.* [69] describe the process of evaluating **inconsistency** in intervention studies. Importantly, this article applies (only) to binary/dichotomous outcomes and relative, not absolute measures of effect. Instead of forest plots, the most common representation format of diagnostic test performance to allow detection of inconsistency is a summary receiver operating characteristic (ROC) curve [70], displaying sensitivity and specificity results from various studies. This can be complemented by a bubble plot of true positive versus false positive rates spread in ROC space. [68]. Reviewers should seek to resolve inconsistency through a critical consideration of possible explanations (differences in populations, interventions or outcomes) for any detected inconsistency.

Post-test probabilities to assess imprecision

Imprecision and **Publication bias** are defined and assessed as for intervention studies [71, 72]. The impact of imprecision on clinical outcomes may be determined by calculating post-test probabilities. In contrast to clinical trials, there is no register for diagnostic accuracy studies. In adaptation of an approach for prognostic studies, one possibility is to generally assume that publication bias is present, except if a diagnostic test has been studied in a large number of diagnostic accuracy studies [49].

A **dose response association** might be observed for tests with continuous outcomes and multiple cut-offs. Inconsistency therefore may arise from test thresholds/cut-offs for positive/negative categorisation varying across test accuracy studies.

Plausible unmeasured confounders. The strength of evidence is increased if, despite plausible confounders that would decrease the diagnostic accuracy, the diagnostic accuracy measured is high.

upgrading possible in case of imperfect reference standard

In intervention studies, strength of association may lead to upgrading of the strength of evidence, if an observed association is large enough that it cannot have occurred solely as a result of bias from confounding factors [68]. For diagnostic accuracy studies, this domain can be applied for upgrading of the evidence when diagnostic accuracy of an index test is measured with an imperfect reference standard and hence may be under-estimated [68].

classification of consequences according to TN, FN, TP, FP categories

The presumed consequences of classification in each of the categories (FP, TP, FN, TN) need to be defined (see example in Table 6.2-2) and the directness of the link assessed. Benefits of correct classification should outweigh the harms of misclassification. This assessment should also include consequences of inconclusive results [30].

Table 6.2-2: Tabular presentation of the changes in classification induced by a new test compared to standard strategy – Example: p16/Ki-67 triage compared with direct referral to colposcopy

Putative benefit of new test	Sensitivity	Specificity	TP	TN	FP	FN	Inconclusive results
Simpler, less time	lower	higher	↘	↗	↘↘	↗	Unknown
Benefits and harms from changes in classification			Benefit from treatment	Benefit from avoiding unnecessary tests	Anxiety and morbidity from unnecessary additional testing and treatment	Possible detriment from delayed diagnosis	Benefit from treatment, anxiety and morbidity from unnecessary testing and treatment
Certainty/Uncertainty of the benefits/harms			No uncertainty	Major uncertainty (not clear if clinicians would trust test results)	No uncertainty	Major uncertainty (impact of negative result on screening attendance unclear	No uncertainty (all cases will be confirmed by colposcopy +/- biopsy)

Prognosis studies

adaptation of GRADE framework to prognosis studies

An adaptation of the GRADE Framework for prognostic studies has been proposed by Huguet *et al.* [49].

Study limitations for prognostic studies are assessed based on the quality appraisal of the single studies (Section 6.1.2).

Reviewers should downgrade for **inconsistency** if estimates of the prognostic factor association with the outcome vary in direction and there is no or minimal overlap of the confidence intervals. If a meta-analysis is conducted, before downgrading for inconsistency, a subgroup analysis in a priori defined subgroups (*e.g.* differences in population, duration of follow-up, study methods) should be performed [49].

Indirectness may be present if the study population does not represent the population of the review question (*e.g.* patients of a headache clinic versus general population) [49]. Similarly, downgrading for indirectness would be justified if the prognostic factor or the outcomes assessed would not represent the full bandwidth of the review question.

Of particular importance in the review of prognostic studies is **publication bias**. In contrast to clinical trials, there is no register for prognostic research studies. One possibility therefore is to generally assume that publication bias is present, except if a prognostic factor has been studied in a large number of cohort studies [49].

Other domains are assessed as for intervention studies, with the exception of **plausible confounders**: in contrast to intervention studies, the effect of inadequate control of confounding on the study effects are unclear and as such cannot be taken into account to estimate the accuracy of the effect estimate. Risk of bias by confounding, however, is covered by the quality appraisal of the single studies [49].

6.2.2 Assessment of co-dependent technologies

framework for evaluating evidence on the clinical benefit of co-dependent technologies for reimbursement decisions

Decisions on reimbursement and/or implementation in clinical practice of new technologies require evidence on the clinical benefit in terms of patient health outcomes.

Biomarker tests are not stand-alone technologies but need to be assessed with regards to patient management changes and treatment incited by the test results. This poses several evidentiary challenges, notably a lack of (direct) evidence from randomised controlled trials (see Sections 3.4 and 5.2) as discussed for example in [73]. To reduce decision-making uncertainty in the absence of direct evidence, a linked-evidence approach for the assessment of medical tests was first proposed in the Australian guidance for the assessment of diagnostic tests [8] and mentioned in several international guidelines on diagnostics evaluation [3, 5, 12, 57]; while other institutions do not include it as an option [7]. Linked evidence describes a chain of arguments linking different types of evidence, provided that a) data are transferrable across different parts of the linkage and b) for each element, evidence is gathered systematically and transparently and is considered internally valid.

A framework for evaluating evidence on the clinical benefit of co-dependent technologies for reimbursement decisions has recently been published [9]. In the following we will present a generalised summary of this approach.

Predictive biomarkers

Predictive biomarkers guide treatment choices: a drug A is expected to perform better than a drug B in biomarker-positive patients, while no differences of treatment effect are expected in biomarker-negative patients.

Table 6.2-3: Differentiation of biomarker types by impact on health outcomes

	Outcome	Biomarker +		Biomarker -	
		Drug A	Drug B	Drug A	Drug B
Predictive	Overall survival	60%	30%	30%	30%
	RR	2.0		1.0	
Prognostic	Overall survival	70%	60%	35%	30%
	RR	1.2		1.2	
Both	Overall survival	60%	40%	20%	20%
	RR	1.5		1.0	

Source: [9]

assess biological plausibility of the relationship between the drug and the biomarker

In a first step, the biological plausibility of the relationship between the drug and the biomarker is evaluated: evidence must be presented that the biomarker is predictive (treatment effect modifier) or prognostic (indicative of disease progression independent of treatment) or both. Only double randomised trials or trials with a marker by treatment interaction design (Section 5.2) allow an answer to this question; the relationship cannot be clarified by a marker strategy design. If the biomarker is a prognostic factor, other treatments will also be likely to have a favourable outcome in the marker positive subgroup and should be included in the comparison [9].

Highest level of evidence on the clinical benefit of a test and treatment strategy is provided by a double randomised controlled trial, with one randomisation to the testing strategy and a second randomisation to treatment and control. Direct evidence, albeit of a lower level, can also be provided by trials with a marker strategy design or a biomarker by treatment design (see Section 5.2) [9].

assess ability of a test and treatment combination to improve patient outcomes: Table 8.2-1

If direct evidence is not available, there are various options to provide linked-evidence. Here we describe a few examples [9]:

Option A: A marker by treatment design is linked to a randomised controlled trial of the same treatment/control in an untested population, thereby allowing comparing the relative effectiveness of treatment versus control in a biomarker stratified versus a biomarker un-stratified population.

Option B: A variation of Option A, where only patients with discordant results in a standard testing strategy vs a biomarker based strategy are randomised to treatment/control and compared to the RCT in an untested population.

Option C: Patients in the standard and the treatment arm of an RCT are retrospectively analysed for marker status or tests are performed in archived samples. Relative effectiveness of treatment versus control is then compared to a biomarker-positive population in an enrichment design. Test accuracy serves as an additional link in this option.

Diagnostic and prognostic biomarkers

need for pragmatic RCT

Diagnostic and prognostic biomarkers are not co-dependent technologies with a specific treatment. Nevertheless, they cannot provide clinical benefit without inducing patient management changes and treatment decisions compared to the established test and treatment strategy. Specific challenges here are to identify (pragmatic) RCT reflecting the population and clinical practice in the country/health system in which the biomarker is intended to be introduced [9].

ⓘ For details on the analytical framework to assess co-dependent technologies, see [9]. To assess the ability of a test and treatment combination to improve patient outcomes
(Qualification 2 in Table 3.4-1), see [9] or Table 8.2-1 in the Annex.

7 References

[1] Schleidgen S, Klingler C, Bertram T, Rogowski WH, Marckmann G. What is personalized medicine: sharpening a vague term based on a systematic literature review. BMC medical ethics. 2013;14:55. Epub 2013/12/24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24359531.

[2] Gartlehner G. Internes Manual. Abläufe und Methoden. Teil 2 (2. Aufl.). Wien: Ludwig Boltzmann Institut für Health Technology Assessment. 2009. Available from: http://eprints.hta.lbg.ac.at/713/.

[3] Chang S. AHRQ. Methods Guide for Medical Test Reviews. AHRQ Publication N012-EHC017. 2012. Available from: http://www.effectivehealthcare.ahrq.gov/reports/final.cfm.

[4] Deeks JJ BP, Gatsonis C (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9.0.The Cochrane Collaboration. 2010. Available from: http://srdta.cochrane.org/.

[5] Derksen J. CVZ. Medical tests (assessment of established medical science and medical practice). 2011;CVZ Report 293. Available from: http://www.zorginstituutnederland.nl/publicaties.

[6] Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy. 2009;64(8):1109-16. Epub 2009/06/06. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19489757.

[7] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen – IQWIG. Allgemeine Methoden Version 4.1. 2013; Available from: https://www.iqwig.de/download/IQWiG_Methoden_Version_4-1.pdf.

[8] Medical Services Advisory Committee. Guidelines for the Assessment of Diagnostic Technologies. Canberra, Australia: Commonwealth of Australia,. 2005. Available from: http://www.msac.gov.au/.

[9] Merlin T, Farah C, Schubert C, Mitchell A, Hiller JE, Ryan P. Assessing personalized medicines in Australia: a national framework for reviewing codependent technologies. Medical decision making: an international journal of the Society for Medical Decision Making. 2013;33(3):333-42. Epub 2012/08/17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22895559.

[10] National Institute of Health and Clinical Excellence. Diagnostics Assessment Programme manual. 2011. Available from: http://www.nice.org.uk/media/8A3/34/DAP_programme_manual_final_for_upload_22_Dec_11.pdf.

[11] Institute of Medicine (IOM). Evaluation of biomarkers and surrogate endpoints in chronic disease. Washington; DC: The National Academies Press,. 2010. Available from: http://www.iom.edu/Reports/2010/Evaluation-of-Biomarkers-and-Surrogate-Endpoints-in-Chronic-Disease.aspx.

[12] Nachtnebel A. Evaluation von Diagnostika – Hintergrund, Probleme, Methoden. HTA Projektbericht Nr 36. 2010. Available from: http://eprints.hta.lbg.ac.at/898/.

[13] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical pharmacology and therapeutics. 2001;69(3):89-95. Epub 2001/03/10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11240971.

[14] Febbo PG, Ladanyi M, Aldape KD, De Marzo AM, Hammond ME, Hayes DF, et al. NCCN Task Force report: Evaluating the clinical utility of tumor markers in oncology. Journal of the National Comprehensive Cancer Network: JNCCN. 2011;9 Suppl 5:S1-32; quiz S3. Epub 2011/12/22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22138009.

[15] Teutsch SM, Bradley LA, Palomaki GE, Haddow JE, Piper M, Calonge N, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. Genetics in medicine: official journal of the American College of Medical Genetics. 2009;11(1):3-14. Epub 2008/09/25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18813139.

[16] van Holten TC, Waanders LF, de Groot PG, Vissers J, Hoefer IE, Pasterkamp G, et al. Circulating biomarkers for predicting cardiovascular disease risk; a systematic review and comprehensive overview of meta-analyses. PloS one. 2013;8(4):e62080. Epub 2013/05/01. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23630624.

[17] Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2005;23(9):2020-7. Epub 2005/03/19. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15774793.

[18] Food and Drug Administration. [4 Nov 2014]; Available from: http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm301431.htm.

[19] Beachy SH, Repasky EA. Using extracellular biomarkers for monitoring efficacy of therapeutics in cancer patients: an update. Cancer immunology, immunotherapy: CII. 2008;57(6):759-75. Epub 2008/01/12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18188561.

[20] Morrison A, Boudreau R. CADTH. Evaluation Frameworks for Genetic Tests. [Environmental Scan issue 36] Ottawa: Canadian Agency for Drugs and Technologies in Health. 2012.

[21] Lijmer JG, Leeflang M, Bossuyt PMM. Proposals for a Phased Evaluation of Medical Tests. Medical Tests-White Paper Series. 2009. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21290784.

[22] Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. American journal of preventive medicine. 2001;20(3 Suppl):21-35. Epub 2001/04/18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11306229.

[23] Tripepi G, Jager KJ, Dekker FW, Zoccali C. Statistical methods for the assessment of prognostic biomarkers (part II): calibration and re-classification. Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association – European Renal Association. 2010;25(5):1402-5. Epub 2010/02/20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20167948.

[24] Tripepi G, Jager KJ, Dekker FW, Zoccali C. Statistical methods for the assessment of prognostic biomarkers (Part I): discrimination. Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association – European Renal Association. 2010;25(5):1399-401. Epub 2010/02/09. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20139066.

[25] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089-92. Epub 2006/05/06. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16675820.

[26] Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Medical decision making: an international journal of the Society for Medical Decision Making. 2009;29(5):E1-E12. Epub 2009/09/24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19773580.

[27] Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Annals of internal medicine. 2013;158(7):544-54. Epub 2013/04/03. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23546566.

[28] Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. BMJ. 2011;343:d4684. Epub 2011/09/10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21903693.

[29] Trikalinos TA, Balion CM. Chapter 9: options for summarizing medical test performance in the absence of a "gold standard". Journal of general internal medicine. 2012;27 Suppl 1:S67-75. Epub 2012/06/08. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22648677.

[30] Hsu J, Brozek JL, Terracciano L, Kreis J, Compalati E, Stein AT, et al. Application of GRADE: making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implementation science: IS. 2011;6:62. Epub 2011/06/15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21663655.

[31] Segal JB. Choosing the Important Outcomes for a Systematic Review of a Medical Test. Methods Guide for Medical Test Reviews. 2012. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22834029.

[32] Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Medical decision making: an international journal of the Society for Medical Decision Making. 2009;29(5):E30-8. Epub 2009/09/04. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19726782.

[33] Berzin TM, Blanco PG, Lamont JT, Sawhney MS. Persistent psychological or physical symptoms following endoscopic procedures: an unrecognized post-endoscopy adverse event. Digestive diseases and sciences. 2010;55(10):2869-73. Epub 2010/04/16.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/20393877.

[34] Centers for Disease Control and Prevention. ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing. [01.07.2014];
Available from: http://www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm.

[35] Hillier S, Grimmer-Somers K, Merlin T, Middleton P, Salisbury J, Tooher R, et al. FORM: an Australian method for formulating and grading recommendations in evidence-based clinical guidelines. BMC medical research methodology. 2011;11:23. Epub 2011/03/02.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/21356039.

[36] Relevo R. Effective Search Strategies for Systematic Reviews of Medical Tests. Methods Guide for Medical Test Reviews. 2012. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22834020.

[37] US Food and Drug Administration. Devices @FDA Catalog.
Available from: http://www.accessdata.fda.gov/scripts/cdrh/devicesatfda/.

[38] Samson D, Schoelles KM. Developing the Topic and Structuring Systematic Reviews of Medical Tests: Utility of PICOTS, Analytic Frameworks, Decision Trees, and Other Frameworks. Methods Guide for Medical Test Reviews. 2012. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22834028.

[39] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. American journal of ophthalmology. 2000;130(5):688. Epub 2000/11/18.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/11078861.

[40] Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ. 1998;317(7167):1185-90. Epub 1998/10/31.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/9794851.

[41] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. The New England journal of medicine. 2000;342(25):1887-92. Epub 2000/06/22.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/10861325.

[42] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. Journal of clinical epidemiology. 2011;64(4):383-94. Epub 2011/01/05. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21195583.

[43] Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. BMC medical research methodology. 2009;9:34. Epub 2009/06/13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19519887.

[44] OCEBM Levels of Evidence Working Group. “The Oxford 2011 Levels of Evidence”. Oxford Centre for Evidence-Based Medicine.; Available from: http://www.cebm.net/index.aspx?o=5653.

[45] Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clinical chemistry. 2005;51(8):1335-41. Epub 2005/06/18.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/15961549.

[46] Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ: Canadian Medical Association journal = journal de l’Association medicale canadienne. 2006;174(4):469-76. Epub 2006/02/16.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/16477057.

[47] Bossuyt P, Leeflang MM. Chapter 6: Developing criteria for including studies. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. 2008;Version 0.4 [updated Sept 2008] The Cochrane Collaboration. Available from: http://srdta.cochrane.org/.

[48] 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? Annals of internal medicine. 2006;144(11):850-5. Epub 2006/06/07. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16754927.

[49] Huguet A, Hayden JA, Stinson J, McGrath PJ, Chambers CT, Tougas ME, et al. Judging the quality of evidence in reviews of prognostic factor research: adapting the GRADE framework. Systematic reviews. 2013;2(1):71. Epub 2013/09/07. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24007720.

[50] Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2009;27(24):4027-34. Epub 2009/07/15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19597023.

[51] Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: a systematic review and classification. Clinical cancer research: an official journal of the American Association for Cancer Research. 2013;19(17):4578-88. Epub 2013/06/22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23788580.

[52] Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. Nature reviews Clinical oncology. 2014;11(2):81-90. Epub 2013/11/28. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24281059.

[53] Andre F, O'Regan R, Ozguroglu M, Toi M, Xu B, Jerusalem G, et al. Everolimus for women with trastuzumab-resistant, HER2-positive, advanced breast cancer (BOLERO-3): a randomised, double-blind, placebo-controlled phase 3 trial. The lancet oncology. 2014;15(6):580-91. Epub 2014/04/20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24742739.

[54] Zujewski JA, Kamin L. Trial assessing individualized options for treatment for breast cancer: the TAILORx trial. Future Oncol. 2008;4(5):603-10. Epub 2008/10/17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18922117.

[55] Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. Journal of clinical epidemiology. 2009;62(4):364-73. Epub 2008/10/24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18945590.

[56] Janatzek S. [The benefit of diagnostic tests--from surrogate endpoints to patient-relevant endpoints]. Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen. 2011;105(7):504-9. Epub 2011/10/01. Nutzen diagnostischer Tests - vom Surrogat zur Patientenrelevanz. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21958609.

[57] Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: design issues. Journal of the National Cancer Institute. 2010;102(3):152-60. Epub 2010/01/16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20075367.

[58] Mandrekar SJ, Grothey A, Goetz MP, Sargent DJ. Clinical trial designs for prospective validation of biomarkers. American journal of pharmacogenomics: genomics-related research in drug development and clinical practice. 2005;5(5):317-25. Epub 2005/10/04. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16196501.

[59] Freidlin B, Korn EL. Biomarker-adaptive clinical trial designs. Pharmacogenomics. 2010;11(12):1679-82. Epub 2010/12/15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21142910.

[60] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). Journal of clinical epidemiology. 2011;64(4):407-15. Epub 2011/01/21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21247734.

[61] Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Annals of internal medicine. 2004;140(3):189-202. Epub 2004/02/06. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14757617.

[62] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36. Epub 2011/10/19. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22007046.

[63] Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 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. Epub 2012/10/26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23097549.

[64] Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ. 2013;346:f2778. Epub 2013/05/18.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/23682043.

[65] Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Annals of internal medicine. 2006;144(6):427-37. Epub 2006/03/22.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/16549855.

[66] Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of internal medicine. 2013;158(4):280-6. Epub 2013/02/20.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/23420236.

[67] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. Journal of clinical epidemiology. 2011;64(4):401-6. Epub 2011/01/07.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/21208779.

[68] Singh S, Chang SM, Matchar DB, Bass EB. Grading a Body of Evidence on Diagnostic Tests. In: Chang SM, Matchar DB, Smetana GW, Umscheid CA, editors. Methods Guide for Medical Test Reviews. Rockville (MD)2012.

[69] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. Journal of clinical epidemiology. 2011;64(12):1294-302. Epub 2011/08/02. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21803546.

[70] Eng, J. (n.d.). ROC analysis: web-based calculator for ROC curves. Retrieved 18.10.2013, from http://www.jrocfit.org.

[71] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. Journal of clinical epidemiology. 2011;64(12):1283-93. Epub 2011/08/16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21839614.

[72] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. Journal of clinical epidemiology. 2011;64(12):1277-82. Epub 2011/08/02. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21802904.

[73] Khoury MJ, Coates RJ, Evans JP. Evidence-based classification of recommendations on use of genomic tests in clinical practice: dealing with insufficient evidence. Genetics in medicine: official journal of the American College of Medical Genetics. 2010;12(11):680-3. Epub 2010/10/27.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/20975567.

[74] Trikalinos TA, Balion CM. Options for Summarizing Medical Test Performance in the Absence of a “Gold Standard”. Methods Guide for Medical Test Reviews. 2012.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/22834025.

8 Annexes

8.1 Formulation of review question: Checklist to assess context of submission and impact on clinical practice

Table 8.1-1: Checklist – rationale for co-dependent relationship between test (T) and drug (D). (O) – overlap between test and drug.

Context for the submission	
1) (T) Who is the test sponsor?	Identify the source of the test (*e.g.* commercial sponsor, research laboratory, widespread pathology practice). This includes clinical sponsors of tests, given that tests not only guide the initiation of therapy but also the cessation of therapy.
2) (D) Who is the drug sponsor?	This enables a different sponsor to be identified if necessary for each component of a pair of co-dependent technologies.
3) (D) What is the proposed drug?	Provide a description of the drug, its background, mechanism of action, etc. Specify the drug's registration status.
4) (O) What is the biomarker?	This initial scenario considers genetic DNA biomarkers only, *i.e.* the assessment of one genetic locus at a time. Note that the Food and Drug Administration (FDA) in the United States of America has provided specific definitions of genomic biomarkers (*i.e.* assessment across the genome, testing hundreds or thousands of loci simultaneously). Genomic testing is beyond this initial scope and would be applicable to more complex scenarios.
5) (T) What is the proposed test?	This relates to a description of a single test or assay. However, often tests are done in series when assessing genetic biomarkers or there may be an algorithm-based computation of the results of a number of tests. Describe the test method in sufficient detail that a laboratory technician would be able to perform it. Specify the range of techniques available to measure the biomarker (*e.g.* polymerase chain reaction (PCR), high resolution melting (HRM)), and indicate which method, if any, is regarded as the reference or 'gold' standard.
6) (T & D) Is the test (or drug) currently reimbursed?	Describe current reimbursement arrangements for the test and the drug. This determines the extent of information needed for the current technology.
7) (T & D) What is the medical condition or problem being managed, *i.e.* the patient indication?	Describe the patient indication being addressed. If different test result thresholds are likely, or if eligibility for the drug is determined subjectively, consider providing alternative indications.
8) (O) Is there a clear definition of the biomarker(s) (*e.g.* specific genetic DNA mutation(s))?	Describe the nature of the genetic DNA biomarker (*e.g.* single nucleotide polymorphisms (SNPs), mutation, or copy number variation (CNV)). Where relevant, include the following elements describing the context for a biomarker: (i) the general clinical area, (ii) the specific use of the biomarker, and (iii) the critical parameters which define when and how the biomarker should be used. Describe exactly what the test is identifying in cases where there is no "specific mutation", *e.g.* an expression micro-array of tumour tissue which identifies cancer with activation of a particular pathway, and susceptibility to a certain drug, but does not identify a specific mutation as such. Categorise the mutation as either a germline or somatic mutation. If the mutation is classified as a germline mutation, then consider issues related to heritability, *e.g.* testing of relatives and genetic counselling would need to be considered and assess the ethical and medico-legal implications of testing.
9) (O) What is the biological rationale for targeting that biomarker(s) with the drug?	Present the initial evidence that was relied on to select the biomarker. Describe and explain the overall approach to the selection of the biomarker including methods and relevant aspects of study design and statistical analysis. Describe the rationale for the selection of the population sample studied in the biomarker qualification. Present the criteria used for selection of candidate genes (*e.g.* candidate by position, by function, based on expression profiling data). Justify, using molecular biological or pharmacological principles, the plausibility of treatment effect modification (or interaction) between the biomarker itself and the drug, or alternatively between the drug and another factor for which the biomarker is a proxy. Advise whether this rationale precedes the specification of the data collection which forms the primary source of evidence.

10) (O) Do any other biomarker(s) predict variation in the comparative treatment effect (between using the drug and not using the drug)? In the case of another biomarker that is a genetic mutation: • Have details on the specific mutation and the nature of the mutation been provided? • Is the effect of treatment on this other mutation consistent with the effect under consideration?	(Note that this may be relevant even if the other biomarker(s) are claimed, but are not proven and/or are not reimbursed.) If testing for other biomarkers is reimbursed, this would move to a more complex scenario.
11) (O) What is the prevalence of a true positive biomarker in the population likely to receive the test?	The source population would be those who are eligible according to the requested reimbursement descriptor and follow the corresponding clinical pathway to the point of being offered the test – or the drug in the absence of the test. An estimate of the prevalence of a true positive biomarker is relevant to calculating the performance of a test in terms of its negative and positive predictive value. Indicate where there is no 'gold' standard to determine this true positive status of the biomarker and use an alternative appropriate methodology to estimate it. Proposed impact on current clinical practice
Proposed impact on current clinical practice	
12) (T & D) What are the relevant clinical pathways? That is, is there a description and comparison of the proposed clinical management of a typical patient up to the point of being offered the proposed test and subsequent therapy with the proposed drug, as compared to the currently existing clinical pathway(s) where the proposed test is not offered and the proposed drug is not available? In these clinical pathways, outline all alternative tests/test strategies (whether in series or occurring concurrently) and all alternative treatments (including non-drug treatments) for the patient indication both with and without knowledge of the patient's biomarker status.	If it is important for patients with a rapidly progressive disease to ensure that a timely test result is available to determine drug eligibility, indicate whether the test is therefore likely to be performed earlier in disease progression in a broader population than might otherwise be considered as potentially eligible for the drug. Identify tests and treatments that are commonly used and likely to be supplemented or replaced by the pair of co-dependent technologies (see Information Requests 13 and 14).
13) (T) Can the proposed test be used with other treatments and/or for other purposes? (Refer to the clinical pathways provided in response to Information Request 12.)	If other treatments or purposes are relevant, this would move to a more complex scenario.
14) (T) Is the test an additional test to other(s) currently defining the condition? Or a replacement test? Or both (*i.e.* depending on the test result, replace some tests or be additional to other tests)? (Refer to the clinical pathways provided in response to Information Request 12.)	Most commonly, the test would be an additional test; although occasionally if the biomarker is a strong predictor, then it could replace another test in the workup.
15) (T) How is it suggested that the test will be offered?	Specify the national registration status of the test. Assess access and quality assurance issues. Identify how many laboratories offering the test have national accreditation for that test. (Note that a way of determining this is not yet available.) Indicate whether the test accessibility is likely to be widespread or only available in a few selected laboratories across the country. Explain how the test would be undertaken in practice and what impact it would have on patient and health professionals. Discuss alternative ways to access the test.
16) (T) Have the following been identified: i) the biospecimen required to perform the test? ii) whether this specimen needs to be collected specifically for the purposes of performing the test or has already been collected for another purpose?	i) For example: blood, tumour material (formalin-fixed paraffin embedded (FFPE) or fresh), bone marrow, cytology specimen, mouth swab. ii) For example: tumour already removed can be tested if archival FFPE is available and the test can identify the biomarker from this tissue. If a new specimen needs to be collected, specify the costs, risks and feasibility of collecting the sample. In some instances, such as a blood sample, the costs and risks would be trivial. In other instances, such as when a new biopsy is required, there may be significant costs as well as safety risks for the patient.

17) (If relevant) (T) What is the potential need for subsequent testing to identify new somatic mutations which may guide dosage or cessation of therapy with the co-dependent drug?	This will impact on the clinical need for the proposed test as well as its potential use to guide drug dosage titration and treatment continuation. If subsequent testing is needed, this would move to a more complex scenario.
18) (T) Are the test results expected to be consistent over time, including over the course of the disease?	Where test results may change over time, provide sufficient detail to clarify the relationship and timeframes between test results and the appropriateness of treatment. For example; Kirsten rat sarcoma viral oncogene homolog (K-RAS) testing of the primary colorectal cancer tumour is usually representative of the findings in metastases. However epidermal growth factor receptor (EGFR) results change with exposure to radiotherapy etc. and so the results of testing the primary tumour may not be representative of what is happening in non-small cell lung cancer metastases.
19) (O) Can the proposed drug be used with other specific tests for that biomarker, other than the test proposed? What methodologies are available to test for the marker?	If other tests are publicly funded, this would move to a more complex scenario.

(Source: [9] with minor adaptations)

8.2 Qualification: assessing the impact of a biomarker test on patient health outcomes

Table 8.2-1: Checklist: Clinical benefit of the pair of co-dependent technologies in terms of patient health outcomes

20) (O) Is there direct evidence of prognostic impact associated with different biomarker status?	This is used to discriminate prognostic impact as an alternative (or in addition) to treatment effect modification. It requires a comparison of outcomes in patients receiving usual care conditioned on the presence or absence of biomarker positive status.
When presenting the body of evidence to address clinical benefit, two different options (Option 1 and Option 2) are provided so that available information can be used to maximum effect to inform a reimbursement decision.	
OPTION 1. Is there 'direct evidence'[11] of the proposed test's impact on patient health outcomes? For example, patients randomised to the proposed test or to no test and followed through to allocation of the proposed drug or usual care and the subsequent impact of that treatment on their health outcomes.	Direct evidence is used to determine whether the pair of co-dependent technologies are (cost-) effective and safe. If randomised to use of the test, then biomarker status would be known and, on that basis, subsequent targeted therapy or usual care could be decided for the patient. If randomised to not using the test, then the patient would receive treatment that is not targeted by the biomarker result. 'Direct evidence' does not exclude the need for an assessment of translational issues.
Level 1: Is a trial available that randomised to use of the test or not, and then randomised to use of the drug or its main drug comparator, and then followed participants to measure clinical outcomes (whether surrogate outcomes or directly patient relevant outcomes)? See Figure 5.2-1– double-randomised controlled trial.	
Level 2: If not, is a trial available that randomised to the use of the test or not, and then followed participants to measure clinical outcomes (whether surrogate outcomes or directly patient relevant outcomes)? See Figure 5.2-1–marker strategy design.	Given that Level 2 direct evidence does not provide information on the test(biomarker)-drug relationship *i.e.* evidence that the biomarker is a treatment effect modifier or prognostic factor, therefore consider supplementing with Level 3 or 4 direct evidence (also see Information Requests 34 and 35).
Level 3: If not, is a trial available that prospectively tested eligible patients, and then randomised test positive or negative patients to use of the drug or its main comparator, and then followed participants to measure clinical outcomes (whether surrogate outcomes or directly patient relevant outcomes)? See Figure 5.2-1– marker by treatment interaction design/ randomised trial of drug only (with the eligibility of all subjects determined by test result).	Given that Level 3 and 4 direct evidence effectively involve uncontrolled study designs (*i.e.* there is no trial arm provided to assess the impact of not testing biomarker status), consider providing a supplementary 'linked evidence' approach (see Option 2 below) so that at least a comparison of the proposed test/test strategy and existing test/test strategy can be made with respect to their relative diagnostic accuracy.
Level 4: If not, is a trial available that randomised eligible patients to use of the drug or its main comparator, and then followed participants to measure clinical outcomes (whether surrogate outcomes or directly patient relevant outcomes), and then analysed results across subgroups of patients defined by whether they are positive for the test (or biomarker) or whether they are negative to the test (or biomarker)? See Figure 5.2-1 – biomarker-stratified design / randomised trial of drug only (with the test result determined through subgroup analysis).	Level 4 direct evidence may use archival tissue/sampling to determine biomarker status. Exercise caution when interpreting results from Level 4 studies where biomarker status might change over time, including where there is evidence that intervening treatment may modify the biomarker.
Level 5: If not, then move to corresponding guidance on 'linked analyses' (see Option 2, below).	

[11] Direct evidence: a trial that compares groups of people receiving either the currently used diagnostic test/test strategy or the proposed diagnostic test/test strategy and measures the differential impact of the diagnostic method on patient health outcomes [8].

21) (O) Is the direct evidence presented and selected in a comprehensive and unbiased manner?	For example, present a systematic review of direct evidence concerning this pair of proposed test and proposed drug for this biomarker with pre-specified inclusion/exclusion criteria and a PRISMA flowchart indicating how trials were selected and the reasons why any potentially relevant trials were excluded.
22) (O) Is the direct evidence of good quality?	Assess bias, confounding, the impact of chance on results and whether the analyses were pre-specified and/or exploratory. Use an intervention study design critical appraisal checklist to cover all issues likely to affect the internal validity of the presented trial results.
23) (O) Does the direct evidence provided show a clinically important and statistically significant impact on patient-relevant health outcomes?	Assess both effectiveness and safety. Describe outcomes in the studies (primary and secondary outcomes) and statistical methods used. Provide an extended assessment of comparative harms. Assess the balance of benefits and harms and interpret findings from the body of evidence.
24) (O) Is the direct evidence provided applicable to the requested populations?	Translation steps (applicability, transformation and extrapolation): ❂ address external validity concerns of trials usually conducted in a different setting or with a different population (i.e. spectrum of disease) ❂ address concerns that usually relate to the length of follow-up of the direct evidence, to the use of surrogate outcomes and most importantly to capture the point estimate and confidence limits of the treatment effect taking into account the impacts of incorporating the test results. Describe patient characteristics in the trials and indicate whether they are relevant to the national situation. Indicate whether the requested technologies were provided in a setting similar to the setting of use.
OPTION 2. Is there 'linked evidence' available of the test's impact on patient health outcomes? In other words, can different types of evidence from different sources be linked in a chain of argument to estimate this impact?	For example, this might involve linking evidence of test accuracy with evidence that the test result changes patient management, and with evidence that the alternative treatments have different effectiveness and safety profiles. Further background is provided in the 2005 Medical Services Advisory Committee (MSAC) Guidelines [8]for the assessment of diagnostic technologies. Note that a full linked evidence approach is only meaningful when the evidence for the proposed test and the evidence for the proposed drug have been generated in similar patient populations and so it is clinically sensible to link the two data sets. If the test identifies patients earlier or with a different spectrum of disease than the patients in whom the drug has been trialled, then it is not clinically sensible to link this evidence. In such circumstances direct evidence is needed.
What is the test effectiveness and safety?	
25) (T) What is the analytical test performance?	Analytical test performance assesses how accurately and how consistently the test identifies biomarker status, *e.g.* the coefficient of variation and other appropriate statistics. Present any differences across laboratories in how they characterise test results (*e.g.* a kappa statistic or other concordance statistic). Identify whether there is an external quality assurance program by which laboratories can benchmark their assays.
26) (T) Is there a clinical reference standard or a 'gold' standard against which test performance can be measured?	Indicate whether this clinical reference standard is also the relevant diagnostic comparator, *i.e.* the current test/test strategy being used in the absence of the proposed test. Note: there are statistical solutions for situations when a reference standard is imperfect or not available or impractical (construct a reference standard, predictive accuracy, adjustment of accuracy estimates for workup bias), which are detailed elsewhere [74].
If a reference standard is available: test performance is determined using diagnostic accuracy measures.	Test performance measures include: sensitivity, specificity, likelihood ratios, positive and negative predictive values, area under curve (AUC). Designate a reference standard and compare the proposed test to the designated reference standard by cross classifying the test results of patients who are representative of the intended use population. Include confidence intervals and significance levels to quantify the statistical uncertainty in these estimates due to the subject/sample selection process. This type of uncertainty decreases as the number of participants in the study increases. Assess whether there is a test performance level below which the test should not be used (for example, either false positives are too great or false negatives are too great) so that other better performing tests are needed.

27) (T) Is the evidence of diagnostic or predictive accuracy presented and selected in a comprehensive and unbiased manner?	For example, present a systematic review of diagnostic accuracy studies for this test with inclusion/exclusion criteria delineated and a PRISMA flowchart indicating how trials were selected and reasons why any potentially relevant trials were excluded.
28) (T) Is the evidence of diagnostic or predictive accuracy of good quality?	See Table 6.1-1 QUADAS~ checklist.
29) (T) Are there any safety considerations that will impact on the entire process of testing?	
30) (T) Is the evidence of test accuracy and safety applicable to the requested populations?	Assess whether test accuracy was determined in the correct population. See Table 6.2-1 QUADAS~ checklist.
31) (If relevant for a comparison of tests) (T) Which test has the best test performance (in terms of accuracy and/or clinical benefit)?	Assess trade-offs in false positives, false negatives, and in positive predictive value and negative predictive value. If other tests are publicly funded, this would move to a more complex scenario.
32) (If relevant for a comparison of tests) (T) Which test is most accessible/ available/ used?	Assess access and quality assurance issues. If other tests are publicly funded, this would move to a more complex scenario.
33) (O) Will knowledge of the test result cause a change in the management of the patient by the treating clinician? Are there instances where management would not change, despite the test indicating the biomarker is present?	There may be 'leakage' issues identified through an assessment of the 'change in management' part of a linkage. Often a test is done to rule out a drug (*e.g.* to avoid potential drug-related adverse events or the development of drug resistance), but the drug is given anyway, or, alternatively, the test is used to select a specific drug, but the drug is not provided. As companion tests in a co-dependent pairing will often be used to guide drug therapy decisions, this would need to be explicitly addressed. Once listed, these issues could be informed by data that compare the numbers of test 'positive' results and prescriptions filled for the drug.
What is the test-drug effectiveness and safety?	
34) (O & D) Is there evidence available of treatment effect modification or significant interaction between biomarker status and treatment outcomes? For example, is there evidence of substantial variation in a measure of relative treatment effect between the proposed drug and usual care trial arms after stratifying for biomarker status?	Treatment effect modification in this setting identifies a relationship between the biomarker and the drug, which is likely to be unique or limited to companion tests assessing a particular biomarker and drugs with a particular mechanism of action. This means that both technologies are required to produce a clinical benefit and the reimbursement decision may need to encompass both technologies.
35) (O & D) Is there evidence available of better targeting to patients likely to respond most by using the prognostic impact of the biomarker to determine the baseline risk of disease progression? For example, is there evidence of minimal variation in a measure of relative treatment effect between the proposed drug and usual care trial arms, but biomarker status helps identify patients at greatest risk of an event which helps maximise the absolute treatment effect?	If a drug's result is due to better targeting to those patients that are likely to respond most, this identifies a relationship between the biomarker and a potentially broader range of existing and future treatment options (potentially including non-drug treatment options) than is likely to apply for treatment effect modification. It is possible for both treatment effect modification and prognostic impact to co-exist. In this case, in order to assess the unique contribution of the drug therapy, an assessment of its effect must be made relative to usual care and adjusted for the background prognostic impact that is operating in both the drug and usual care arms and which is also flagged by that particular biomarker. By contrast, if the drug's apparent improvement in result is simply due to the fact that a certain patient subgroup (flagged by a specific biomarker) will always do better, then the level of co-dependency between the technologies is low. This may allow reimbursement of either test or drug or both technologies.
36) (O & D) Is the drug effectiveness evidence, as conditioned by the test or biomarker result, obtained in a comprehensive and an unbiased manner?	For example, present a systematic review of randomised trials of the proposed drug targeting this biomarker with inclusion/exclusion criteria delineated and a PRISMA flowchart indicating how trials were selected and reasons why any potentially relevant trials were excluded.
37) (O & D) Is this drug effectiveness evidence, as conditioned by the test or biomarker result, of good quality?	Assess bias, confounding, the impact of chance on results and whether the analyses were pre-specified and/or exploratory. Use an intervention study design critical appraisal checklist to cover all issues likely to affect the internal validity of the presented trial results. Confounding may occur as a consequence of imbalance in biomarker status in the drug and usual care trial arms in the case where biomarker status is also a prognostic factor.

38) (O & D) Does this drug effectiveness evidence, as conditioned by the test or biomarker result, show a clinically important and statistically significant impact on patient-relevant health outcomes (both safety and effectiveness)?	Relate this to factors intrinsic to the proposed drug: i) treatment effect modification when prognostic impact is not present in the drug/biomarker relationship, and/or ii) absolute treatment effect when prognostic impact is present in the drug/biomarker relationship (see Information Request 35). And to the factor intrinsic to the proposed test: iii) identification of true biomarker status given test result status (*i.e.* positive predictive value and negative predictive value) or evidence that there is complete agreement on an individual patient level between test outcomes across the proposed test and the test used to identify patients in the evidence provided.
39) (O) Is the evidence supporting the pairing of the co-dependent technologies applicable to the intended populations?	

(Source: [9] with minor adaptations)