The Health Technology Assessment Approach of the Economic Value of Diagnostic Tests - A Literature Review

David Bardey, Philippe de Donder, Vera Zaporozhets

WORKING PAPERS

N° 1508

February 2024

"The Health Technology Assessment Approach of the Economic Value of Diagnostic Tests A Literature Review"

David Bardey, Philippe De Donder and Vera Zaporozhets

Toulouse School of Economics

Toulouse School of Economics

TSE HEALTH CENTER

THE HEALTH TECHNOLOGY ASSESSMENT APPROACH OF THE ECONOMIC VALUE OF DIAGNOSTIC TESTS

A LITERATURE REVIEW[1]

David Bardey (U Los Andes and TSE, d.bardey@uniandes.edu.co)

Philippe De Donder (TSE and CNRS, philippe.dedonder@tse-fr.eu)

Vera Zaporozhets (TSE and INRAe, vera.zaporozhets@tse-fr.eu)

February 2024

Economics for the Common Good

[1] The authors acknowledge financial support from the French Agence Nationale de la Recherche under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

ABSTRACT

We review the medico-economic literature assessing the economic value of diagnostic tests. We first present the health technology assessment methods, as applied to generic health interventions. We then define our object of study, diagnostic and prognostic tests, and relate them to various definitions of personalized medicine. We then review the empirical assessments of diagnostic tests related to personalized medicine andof companion tests. We summarize systematic reviews which are not performing quantitative meta-analyses, but rather provide a descriptive synthesis of the results reviewed. We find no evidence that such tests perform better than more traditional approaches, such as pharmaceutical interventions. At the same time, there is a lot of heterogeneity in the cost per QALY (Quality-Adjusted Life Year) gained, so that some genetic testing procedures may perform better than non-genetic ones. Finally, we focus on imperfect tests and show how to optimize, from an economic perspective, their accuracy levels, and how to take accuracy levels into considerations when assessing their economic value.

Keywords: genetic tests, companion tests, cost-benefit analysis (CBA), cost-effectiveness analysis (CEA); cost-utility analysis (CUA); and cost-minimization analysis (CMA), personalized medicine, Receiver-Operator (ROC) curve, Incremental cost-effectiveness ration (ICER)

JEL Codes: H51, I18, J17

1.Introduction

Diagnostics serve essential functions in health systems, enabling epidemic response, health surveillance, and screening programs. They are also critical for achieving universal health coverage and the United Nations' Sustainable Development Goal 3: "Ensure healthy lives and promote well-being for all at all ages." (United Nations, 2015). Hereafter, we refer to diagnostics as any equipment, method, or system used for determining a medical diagnosis (*e.g.,* World Health Organization, 2011 or McNerney, 2015).[2] While for some medical tests only minimal equipment is necessary, such as auscultation, other medical tests require extremely expensive medical equipment such as magnetic resonance imaging (Snowsill, 2023).

During the last decades technological innovation has led to remarkable developments in health care. The breakthroughs are not only limited to the invention of new drugs and vaccines, targeted cancer therapies, innovative diagnostic imaging, and minimally invasive surgery. Advances in genomics (Human Genome Project) have allowed to identify different diseases subtypes based on genetics. Such knowledge helps to determine whether patients with certain disease subtypes are more likely than others to be responsive to a particular drug. Innovative treatment is often coupled together with a specialized diagnostic test (called companion diagnostics). Nowadays, the diagnostic testing has become essential not only in establishing a diagnosis, but also in taking decisions on management strategies providing the information on how patients are stratified into the most appropriate treatments. Given this importance, the question is how to evaluate a diagnostic test?

Historically, the primary focus of diagnostic tests' evaluation was on their clinical accuracy, *i.e.* how good they are at categorizing patients as having or not having the disease. Although test accuracy is an important component of test evaluation, the effectiveness of a diagnostic test cannot be narrowed down to clinical accuracy. Thus, the WHO has published criteria for an ideal test that can be used at the point of care. These criteria are known by the acronym ASSURED[3] and embody three key characteristics: accuracy, accessibility and affordability. As no test is perfect, the tradeoffs between the three criteria need to be considered for the different levels of the health care systems. The ideal diagnostic test would have an accuracy of 100%, but such perfection is not achievable in routine clinical practice and compromises may be needed between accuracy and accessibility.

---

[2] The UK Faculty of Public Health stresses that "screening tests are not diagnostic tests" (see https://www.healthknowledge.org.uk/public-health-textbook/disease-causation-diagnostic/2c-diagnosis-screening/screening-diagnostic-case-finding). The primary purpose of screening tests is to detect early disease or risk factors for disease in large numbers of apparently healthy individuals. The purpose of a diagnostic test is to establish the presence (or absence) of disease as a basis for treatment decisions in symptomatic or screen positive individuals (confirmatory test).

[3] "ASSURED" - (a) affordable, (b) sensitive, (c) specific, (d) user friendly, (e) rapid and robust, (f) equipment-free, and (g) deliverable to end-users.

Clinical accuracy is an important component of test evaluation, but it does not capture the impact of test on the patient outcome. Ideally, a new test should be introduced into clinical practice if it has a better chance of improving patient health than existing tests (di Ruffano *et al*., 2023). When a new diagnostic test is introduced to healthcare system, economic evaluations assess the comparative effectiveness of a new diagnostic technology and balance it against its expected costs. Economic evaluation can be challenging as the relationship between its use and health outcomes and total costs is indirect. Moreover, each test needs to be matched to its testing environment, which includes population characteristics, prevalence of target diseases, health system characteristics.

In the last decades, population aging, increases in chronic disease and in healthcare costs have become prominent challenges all over the world (ex., Nimmesgern *et al.*, 2017). In order to address these challenges, health authorities in various countries are looking for instruments helping to implement new effective health technologies while controlling for health expenditures. This can be achieved through health economic assessments to assist in informed decision-making on allocation of limited health-care resources and on pricing and reimbursements.

This survey reviews the medico-economic literature assessing the economic value of diagnostic tests for individual medical decisions (as opposed to public health), using the health technology assessment (HTA henceforth) approach. HTA plays an important role in assessing "the value of money" of health technologies and interventions. Section 2 describes and compares the four main valuation methods used for generic treatments, namely the cost-benefit analysis (CBA), cost-effectiveness analysis (CEA); cost-utility analysis (CUA); and cost-minimization analysis (CMA). These methods differ in how the health effects of the procedure are measured, and in how they are compared to their costs. The main result of this section is that the CEA/CUA are recommended by many health authorities across the world, although their economic foundations are still being assessed (and criticized) by theoretical economists.

In Section 3, we define more precisely our subject of research, the diagnosis and prognostic tests, which reveal what the patient suffers from and what treatment is best suited for him/her. We focus more specifically throughout the review on companion tests, which come together with a specific treatment. We then link these companion tests with the ubiquitous term of "personalized medicine''.

We then turn in section 4 to the empirical assessment of innovative tests. In the case of diagnostic technologies, the HTA methodology is not as established as for treatments (*e.g*., van der Pol *et al*., 2021). One of the main reasons for that is that, contrary to pharmaceuticals, which directly influence the patient's health outcome, the impact of diagnostic technologies is indirect and only takes effect when diagnostic results change downstream clinical interventions. Medical tests have the potential to improve patient outcomes if improvements in accuracy are translated into more appropriate diagnoses and more appropriate treatments. Furthermore,

tests may offer similar accuracy at reduced cost, simplify healthcare delivery, improve diagnostic confidence and improve diagnostic yield, reduce time to diagnosis, lower patients' anxiety, reduce uncertainty or improve safety (*e.g.*, di Ruffano *et al*., 2023).

We then review the recent empirical literature on the economic value of innovative tests, mostly dubbed precision medicine tests here. There are still quite few such tests, both because the technology is still in its infancy, despite the high hopes raised by the Human Genome Project, and because there are few results from clinical studies. Rather, most studies are based on analytic modelling, sometimes called indirect evidence for the clinical assessment of the test. The assessment method most used recently is the CUA, and studies show a small fraction (of around one fifth to one quarter) of genetic tests resulting in cost savings (because they allow to skip costly treatments for non-responsive patients), with the bulk generating health improvements with higher costs, including for a ratio of cost to benefit that looks effective by today's standards.

Up to now, we have taken the characteristics of the tests as given. But, except may be for some of the genetic tests reviewed in section 4, tests are imperfect in the sense that they make wrong predictions for a subset of the tested population. It is then most important to take this accuracy problem explicitly into account when assessing the economic value of these tests. Moreover, their accuracy degree is often endogenous, the result of a trade-off between false positives and false negatives (or specificity and sensitivity respectively, as they are called in the health literature).

In section 5, we first present the canonical framework used to determine the trade-off between sensitivity and specificity in the design of a test. We then show how authors such as Laking *et al.* (2006) employ this framework (making use of the central concept of the Receiver-Operator, or ROC, curve) to assess the value of the information brought by the tests.

In the second part of section 5, we go back to the empirical approach, presenting with the help of Sutton *et al*. (2008) the methodology to be used to proceed to meta-analyses of tests which differ in their sensitivity and specificity. We then survey the recent methodological and/or empirical literature adopting this methodology.

Section 6 concludes this document, while we summarize at the end of each section its main key messages.

2. Health Technology Assessment Methods

Health technology assessment (HTA) is designed to provide a coherent framework for informing choices of treatment interventions based on maximizing health outcomes under limited available resources. HTA can be conducted for all types of

interventions: diagnostic, surgical, medical, behavioral or complex, which can include pharmaceutical and medical devices (European Network for Health Technology Assessment, 2015; National Institute for Health and Care Excellence, 2013). In the analysis below, an intervention is compared to one or more alternative interventions called comparator(s).

The Diagnostics Assessment Programme (DAP) was set up by NICE in 2009 to evaluate innovative medical diagnostic technologies. There are four main types of economics evaluation methods, according to how health outcomes are measured and valued: cost-benefit analysis (CBA), cost-effectiveness analysis (CEA); cost-utility analysis (CUA); and cost-minimization analysis (CMA).[4] The table below summarizes how the effects are measured in each method. The literature also sometimes references at the fifth method, called cost consequences analysis (CCA).

Table 1: the different types of economic evaluation (Adapted from Abbott *et al.*, 2022)

Cost–benefit analysis (CBA)	Effects are measured in monetary units
Cost-effectiveness analysis (CEA)[i]	Effects are measured in any other unit of effect, *e.g.*, life-years gained, deaths averted, jobs saved, treatment responders, units of a patient-reported outcome measure, …
Cost-utility analysis (CUA)	Effects are measured in QALYs (or less commonly DALYs), which are utilities summed over time
Cost-minimisation analysis (CMA)	Effects are not considered, just costs alone

One QALY (or Quality Adjusted Life Year) is equal to one year in perfect health. It ranges from 1 (perfect health) to 0 (dead). For example, a person with chronic disease in which they experience utility of 0.5, will have ½ a QALY in one year, and 1 QALY over two years. QALYs are thus a measure of the total amount of (quality-adjusted) health experienced by an individual over a period of time; so even though utility is measured on a scale from 0 to 1, the QALYs reported in a given study can range from 0 to the length of follow-up (in years).

We are now going to cover each method sequentially.

[4] Often the terminology can be confusing, as these terms are used in various ways by different authors and do not always accurately describe the nature of the research (*e.g.*, Drummond *et al.*, 2015).

Cost-benefit analysis (CBA).

CBA is a form of comparative analysis of interventions with the distinguishing characteristic that it places monetary value on the consequences. Therefore, the resulting benefits (consequences) and the costs are measured in the same units. The main principle of CBA states that the health intervention is desirable if the benefit is greater than the costs (cost-benefit criterion). This criterion can be alternatively formulated in terms of the benefit-cost ratio: if it exceeds unity the intervention should be approved. If there is a choice between several interventions, on top of this requirement, the best alternative should have the highest benefit-cost ratio. It means that the chosen intervention has higher benefit per monetary unit spent on costs than under the alternative use of funds.

In the health care evaluation field, there has been a general difficulty to express the health outcomes (such as survival) in monetary units. For this reason, CBA is not currently widely used as a type of health-economic evaluation while it is popular in other fields. Guidelines of several EU countries (Finland, Portugal, Russia, Spain and Sweden) include CBA as a possible type of analysis, while some others (Belgium, Hungary, Italy, Norway) state that CBA is not a recommended type of analysis or that it should be used as a complementary analysis. It has been argued that QALYs are not appropriate when the health condition is acute, and that in this case an alternative approach such as willingness to pay should be employed. For example, the guidelines for Sweden state that a CBA may be used in cases of difficulties to use QALYs (*e.g.*, when an intervention is associated with severe pain over a short period of time) (European Network for Health Technology Assessment, 2015).

Cost-effectiveness analysis (CEA).

In contrast to CBA, CEA compares the relative costs and effects of different interventions without pricing the effects. Each intervention can then be summarized as a point on a bi-dimensional graph, with the net input cost (in monetary units) on the vertical axis, and the effect (measured in life-years gained, deaths averted, jobs saved, *etc.*) on the horizontal axis. In order to compare different interventions, it is primordial to use the same measure of the health effects throughout. Also, this approach is limited to a single health effect measure per study.

The CEA is often performed compared to an existing intervention, or standard-of-care. In that sense, one measures the *incremental* cost of the assessed procedure (compared to this standard) and the *incremental* health result. The costs include the price and associated medical costs of the new treatment and standard of care. An important element to consider is savings from avoiding adverse effects associated with the standard of care, which especially applies in personalized medicine (see

next sections). Each of these measures (incremental cost and effect) can then be positive or negative, resulting in the four quadrants depicted in Figure 1.

The CEA may be expressed in terms of *incremental cost-effectiveness ratio* or ICER. This is typically the net input costs (in monetary units) to achieve each unit of health outcome:

$$\text{ICER} = \frac{\text{the change in costs}}{\text{the change in effect}}.$$

The ICER then corresponds to the slope of the line linking the studied procedure to the (0,0) point in Figure 1. Interventions can be ranked by ICER from lowest to highest. The most cost-effective intervention has the lowest ICER. Note that the ICER can be negative, when a better outcome is reached for a lower cost (Quadrant 4 on Figure 1) for instance.

Under CEA, an intervention is considered cost-effective if the ICER is lower than a given value called the *cost-effectiveness threshold* (*e.g.*, it can be measured through the maximum willingness to pay for one additional unit of health outcome). There is a lot of controversy around the use of a cost-effectiveness threshold, about what the threshold should represent, and about the appropriate methodology to arrive at its value (*e.g.*, Vallejo-Torres *et al*., 2016 and Nimdet *et al*. 2015). We will come back to this point when we discuss the Cost Utility Approach below.

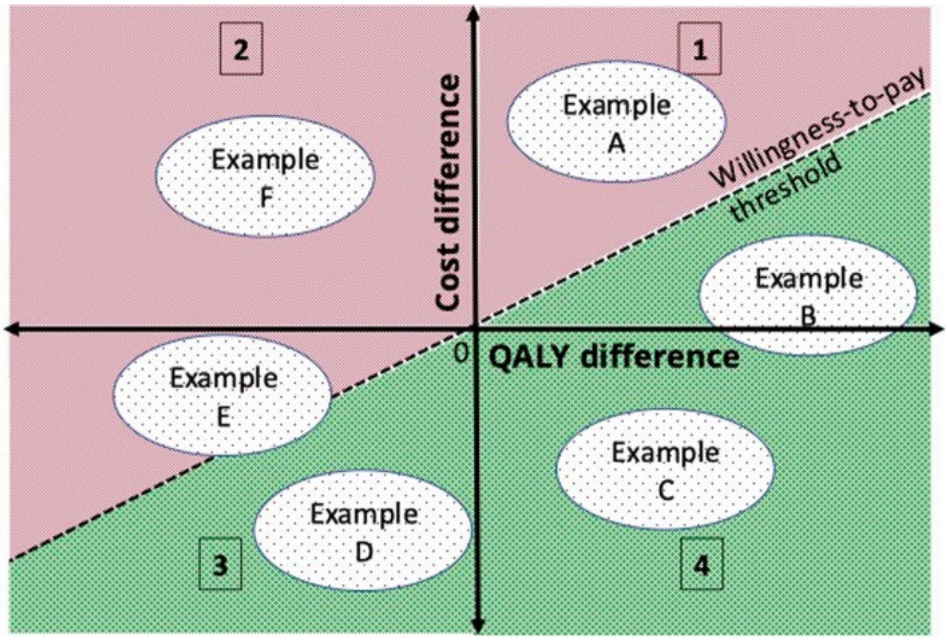
An alternative measure to the ICER that is also often used in the literature is the incremental net monetary benefit or NMB. This measure has a more straightforward interpretation as it allows to avoid the ambiguity of what a positive or a negative ICER means (Abbott *et al.*, 2022).[5] The NMB is calculated as the product of incremental effects and willingness-to-pay (or cost-effectiveness threshold) minus incremental costs. A positive MNB means that the treatment is cost-effective at the given willingness-to-pay threshold and hence is worthwhile compared to the comparator.

We are now in a position to discuss the various examples depicted in Figure 1. The cost effectiveness plane depicted in Figure 1 has 2 axes that illustrate the incremental difference between the intervention group(s) and the comparator group. Differences in effects are measured on the horizontal axis, and difference in costs on the vertical axis. The plane is divided into 4 quadrants. All procedures in Quadrant 2 (such as Example F) are obviously dominated by the comparator, since they cost more for a worse result. Likewise, all procedures in Quadrant 4 dominate the comparator, since they cost less for a better result. To determine whether examples in Quadrants 1 are 3 should be chosen, we must depict how much society is ready to pay for each additional unit of health outcome. This corresponds to the "Willingness-to-pay" (WTP hereafter), or cost-effectiveness threshold on the figure.

[5] Obviously, better effects at lower costs are always desirable, while worse effects at higher costs never are, while both situations exhibit a negative ICER.

All examples below this line (in the green area) are deemed worth implementing (and have a positive MNB), while all examples above are not (and exhibit a negative MNB). Note that positive NMB examples are composed both of procedures where the added benefit is worth its cost (such as Example B), but also of procedures with worse health outcome if they decrease sufficiently the cost incurred (Example D). Symmetrically, procedures deemed not cost-effective can increase health outcomes (but at too high a cost, see Example A) or decrease them, if the costs decreases are too low (Example E).

Figure 1: The cost-effectiveness plane



Cost-utility analysis (CUA).

CUA is a form of CEA, where the health outcomes are measured in terms of quality-adjusted life years (QALY). Alternatively, CUAs sometimes use the disability-adjusted life year (DALY) as another measure of disease burden, expressed in terms of the number of years lost due to ill-health, disability or early death. DALYs are calculated as a sum of years lived with disability and the years of life lost. This measure combines measures of life expectancy as well as the adjusted quality of life during a disease or disability for a population, therefore it is viewed as a societal measure of the disease burden to the contrary to QALY which tend to be an individual measure.

It is generally considered best practice to design a CUA, a single summary ratio which provides information on the incremental cost *per* QALY gained of a new

technology compared to the current best practice. A CUA is a frequently used and recommended approach due to its ability to compare the results across different health programs and policies using a common unit of measure (Drummond *et al.*, 2015). A CUA though is not always possible or practical, particularly when information about morbidity is not available to calculate QALYs.

As for the value to be put for each unit of QALY (the cost-effectiveness ratio), it can be based on different theoretical or methodological approaches, namely the opportunity cost approach (supply-sided approach) and the willingness to pay (WTP) approach, or demand sided approach (*e.g.*, Baker *et al.*, 2011, Ryen and Svensson, 2015, Neumann *et al.*, 2015). The opportunity cost/supply-side threshold reflects the opportunity cost associated with devoting health system resources to a particular use and hence, a forgone benefit that could have been achieved if the same resources were used for other activities. This implies that the threshold value represents the shadow price of the budget constraint.

The second approach, the WTP or demand-side approach, relates to the willingness to pay for health improvement of a relevant group of individuals (*e.g.*, the general public), or less often of the patients. The WTP approach may be traced back to attempts to link CEA with CBA and welfare economics. If QALYs satisfy certain conditions such that they represent utility, and there is one societal WTP for a QALY, then CEA can be reformulated in a way that is equivalent to the CBA (Phelps and Mushlin, 1991).

Cost-minimization analysis (CMA).

CMA does not take health outcomes into account and only focus on costs. Therefore, CMA can be conducted when it is demonstrated that there is no difference in the effect between an intervention and its relevant comparator (*e.g.*, European Network for Health Technology Assessment, 2015).

Cost consequences analysis (CCA).

CCA is an economic evaluation, in which disaggregated costs and a range of outcomes are presented to allow decision-makers to form their own opinion on the relevance and relative importance to their decision-making context (Drummond *et al.*, 2015). CCAs have been recommended for complex interventions that have multiple effects, and public health interventions, which have multiple health and non-health benefits that are difficult to measure in a common unit (NICE, 2013). The outcomes are not restricted to health outcomes such as QALYs and can include other measures, for example non-health considerations relevant for decision-makers. CCA may be of particular value to funders that are more concerned with patient-oriented outcomes and intervention costs. CCAs may also be useful in

feasibility or pilot studies when it is not clear which costs and outcomes will be most relevant to future definitive trials. CCA is considered as an underused method of economic evaluation (Hunter and Shearer, 2014).

CEA/CUA as the preferred assessment methods

Since the seminal work of Weinstein and Stason (1977), CEA have been widely published in the US medical literature covering a diverse set of drugs, devices and medical procedures. However, they have received mixed welcome in the US medical healthcare (Neumann *et al.*, 2015).[6] Due to concerns on the methodological standards, in 1993, the US Public Health Service convened a panel of 13 non-governmental scientists and scholars to review the state of the field of CEA and to provide recommendations for the use and conduct of the CEA in health and medicine. The primary goals were to improve quality of CEA and to promote comparability across studies as many published studies described as CEA used "surprisingly different methods" (*e.g.*, Roberts, 2016). In 2016, the Second Panel on cost-effectiveness in health and medicine updated the work of the original panel by reflecting on the evolution of CEA and its perspectives (see, for example, Neumann *et al*., 2015). It continues to recommend QALY as the best societal outcome measure although with caveats.[7]

Most EU countries recommend using the CUA as the main type of analysis. In some cases (*e.g.*, France, Ireland and the Netherlands), to enhance the usability of the economic evaluations, it is also recommended that the results of the CUA be accompanied by a CEA with the costs per life-year gained (LYG) as the outcome measure (European Network for Health Technology Assessment, 2015). Other guidelines (in Belgium, Norway and Sweden) state that the CUA should be always accompanied by a CEA with costs per life-years gained as the outcome measure (European Network for Health Technology Assessment, 2015).

There are some controversial issues in the CEA (see Garber [2000] for a discussion) such as:

(i) How to include the indirect time-related costs of treatment or benefits?
(ii) Should CEA include future medical costs incurred during years of life "extended" by a current medical intervention?

[6] There is a certain reluctance towards using CEA within the US legislative system such as the specific prohibition in the Affordable Care Act about the use of QALY and cost effectiveness in allocation decisions, or the congressional prohibition on funding CEA. Those restrictions have led to limited public funding for these analyses (*e.g*, Roberts, 2016).

[7] The panel members specifically acknowledge the problem with aggregating the outcomes across individuals. For instance, it is unlikely that many would consider the saving of 1 minute of life among 525 600 people as being equivalent to the saving a year of life in a single individual (Roberts, 2016).

(iii) Does the applied in CEA measures of effectiveness (*e.g*., incremental life years) discriminate against older patients?
(iv) Is it possible to find an "optimal" threshold for cost-effectiveness ratios?

As for the latter point, the values taken by this threshold vary a lot between countries. Decision-makers may use either implicit or explicit threshold values. Explicit threshold values mean that decision-makers have formally adopted and made the threshold public, and their decisions on resource allocation are based on these values. By contrast, implicit thresholds are not official or public, but may be inferred retrospectively by analysis of the decision-making pattern in a given health-care system. Thus, for the National Health Service (NHS) in the UK, they explicitly set the threshold used on behalf of the NHS at £20,000/QALY, ranging up to £50,000 for life-threatening conditions (Garrison and Towse [2017]; Chen *et al*. [2020]). In a US context, the cost-effective ratio may vary from $50,000/QALY up to $150,000/QALY—or more depending on individual or disease. Hirth *et al*. (2000) report that the value of $50,000/QALY was originally based on the supposed annual cost per QALY for the Medicare program for patients with chronic renal failure. In Sweden and the Netherlands, relevant government authorities have recommended the thresholds of 500,000 SEK (approx. € 57,000) (see, *e.g.,* Ryen and Swensson, 2015) and € 80,000 (Bobinac *et al.*, 2010), respectively.

In 2011-2012 the French HAS (*Haute Autorité de Santé*) published a CEA guideline and a subsequent law enacted making CEA mandatory for determining pricing and reimbursement for new drugs and medical therapies (Haute Autorité de Santé, 2012). However, until recently no cost-effectiveness threshold has been officially proposed to qualify ICERs. The study by Téhard *et al.* (2020) proposes a method for estimating a value for statistical QALY that can be used as reference values for ICERs in health assessment in France. The estimated reference values of €147 093 to €201 398 for a QALY are provided as appropriate thresholds. One of the big limitations to the widespread adoption of CEA is the absence of a worldwide accepted rule to set the relevant cost-effectiveness thresholds. The best-known recommendation is the WHO rule that considers an intervention to be cost-effective if a healthy year is gained at less than three times the GDP *per* capita. In the past decade, several studies have challenged this rule by showing substantially lower cost-effectiveness thresholds (less than 1 GDP *per* capita) in different countries (especially low and middle-income countries). They argue that higher thresholds may boost health expenditure *per* capita. Thus, the recent study by Pichon-Riviere *et al*. (2023) presents a conceptual framework to estimate cost-effectiveness thresholds, and then empirically derives them for 174 countries, using World Bank data on country specific health expenditures and health outcomes for the period 2010-2019. The findings suggest that the cost-effectiveness thresholds *per* QALY should vary between US$87 for the Democratic Republic of Congo and $95 958 for the US and are less than 0.5 GDP *per* capita in 96% of low-income countries, 76% of lower-middle countries, 31% of upper-middle countries, and 26% of high-income

countries. Cost-effectiveness thresholds *per* QALY are less than 1 GDP *per* capita in 168 (97%) of the 174 countries. Cost-effectiveness thresholds *per* life-year range between $78 and 80 529 and between 0.12 and 1.24 GDP *per* capita and are less than 1 GDP *per* capita in 171 (98%) countries.

Recently, in healthcare the focus on societal perspective has been put forward, reflecting the viewpoint of the decision-maker with an intention to allocate optimally health resources across entire population. Therefore, when evaluating interventions, the following health and non-health consequences should be included: consequences on economic productivity, education, social services, criminal justice, housing or environment (Neumann *et al*., 2015). In this context, CBA is argued as a more relevant economic evaluation method (*e.g*., Brent [2023]). This is because it ensures that outputs will be valued in monetary terms, and therefore made comparable to the costs, and determine whether the expenditure is socially worthwhile or not. Also, it provides a social perspective by including effects on everyone affected by an intervention both directly and indirectly. In contrast, according to CEA, an intervention can be cost effective and not socially worthwhile, or it can be the least cost-effective intervention yet, none-the-less, be socially worthwhile. Another limitation of CEA is that it does not consider externalities. Finally, it is worth mentioning that CUA uses a single threshold price that is irrespective of the preference of the persons who are receiving the benefits of the interventions.

Roberts (2016) points out that the most important problem with CEA is not the accuracy and consistency but rather “it is the misunderstanding and subsequent prohibitions of CEA use to improve resource allocation in US health care…Trade-offs are unavoidable in the allocation of resources, and the methods of CEA render those trade-offs explicit and debatable.”

To conclude this section, we note that, although CEA/CUA are the assessment methods recommended by many health authorities across the world, their economic foundations are still being assessed by (theoretical) economists (see *e.g*., Garber and Phelps [1997]; Brouwer and Koopmanschap [2000]; Meltzer *et al*. [2016]). For instance, Garber (2000) demonstrate that CEA can provide an appropriate tool for choosing among health interventions within a standard utility maximization framework. Meltzer *et al*. (2016) consider theoretical grounds of CEA in constrained optimization, highlighting issues such as objectives to be maximized, constraints to be considered, resources consumed, and opportunities foregone. Even when the objective is to maximize health, there are multiple questions to consider as how to measure and combine effects on survival and health-related quality of life; how to measure costs; and how to treat effects that are uncertain or that occur over time. Other topics covered include theoretical issues that arise in the QALY model and their links to individual utility.

The following box contains a summary of the main points made in this section.

KEY TAKEAWAYS

There are several approaches which can be used, which mostly differ in how the benefits of the treatments are measured. Many health authorities recommend using the Cost Utility Approach (CUA) where health benefits are measured by the number of QALYs (Quality-Adjusted Life-Years) gained thanks to the evaluated procedure. Recommended treatments are those exhibiting a cost *per* QALY lower than society's willingness to pay for such improvements (also called the cost-effectiveness threshold). We nevertheless note that academic economists developing formal models are still assessing and debating the normative properties of this approach.

The next section introduces the type of health intervention whose value we want to assess, namely the diagnostic tests.

3. Diagnostic and companion tests

This literature review covers the so-called diagnostic (and prognostic) tests, encompassing all the procedures that allow to reveal either what the patient suffers from, and/or the treatment(s) that are best suited to tackle what ails them (including the likeliness of developing adverse effects, and the optimal doses to be used).

An important part of these diagnostic tests is constituted of so-called companion tests, which come together with treatment(s), with the objective of determining the adequacy between the treatment(s) and the patient. These companion tests are playing an increasingly important role, both in enhancing the use of existing treatments and in the authorization of new ones. For instance, a review by the European Medicines Agency shows that approximately half of cancer drugs authorized over the 2015-2018 period required patients to be screened by a genetic test before determining their treatment (Antoñanzas *et al*., 2019).

Before moving to the literature review proper, it is important to link diagnostic tests with the emerging field of personalized medicine. Unfortunately, as we explain in the Annex, there is no consensus on the definition of personalized medicine, with different jurisdictions (such as Europe and the U.S. for instance) using different meanings, and with a profusion of similar (but not identical) terms (such as personalized, precision, stratified or individualized medicine) being used in the literature.

At one extreme, all medicine can be dubbed as "personalized" since medicine has always been concerned with the individual needs, with an aim of diagnosing the specific ailment of a given individual, and the best treatment to be applied in that particular case. At the other extreme, the narrowest definition of personalized medicine requires the creation of drugs or medical devices that are unique to a

patient.[8] The typology proposed by Trusheim *et al.* (2007) is helpful in understanding this variance. They propose the patient therapeutic continuum illustrated on Figure 1 below. Most medicines have been prescribed empirically, meaning that either they work for all patients, or, if response rates are variable, there is no way to identify patients who are likely to respond well to a specific treatment. Advances in understanding the mechanisms underlying both the diseases and the drug responses have allowed to better match patients with treatments. At the extreme, the treatment is individualized, custom produced using the patient's own fluid, cells or tissue to seed production. This is what they call "individual medicine". In between these extremes, there is a growing number of cases where some marker exists which indicates whether a given patient is likely to react to a therapy. They propose to call these situations "stratified medicine" (rather than the more ambiguous term of personalized medicine). More precisely, "in stratified medicine, a patient can be found to be similar to a cohort that has historically exhibited a differential therapeutic response using a biomarker that has been correlated to that differential response" (p1).

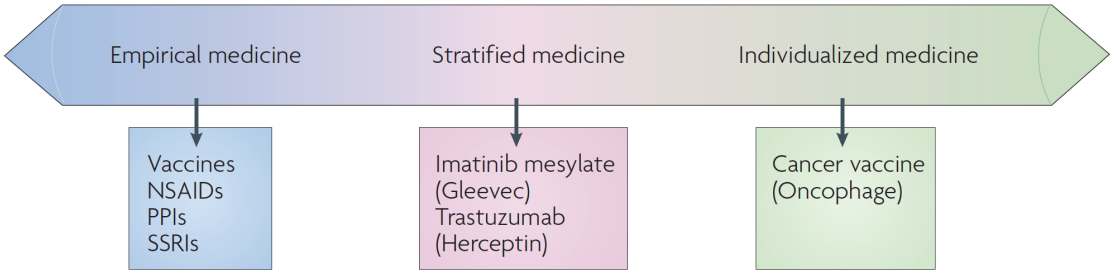


Figure 1 | **The patient therapeutic continuum.** Individualized medicines, such as cancer vaccines that are based on a particular patient's tumour, represent one end of a continuum of patient therapy. Empirical medicine is at the other end of this continuum: some agents work for almost all relevant patients, such as non-steroidal anti-inflammatory drugs (NSAIDs), whereas others may only work for a subset of patients but no method is available to identify these patients, such as with antidepressants. In between lies the field of stratified medicine, in which a patient can be found to be similar to a cohort that has historically showed a differential therapeutic response to a particular therapy using a clinical biomarker that has been correlated to that differential response. For example, the anticancer drug trastuzumab (Herceptin) shows superior efficacy in breast cancer patients with HER2/*neu*-positive cancer. PPIs, proton-pump inhibitors; SSRIs, selective serotonin-reuptake inhibitors.

Source: Trusheim *et al.* (2007).

[8] The cancer vaccine Oncophage is an example of individualized medicine. To produce this vaccine, tumor cells are taken from a patient during surgery. A heat-stock protein and its associated peptides, which represent a unique 'signature' of the patient cancer, are then isolated from the tumor cells and formulated into a vaccine for administration after the recovery of the patient from surgery. This vaccine, which is only suitable for the patient from whom it is derived, stimulates an immune response that attacks tumor cells remaining after the surgery (Trusheim *et al.*, 2007).

Equipped with this terminology, Figure 2 allows us to make the link with diagnostic tests, as they appeared both within the realm of empirical medicine (to confirm a diagnostic) and of stratified medicine (to test for the treatment response). We will cover both tests in this survey.[9] Note that the treatment offered to the patient may or may not be individualized, and that we are not going to cover susceptibility tests, which allow to assess the probability that an individual may develop a disease in the future (see Bardey and De Donder [2013], for instance).

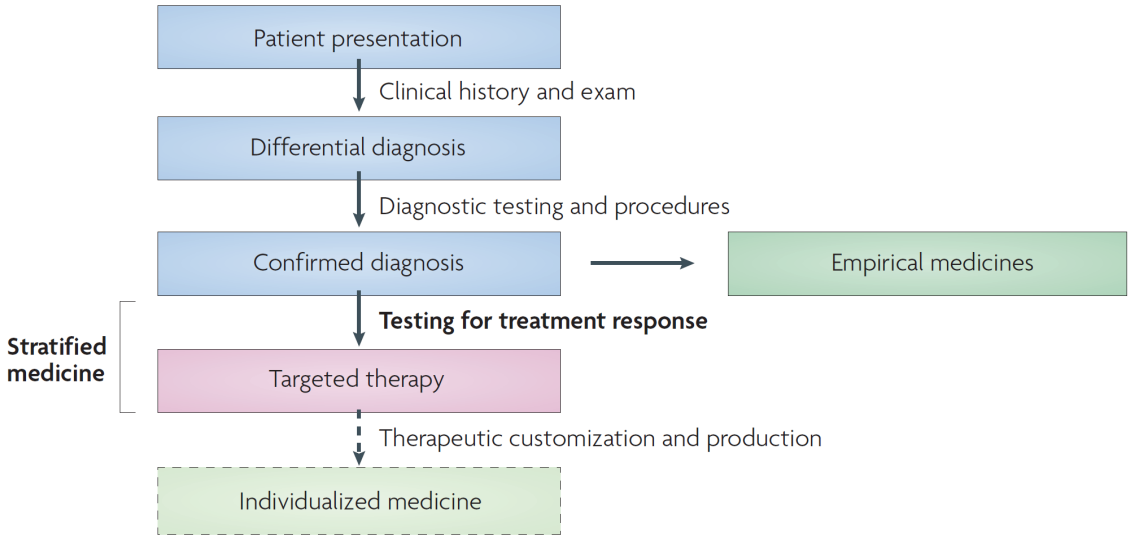


Figure 2 | **Stratified medicine in the clinical context.** In empirical medicine, a differential diagnosis is made on the basis of patient history and physical assessment, and following diagnosis confirmation from laboratory tests and clinical observation, a therapy is prescribed. Stratified medicine involves a further step in which a clinical biomarker is evaluated to associate a patient with a specific therapy. Extending this further, individualized medicine involves the customized production of the therapy (for example, using the patient's own cells).

Targeted therapy is the combination of test/treatment.

Source: Trusheim *et al.* (2007).

[9] See the Annex for more on how to link the various types of tests to several definitions of personalized medicine used in practice (especially in the U.S.A).

KEY TAKEAWAYS

The innovative tests that we survey in this document are the diagnosis and prognostic tests, namely all the procedures that allow to reveal either what the patient suffers from, and/or the treatment(s) that are best suited to them. We sometimes focus on companion tests, which come together with a treatment. We make the link between such tests and the so-called personalized medicine (PM). We review several definitions of personalized medicine co-existing in the literature, from empirical to stratified to individual medicine. All definitions of personalized medicine include the companion diagnostic tests (but of course, companion DT need not be genetic tests), while only the larger definitions include also the susceptibility tests, which we do not cover here.

We now move to the application of the HTA methods seen in section 2 to the diagnostic tests defined in this section.

4. The empirical assessments of diagnostic tests

Many recent systematic economic evaluations of medical tests are closely related to personalized medicine (PM). This is due to two related factors. First, the advent of PM has generated a lot of promises in terms of better understanding of diseases and of treatments. The recent empirical literature has then tried to see whether these promises have been translated into reality. Second, PM has generated two different predictions. The first one is that tests will allow to prescribe specific treatments only to those individuals with the higher probability of benefiting from them, the lower probability of developing harsh side effects, and in the right doses. This rosy picture then predicts both better health effects, and lower costs.[10] The second one recognizes the advent of very costly treatments, whose cost is made acceptable to society only because they are accompanied by tests restricting them to the individuals best suited to them. In that case, although the literature still predicts better effects, it should be accompanied by larger costs.[11]

It is important to recall one important difference between general diagnostic tests and tests of personalized medicine that involve biomarkers and that may convey information to develop a personalized therapy later on. Not only the price of the test matters, but also the price of treatments recommended by the test! As we have explained previously, most diagnostic tests allow physicians to choose the best

[10] This would correspond to Example C in Quadrant 4 of Figure 1 in Section 2.

[11] This would correspond to Quadrant 1 of Figure 1 in Section 2. Note that we could also observe that tests are used not to treat individuals with expensive drugs when the likelihood of success is low enough, generating both somewhat lower health results at much lower costs (Quadrant 3). We indeed have a few empirical observations corresponding to that situation, as we will point out below.

therapeutic alternative among different treatments whose cost may vary according to its patent status. Thus, the economic value of the diagnostic test also depends on the fact that some of the therapeutic alternatives may be relatively cheap, especially when some of them are off patent. Concerning personalized medicine, the cost structure is pretty different since not only the biomarker test may be more expensive but also the personalized treatment developed, even though in some cases such personalized treatments, especially when they are one shot, allow patients to get rid of chronic and costly treatments.

Historically, until recently the assessment of new diagnostic techniques has mainly focused on clinical validity such as test sensitivity and specificity. Test accuracy is only one component of test evaluation, but it does not capture the impact of test on the patient outcome. Ideally, a new test should be introduced into clinical practice if it has a better chance of improving patient health than existing tests (di Ruffano *et al*., 2023). One possibility to compare tests is by evaluating the downstream consequences of testing directly in a randomized controlled trial (RCT). Test-treatment trials randomly allocate patients to tests, follow up subsequent management, and measure outcome after treatment has been received. Test-treatment RCTs are however, are not very common due to many challenges in conducting such trials and, to deliver robust results (Yang *et al.*, 2019). A key issue for trials is that it is impossible to estimate the full effect of a test on costs and health without the use of modelling assumptions (Snowsill, 2023). Therefore, decision analytic modelling (called indirect evidence for the clinical assessment of the test) is recognized as a practical alternative. It facilitates the evaluation of both economic and clinical impacts simultaneously.

Since 2022, the European “in vitro diagnostic regulation” law has come into effect, making it mandatory for companies to prove clinical effectiveness of new diagnostics before they enter the market (van der Pol *et al*., 2021). In addition, recent systematic reviews of test–treatment randomized controlled trials demonstrates that improvements in test accuracy are rarely an indicator of patient health benefit (*e.g*., Yang *et al*. [2019]; Siontis *et al*. [2014]). It is also important to understand that a decision to introduce a new test cannot be restricted only to its accuracy but should also take into consideration other factors such as time to diagnosis and acceptability for patients. Therefore, to evaluate the impact of a new diagnostic test on patient health outcomes, it must be examined as part of a broader test–treatment management strategy (*e.g*., Snowstill [2023] and di Ruffano *et al*. [2023]).

Most studies we have uncovered take the form of systematic reviews, identifying relevant papers by querying databases (mainly but not exclusively PubMed) with terms related to economic evaluations (mostly the four types detailed in section 2) and tests (or versions of PM). Given the heterogeneity of the evaluations reviewed, the survey papers do not combine results to perform quantitative meta-analyses, but rather provide a descriptive synthesis of the results reviewed. The advent of PM

being relatively recent, most studies have been published in the last decade. We now present their results, starting with the oldest ones within this period.

Berm *et al*. (2016) proceed to a systematic review of the economic evaluations of pharmacogenetic and pharmacogenomic screening tests (the first term covering the study of single genes, the latter of several genes, both covered by the generic abbreviation PGx). They note that "PGx is nowadays often used as a synonym for personalized medicine, although personalized medicine is a much broader concept." (p.2). Their literature search on PubMed identifies a total of 80 studies ranging from 2000 to 2014. On methodology, they point that CEA (with results expressed in other dimensions than QALYs) was the most frequently applied study type before 2008, while CUA (expressed in QALYs) has been performed in most applications since 2008. They also note a bifurcation in the nature of the economic evaluations of PGx testing in, on the one hand, studies assessing the intrinsic value of a test and, on the other hand, studies assessing the value of the test in combination with an active compound (targeted therapies). While the evaluation of the test only (such as KRAS for colorectal cancer) results in both cost savings and in better health, the evaluation of targeted therapies generates better health but at a higher cost. Once the targeted therapy becomes the usual care, it is compared with new treatments, so that the economic evaluation of tests alone represents a dwindling fraction of the studies reported as time passes.

A quarter of the studies surveyed conclude that PGx testing is dominant, resulting in both clinical benefits and costs savings. Several recent studies further provide the specific conditions under which genetic testing might be cost-effective, for instance as a function of the patient population or the disease. Interestingly, three studies found that the GPx testing strategy was cost-saving, but with a small health loss (compared to the non-testing strategy) because of misclassification and thus suboptimal treatment of some patients. As for the studies comparing a PGx test treatment combination (targeted therapy) with an alternative (independent of pharmacogenetics) treatment, the latter were found to be cost-effective.

Although the authors document an increase in the quality of the studies as time passes, they mention two areas of concern to us. First, most studies lack solid clinical evidence of the testing strategy and have recourse to assumptions or experts' opinions. They also lack data with respect to heterogeneity in patient populations, hampering extrapolation of results to patients of different ethnicities, subpopulations and/or country specific populations. Second, they document both an increase in the proportion of studies funded by pharmaceutical companies (from none before 2008 to 24% after 2010), and the fact that, while all such studies conclude that PGx tests are dominant, 14% of the studies not funded by pharmaceutical firms find that PGx tests are not cost-effective. This suggests at the very least a publication bias for industry sponsored studies, as the positive biased results do not seem related to the quality of the studies.

Finally, two remarks are in order. First, there is a lot of heterogeneity in tests costs across countries, with costs (for the same tests) ranging from instance from £20 to US$575. Second, most studies assume that tests results are immediately available. Considering the turnaround time of the tests would then decrease their cost effectiveness.

Both Berm *et al.* (2016) and D'Andrea *et al.* (2015) share the observation that very few potential genetic/genomic applications (tests or interventions) have been implemented into clinical practice.[12] D'Andrea *et al.* (2015) mention as one barrier to implementation a lack of appreciation of the cost-benefit of new testing regimes, and thus proceed to a systematic review of 128 primary economic evaluations (EEs) of predictive genetic and pharmacogenetic testing programs, as well as to an overview of 11 previously published systematic reviews of such economic evaluations (economic reviews, ERs). All were published up to the end of 2012.

Cost-utility analysis (CUA) was the methodology most frequently used (73, 57%), followed by cost-effectiveness analysis (CEA) (67%), and most studies were performed either in the U.S. (48%) or the EU (36%). In terms of effectiveness, outcome measures were different according to the test category: for predictive genetic testing programs the results were mainly presented as LYGs (Life Years Gained), while for pharmacogenetic testing programs the outcomes most frequently used were QALYs. Predictive genetic testing programs were mainly concerned with prevention of oncological diseases (40%).

The key findings are as follows. A total of 138 incremental cost-effectiveness ratios were extracted from 66 CUAs and expressed as 2013 Euros per QALY gained. Only 12% of predictive genetic tests and 21% of pharmacogenetic tests are cost-saving. The majority (68%) of cost/QALY ratios indicate that genetic testing programs provide better health outcomes although at higher cost (corresponding to quadrant 1 on Figure 1, section 2), with almost half the ratios falling below €37,000 per QALY, a commonly used threshold (hence corresponding to example B in this figure). Seventeen percent of genetic testing programs are cost-saving (quadrant 4). Pharmacogenetic testing programs are more likely to be cost-saving, but predictive genetic tests more frequently result in cost-effectiveness ratios below the threshold of €37,000 *per* QALY.

This being said, D'Andrea *et al*. (2015) share the concern of Berm *et al.* (2016) of the absence of demonstrated clinical utility for a significant proportion of genetic tests, leading to their being not cost-effective.[13] Philipps *et al.* (2014) identified 59

[12] D'Andrea *et al.* (2015) cite the proportion of 3% of published research focused on the translation from experimental genetic/genomic applications to evidence-based guidelines and health care practice.

[13] D'Andrea *et al.* (2015)'s paper is motivated by the adoption in Italy of the National Prevention Plan 2014-2018 that has introduced genetic testing for BRCA as a preventive strategy aimed at reducing the incidence of inherited breast and ovarian cancer. It then reports only those details of incremental cost-effectiveness ratios for BRCA testing strategies, which thus serve as an example and case study.

cost–utility analyses studies that examined personalized medicine tests (1998–2011). A majority (72%) of the cost/QALY ratios indicate that personalized medicine tests provide better health, although at higher cost, with almost half of ratios falling under $50,000 *per* QALY gained, a commonly used threshold, and 80% falling under $100,000 *per* QALY gained. Twenty percent of the results indicate that tests are cost saving and 8% of the results that tests may cost more without providing better health.

Figure 2 in Philipps *et al.* (2014), reproduced below, provides a histogram of cost *per* QALY in these studies:

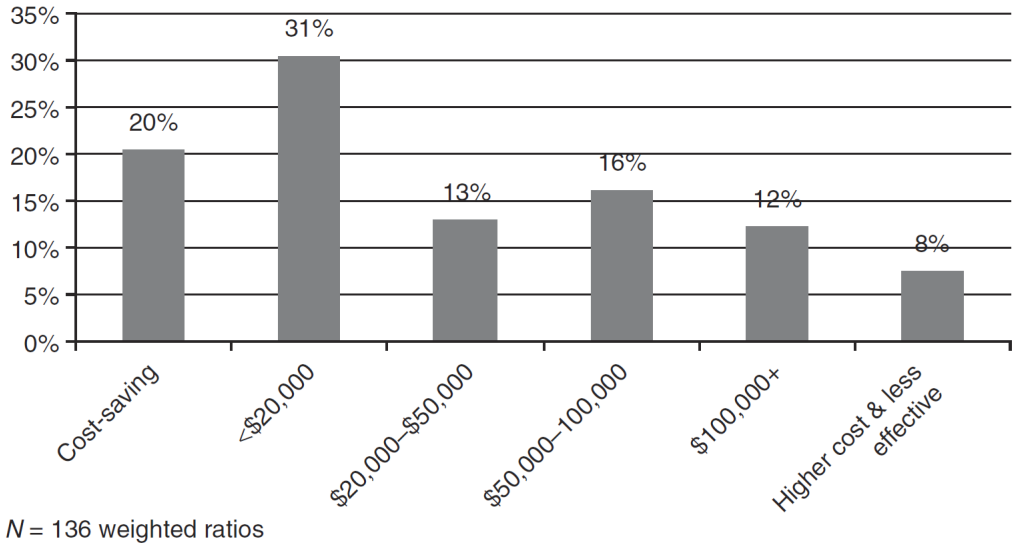


Figure 2 Distribution of ratios of cost per QALY gained for personalized medicine tests. QALY, quality-adjusted life year.

Philipps *et al.* (2014) also compared the CUAs of personalized medicine tests with CUAs of pharmaceuticals. They choose pharmaceuticals for comparison because these interventions are closely related to personalized medicine tests and because there are a large number of studies for analysis. Indeed, there are vastly more CUAs of pharmaceuticals (n = 1,385) than CUAs of personalized medicine tests (n = 59). Although the number of CUAs of personalized medicine tests are increasing over time, in 2011 there were still far more published CUAs of pharmaceuticals (n = 148) than of personalized medicine tests (n = 10). The distribution of cost/QALY ratios for

A systematic review of cost-effectiveness analyses clearly illustrates that there is evidence of cost-effectiveness only for genetic testing targeted to populations at high risk, such as the close relatives of carriers (cascade genetic screening programs).

somatic (acquired) *versus* germline (inherited) mutations and for personalized medicine tests versus pharmaceuticals was similar.

Summarizing the results provided in Philipps *et al*. (2014), Grosse (2014) stresses that just 6 of the 59 tests reviewed were classified by the Centers for Disease Control and Prevention as supported by evidence-based recommendations and concludes that "the primary constraint in understanding the economic value of genetic testing in medicine may not be lack of formal economic evaluations, but rather the unmet need for reliable, reproducible data on clinical outcomes." (p. 226).

Hatz *et al*. (2014) perform a literature search of MEDLINE database for cost-effectiveness analyses of Individualized Medicine (or IM) defined as a "therapeutic approach tailoring therapy for genetically defined subgroups of patients'' and including gene tests, chromosomal tests and biochemical tests.

They report results on 84 studies, mostly performed in the U.S. (51%) or Europe (32%). 79% of the studies performed a CUA (*i.e.*, expressed outputs in QALYs, the rest being expressed in Life Years Gained, LYGs). 71% of studies covered the period 2005 to 2012. Thirty-one different diseases were subject to analysis in the publications, with cancer diseases studied in 46 % of the articles.

Overall, 53 (63 %) studies found the ICER of individualized strategies to be acceptable in relation to their assumed thresholds. Dominance of the IM strategy was reported in six (7 %) studies. Twenty-one studies (25 %) presented an equivocal result, and four studies (5 %) stated that genetically guided care was not the favorable option. Interestingly, the cost-effectiveness of IM differed depending on the type of test. The median values of IM base-case ICERs for studies that included tests for disease prognosis ($US10,150/QALY gained) or screening ($US8,497/QALY gained) were lower than the medians for studies including tests to stratify patients experiencing adverse effects ($US39,196/QALY gained) and studies including tests to stratify patients for responders and non-responders ($US37,308/QALY gained).

Their conclusion is then that "generally, the existing evidence confirms neither the vision that IM is highly cost-effective nor the fear that it is associated with low benefit at high costs. Instead, the median of ICERs of IM CUAs ($US21,529/QALY gained) was in line with the value calculated by Neumann *et al.* (2009) in their review of CUAs from 30 years of cost-effectiveness analysis, which was $US22,000/QALY gained." (p.8) They also stress the heterogeneity between different test strategies. For instance, "tests for screening asymptomatic patients and tests for assessing the prognosis of a disease appeared to yield lower median IM base-case ICERs than tests for detecting responders or patients likely to incur adverse drug reactions." (p9)

Vellekoop *et al.* (2022) constitute both the most recent and the most complete assessment of the cost effectiveness of personalized medicine (PM), as they contain both a systematic literature review and a regression analysis. More precisely, they

investigate the net monetary benefit (NMB) of PM interventions instead of their ICERs and perform regression analyses in which they explore the heterogeneity in the cost-effectiveness of PM interventions.

They perform a systematic literature review to identify all published economic evaluations of PM between 2009 and 2019. PM was defined as "a medical model that bases therapeutic choice on the result of gene profiling or aims to correct pathogenic gene mutations," based on a study by Hatz *et al.* (2014) (surveyed above). Studies were included if they fell within this definition of PM, presented a cost-effectiveness model, provided patient-level cost and quality adjusted life-year (QALY) outcomes, extrapolated outcomes beyond short-term clinical trial data, and described an existing (*i.e.*, non-hypothetical) intervention. Studies also had to compare a PM intervention with a non-PM intervention.

A total of 128 studies were selected, providing cost-effectiveness data for 279 PM interventions. Most interventions are evaluated in the United States and the United Kingdom (48% and 16%, respectively). All included countries are upper-middle or high-income economies according to the World Bank country classification. The most frequently occurring cases were cancer treatments (60%) and pharmaceutical interventions (72%). Prognostic tests (19%) and tests to identify (non)responders (37%) were least and most common, respectively.

Regression analysis was conducted to explore the heterogeneity in the reported cost-effectiveness of PM in the included studies, aiming to identify characteristics of PM that may be associated with higher (or lower) health benefits, costs, and NMB. The paper performed separate evaluations of the QALYs, the costs, and the (incremental) net medical benefit (NMB) of the procedures, with the latter obtained by multiplying the gain in QALYs by the cost-effectiveness threshold of the corresponding country, and then by subtracting the cost (see section 2). The cost-effectiveness threshold used corresponds to the opportunity cost of healthcare spending (rather than to society's willingness-to-pay for increases in health), because of the availability of national estimates for all countries included in the data set.

The median amount of gains in QALYs of PM interventions relative to their non-PM comparators was 0.03, whereas the mean was 0.26. Most (incremental) QALY values are just above 0, with 0.00 and 0.16 at the 25th and 75th percentile, respectively. These figures are comparable with the QALY gains found by a literature review of cost-utility analyses for all types of healthcare, which identified a median QALY increase of 0.06 (mean 0.31). The health benefits of PM then tend to be similar to (or possibly slightly lower than) the health benefits of other (new) healthcare interventions. The regression analysis is suggesting large QALY gains for gene therapies. This may be because most of the gene therapies included in the review focus on early onset conditions with high morbidity and mortality.

Median costs were Int$575, whereas mean costs were close to Int$ 100,000. A small number of interventions have notably higher costs than the rest. On average, the cost for gene therapies is more than 1 million Int$ higher than for PM interventions that are not gene therapies. Median NMB across the included interventions was Int$18, and mean NMB was Int$277 072. NMB centers around 0, with a value of Int$ -22,665 at the first quantile and Int$3,538 at the third quantile. Extreme negative values are more common than extreme positive values for NMB. The median NMB of PM close to 0 implies that any QALY gains of PM interventions tend to be counterbalanced by their costs to the healthcare system.

On average, gene therapies bring Int$868,759 less net benefit compared with non-PM interventions, despite offering higher QALY gains. This implies that the costs associated to gene therapies are higher than the monetary value of the QALY gains, leading to a net loss.

PM interventions in neoplasms (cancers) have lower costs and higher NMB than other procedures. The regression coefficient for pharmaceutical interventions is positive in the QALY and costs models and negative in the NMB model. This means that although PM pharmaceuticals have higher health gains than non-pharmaceuticals, PM pharmaceuticals come at a higher cost than non-pharmaceuticals, causing lower net value (NMB). Finally, the positive coefficient for "industry sponsorship" in the NMB means that reported industry-sponsored studies are more likely to have positive cost-effectiveness outcomes. This in line with the concern stressed by Berm *et al.* (2016) (see above) about the publication biases linked to the sponsorship of the studies.

The following table summarizes the results obtained by these papers.

Study	**Purpose and Methods**	**Main Conclusions**
Philipps *et al.* (2014)	-CUA of personalized medicine tests during 1998-2011.	- A majority (72%) of studies show that personalized medicine tests lead to better health although at higher cost. -20% of studies indicate that the tests are cost saving; -8% of the results demonstrate that tests may cost more without providing better health.
Hatz *et al.* (2014)	-84 Cost effectiveness studies of individualized medicine (IM) performed in the US (51%) or Europe (32%). 71% of studies covered the period 2005-2012. -79% of studies performed CUA.	- 63% of studies found the IM strategies cost effective. -25% presented an equivocal result. - 5% of studies stated that genetically guided care was not the favorable option.

		-The existing evidence confirms neither the vision that IM is highly cost-effective nor the fear that it is associated with low benefit at high costs.
D'Andrea *et al*. (2015)	- A systematic review of economic evaluations (EE) of predictive genetic and pharmacogenetic testing programs up to the end of 2012. Most studies were performed in the US (48%) or the EU (36%). - CUA was the most frequently used (73,5%). - CEA is the second most frequent methodology (67%).	- The majority of studies (68%) indicate that genetic testing programs provide better health outcomes although at higher cost. -17% of genetic testing programs are cost saving. - Predictive genetic tests (contrary to pharmacogenetic testing programs) more frequently result in cost-effectiveness below the threshold.
Berm *et al.* (2016)	- A systematic review of the economic evaluations of pharmacogenomic screening tests (PGx). - CEA was the most frequently applied type of analysis before 2008. - CUA had been performed in most applications since 2008.	- The evaluation of test only results in both cost savings and better health. - The evaluation of targeted therapies (combination of a test with an active compound) generate better health but at a higher cost. - A quarter of studies conclude that PGx testing is dominant and results in both clinical benefits and cost savings. -Several studies provide the specific conditions for genetic testing being cost-effective. -Possibly, there is a bias for industry sponsored studies.
Vellekoop *et al.* (2022)	-Most recent and most complete assessment of the cost effectiveness of personalized medicine based on Hatz et al. (2014). Most intervention are evaluated in the US (48%) and the United Kingdom (16%). - They investigate NMBs of PM interventions instead of ICERs and perform regression analysis in order to explore the heterogeneity in the cost-effectiveness of PM interventions.	-PM interventions in cancer have lower costs and higher NMB than other procedures. -Although PM pharmaceuticals have higher health gains than non-pharmaceuticals, PM pharmaceuticals come at a higher cost than non-pharmaceuticals. -Industry-sponsored studies are more likely to have positive cost-effectiveness outcome (in line with Berm *et al*., 2016).

	-Studies also compare a PM intervention with a non-PM intervention. -Most frequently occurring cases were cancer treatments (60%) and pharmaceutical interventions (72%). Prognostic tests (19%) and tests to identify (non)responders (37%) were least and most common respectively.	

Luis and Seo (2021) mention that two important reasons why progress has been slower than expected with few biomarkers reaching clinical practice are (1) the limitations of genetic prediction due to biological complexity, and (2) the lack of appropriate incentives for pharmaceutical firms. They stress the need for an economic evaluation of biomarker tests with real world longitudinal and/or patient data, while most existing evaluations rather use clinical trial data or simulations based on such data. They start by pointing out the existing literature on the effect of pharmaceutical innovation for cancer on increasing survival or reducing mortality.

The aim of the study is to determine the effect of the utilization of biomarkers for cancer therapies on premature mortality and survival using Norwegian data from 2000 to 2016.[14] Their empirical strategy consists in regressing health outcomes (potential years of life lost before age 75 and 65, and a 3-year survival dummy variable) on the number of cancer drugs and the availability of biomarker tests to treat the specific cancer each patient is diagnosed with. An advantage of premature mortality over survival probability is that the former is not subject to lead-time bias.[15] They document that having at least one biomarker test available decreases premature mortality on average. More surprisingly, they show that the total effect of biomarker testing on survival decreases as the number of cancer drugs available increases. This suggests that biomarker tests improve health by better matching patients to treatments, but that matching is better when fewer drugs are available. They mention a few reasons for the latter, related to the time it takes to test patients for multiple biomarkers, to the bias exhibited by doctors who prefer to first use well known drugs, and more generally to the fact that having access to more biomarkers and drugs increases the complexity of the treatment decisions and makes it more difficult to “match the right patient to the right drug”.

[14] An earlier paper by Oosterhoof *et al.* (2016) reviews 33 studies assessing diagnostic biomarkers for the main non-communicable diseases in middle-income or high-income countries, over the period 2010 to 2015. It focuses on biomarkers for diagnosing, staging, and guiding the selection of therapeutic strategies for noncommunicable diseases. Its goal is methodological, reporting the factors that affect the economic evaluations in practice, rather than reporting the empirical results themselves.

[15] Lead-time bias occurs if improvements in screening tests for some cancer types lead to earlier diagnosis.

They also find that nonguided therapies (those not requiring biomarker testing) are associated with an increased probability of being alive 3 years after diagnosis, while biomarker-guided drugs are associated with a reduction of premature mortality before age 75 and 65. They attribute these differences to variations in the samples for the regression on premature mortality and on survival, together with the plausible assumption that cancer patients at the end of life benefit more from new drugs compared to patients who have just been diagnosed.

Finally, their estimates of the cost *per* life-year gained before ages 75 and 65 in 2016 from biomarker-guided drugs introduced during 2000–2015 are well below the EUR 30,000 per QALY often mentioned in the literature as the threshold value at which an intervention is considered cost-effective (note that the authors compare their estimate of cost *per* LYG to cost thresholds for QALYs!). As should be clear from this summary, the main limitation of their analysis is that the lack of data does not permit a deeper analysis of the mechanisms at play.

KEY TAKEAWAYS

We focus in this section on the empirical assessment of PM tests. PM has generated two opposite types of predictions: (i) that such tests would restrict existing costly treatments to those who would benefit from them, allowing to save costs and reach better (or at most slightly deteriorating) health outcomes, and (ii) that very costly procedures would be found worthwhile for some patients, generating higher health costs for better health.

We summarize systematic reviews which are not performing quantitative meta-analyses, but rather provide a descriptive synthesis of the results reviewed. Empirically, a small fraction (of around one fifth to one quarter) of the studied PM tests result in cost savings. The bulk of CUAs (increasingly the favored type of assessment) of targeted therapies (*i.e.*, joint evaluation of tests and therapies) generate improvements in health but with higher costs, with a large fraction (although far from the totality) for a cost-*per*-QALY that looks effective by today's standards.

More generally, there is no evidence that PM performs better in terms of cost *per* QALY than more traditional approaches, such as pharmaceutical interventions. But there is a lot of heterogeneity in the cost *per* QALY gained, so that some genetic testing procedures may perform better than non-genetic ones.

The studies reviewed raise two red flags. First, most are based on simulations or experts' opinions rather than on solid clinical evidence (due to the lack of the latter). Second, there is a suspicion of publication bias, with an increasing share of industry-sponsored studies, which all conclude to the effectiveness of the test under consideration (by opposition with the other studies).

The next section focuses on the fact that tests are generically imperfect, and on how to take their accuracy degree into account when assessing their economic value.

5. Health Technology Assessments of imperfect tests

Diagnostic tests enable clinicians to allocate the right treatment to the right patient. But, with very few exceptions, they are imperfect in the sense that they make wrong predictions for a subset of the tested population. It is important to take this accuracy problem into account when assessing the economic value of these tests. Moreover, as we will see, the accuracy degree is often endogenous, and economic analysis can help determining the optimal accuracy degree, as well as the corresponding economic value of the optimal test.

We first present the canonical framework used to determine the trade-off between *sensitivity* and *specificity* in the design of a test. We then show how to employ this framework to assess the value of the information brought by the tests. We then go back to the empirical approach, presenting first the methodology to be used to proceed to meta-analyses of tests which differ in their sensitivity and specificity, and then surveying the recent methodological and/or empirical literature adopting this methodology.

5.1. The analytical approach

The canonical example is the one where the population is divided into two groups, one with a disease and the other one without. The diagnostic tests are then used to sort the tested population into these two groups. As we shall see, this requires setting a threshold value of the test result that separates those deemed disease-positive from the others. The threshold value chosen simultaneously determines the fraction of false positive and of false negative individuals. This has to be taken into account when assessing the economic value of the test.

Laking *et al.* (2006) mention a “schism” between two schools in the evaluation of diagnostic tests. One school focuses on the test’s ability to classify individuals into those affected by one disease, and those who are not. The other school rather focuses on the value of information brought by the tests, using an approach that is more familiar to economists. At the time of the writing of this literature review, the second approach still had to make its way into the mainstream practice of health technology assessment. These authors then propose a way to bridge the gap between these two approaches, starting with the principal analytical tool of the first (the soon-to-be-defined receiver-operator curve, or ROC) and integrating diagnosis into conventional cost-effectiveness analysis.

Most diagnosis tests result in a continuous measure. To be of interest, the distribution of this measure must differ between those who have the disease, and those who do not. We illustrate this reasoning with the following Figure 2 taken from Sutton *et al.* (2008). In part a, the yellow (resp., grey) curve represents the distribution function of the test results among the healthy (resp. diseased) portion of the population.[16] A threshold is needed to determine who will be considered diseased-positive. On the figure, all individuals whose test results lie above (resp., below) $\mathbf{D_T}$ are considered disease-positive (resp., -negative).

The test result is then not perfectly predictive of the disease status, and the use of a threshold generates two types of errors: false positives (whose fraction corresponds to the yellow area to the right of $\mathbf{D_T}$) and false negatives (whose fraction corresponds to the grey area to the left of $\mathbf{D_T}$). The table in part b of Figure 2 reports the fractions of false positive, true positive, false negative and true positive obtained from part a of the figure.

The medical literature prefers employing the terms of sensitivity and of specificity rather than false positives or negatives. They are defined as follows (see also part c of Figure 2):

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{Total with disease}}, \qquad \text{Specificity} = \frac{\text{True negatives}}{\text{Total without disease}}.$$

[16] The exact same reasoning applies to individuals who are receptive to a drug *versus* those who are not, or those who will develop side-effects from the drug and those who will not, or those who require a low dose of the drug *vs* those who require a high dose. The groups of diseased *vs* non-diseased is determined by a so-called "gold standard" test.

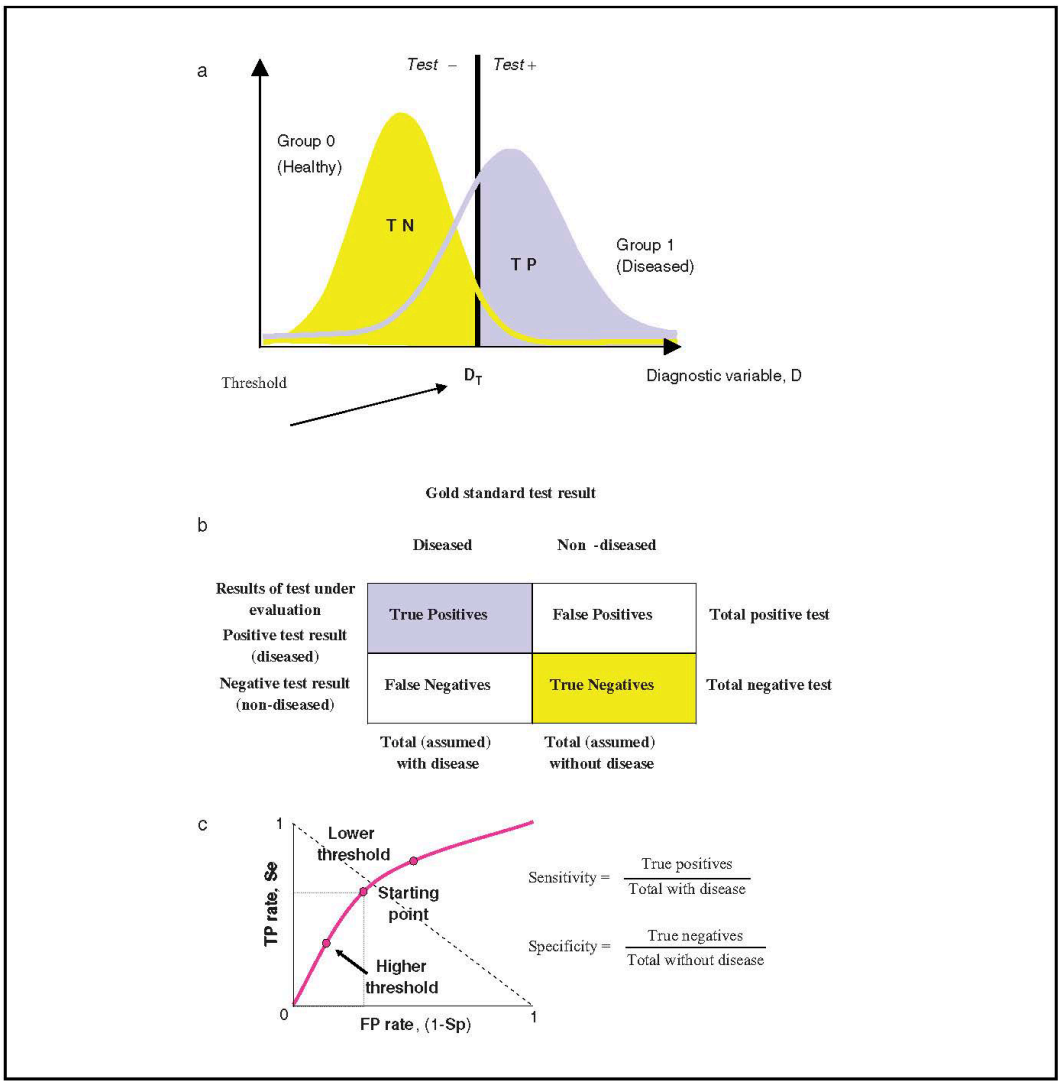


Figure 2 Evaluation of a diagnostic test using data from a single study. (a) Distributions of test results for diseased and nondiseased populations with categorizations defined by threshold (D_T). (b) A 2 × 2 table indicating test categorization of individuals from (a). (c) Receiver operating characteristic curve derived from changing test threshold.

The crucial points are that (i) specificity and sensitivity are not set exogenously, but vary with the threshold value D_T chosen, and that (ii) there is a trade-off at play, with one measure increasing at the expense of the other as D_T is moved.

By changing the value of DT, we change the corresponding sensitivity and specificity levels of the test. Part c of Figure 2 depicts the set of those levels that can be attained, which is called in the literature the receiver-operator characteristic (ROC) curve. It is often expressed in the (false positive rate, true positive rate) space, or equivalently, the (specificity, sensitivity) space. The top right point on this curve corresponds to the minimal value of D_T, where all tested agents are deemed disease-positive, so that both the true positive and false positive rates are equal to one. Increasing the threshold D_T then reduces the fraction of false positives, but at the expanse of the fraction of true positives. When the threshold D_T is set at its maximal

level, all tested agents are deemed disease-negative, resulting in zero true positive and true negative rates. To each (imperfect) test corresponds a ROC curve. It is at this point that the two approaches mentioned by Laking *et al.* (2006) in the schism diverge. The first approach tries and summarizes the accuracy level reached by each test with measures such as the "area under the curve" (AUC), with the larger area being preferred.

This approach is not the one favored by economists, for (at least) two reasons. First, and less importantly, the ROC curves of two tests may cross each other, in which case it is far from obvious that the AUC criterion is the relevant one. A test with larger true positive rate and smaller false positive rate is obviously more attractive, so that the health decider's welfare increases as we move to the northwest on Figure 2 c. So, if the ROC curve of one test is everywhere above the ROC curve of the other test, it is certainly more desirable (provided that both tests have the same cost) and has a larger AUC. But a larger AUC does not imply that the corresponding test is better for society when the ROC curves cross each other. Second, and more importantly, this approach does not consider the health and/or economic consequences of mis-allocating patients. For instance, there is *a priori* no reason to maximize accuracy, or the correct number of correctly diagnosed patients,[17] because there is no reason to impose that the medical/economic consequences of the two types of misdiagnoses (false positive and false negative) are equivalent.

The economically sound way to proceed (when comparing two tests) is to specify the objective function that we wish to maximize, to determine what is the optimal threshold $\boldsymbol{D_T}$ to be used for each test (often called the Optimal Operating Point, or OOP), and to favor the test for which the value of the objective function reached at the optimal threshold is the highest. The approach advocated is closely linked to the CUA detailed in section 2, since it consists in maximizing the value (net monetary benefit NMB) of the information brought by the test (by choosing the threshold $\boldsymbol{D_T}$ optimally), and to choose the test bringing the highest NMB at its optimal threshold. We follow the setting proposed in in the seminal paper by Laking *et al.* (2006).

Assume that there are two groups in the population (called *x* and *y*) and two potential treatments (*A* and *B*) for the disease suffered by this population. Assume that treatment *A* is better suited to group *x* (in the sense that $NMB_A > NMB_B$ for group *x*), while *B* is better suited to group y (since $NMB_A < NMB_B$ for group *y*). The test is then used to diagnose patients as belonging to group *x* or *y*, to prescribe them the treatment most suited to them. To find the optimal test threshold $\boldsymbol{D_T}$, we compute, for each point on the ROC curve, its corresponding position in the cost-effectiveness space, with expected QALYs gained (compared to no treatment) on the horizontal axis, and cost on the vertical axis. This requires knowing the costs of the two treatments *A* and *B*, the prevalence of subgroup *x* in the population, and the health

[17] This corresponds to the point at which the ROC curve crosses the line sensitivity=specificity (the dotted line of Figure 2c).

effects (measured in QALYs) of the two treatments in the two groups (see for instance Table 3 in Laking *et al.* [2006]).[18]

Setting the test threshold at its minimal level means that all agents are assumed to belong to group *x* (for instance) and must thus be treated with *A*, resulting in an expected QALY gain and corresponding (treatment) costs. This point is labeled as *R* on Figure 5 below (imported from Laking *et al.*, 2006). Likewise, if we set the test threshold to its maximal level, all patients are assumed to belong to group *y* and are then treated with *B*, resulting in another combination of expected QALY gains and cost to obtain point *S*. We can compute the ICER (see section 2) of, say, treatment *B* compared to *A* to determine which of the two treatments should be the default one (it is the treatment with the highest NMB among the two). The straight lines on Figure 5 correspond to "net benefit isoquants", linking all the points in the cost-effectiveness space with the same NMB (*i.e.*, where the ICER corresponds to the cost-effectiveness ratio/WTP for QALY). NMB increases to the southeast (higher QALY/lower cost) so that we see on the Figure that treatment *A* is the default treatment in the absence of a test (since point *R* corresponds to higher NMB/lower isoquant than point *S*).

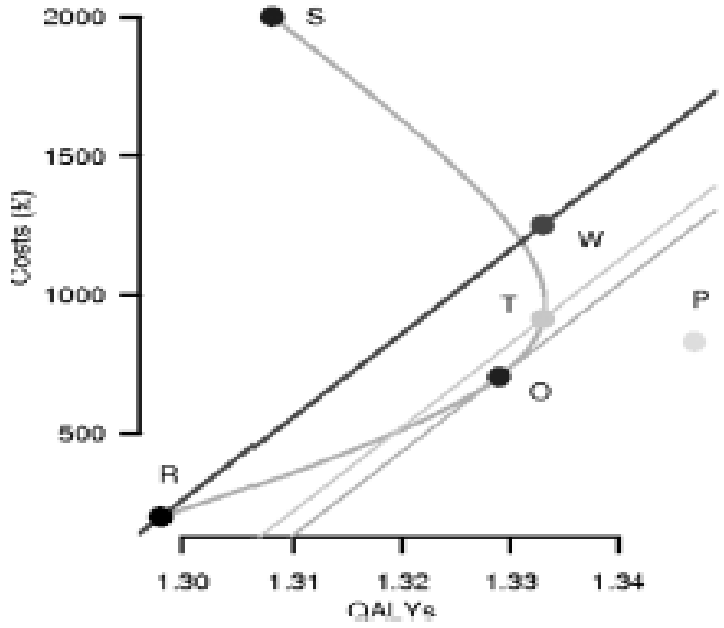


Figure 5 from Laking *et al.* (2006). ROTS curve (without costs of testing). The lines are net benefit isoquants.

[18] We abstract for the moment from tests costs. We introduce them later in the analysis.

Points *R* and *S* of course do not make any use of the test information (since all patients receive the same treatment, whether *A* at point *R* or *B* at point S).[19] Increasing the test threshold $\boldsymbol{D_T}$ from its minimum level, more and more patients are identified as belonging to group *y* and thus treated with *B*. As we saw above, this change simultaneously increases the fraction of true negative (*i.e.*, here truly *y*) and of false negative, but by different magnitudes. For each threshold level and corresponding sensitivity and specificity levels, we then compute the expected QALYs gained (when treating all agents revealed -truly or falsely- to be *x* with *A* and the others with *B*) and the corresponding treatment cost. Two examples (corresponding to different values of $\boldsymbol{D_T}$) on Figure 5 are points O and T, and varying $\boldsymbol{D_T}$ in a continuous way generates the so-called ROTS curve.[20]

Finding the optimal $\boldsymbol{D_T}$ threshold (the value maximizing the NMB of the test) then consists in finding the point on the ROTS curve where the slope of the curve equals the willingness to pay for QALY. This corresponds to point O on Figure 5, since it sits on the lowest net benefit isoquant attainable with the test (*i.e.*, on the ROTS curve). Observe also that the NMB of the test at its optimal threshold (corresponding to point O) is given by the vertical distance between O and the net benefit isoquant through the default treatment point *R*.

Comparing two (free) tests, the technically superior one is the one allowing to attain the lowest net benefit isoquant. Finding it requires drawing the ROTS curve of each test, the optimal point O on each ROTS curve, and finding the one corresponding to the lowest net benefit isoquant (see Figure 6 below, where the test corresponding to the dark ROTS curve is technically superior to the other one).

Introducing heterogenous test costs can easily be done by shifting the net benefit isoquant passing through the optimal point O upward by the amount of the cost per patient, and then selecting the test with the lowest such isoquant. Figure 6 shows an example where the technically superior test (in black) is not the one maximizing NMB once tests costs are included, because its cost is much higher than the cost of the other test. The adverse effects of the tests can be introduced in a similar way (once the adverse effects have been quantified, both in probability of occurring in the different groups and in their QALY losses consequences).

[19] Just as in the ROC curve, point *R* corresponds to false positive and true positive rates of 0, and point *S* to false positive and true positive rates of 1.

[20] ROTS is not an acronym, but is the term used throughout the literature for this curve following Figure 5 in Laking *et al.* (2006)!

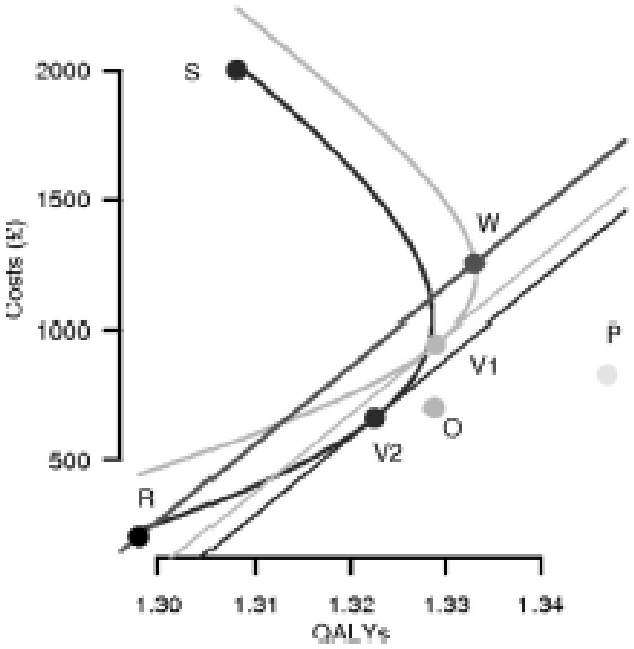


Figure 6 from Laking *et al.* (2006). ROTS curve (with costs of testing). The lines are net benefit isoquants. The black curve touches the most favorable isoquant.

Often, the scope for diagnostic testing is analyzed considering the benefits and the costs of both testing and treatment. The tradeoffs influence the decision whether to withhold the therapy, or to perform the test and then, based on the results, to administer the therapy. Thus, Pauker and Kaussirer (1980) derive two thresholds, “testing” threshold and “test-treatment” threshold, which should guide the medical decision-making. The test threshold defines the probability of the disease above which the test should be administered, while the test-treatment threshold defines the disease probability above which the treatment without prior testing is preferable. Between the two thresholds the test should be performed, and depending on the test outcome, the treatment should follow. The values of the thresholds are based on the accuracy and potential risk of the test and the risk and the benefits of a particular treatment. The authors highlight the tradeoff of a treatment between the utility gain for diseased patients and utility loss for healthy patients.

In the last decades, genetic tests[21] have become more common. The accuracy (sensitivity and specificity) for genetic tests is often very close to 100%. On top of that, the costs have been considerably decreased due to the revolutionary advances in DNA-sequencing technologies. Given these considerations, Felder (2022) apply the insights of the threshold analysis for the genetic testing. In this context, the thresholds do not refer to the probability of the disease but rather to the probability of genetic mutation. Consequently, preventive treatment may become more relevant relative to curative measures. A positive diagnostic test outcome reveals a mutation

[21] A genetic test is a test for the presence or absence of a genetic mutation.

associated with the increased risk for a disease. In such a case, if the penetrance rate[22] is sufficiently high, a patient may choose preventive treatment.[23]

The analysis above can be modified to introduced uncertainty, both in the technical characteristics of the test (the ROC curve) and its economic consequences (the ROTS curve), as we show in the next section, before turning to the empirical analyses.

A final methodological word is in order, before turning to the empirical assessment of imperfect tests. Even though the cost-benefit and cost-utility analysis presented in section 2 are important to determine the optimal use of diagnostic tests, it is worth recognizing that in medical practice, decisions are often based on limited information, in particular to establish patients' diagnostics, but also to anticipate what is supposed to be the best treatment for patients according to their corresponding diagnostics. This uncertainty may come from the fact that data are either incomplete or not representative of the patient population so that the prevalence of a certain disease or the probability of success of a specific treatment remain uncertain.

To tackle this issue, decision theory highlights the difference between risk and uncertainty that yields to the notion of ambiguity.[24] The concept of uncertainty differs from the one of risk, which refers to an objectively known probability distribution. By contrast, uncertainty is characterized by both an unknown outcome and an unknown probability distribution. Thus, the preferences of patients or their physicians toward ambiguity concerning the correct probability of disease and treatment success are crucial to understand their decisions. The threshold values used to determine when treatments and diagnostic tests have to be reimbursed are also influenced by these attitudes.

[22] The probability of developing the disease given the mutation.

[23] For example, for the prevention of breast or ovarian cancer, women positively tested for the BRCA1 and BRCA2 genes might undergo intensive surveillance, bilateral salpingo-oophorectomy, or mastectomy, whereas a curative chemotherapy would not be indicated if the cancer has not penetrated. The results highlight that a low penetrance rate narrows the scope for genetic testing because the carrier probability threshold is high when penetrance of the disease is low. A low penetrance rate comes with such a low expected monetary value that it might fall short of the cost of preventive treatment. These factors may lead to too high carrier probability threshold for justifying the use of genetic testing.

[24] See Klibanoff *et al.* (2005)

KEY TAKEAWAYS

Economists assess the economic value generated by a test, which depends on its accuracy. The latter is determined endogenously by trading off sensitivity and specificity, as described for instance by the Receiver-Operator Curve (or ROC). One should then determine the optimal point on this ROC (called the Optimal Operating Point, or OOP) as a function of the objective one wishes to maximize, such as the net marginal benefit brought by the test, as detailed by the seminal paper of Laking *et al.* (2006). The analysis requires drawing, for each point on the ROC, its corresponding position in the cost-effectiveness space introduced in section2., and then optimizing taking into account both the test cost, and the willingness to pay for each additional QALY.

We now review the literature assessing empirically imperfect diagnostic tests.

5.2. The empirical approach

Sutton *et al*. (2008) explain how to proceed to meta-analyses of tests while taking into account of the thresholds used in different studies. The main methodological problem here is that different studies report different sensitivities and specificities, and it does not make sense to assume that they are independent from each other, except in the very unlikely case where all studies have used the same threshold D_T. Unfortunately, Sanghera *et al.* (2013) observe that "most economic evaluations of diagnostic tests consider sensitivity and specificity to be independent" (p. 54).

To illustrate this problem, Sutton *et al.* (2008) use as a case study some 198 studies of a (so called d-dimer) test for deep vein thrombosis. The first step is to construct a summary ROC (or SROC) curve from the (sensibility, sensitivity) pairs reported in all studies. They report these pairs on their Figure 3 below. A meta-analysis assuming that sensitivity and specificity are independent from each other generates Figure 3a where mean estimates plus 95% credible intervals (denoted by "CrIs" in Figure 3 below, Bayesian equivalent of confidence intervals) are depicted by the sets of 3 horizontal and vertical lines. This is obviously not the way to go.

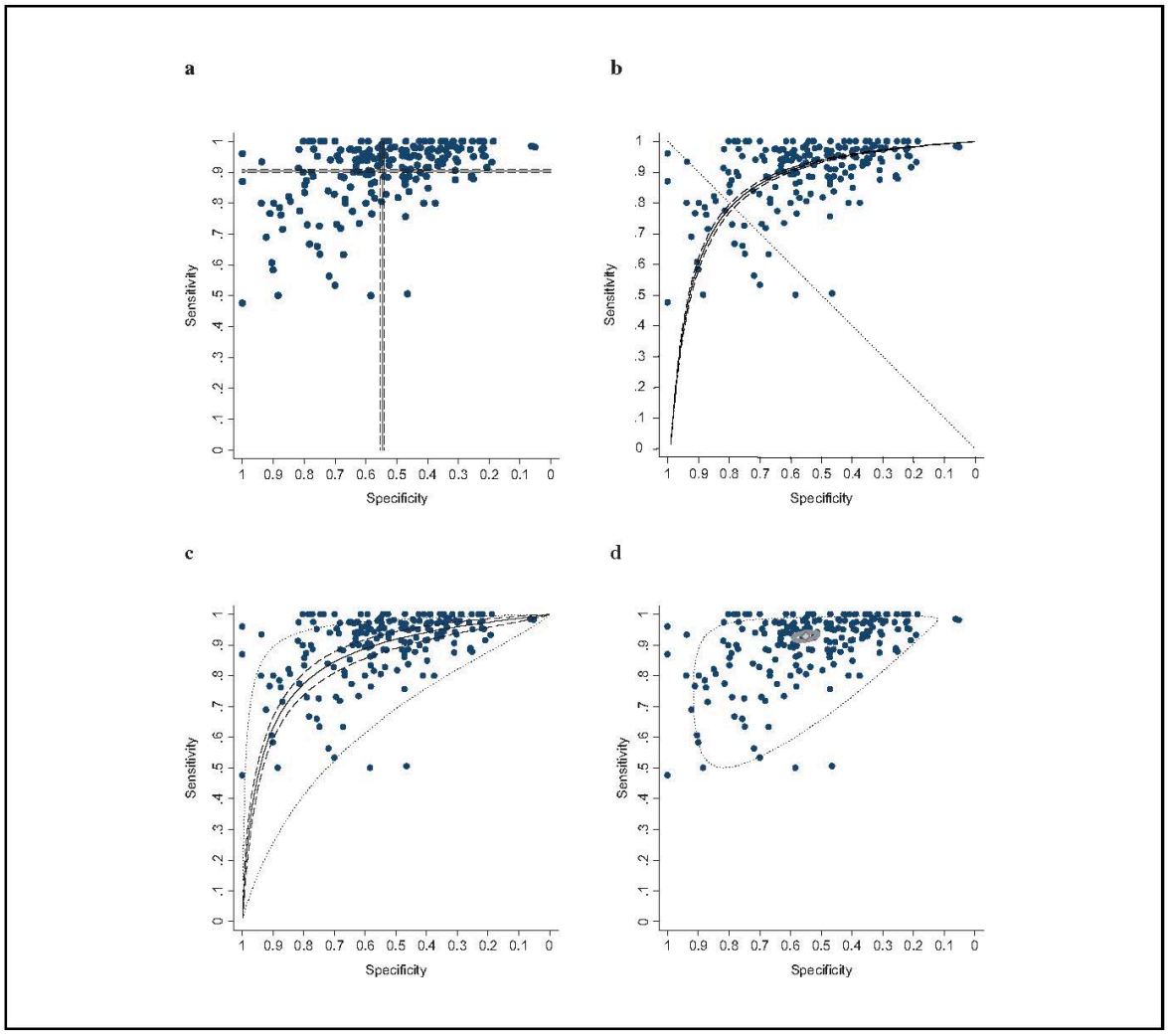


Figure 3 Graphical displays of d-dimer for deep vein thrombosis meta-analyses results. (a) Fixed effect, independent sensitivity and specificity. Solid lines = pooled estimate; dashed lines = 95% credible intervals (CrIs). (b) Fixed effect, constant odds ratio. Solid line = pooled summary receiver operating characteristic (SROC) estimate; dashed lines = 95% CrIs; dotted line = line of SROC symmetry. (c) hierarchical SROC (random effect, no symmetry). Solid line = pooled hierarchical SROC estimate; dashed lines = 95% CrIs; dotted lines = 95% prediction intervals. (d) Bivariate normal (random effect). Diamond = mean sensitivity and specificity; thick dashed line = 95% credible region; dotted line = 95% prediction region.

Some studies go one step further and assume that specificity and sensitivity can be combined into a single measure defined as the diagnostic odds ratio, defined as

$$\text{diagnostic odds ratio} = \frac{\text{sensitivity}}{1-\text{sensitivity}} \Big/ \frac{1-\text{specificity}}{\text{specificity}},$$

or the ratio of the odds of a positive result in a patient with disease compared to a patient without disease. Assuming a constant diagnostic odds ratio across studies implies that the summary ROC curve (SROC) is symmetrical around the line with sensitivity=specificity. This SROC is depicted in Figure 3b, with credible intervals quite small, thanks to the large number of studies.

Alternatively, one can assume away the constant diagnostic odds ratio,[25] and assume rather that there is heterogeneity across studies beyond those attributable to varying thresholds and construct a hierarchical summary ROC curve (HSROC) by means of a regression using study-level covariates. The result is reported on Figure 3c, with much larger credible intervals because of the incorporation of between-study heterogeneity.[26] This method can also be used in the case where accuracy data relating to more than one test threshold are available from each study.

Once (H/S) ROC curves have been constructed, we can move to their economic evaluation. Sutton *et al.* (2008) proceeds slightly differently from Laking *et al.* (2006), in that they introduce explicitly a medical decision tree model (see their Figure 1), describing in detail the medical treatment decisions and expected health outcomes (and corresponding QALYs) as a function of test results and accuracy. As the test threshold is varied, the proportion of patients entering each branch of the decision tree also varies, and so does the net monetary benefit of the test, defined, as in section 2, with NMB= λ*effect-cost, where the effect is measured in incremental QALYs, and where λ measures the willingness to pay for one additional QALY. This NMB can be compared with a no-test strategy, and one can even perform some comparative statics with respect to λ.

In theory, this approach allows to answer simultaneously three questions: 1) is any test worth doing?, 2) what is a test's optimal threshold? and 3) if more than one test is available, which one is the best? Empirically, answering all these three questions may be impossible due to a lack of reported data. For instance, if studies differ in the threshold $\boldsymbol{D_T}$ they use, one can identify the optimal operating point (OOP, see section 5.1 above) on each test's SROC curve as a function of λ, but it is impossible to infer from this optimal point the corresponding optimal threshold, because none of the meta-analysis methods incorporate threshold value data from each of the primary studies.

Sutton *et al.* (2008) then provide an application to the meta-analysis of the cost-effectiveness of two tests used for deep vein thrombosis (DVT), namely d-dimer and ultrasound. They also compare those two tests with both no treatment and treatment of everyone (without prior testing). They estimate cost-effectiveness using 3 models corresponding to Figure 3a (based on mean fixed sensitivity and specificity, ignoring threshold effects and heterogeneity, 3d (using the prediction region around mean sensitivity and specificity based on the bivariate meta-analysis model) and 3c (as a function of the mean HSROC curves, using the threshold value that maximizes net benefit for each value of willingness to pay λ). Note that that the discharge without test and treat without test options are treated as points corresponding to tests

[25] Asymmetry of the ROC curve occurs if the distribution of test results in the diseased and non-diseased populations have different variances.

[26] Alternatively, one can model (logit) sensitivity and specificity as bivariate normally distributed, as reported on Figure 3d, with the dashed line being the credible region around the mean sensitivity and specificity.

(costing no money) operating at the 2 anchored extremes of an ROC curve with 0 sensitivity and specificity of 1 and sensitivity of 1 and specificity of 0, respectively.

They compute, for each model, the probability that the test is cost-effective as a function of λ.[27] Focusing on their third model, they obtain that "as willingness to pay increases, the optimal test performance [for both tests considered] point moves left along the SROC, indicating a lower threshold should be used, which makes the tests less specific but more sensitive.[28] This implies that the benefits of identifying and treating DVT increasingly outweigh the risks of treating those without DVT as willingness to pay increases" (p. 662). Comparing the two tests, they conclude that ultrasound is almost certainly the strategy with the greatest chance of being optimal for all values of willingness to pay greater than £5000 per QALY. Also, decisions regarding whether just to discharge without any testing or not depends on a decision maker's willingness to pay.

To conclude, Sutton *et al.* (2008) mention that "although we are advocates of systematic review and meta-analysis methods generally, in the diagnostic test decision-modeling context, because of the limitations of most studies and the data their reports contain, we question whether there are better ways of informing decision models than initially conducting exhaustive (and very time consuming) meta-analyses of the published literature" (p665). They stress that "even if only 1 study were available with IPD [individual patient data] that compared all tests of interest with a reference standard, this could be more reliable and could contain more information than single-point summaries from numerous studies, which evaluate only a single test".

Sanghera *et al.* (2013) provide a step-by-step guide of the approach proposed by Sutton *et al.* (2008) and use a case study of fetal anemia in which data from a screening test are used in combination with a confirmatory test.[29] They contrast results obtained when the same test threshold is used in several studies, and when data from several studies that use different test thresholds are employed. They stress that the first scenario can under-estimate the cost effectiveness of the test studied if the test threshold used by the studies is not the optimal one. The second scenario is superior since it allows estimating the optimal test threshold. In their case study, both scenarios conclude to the cost-effectiveness of using the screening test before the confirmatory test.

[27] The so-called cost effectiveness acceptability curve, or CEAC, proposed by Fenwick *et al.* (2001), depicts the probability that a test is optimal as a function of λ.

[28] Since ultrasound has an implicit threshold, obtaining performance on the SROC at the points indicated by the model may not be possible in practice, limiting the usefulness of this analysis in this context.

[29] Kohn *et al.* (2001) stress that the optimal threshold point differs with the pre-test probability of disease (the larger the latter, the smaller the former).

Jones *et al.* (2019) propose a statistical method to be used in meta-analyses and making use of the facts that many studies report sensitivity and specificity obtained at different test thresholds. Their model assumes that some prespecified or Box-Cox transformation of test results in the diseased and disease-free populations has a logistic distribution. The Box-Cox transformation parameter can then be estimated from the data, allowing for a flexible range of underlying distributions. They show how their model works by applying it to two case study meta-analyses, studying the accuracy of tests for heart failure and preeclampsia.

Rautenberg *et al.* (2020) provide a pictorial primer on how to make the link between accuracy measures (such as sensitivity and specificity) and a decision tree, including when several tests are undertaken sequentially. They point out the two main mistakes observed in the empirical literature: not including diagnostic test accuracy in the structure of decision trees and treating sequential diagnostics as independent. "For example, a review of thirty economic evaluations for diagnostics in oncology showed that only twelve evaluations modelled diagnostic test accuracy (DTA); the remaining eighteen models only considered the cost of diagnostics and not DTA. (…) It has been shown that models that (correctly) include DTA have higher incremental cost effectiveness ratios and are therefore less likely to be cost-effective when compared to models that do not include DTA" (p.1).

Doble *et al*. (2014) systematically assess the published model-based economic evaluations, in which a targeted oncology therapy has been evaluated alongside a companion diagnostic. They contrast the results obtained from economic evaluations including model parameters for the sensitivity and specificity of the companion diagnostic to economic evaluations of targeted therapies that limited model parameters for the companion diagnostic to only its cost. They show that the latter provide significantly different results.

Drakopoulos *et al.* (2021) show that, when there are constraints on the availability of a test, the optimal combination of sensitivity and specificity may not lie on the frontier of the ROC curve, if agents choose whether to get tested or not. The reason runs as follows: when a test's accuracy is low, this discourages some agents from taking it. This is actually a good thing if the social planner would not prioritize the testing of these agents anyway. Then if the test accuracy increases, these agents may start testing, preventing higher-priority agents (in the eyes of the social planner) from having access to the test. This study was of course inspired by the lack of testing at the beginning of COVID 2019.

KEY TAKEAWAYS

Meta-analyses should take into account that the tests studied may differ in the pairs (specificity, sensitivity) that they report, and that the latter are not independent from each other. A partial move in the right direction consists in assuming a constant diagnostic odd ratio, implying a ROC symmetrical around the sensitivity=specificity line. A better but more demanding approach introduces a medical decision tree model specifying the health consequences of the test accuracy. Exploiting a single study with individual patient data then results in more reliable estimates than performing a meta-analysis of single-point summaries. Several studies combining a methodological with an empirical approach show how estimates of cost-effectiveness can be biased when once does not take these aspects into account.

6. Conclusion

This review provides a global overview of the existing methods of assessing the economic value of diagnostic tests. Health technology assessment is indeed central for making informed social decisions on the optimal allocation of limited health resources. When assessing the value of innovative tests, the approach advocated by economists may differ from the one(s) used by other health researchers in several dimensions. Economists assess the value of the information brought about by the test. The design of the test, its characteristics, do matter. Regarding its accuracy, there is in general a trade-off between false positives and false negatives (or, to use medical terminology, sensitivity and specificity). Economists first optimize over the best trade-off before measuring the value of information at this point. Note that the optimal trade-off depends on the consequences of each type of error as well as on the health benefits obtained when the test recommends the correct treatment. These in turn depend on society's willingness-to-pay for health improvements (such as the maximum cost per QALY, for instance), on attitudes towards uncertainty (or towards ambiguity when it is difficult to give precise probabilities to various events) and on treatment and test costs.

Rather than stating anew the key takeaways summarized at the end of each section, we would like to conclude by singling out one result, and by mentioning one important criticism to the approach we have surveyed here.

The development of innovative tests within the realm of PM has generated two types of (opposite) predictions. On the one hand, in as far as those tests allow not to treat some agents with costly and ineffective drugs, they may result in better (or at least not too detrimental) health outcomes at a lower cost. On the other hand, those tests are often associated with very costly treatments which would not be approved in the

absence of companion tests. In this latter case, one would expect to observe health improvements associated to high costs.

The empirical literature finds evidence of both effects. In the studies surveyed in section 4, we find that roughly one fifth to one quarter of the results are consistent with the first hypothesis (lower costs for maintained or even better health), with a larger fraction of studies confirming the second hypothesis, but with a cost *per* measure of health benefit (usually, QALYs) acceptable by society's standard. Also, there is little evidence that PM tests perform better (per QALY) than non-genetic tests. But there is a lot of heterogeneity in the cost *per* QALY gained, so that some genetic testing procedures may perform better than non-genetic ones.

Since 1970s QALY has been recognized as the most rigorous standardized metrics for valuing health economic outcomes across different health care interventions for very different health conditions. However, a growing literature identifies several limitations from methodological, ethical or context-specific grounds (*e.g.*, Pettitt *et al*. [2016]; Rand and Kesselheim [2021] or Schneider [2022]). There is a long-standing criticism of the QALY based on ethical considerations. QALY is argued to discriminate against elderly people and those with disabilities or chronic illnesses: extending the lives of individuals with underlying health conditions generates fewer QALYs than extending the lives of 'healthier' individuals. Such ethical criticism can be considered beyond the scope of economists, since it requires political decisions on social values. Economists can at most shed light on the consequences and trade-offs of favoring various metrics.

From a methodological point of view, the main criticisms are related to whether the theoretical assumptions required for the QALY to be a valid metric are satisfied in practice, as related for instance to the measurement techniques and the source of sample used to value health states. For example, some studies argue that different populations may evaluate conditions differently: utility values for a physician and for general population are likely to differ. QALYs have also been criticized for not considering non-health benefits and in particular, societal benefits (*e.g.*, faster return to work or better school performance).[30] So, even though QALY has been the main measure of health benefits in the literature surveyed here, it is far from the optimal (or even the only) measure that should be considered.

[30] For more details, please refer to Pettitt *et al.* (2016).

Annex: Definition(s) of personalized medicine

The term personalized medicine is used regularly but interpreted in different ways. Various definitions of the term have been proposed, with no clear consensus on which questions (*e.g.*, what is the diagnosis?), methods used to answer them (*e.g.*, a test) and actions (*e.g.*, to give or not a specific drug) fall within its domain. As explained in the text, the broader definition encompasses basically all of medicine, while narrower definitions refer to the use of a diagnostic test to predict drug response of a patient, and are often associated with the fields like genetics, genomics, *etc*. An often-cited example is the HER2/neu test to predict the effectiveness of trastuzumab in breast cancer. Many of those who adopt this definition then associate personalized medicine with the genetics, genomics, and other "-omics" fields.

Indeed, the remarkable developments in many research areas such as genomics (*e.g.*, Human Genome Project), have allowed to identify different diseases subtypes based on genetics. Thus, the knowledge of genetics can help determine whether patients with certain disease subtypes are more likely than others to be responsive to a particular drug. In this sense, personalized medicine has changed the paradigms in oncology, because it is based on understanding molecular carcinogenesis, pharmacogenomics, and individual genetic differences that determine the response to chemotherapy. The transition to molecular biomarker-driven therapeutic decision process is still evolving, however new classes of drugs and companion diagnostics are already emerging. These advances are changing the landscape for the management of many advanced-stage cancers (*e.g.*, Kalia, 2015).

There are geographical discrepancies in how the term is defined, for instance between Europe and the US. In Europe, the aim of personalized medicine is generally perceived to be the "right treatment for the right person at the right time." However, in the US the term personalized medicine "does not literally mean the creation of drug or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differs in their susceptibility to a particular disease or their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not" (Report of the President's Council of Advisors on Science and Technology, 2008).

European Commission

https://ec.europa.eu/info/research-and-innovation/research-area/health-research-and-innovation/personalised-medicine_en

The Horizon 2020 Advisory Group defines personalized medicine as "a medical model using characterization of individuals" phenotypes and genotypes (*e.g.* molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention." (see https://research-and-innovation.ec.europa.eu/research-area/health/personalised-medicine_en)

US National Library of Medicine

https://medlineplus.gov/genetics/understanding/precisionmedicine/precisionvspersonalized/

There is a lot of overlap between the terms “precision medicine” and “personalized medicine.” According to the National Research Council, “personalized medicine” is an older term with a meaning similar to “precision medicine.” However, there was concern that the word “personalized” could be misinterpreted to imply that treatments and preventions are being developed uniquely for each individual (corresponding to the “individualized medicine” concept presented in the text following Trusheim *et al.* (2007). By contrast, in precision medicine the focus is on identifying which approaches will be effective for which patients based on genetic, environmental, and lifestyle factors. The Council therefore preferred the term "precision medicine" to “personalized medicine.” However, some the two terms are often used interchangeably.

To illustrate how much existing definitions of personalized medicine in the literature may vary, Redekop and Mladsi (2013) propose a bottom-up approach to define personalized medicine, starting from the frequently asked questions that can be answered using medical tests. They identify the following chronological questions (starting well before the onset of a disease), and the tests used to answer them:

- What is the risk of developing a specific disease for a given patient? Use of *susceptibility tests*.
- How to detect a disease in its early stage, before symptoms have occurred? Use of *disease screening tests*.
- How to detect a disease after the first symptoms have occurred? Use of *diagnostic tests*.
- What treatment to prescribe (including potentially no treatment at all)? *Prognostic tests*, or *companion diagnostic tests (*sometimes called *predictive biomarkers)* if linked to specific treatment. The combination of test-treatment is referred to as *targeted therapy*.
- How is the treatment working?
- What are the prospects of disease recurrence, and what can be done about it?

They stress that prognosis tests may be helpful either in determining the likely effectiveness of a treatment before it is started, but also in assessing the likeliness of side effects or the dose to be used.

They propose three different definitions of personalized medicine, on the basis of which tests are included. All definitions of personalized medicine include the companion diagnostic tests, so that their first proposed definition is based on them:

“the use of combined knowledge (genetics or otherwise) about a person to predict treatment response and thereby improve that person's health.”

A larger definition also included the prognostic tests, resulting in the enlarged definition: “the use of combined knowledge (genetics or otherwise) about a person to predict disease prognosis or treatment response and thereby improve that person's health.” Finally, others (such as the US National Cancer Institute, or the US President’s Council) also include the susceptibility tests when describing personalized medicine, resulting in the even broader following definition: “the use of combined knowledge (genetics or otherwise) about a person to predict disease susceptibility, disease prognosis, or treatment response and thereby improve that person's health.”

Table 2 from Redekop and Mladsi (2013), reproduced below, provides concrete examples of what could then be viewed as personalized medicine.

As mentioned in the main test, for the sake of this survey we will exclude the susceptibility tests.

Table 2 – Specific examples of personalized medicine or personalized health care.

Type of test	Disease	Test	Function	Implications for treatment
Disease susceptibility test	Breast cancer	BRCA1	Individuals with a deleterious BRCA1 or BRCA2 mutation are at increased risk of breast and ovarian cancer.	Surveillance, risk modification, chemoprevention, prophylactic surgery
Prognostic test	Breast cancer	Mammaprint	Test predicts the risks of cancer recurrence within 5–10 y after the initial event.	Adjuvant chemotherapy (yes or no)
Companion diagnostic — effectiveness-oriented	Breast cancer	HER2	Trastuzumab (Herceptin) is beneficial only for tumors with an HER2 overexpression.	Trastuzumab (yes or no)
Companion diagnostic — safety-oriented	Epilepsy and other indications for carbamazepine	HLA-B*1502	Patients with HLA-B*1502 are more likely to have dangerous skin reactions following carbamazepine therapy than other patients.	Carbamazepine (yes or no)
Companion diagnostic	Atrial fibrillation and other indications for warfarin and other coumarin derivatives	CYP2C9, VKORC1	Optimal maintenance dose for coumarin therapy is partly dependent on CYP2C9 and VKORC1 genotypes.	Warfarin dosage
Treatment response monitoring test	Hepatitis C	HCV RNA test	The test measures viral RNA levels after starting treatment with pegylated interferon alfa and ribavirin.	Length of treatment

HCV, hepatitis C virus.

References

Abbott JH, Wilson R, Pryymachenko Y, Sharma S, Pathak A, Chua JYY, 2022. “Economic Evaluation: A Reader's Guide to Studies of Cost-Effectiveness, *Archives of Physiotherapy* 12(1):28.

Antoñanzas, F., Rodríguez-Ibeas, R., and C. Juárez-Castelló, 2019, “Pre-approval incentives to promote adoption of personalized medicine: a theoretical approach”, *Health Economics Review*, 9:28, 2-10.

Baker, R., Chilton, S., Donaldson, C., Jones-Lee, M., Lancsar, E., Mason, H., Metcalf, H., Pennington, M. and Wildman, J., 2011. Searchers vs surveyors in estimating the monetary value of a QALY: resolving a nasty dilemma for NICE. *Health Economics, Policy and Law*, *6*(4), pp.435-447.

Bardey D. and P. De Donder, “Genetic Testing with Primary Prevention and Moral Hazard”, *Journal of Health Economics*, 2013, 32 (5), 768-779 & 1007-1012.

Berm EJJ, Looff Mde, Wilffert B, et al. Economic evaluations of pharmacogenetic and pharmacogenomic screening tests: a systematic review. Second update of the literature. *PLoS One*. 2016;11(1).

Bobinac A, Van Exel NJ, Rutten FF and WB. Brouwer, 2010. “Willingness to Pay for a Quality-Adjusted Life-Year: The Individual Perspective”, *Value Health* 13(8):1046-55.

Brent, R.J., 2023. “Cost-Benefit Analysis versus Cost6effectiveness Analysis from a Societal Perspective in Healthcare”, *International Journal of Environmental Research and Public Health* 20, 4637.

Brouwer, W.B. and Koopmanschap, M.A., 2000. On the economic foundations of CEA. Ladies and gentlemen, take your positions!. *Journal of health economics*, *19*(4), pp.439-459.

Chen G. and V. Peirce, 2020. “Evaluation of the National Institute for Health and Care Excellence Diagnostics Assessment Program Decisions: Incremental Cost-Effectiveness Ratio Thresholds and Decision-Modifying Factors”, *Value in Health* 23(10), 1300-1306.

D’Andrea E, Marzuillo C, Pelone F, De Vito C and P. Villari. 2015, Genetic testing and economic evaluations: a systematic review of the literature. *Epidemiol Prev*.; 39(4 suppl 1):45–50.

Di Ruffano F. L, Harris IM, Zhelev Z, Davenport C, Mallett S, Peters J, Takwoingi Y, Deeks J and C. Hyde., 2023. “Health technology assessment of diagnostic tests: A state of the art review of methods guidance from international organizations”, *Int Journal Technol Assess Health Care*; 39(1).

Doble, B., Tan, M., Harris, A. and Lorgelly, P., 2014. Modeling companion diagnostics in economic evaluations of targeted oncology therapies: systematic review and methodological checklist. *Expert review of molecular diagnostics*, *15*(2), pp.235-254.

Drakopoulos, K. and Randhawa, R.S., 2021. Why perfect tests may not be worth waiting for: Information as a commodity. *Management Science*, *67*(11), pp.6678-6693.

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L. and G. W. Torrance, 2015. Methods for the Economic Evaluation of Health Care Programmes, Oxford University Press, Fourth Edition.

Eeckhoudt, L., 2002. *Risk and medical decision making* (Vol. 14). Springer Science & Business Media.

European Network for Health Technology Assessment, 2015. *Methods for Health Economics Evaluations. Guideline Based on Current practices in Europe*. Available at https://www.eunethta.eu/wp-content/uploads/2018/01/Therapeutic-medical-devices_Guideline_Final-Nov-2015.pdf.

Felder, S., 2022. "Decision Thresholds with Genetic Testing", *The European Journal of Health Economics* 23(6):1071-1078.

Fenwick, E., Claxton, K., and M Sculpher, 2001. "Representing uncertainty: the role of cost-effectiveness acceptability curves", *Health economics*, 10(8), 779-787.

Garber, A.M., 2000. Advances in cost-effectiveness analysis of health interventions. In *Handbook of health economics* (Vol. 1, pp. 181-221). Elsevier.

Garber, A.M. and Phelps, C.E., 1997. Economic foundations of cost-effectiveness analysis. *Journal of health economics*, *16*(1), pp.1-31.

Garrison Jr, L.P. and Towse, A., 2017. Value-based pricing and reimbursement in personalised healthcare: introduction to the basic health economics. *Journal of personalized medicine*, *7*(3), p.10.

Grosse SD., 2014, Economic analyses of genetic tests in personalized medicine: clinical utility first, then cost utility. *Genet Med*, 16:225-7.

Hatz, M.H.M., K. Schremser and W.H. Rogowski, 2014. "Is individualized medicine more cost-effective? A systematic review". *Pharmacoeconomics* 32 (5), 443-455.

Haute Autorité de Santé 2012. Choices in methods for economic evaluation. A methodological guide. https://has-sante.fr/upload/docs/application/pdf/2012-10/choices_in_methods_for_economic_evaluation.pdf, accessed on Jan 21, 2024.

Hirth, R.A., Chernew, M.E., Miller, E., Fendrick, A.M. and Weissert, W.G., 2000. Willingness to pay for a quality-adjusted life year: in search of a standard. *Medical decision making*, *20*(3), pp.332-342.

Hunter, R. and Shearer, J., 2014. Cost-consequences analysis-an underused method of economic evaluation. *National Institute for Health Research*, pp.4-5.

Jones, H. E., Gatsonsis, C. A., Trikalinos, T. A., Welton, N. J., and A.E. Ades, 2019. Quantifying how diagnostic test accuracy depends on threshold in a meta analysis. *Statistics in Medicine*, 38(24), 4789-4803.

Klibanoff P, Marinacci M and S. Mukerji, 2005, “A smooth model of decision making under ambiguity.” *Econometrica*. 73:1849–92.

Kohn, M. A., and Newman, T. B., 2001. What white blood cell count should prompt antibiotic treatment in a febrile child? Tutorial on the importance of disease likelihood to the interpretation of diagnostic tests. *Medical Decision Making*, 21(6), 479-489.

Laking, G., Lord, J. and Fischer, A., 2006. The economics of diagnosis. *Health economics*, *15*(10), pp.1109-1120.

Lucas, F.L., Siewers, A.E., Malenka, D.J. and Wennberg, D.E., 2008. Diagnostic-therapeutic cascade revisited: coronary angiography, coronary artery bypass graft surgery, and percutaneous coronary intervention in the modern era. *Circulation*, *118*(25), pp.2797-2802.

Luıs, A. B., and M.K. Seo, 2021. “Has the development of cancer biomarkers to guide treatment improved health outcomes?” *The European Journal of Health Econ*omics 22, 789-810.

McNerney R., 2015. “Diagnostics for Developing Countries”, *Diagnostics (Basel)* 5(2), 200-9.

Meltzer, D.O., Basu, A. and Sculpher, M.J., 2016. Theoretical foundations of cost-effectiveness analysis in health and medicine. In *Cost-effectiveness in health and medicine* (pp. 39-66). New York: Oxford University Press.

National Institute for Health and Care Excellence, 2013. *Guide to the methods of technology appraisal*. https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781.

Neumann PJ, Fang CH and JT Cohen, 2009. 30 years of pharmaceutical cost–utility analyses: growth, diversity and methodological improvement. *PharmacoEconomics*; 27(10):861–72.

Neumann, P.J., Gillian D. Sanders, L Russell, J. Siegel and T. Ganiats, 2015. *Cost-Effectiveness in Health and Medicine*. Oxford University Press.

Nimdet, K., Chaiyakunapruk, N., Vichansavakul, K. and Ngorsuraches, S., 2015. A systematic review of studies eliciting willingness-to-pay per quality-adjusted life year: does it justify CE threshold?. *PloS one*, *10*(4), p.e0122760.

Nimmesgern, E., Norstedt, I. and Draghia-Akli, R., 2017. Enabling personalized medicine in Europe by the European Commission's funding activities. *Personalized medicine*, *14*(4), pp.355-365.

Oosterhoff, M., van der Maas, M.E. and L.M. Steuten, 2016: A systematic review of health economic evaluations of diagnostic biomarkers. *Appl. Health Econ. Health Policy*, 14(1), 51–65.

Pauker, S. G. and JP Kaussirer, 1980. “The Threshold Approach to Clinical Decision Making”, *N. Engl. J. Med.* 302(20), 1109-1117.

Pettitt DA, Raza S, Naughton B, Roscoe A, Ramakrishnan A, Ali A, Davies B, Dopson S, Hollander G, Smith JA and Brindley DA, 2016. “The Limitations of QALY: A Literature Review”, *J Stem Cell Res Ther* 2016, 6:4.

Phelps CE and AI Mushlin, 1988, “Focusing technology assessment using medical decision theory”, *Med Decis Making*; 8(4): 279–289.

Phillips KA, Ann Sakowski J, Trosman J, Douglas MP, Liang SY and P. Neumann, 2014, “The economic value of personalized medicine tests: what we know and what we need to know”, *Genet Med;*16:251-7.

Pichon-Riviere, A., Drummond, M., Palacios, A., Garcia-Marti, S. and F. Augustovski, 2023. “Determining the efficiency path to universal health coverage: cost-effectiveness thresholds for 174 countries based on growth in life expectancy and health expenditures”, *The Lancet Global Health* 11(6), pp. 833-842.

Rand, L.Z. and A.S. Kesselheim, 2021. “Controversy Over Using Quality-Adjusted Life-Years in Cost-Effectiveness Analyses: A Systematic Literature Review,” *Health Affairs* 40(9):1402-1410.

Rautenberg, T., Gerritsen, A. and Downes, M., 2020. Health economic decision tree models of diagnostics for dummies: a pictorial primer. *Diagnostics*, *10*(3), p.158.

Redekop, W. and D. Mladsi, 2013, "The faces of personalized medicine: a framework for understanding its meaning and scope." *Value in Health*, 16.6: S4-S9.

Report of the President’s Council of Advisors on Science and Technology, 2008. Available at https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/archives

Roberts, M.S., 2016. “The Next Chapter in Cost-effectiveness Analysis”, *Journal of the American Medical Association* 316(10), 1049-1050.

Ryen L and M. Svensson, 2015. “The Willingness to Pay for a Quality Adjusted Life Year: A Review of the Empirical Literature”, *Health Economics* 24(10):1289-1301.

Sanghera, S., Orlando, R., and T. Roberts, 2013. Economic evaluations and diagnostic testing: an illustrative case study approach*. International Journal of Technology Assessment in Health Care*, 29(1), 53-60.

Sevim, D. and Felder, S., 2022. Decision thresholds for medical tests under ambiguity aversion. *Frontiers in Health Services*, *2*, p.825315.

Siontis, K.C., Siontis, G.C., Contopoulos-Ioannidis, D.G. and Ioannidis, J.P., 2014. Diagnostic tests often fail to lead to changes in patient outcomes. *Journal of clinical epidemiology*, *67*(6), pp.612-621.

Schneider, P., 2022. “The QALY is ableist: on the unethical implications of health states worse than dead,” *Quality of Life Research* 31:1545–1552.

Snowsill, T., 2023. “Modelling the Cost-Effectiveness of Diagnostic Tests,” *PharmacoEconomics* 41:339–351

Sutton, A.J., Cooper, N.J., Goodacre, S. and Stevenson, M., 2008. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. *Medical Decision Making*, *28*(5), pp.650-667.

Téhard, B., Detournay, B., Borget, I., Roze, S. and De Pouvourville, G., 2020. Value of a QALY for France: a new approach to propose acceptable reference values. *Value in Health*, *23*(8), pp.985-993.

Trusheim, Mark R., Ernst R. Berndt, and F. Douglas., 2007, "Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers." Nature reviews Drug discovery 6.4: 287-293.

United Nations; 2015. United Nations The General Assembly. Resolution adopted by the General Assembly on 25 September 2015.

Vallejo-Torres, L., García-Lorenzo, B., Castilla, I., Valcárcel-Nazco, C., García-Pérez, L., Linertová, R., Polentinos-Castro, E. and Serrano-Aguilar, P., 2016. On the estimation of the cost-effectiveness threshold: why, what, how?. *Value in Health*, *19*(5), pp.558-566.

Van der Pol, S., Garcia P.R., Postma, M.J., Villar, F.A. and A.D.I. van Asselt, 2021. "Economic Analyses of Respiratory Tract Infection Diagnostics: A Systematic Review", *PharmacoEconomics* 39, pp. 1411-1427.

Vellekoop, H, et al., 2022 "The Net Benefit of Personalized Medicine: A Systematic Literature Review and Regression Analysis." *Value in Health* 25 (8): 1428-1438.

Weinstein, M.C. and Stason, W.B., 1977. Foundations of cost-effectiveness analysis for health and medical practices. *New England journal of medicine*, *296*(13), pp.716-721.

World Health Organization, 2011. Increasing access to diagnostics through technology transfer and local production. Available at https://www.who.int/publications/i/item/9789241502375

World Health Organization, 2021. "The selection and use of essential in vitro diagnostics", Report of the third meeting of the WHO Strategic Advisory Group of Experts on In Vitro Diagnostics. Available from: https://iris.who.int/bitstream/handle/10665/339064/9789240019102-eng.pdf?sequence=1

Yang Y, Abel L, Buchanan J, Fanshawe T and B. Shinkins, 2019. "Use of Decision Modelling in Economic Evaluations of Diagnostic Tests: An Appraisal and Review of Health Technology Assessments in the UK", *Pharmacoecon Open* Sep;3(3):281-291.