Discover Health Systems

Review

The health technology assessment approach of the economic value of diagnostic tests: a literature review

David Bardey[1] · Philippe De Donder[2] · Vera Zaporozhets[3]

Received: 15 May 2024 / Accepted: 17 February 2025
Published online: 23 March 2025

Abstract
We review the medico-economic literature assessing the economic value of diagnostic and prognostic tests, with a focus on innovative and, more specifically, companion tests. Our analysis begins with a summary of systematic reviews that provide a descriptive synthesis of existing findings rather than conducting quantitative meta-analyses. These reviews reveal no consistent evidence that such tests outperform traditional approaches, such as pharmaceutical interventions. However, the cost-effectiveness of these tests, often measured in cost per QALY (Quality-Adjusted Life Year) gained, exhibits considerable heterogeneity. Notably, some genetic testing procedures may demonstrate superior performance compared to non-genetic alternatives. We then examine the economic implications of imperfect test features, exploring strategies to optimize their accuracy levels and integrating these considerations into the assessment of their economic value. Lastly, we review recent methodological and empirical studies employing these approaches, highlighting advancements in evaluating the economic impact of diagnostic and prognostic tests.

Keywords Genetic tests · Innovative tests · Companion tests · Health Technology Assessment (HTA) · Personalized medicine · Receiver-operator (ROC) curve · Incremental cost-effectiveness ration (ICER)

JEL Classification H51 · I18 · J17

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44250-025-00196-9.

✉ David Bardey, d.bardey@uniandes.edu.co; Philippe De Donder, philippe.dedonder@tse-fr.eu; Vera Zaporozhets, vera.zaporozhets@tse-fr.eu | [1]U Los Andes and TSE, Bogotá, Colombia. [2]Toulouse School of Economics, CNRS, University Toulouse Capitole, Toulouse, France. [3] Toulouse School of Economics, INRAe, University Toulouse Capitole, Toulouse, France.

1 Introduction

Diagnostics play a pivotal role in health systems, supporting epidemic response, health surveillance, and screening programs. They are also essential for achieving universal health coverage and contributing to the United Nations' Sustainable Development Goal 3: "Ensure healthy lives and promote well-being for all at all ages" [59]. In this context, diagnostics encompass any equipment, method, or system used to establish a medical diagnosis [32, 64].[1] While some tests, like auscultation, require minimal equipment, others rely on highly sophisticated and expensive technologies, such as magnetic resonance imaging (MRI) [55].

Technological innovation over recent decades has driven remarkable advancements in healthcare. These breakthroughs extend beyond the development of new drugs and vaccines to include targeted cancer therapies, cutting-edge diagnostic imaging, and minimally invasive surgical techniques. Advances in genomics, such as those stemming from the Human Genome Project, have enabled the identification of disease subtypes based on genetic profiles. This knowledge facilitates the stratification of patients according to genetic mutations, allowing clinicians to identify individuals most likely to respond to specific treatments. Consequently, innovative treatments are increasingly paired with specialized diagnostic tests, known as companion diagnostics.

Today, diagnostic testing is integral not only to establishing a diagnosis but also to guiding management strategies by providing critical information for stratifying patients into the most appropriate treatment pathways.[2] Given the growing importance of diagnostics in modern medicine, a key question emerges: how should we evaluate the effectiveness and value of a diagnostic test?

Historically, the evaluation of diagnostic tests has primarily focused on their clinical accuracy—how effectively they categorize patients as having or not having a disease. While clinical accuracy is an essential aspect of test evaluation, it alone does not capture the broader impact of diagnostic tests on patient outcomes. Recognizing this, the World Health Organization (WHO) has outlined criteria for an ideal diagnostic test, particularly for point-of-care applications. These criteria, summarized by the acronym ASSURED,[3] emphasize three core attributes: accuracy, accessibility, and affordability. However, no test is perfect, and trade-offs among these attributes must be considered at different levels of the healthcare system. For example, while a test with 100% accuracy is theoretically ideal, achieving such precision in routine clinical practice is unrealistic. Therefore, practical compromises are often made between accuracy and accessibility to ensure diagnostic tools meet the needs of diverse healthcare settings.

Although clinical accuracy is crucial, it fails to fully reflect the influence of diagnostic tests on patient health outcomes. Like pharmaceuticals, a new diagnostic test should ideally be introduced into clinical practice only if it offers a higher likelihood of improving patient health compared to existing alternatives [11]. When integrating new diagnostic technologies into healthcare systems, economic evaluations play a critical role by comparing the effectiveness of these innovations against their expected costs. However, such evaluations are inherently complex because the relationship between a diagnostic test's use, its impact on health outcomes, and its contribution to total healthcare costs is often indirect. Furthermore, each test must be carefully matched to its testing environment, taking into account factors such as population characteristics, disease prevalence, and the structural features of the health system.

In recent decades, population aging, the rising prevalence of chronic diseases, and escalating healthcare costs have emerged as significant global challenges [38]. To address these issues, health authorities worldwide are increasingly seeking tools to facilitate the implementation of effective new health technologies while simultaneously managing healthcare expenditures. One key approach involves the use of health economic assessments, which provide valuable insights to support informed decision-making. These assessments help allocate limited healthcare resources efficiently and guide pricing and reimbursement decisions, ensuring that innovations deliver maximum value within constrained budgets.

We narrow the scope of this manuscript in two key ways. First, we focus exclusively on diagnostic and prognostic tests, which encompass procedures designed to identify a patient's condition and determine the most appropriate treatment

[1] The UK Faculty of Public Health stresses that "screening tests are not diagnostic tests" (see HealthKnowledge). The primary purpose of screening tests is to detect early disease or risk factors for disease in large populations of seemingly healthy individuals. In contrast, diagnostic tests aim to confirm or rule out the presence of disease, typically as a follow-up to symptoms or a positive screening result, to guide treatment decisions.

[2] The growing adoption of point-of-care tests across modern health systems reflects an increasing awareness among decision-makers of the critical role that diagnostic tests play in optimizing and rationalizing medical services. These tests not only support timely and accurate diagnoses but also contribute to more efficient resource allocation and improved patient outcomes.

[3] "ASSURED"—Affordable, Sensitive, Specific, User friendly, Rapid and robust, Equipment-free, and Deliverable to end-users.

options. This includes assessing the likelihood of adverse effects and optimizing dosage requirements. By taking this approach, we deliberately exclude susceptibility tests, which estimate an individual's probability of developing a disease in the future.

Second, our umbrella review specifically emphasizes recent, innovative tests, with particular attention to companion diagnostics. These tests are instrumental in aligning treatments with individual patient characteristics, playing an increasingly significant role in both optimizing the use of existing therapies and facilitating the approval of new ones. For example, a review by the European Medicines Agency found that nearly half of the cancer drugs authorized between 2015 and 2018 required patients to undergo genetic testing to determine the most suitable treatment [2].

The objective of this manuscript is twofold. First, in Sect. 2, we provide an overview of reviews summarizing the empirical evidence on the features and outcomes of innovative diagnostic tests. Second, we offer theoretical insights into the assessment methodologies for diagnostic tests when their accuracy is endogenous, determined by the trade-off between false positives and false negatives (or specificity and sensitivity, respectively).

Section 3 begins by presenting the canonical framework used to address the trade-off between sensitivity and specificity in test design. We then discuss how authors, such as Laking et al. [30], utilize this framework, particularly through the central concept of the Receiver Operating Characteristic (ROC) curve, to evaluate the value of the information provided by diagnostic tests. In the second part of Sect. 3, we return to the empirical approach, outlining, with the support of Sutton et al. [56], the methodology for conducting meta-analyses of tests with varying sensitivity and specificity. This is followed in a third subsection by a review of the recent methodological and empirical literature employing this approach.

Section 4 concludes the manuscript, and we summarize the key takeaways at the end of each section for clarity and emphasis.

2 The empirical assessments of diagnostic tests

Historically, the evaluation of new diagnostic techniques has predominantly focused on clinical validity metrics, such as test sensitivity and specificity. While test accuracy is an important component of assessment, it alone does not capture the impact of a diagnostic test on patient health outcomes. Ideally, a new diagnostic test should be introduced into clinical practice only if it offers a greater likelihood of improving patient health compared to existing alternatives [11]. One method for comparing tests involves directly evaluating their downstream consequences through randomized controlled trials (RCTs). These test-treatment RCTs randomly assign patients to different diagnostic tests, monitor subsequent management decisions, and measure outcomes after treatment. However, such trials are relatively uncommon due to the significant challenges involved in their design and execution, as well as the need to deliver robust and generalizable results [66].

A critical limitation of test-treatment trials is their inability to fully estimate the overall effect of a diagnostic test on costs and health outcomes without relying on modelling assumptions [55]. As a practical alternative, decision-analytic modelling—referred to as indirect evidence for clinical assessment—is widely recognized. This approach facilitates the simultaneous evaluation of both economic and clinical impacts, offering a comprehensive framework for assessing diagnostic tests.

Since 2022, the European in vitro *diagnostic regulation* law has mandated that companies demonstrate the clinical effectiveness of new diagnostics before they can be brought to market [61]. Additionally, recent systematic reviews of test-treatment RCTs indicate that improvements in test accuracy rarely correlate with measurable benefits to patient health [53, 66]. Moreover, the decision to adopt a new diagnostic test should not be based solely on accuracy metrics. Other critical factors, such as time to diagnosis and patient acceptability, and the practicality of point-of-care tests, must also be considered. Evaluating the impact of a diagnostic test on patient health outcomes requires an integrated approach, considering the test as part of a broader test-treatment management strategy [11, 55].

All the reviews surveyed here utilize Health Technology Assessment (HTA) methods to evaluate the economic value of diagnostic tests. These methods are classified into four main types, based on how health outcomes are measured and valued. Cost-Benefit Analysis (CBA) quantifies both costs and benefits in monetary terms to assess whether benefits outweigh costs. Cost-Effectiveness Analysis (CEA) compares interventions by calculating costs per unit of health outcome, such as lives saved or cases of disease prevented, making it particularly suitable for evaluating relative efficiency. Cost-Utility Analysis (CUA) is a specialized form of CEA that incorporates patient preferences by measuring outcomes in terms of Quality-Adjusted Life Years (QALYs) or Disability-Adjusted Life Years (DALYs), which allows for comparisons

across a wide range of health interventions. Cost-Minimization Analysis (CMA), applicable only when interventions yield equivalent outcomes, focuses exclusively on identifying the least costly option. These HTA methods collectively provide a framework for evidence-based decision-making in the allocation of healthcare resources. For a detailed comparison of these methods, readers can refer to the Online Appendix, which also highlights that CEA and CUA are the most widely recommended approaches by health authorities worldwide for assessing various health interventions.

That said, the application of HTA methodologies to diagnostics is less established than their use for treatments [61]. This is primarily because, unlike pharmaceuticals, which have a direct and immediate impact on patient health outcomes, the benefits of diagnostic technologies are indirect, materializing only when test results lead to changes in downstream clinical interventions. Diagnostic tests can enhance patient outcomes if improvements in accuracy lead to better diagnoses and more appropriate treatments. Additionally, tests may deliver benefits such as comparable accuracy at reduced costs, streamlined healthcare delivery, enhanced diagnostic confidence, improved diagnostic yield, faster diagnoses, reduced patient anxiety, greater certainty, or improved safety [11].

We review the recent empirical literature on the economic value of innovative tests in general, often referred to as precision medicine tests.[4]

Most of the studies we identified[5] are systematic reviews that gather relevant papers by querying databases, primarily PubMed but also others, using terms related to economic evaluations and diagnostic tests (or personalized medicine variants). Due to the heterogeneity of the evaluations included in these reviews, our synthesis does not attempt to combine results into quantitative meta-analyses. Instead, we provide a descriptive summary of the findings presented in these studies. Below, we outline their results, beginning with the earliest studies within the selected time frame.

2.1 Berm et al. [5]

Berm et al. [5] 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 out 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. The authors also highlight a divergence in the focus of economic evaluations of pharmacogenomic (PGx) testing. On one hand, some studies assess the intrinsic value of the test itself, while others evaluate the value of the test in conjunction with an active compound, such as targeted therapies. For instance, the evaluation of standalone tests (e.g., KRAS testing for colorectal cancer) often demonstrates both cost savings and improved health outcomes. In contrast, the evaluation of targeted therapies typically shows improved health outcomes but at a higher cost. Over time, as targeted therapies become standard care and are subsequently compared with newer treatments, the proportion of studies exclusively evaluating the economic impact of standalone tests declines.

A quarter of the studies surveyed conclude that PGx testing is dominant, resulting in both clinical benefits and cost savings. Several recent studies further provide the specific conditions under which genetic testing might be cost-effective, for instance as a function of the share of the population at risk. 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. Studies comparing a pharmacogenomic (PGx) test-treatment combination (targeted therapy) with an alternative treatment that does not rely on pharmacogenetics consistently found the latter approach to be cost-effective.

While the authors note an improvement in the quality of studies over time, they also highlight two areas of concern that warrant attention. 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

[4] We refer the reader to the Online Appendix for a discussion of the link between diagnostic tests and the emerging field of personalized and precision medicine.

[5] We utilized search engines such as Google, ResearchGate, and particularly Google Scholar to identify recent publications (i.e., from the past 10 years) related to our topic. We employed search terms like "economic assessments", "diagnostics", "individualized or personalized medicine", "biomarkers" (along with their synonyms) and leveraged Google Scholar to explore both papers cited by and those citing the research articles we initially identified. This iterative approach allowed us to uncover a comprehensive and up-to-date selection of relevant studies.

Table 1 Distribution of ratios of cost per QALY gained for personalized medicine tests. *Source:* Fig. 2 of Philipps et al. [43]

Ratios of cost per QALY	Percentage of the studies (%)
Cost-saving%	20
< $20,000	31
$20,000–$50,000	13
$50,000–$100,000	16
>$100,000	12
Higher cost and less effective	8

QALY, quality-adjusted life year. $N = 136$ weighted ratios

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 in the case of industry-sponsored studies, as the positive biased results appear unrelated 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 reviewed in [5] assume that tests results are immediately available. Considering the turnaround time of the tests would then decrease their cost effectiveness.

2.2 D'Andrea et al. [10]

Both [5] and [10] share the observation that very few potential genetic/genomic applications (tests or interventions) have been implemented into clinical practice.[6] D'Andrea et al. [10] identify a lack of appreciation for the cost–benefit of new testing regimes as a key barrier to their implementation. To address this, they conducted a systematic review encompassing 128 primary economic evaluations (EEs) of predictive genetic and pharmacogenetic testing programs, along with 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 evaluations, 57% of the total), 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 differ with 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 the 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 a higher cost, with almost half the ratios falling below €37,000 per QALY, a commonly used threshold. Seventeen percent of genetic testing programs are cost-saving. Pharmacogenetic testing programs are more likely to yield cost savings, whereas predictive genetic tests more often achieve cost-effectiveness ratios below the commonly accepted threshold of €37,000 per QALY.

That said, D'Andrea et al. [10] echo the concerns raised by Berm et al. [5] regarding the lack of demonstrated clinical utility for a substantial proportion of genetic tests, which often renders them not cost-effective.

2.3 Philipps et al. [43]

Philipps et al. [43] reviewed 59 cost–utility analyses of personalized medicine tests conducted between 1998 and 2011. The majority (72%) of cost-per-QALY ratios suggest that personalized medicine tests improve health outcomes, albeit at a higher cost. Nearly half of these ratios fall below the commonly accepted threshold of $50,000 per QALY gained, while 80% are below $100,000 per QALY gained. Additionally, 20% of the studies indicate that these tests are cost-saving,

[6] D'Andrea et al. [10] 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.

while 8% report that the tests may incur higher costs without delivering improved health outcomes. Table 1 shows the distribution of cost *per* QALY in the studies reviewed in Philipps et al. [43].

Philipps et al. [43] conducted a comparison of the cost-utility analyses (CUAs) of personalized medicine tests with those of pharmaceuticals. The authors chose pharmaceuticals as a point of comparison due to their close relationship with personalized medicine tests and the availability of a larger body of research in this area. Indeed, there are significantly more CUAs related to pharmaceuticals (n = 1385) than to personalized medicine tests (n = 59), which provides a broader dataset for analysis. Although the number of cost-utility analyses (CUAs) for personalized medicine tests has been increasing over time, in 2011, there were still significantly more published CUAs for pharmaceuticals (n = 148) than for personalized medicine tests (n = 10). The distribution of cost/QALY ratios was found to be similar for both somatic (acquired) versus germline (inherited) mutations, as well as for personalized medicine tests versus pharmaceuticals.

Summarizing the results provided in [43], Grosse [22] 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).

2.4 Hatz et al. [23]

Hatz et al. [23] 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 in [35] 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." (p. 9).

2.5 Vellekoop et al. [62]

Vellekoop et al. [62] provide the most recent and comprehensive assessment of the cost-effectiveness of personalized medicine (PM), as their study includes 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 conduct regression analyses in which they explore the heterogeneity in the cost-effectiveness of PM interventions.

They undertake 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 study [23]. 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 the 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 our online appendix). The cost-effectiveness threshold used reflects the opportunity cost of healthcare spending, rather than society's willingness to pay for improvements in health, due to the availability of national estimates for all countries included in the dataset.

The median gain in QALYs for personalized medicine (PM) interventions compared to their non-PM counterparts was 0.03, while the mean gain was 0.26. Most incremental QALY values were just above 0, with the 25th and 75th percentiles at 0.00 and 0.16, 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 suggests 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.

The median cost was Int$575, while the mean cost approached Int$100,000. A small number of interventions exhibited notably higher costs than the others. On average, the cost of gene therapies was more than 1 million Int$ higher than that of non-gene therapy PM interventions. The median NMB across the included interventions was Int$18, while the mean NMB was Int$277,072. NMB values were centered around 0, with a value of Int$-22,665 at the first quartile and Int$3,538 at the third quartile. Extreme negative values were more common than extreme positive values for NMB. The median NMB of personalized medicine (PM) interventions being close to 0 suggests that any QALY gains from these interventions are often offset 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 with 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 is in line with the concern stressed in [5] (see Sect. 2.1) about the publication biases linked to the sponsorship of the studies.

The following Table 2 summarizes the results obtained by the papers summarized so far.

2.6 Luis and Seo [31]

Luis and Seo [31] provide two key reasons for the slower-than-expected progress in the clinical application of biomarkers, with few reaching clinical practice. First, they highlight the limitations of genetic prediction due to the inherent biological complexity. Second, they emphasize the lack of appropriate incentives for pharmaceutical companies. The authors advocate for the economic evaluation of biomarker tests using real-world longitudinal and/or patient data, contrasting this with the majority of existing evaluations, which rely on clinical trial data or simulations based on such data. They begin by reviewing the literature on the impact of pharmaceutical innovation in cancer, particularly in terms of increasing survival or reducing mortality.

The aim of Luis and Seo [31] 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.[7] 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

[7] An earlier paper [39] 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.

Table 2 Results from Sects. 2.1 to 2.5

Study	Purpose and methods	Main conclusions
Philipps et al. [43]	—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. [23]	—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. [10]	—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%)	—Most 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. [5]	—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 tests only results in both cost savings and better health —The evaluation of targeted therapies (combination of a test with an active compound) generates 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 specific conditions for genetic testing being cost-effective —Possibly, there is a bias for industry sponsored studies
Vellekoop et al. [62]	—Most recent and most complete assessment of the cost effectiveness of personalized medicine based on [23]. 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 —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 the least and most common respectively	—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)

of premature mortality over survival probability is that the former is not subject to lead-time bias.[8] They document that having at least one biomarker test available decreases premature mortality on average. Surprisingly, they demonstrate that the total effect of biomarker testing on survival diminishes as the number of available cancer drugs increases. This suggests that while biomarker tests improve health by better matching patients to treatments, this matching is more effective when fewer drugs are available. The authors provide several reasons for this, including the time required to test patients for multiple biomarkers, the bias among doctors who tend to prioritize well-known drugs, and, more broadly, the increased complexity of treatment decisions when more biomarkers and drugs are available. This complexity makes it more challenging to "match the right patient to the right drug."

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 65 and 75. They attribute these differences to variations in the samples used for the regression analyses on premature mortality and survival, along with the plausible assumption that cancer patients nearing the end of life benefit more from new drugs than those who have just been diagnosed.

Finally, their estimates of the cost per life-year gained (LYG) before ages 65 and 75 in 2016 for biomarker-guided drugs introduced between 2000 and 2015 are well below the EUR 30,000 per QALY often cited in the literature as the threshold for cost-effectiveness (it is important to note that the authors compare their estimate of cost per LYG to cost thresholds for QALYs). As highlighted in this summary, the main limitation of their analysis is the lack of data, which prevents a more detailed exploration of the underlying mechanisms.

KEY TAKEAWAYS

In this section, we focus on the empirical assessment of personalized medicine (PM) tests. PM has led to two opposing predictions: (i) that such tests would limit costly treatments to those who would benefit, enabling health systems to save costs while achieving better (or at most slightly deteriorating) health outcomes, and (ii) that some patients would find very costly procedures worthwhile, resulting in higher health costs for improved health.

We summarize systematic reviews that do not conduct quantitative meta-analyses but instead provide a descriptive synthesis of the reviewed results. Empirically, only a small fraction—around one fifth to one quarter—of the studied PM tests result in cost savings. The majority of cost-utility analyses (CUAs) of targeted therapies (i.e., the joint evaluation of tests and therapies) show improvements in health, but at higher costs, with a significant proportion (though not the entirety) achieving a cost-per-QALY that would be considered cost-effective by today's standards.

More generally, there is no evidence that PM performs better in terms of cost *per* QALY compared to more traditional approaches, such as pharmaceutical interventions. However, there is considerable heterogeneity in the cost per QALY gained, meaning that some genetic testing procedures may outperform non-genetic alternatives.

The studies reviewed raise two key concerns. First, most are based on simulations or expert opinions rather than solid clinical evidence, due to the lack of the latter. Second, there is a potential for publication bias, as an increasing share of industry-sponsored studies all conclude the effectiveness of the test under consideration, in contrast to the findings of other studies.

The second part of this manuscript delves into the inherent imperfections of diagnostic tests and explores the critical challenge of accounting for their varying degrees of accuracy when assessing their economic value. Given that no test is flawless, understanding the trade-offs between false positives and false negatives is essential for accurately evaluating a test's impact. This part will discuss how these imperfections influence not only the test's ability to correctly identify disease but also the broader implications for healthcare resource allocation, patients' outcomes, and cost-effectiveness. By integrating the test's diagnostic performance with economic evaluations, we can better understand how its accuracy should be weighed against the costs and benefits of its use in healthcare settings. This comprehensive approach is necessary to ensure that tests are deployed in a manner that maximizes societal welfare, considering both clinical and economic outcomes.

3 Health technology assessments of imperfect tests

Diagnostic tests enable clinicians to match the right treatment with the right patient. However, with very few exceptions, these tests are imperfect, as they yield incorrect predictions for a subset of the tested population. It is crucial to account for this accuracy issue when evaluating their economic value. Furthermore, as we will discuss, the degree of accuracy

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

is often endogenous. Economic analysis can play a key role in determining the optimal level of accuracy and the corresponding economic value of an optimized test.

We first present in Sect. 3.1 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 return to the empirical approach, beginning in Sect. 3.2 with a presentation of the methodology used for conducting meta-analyses of tests that vary in sensitivity and specificity. This is followed in Sect. 3.3 by a review of recent methodological and empirical studies that have adopted this approach.

3.1 The analytical approach

The canonical example involves dividing the population into two groups: those with the disease and those without. Diagnostic tests are then used to classify individuals into these two groups. As we will explore, this process requires establishing a threshold value for the test results to distinguish disease-positive individuals from others. The chosen threshold simultaneously determines the proportions of false positives and false negatives, which must be considered when evaluating the test's economic value.

Laking et al. [30] highlight a "schism" between two schools of thought in the evaluation of diagnostic tests. The first school focuses on the test's ability to classify individuals as either affected by a disease or not, emphasizing diagnostic accuracy. The second school, more aligned with economic principles, evaluates the value of the information provided by the tests. According to Laking et al. [30], the second approach had yet to gain widespread adoption in mainstream health technology assessment practices. To address this divide, the authors propose a framework that bridges the two perspectives, starting with the principal analytical tool of the first—the receiver operating characteristic (ROC) curve, to be defined shortly—and integrating diagnostic evaluation into conventional cost-effectiveness analysis.

Most diagnostic tests produce continuous measures. For these tests to be useful, the distribution of the measure must differ between individuals with the disease and those without. This concept is illustrated in Fig. 1, adapted from [56]. In part (a), the yellow curve represents the distribution of test results for the healthy population, while the gray curve represents the distribution for the diseased population.[9] To classify individuals, a threshold value is required. In the figure, individuals with test results above the threshold (D_T) are classified as disease-positive, while those with results below D_T are classified as disease-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 D_T) and false negatives (whose fraction corresponds to the grey area to the left of D_T). The table in part (b) of Fig. 1 reports the fractions of false positive, true positive, false negative and true positive obtained from part (a) of the figure.

In medical literature, the terms *sensitivity* and *specificity* are commonly used instead of false positive and false negative rates. These concepts are defined in part (c) of Fig. 1.

The key points are that (i) specificity and sensitivity are not fixed but depend on the chosen threshold value (D_T), and (ii) there is an inherent tradeoff between the two, as increasing one comes at the expense of the other when the threshold is adjusted.

By changing the value of D_T, we change the corresponding sensitivity and specificity levels of the test. Part (c) of Fig. 1 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 in [30] in the schism diverge.

The first approach attempts to summarize the accuracy of each test using measures such as the "area under the curve" (AUC), where a larger area is considered preferable. However, this approach is not favored by economists for at least two reasons. This approach is not the one favored by economists, for (at least) two reasons.

[9] 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 versus those who require a high dose. The groups of diseased *vs* non-diseased is determined by a so-called "gold standard" test.

First, and less importantly, the ROC curves of two tests may intersect, making it unclear whether the AUC is the most appropriate criterion. A test with a higher true positive rate and a lower false positive rate is inherently more desirable, as health decider's welfare increases as we move closer to the northwest corner of Fig. 1c. Thus, if the ROC curve for one test is entirely above that of another, the former is clearly preferable (assuming both tests have the same cost) and will also have a larger AUC. However, when ROC curves intersect, a larger AUC does not necessarily indicate that the corresponding test is better for society.

Second, and more importantly, this approach fails to account for the health and economic consequences of misallocating patients. For instance, there is no inherent reason to prioritize maximizing accuracy or the total number of correctly diagnosed patients,[10] as it cannot be assumed that the medical and economic consequences of the two types of misdiagnoses—false positives and false negatives—are equivalent.

The economically sound approach when comparing two tests is to first specify the objective function we aim to maximize, then determine the optimal threshold (D_T) for each test, often referred to as the Optimal Operating Point (OOP). The test that yields the highest value for the objective function at its optimal threshold should be preferred. This approach is closely aligned with the concept of cost-utility analysis (CUA), as it involves maximizing the net monetary benefit (NMB) of the information provided by the test through optimal threshold selection, and then choosing the test with the highest NMB at its optimal threshold. We follow the framework outlined in the seminal paper by Laking et al. [30].

Consider a population divided into two groups, labeled *x* and *y*, and two potential treatments, *A* and *B*, for the disease affecting this population. Treatment *A* is better suited to group *x* (in the sense that $NMB_A > NMB_B$ for group *x*), while treatment *B* is better suited to group *y* ($NMB_A < NMB_B$ for group *y*). A diagnostic test is then used to classify patients as belonging to either group *x* or *y*, allowing for the prescription of the most appropriate treatment.

To determine the optimal test threshold D_T, we compute, for each point on the ROC curve, its corresponding position in the cost-effectiveness space. The horizontal axis represents expected QALYs gained (compared to no treatment), while the vertical axis represents the cost. This requires knowledge of the costs of treatments *A* and *B*, the prevalence of subgroup *x* in the population, and the health effects (measured in QALYs) of both treatments in each group (see for example [30]).[11]

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 Fig. 2 below (adapted from [30]). 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 incremental cost-effectiveness ratio (ICER)[12] 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 Fig. 2 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 Fig. 2 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*).

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).[13] Increasing the test threshold 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 its corresponding sensitivity and specificity levels, we then calculate 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 D_T) on Fig. 2 are points O and T, and varying D_T in a continuous way generates the so-called ROTS curve.[14]

[10] This corresponds to the point at which the ROC curve crosses the line where sensitivity equals specificity (the dotted line in Fig. 1c).

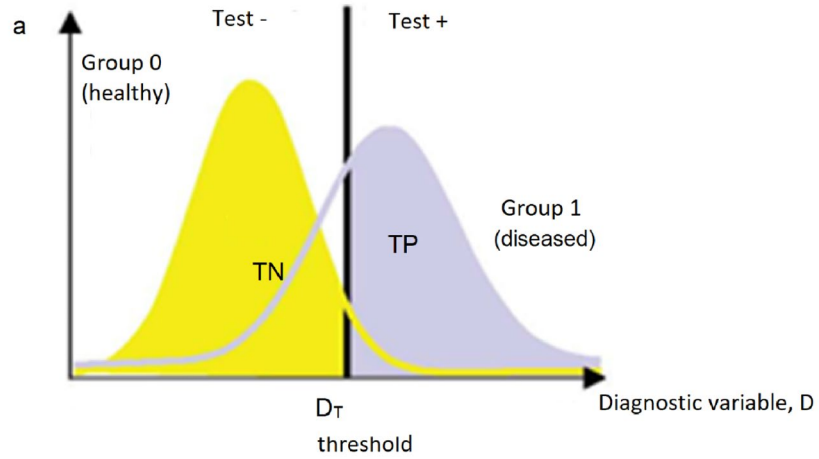
[11] We abstract for the moment from tests costs. We introduce them later in the analysis.

[12] 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. More details are to be found in the online appendix.

[13] 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.

[14] ROTS is not an acronym, but is the term used throughout the literature for this curve following Fig. 2 as represented in [30]!

(a)



(b). Gold standard test result

	Diseased	Non-diseased
Results of test under evaluation	True Positives	False Positives
Positive test result (diseased)	Negatives	True Positives
Negative test result (non-diseased)	Total (assumed) with disease	Total assumed without disease

(c).

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

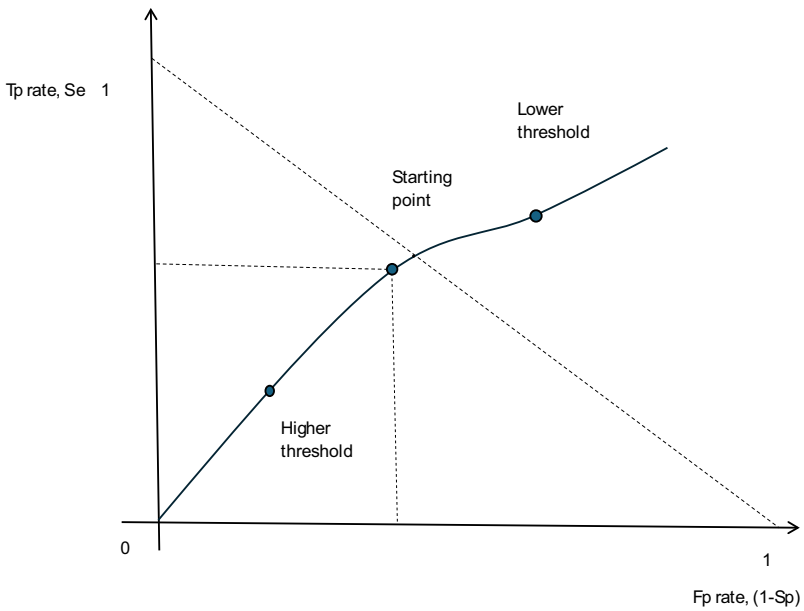


Fig. 1 Evaluation of diagnostic test using data from a single study. **a** Distributions of test results for diseased and non-diseased population with categorization defined by threshold D_T. **b**, **c** A table indicating test categorization of individuals from (**a**, **c**). Receiver operating characteristic curve from changing test threshold. *Source:* Fig. 2 of Sutton et al. [56]

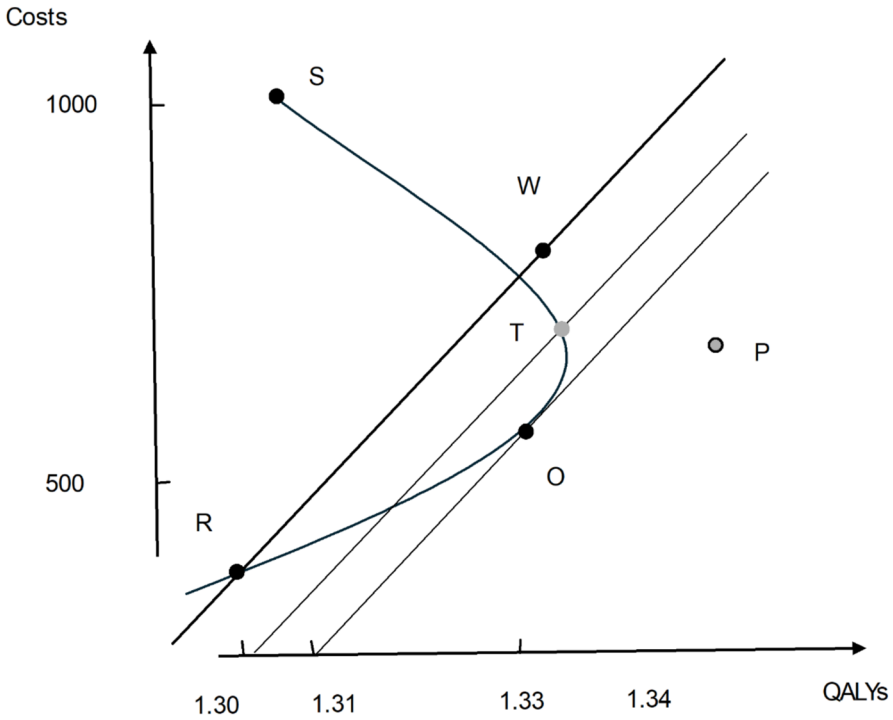


Fig. 2 ROTS curve (without the costs of testing). The lines are net benefit isoquants. *Source:* Fig. 5 of Laking et al. [30]

Finding the optimal threshold D_T (the value that maximizes the NMB of the test) involves locating the point on the ROTS curve where the slope equals the willingness to pay for a QALY. This corresponds to point O on Fig. 2, as it lies on the lowest net benefit isoquant attainable with the test (i.e., the ROTS curve). Additionally, the NMB of the test at its optimal threshold (point O) is represented by the vertical distance between point O and the net benefit isoquant passing through the default treatment point R.

Comparing two (free) tests, the technically superior one is the one that allows us 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 Fig. 3 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 3 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 incorporated in a similar manner, once they are quantified in terms of their probability of occurrence across different groups and the resulting QALY losses.

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 [40] 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 recent decades, genetic tests that assess the presence or absence of genetic mutations have become increasingly common. The accuracy (sensitivity and specificity) of these tests is often very close to 100%. Additionally, the costs of genetic testing have significantly decreased due to revolutionary advances in DNA sequencing technologies. Building on these developments, Felder [17] applies the insights from threshold analysis to genetic testing. In this context, the thresholds refer not to the probability of the disease but to the probability of a genetic mutation. As a result, preventive treatments may become more relevant compared to curative ones. A positive diagnostic test outcome indicates the

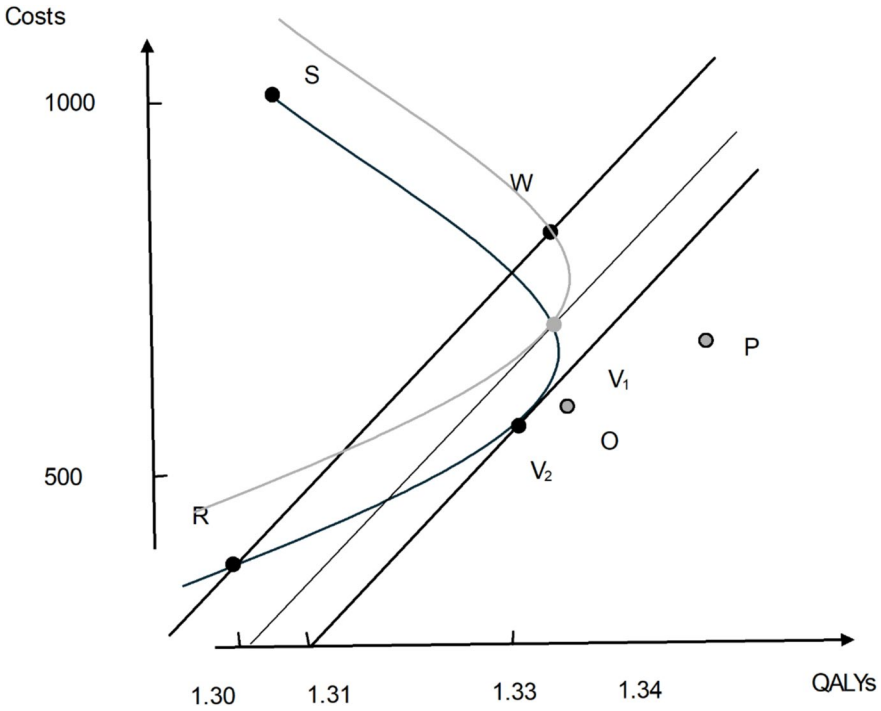


Fig. 3 ROTS curve (with costs of testing). The lines are net benefit isoquants. The black curve is tangent to the most favorable isoquant. *Source:* Fig. 6 of Laking et al. [30]

presence of a mutation associated with an increased risk of disease. If the penetrance rate (the probability of developing the disease given the mutation) is sufficiently high, the patient may opt for preventive treatment.[15]

Finally, the analysis above can be modified to introduce uncertainty, both in the technical characteristics of the test (the ROC curve) and its economic consequences (the ROTS curve).

KEY TAKEAWAYS

Economists evaluate the economic value generated by a test, which depends on its accuracy. This accuracy is determined endogenously through a trade-off between sensitivity and specificity, as illustrated by the Receiver-Operating Characteristic (ROC) curve. One should then identify the optimal point on this ROC, known as the Optimal Operating Point (OOP), based on the objective to be maximized, such as the net marginal benefit provided by the test, as discussed in the seminal paper by [30]. This analysis involves mapping each point on the ROC to its corresponding position in the cost-effectiveness space, then optimizing the decision by considering both the test's cost and the willingness to pay for each additional QALY.

We now review the literature assessing empirically imperfect diagnostic tests.

3.2 The empirical approach - Methodology

Sutton et al. [56] outline how to conduct meta-analyses of diagnostic tests while accounting for the thresholds used across different studies. The main methodological challenge is that studies often report varying sensitivities and specificities, and it is unreasonable to assume these values are independent, except in the rare case where all studies have used the same threshold (D_T). Unfortunately, Sanghera et al. [51] note that "most economic evaluations of diagnostic tests consider sensitivity and specificity to be independent" (p. 54).

To illustrate this issue, Sutton et al. [56] use a case study of 198 studies on the so-called d-dimer test for deep vein thrombosis. The first step involves constructing a summary ROC (SROC) curve based on the sensitivity and specificity

[15] For example, in the prevention of breast or ovarian cancer, women who test positive for the BRCA1 and BRCA2 genes may undergo intensive surveillance, bilateral salpingo-oophorectomy, or mastectomy. However, curative chemotherapy would not be indicated if the cancer has not yet developed. The findings highlight that a low penetrance rate limits the applicability of genetic testing, as the carrier probability threshold becomes high when the disease's penetrance is low. A low penetrance rate results in a low expected monetary value, which may not justify the cost of preventive treatment. These factors could lead to an excessively high carrier probability threshold, making it difficult to justify the use of genetic testing.

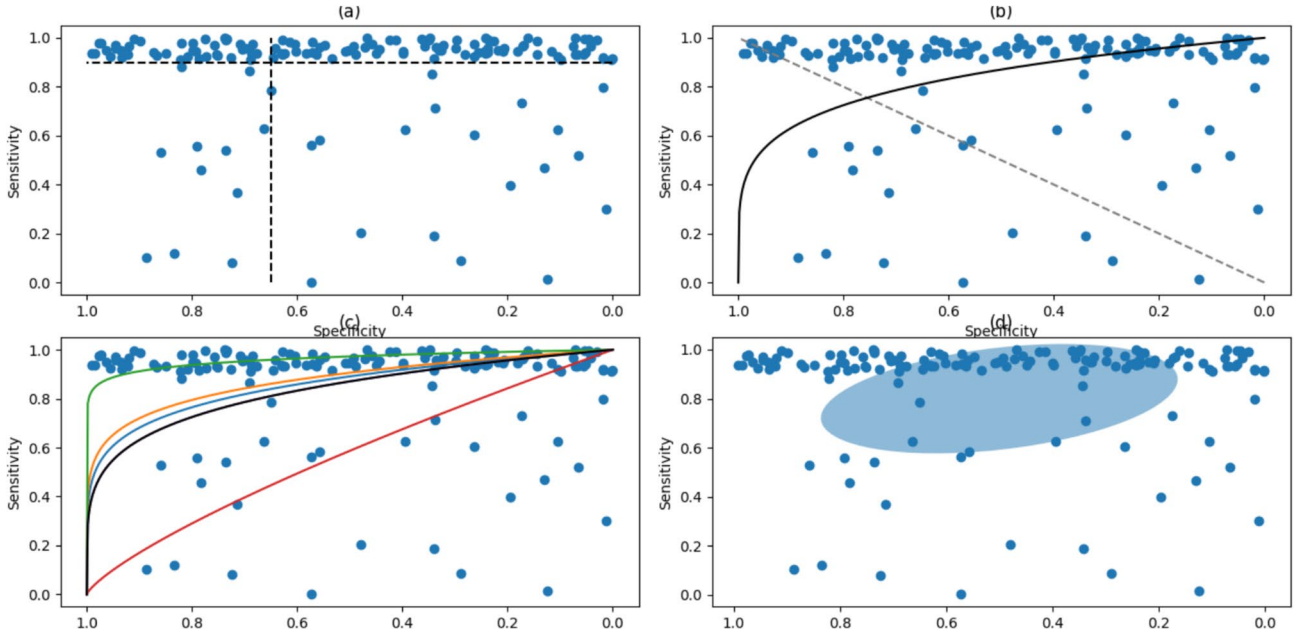


Fig. 4 Graphical display of d-dimer for deep vein thrombosis meta-analyses results. *Source:* Fig. 1 of Sutton et al. [56]

pairs reported across all studies. These pairs are shown in Fig. 4 below. A meta-analysis that assumes independence between sensitivity and specificity leads to the results depicted in Fig. 4a, where the mean estimates and their 95% credible intervals (denoted as "CrIs" in Fig. 4, the Bayesian equivalent of confidence intervals) are represented by sets of eight horizontal and vertical lines. This approach, however, is clearly not appropriate.

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) should be symmetrical around the line where sensitivity equals specificity. This SROC is shown in Fig. 4b, with relatively small credible intervals, reflecting the large number of studies included in the analysis.

Alternatively, one can relax the assumption of a constant diagnostic odds ratio,[16] and instead account for heterogeneity across studies beyond variations in thresholds. This can be achieved by constructing a hierarchical summary ROC curve (HSROC) through regression using study-level covariates. The result is shown in Fig. 4c, where the credible intervals are much wider due to the inclusion of between-study heterogeneity.[17] This approach is also applicable when accuracy data from multiple test thresholds are available within each study.

Once the (H/S) ROC curves have been constructed, the next step is their economic evaluation. Sutton et al. [56] take a slightly different approach compared to [30] by explicitly introducing a medical decision tree model (see their Fig. 1). This model outlines in detail the medical treatment decisions and expected health outcomes (including corresponding QALYs) based on test results and test accuracy. As the test threshold is varied, the proportion of patients entering each branch of the decision tree changes, which in turn alters the net monetary benefit (NMB) of the test. The NMB is defined, as in Sect. 2, as NMB = λ * effect—cost, where the effect is measured in incremental QALYs and λ represents the

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

[17] Alternatively, one can model (logit) sensitivity and specificity as bivariate normally distributed, as reported on Fig. 4d. The dashed line represents the credible region surrounding the mean sensitivity and specificity.

willingness to pay for one additional QALY. This NMB can then be compared to a no-test strategy, and one can perform comparative statics with respect to λ.

In theory, this approach allows us to simultaneously address three key questions: (1) Is any test worth performing? (2) What is the optimal threshold for a test? (3) If multiple tests are available, which one is the best? However, answering all three questions empirically can be challenging due to the lack of comprehensive data. For example, when studies use different thresholds (D_T), one can identify the optimal operating point (OOP, as discussed in Sect. 3.1) on each test's summary ROC (SROC) curve as a function of λ. Nevertheless, it is impossible to infer the corresponding optimal threshold from this point, since none of the meta-analysis methods account for threshold value data from each of the primary studies.

Sutton et al. [56] provide a practical application of their methodology by conducting a meta-analysis of the cost-effectiveness of two diagnostic tests for deep vein thrombosis (DVT): d-dimer and ultrasound. They compare these tests with two strategies: no treatment and treating all patients without prior testing. They estimate the cost-effectiveness of each strategy using three different models: (i) a model based on mean fixed sensitivity and specificity, ignoring threshold effects and heterogeneity (Fig. 4a); (ii) a model that incorporates the mean HSROC curves and optimizes the threshold for each willingness to pay value (λ) to maximize net benefit (Fig. 4c); and (iii) a third model that uses the prediction region around the mean sensitivity and specificity from the bivariate meta-analysis. It's important to note that the "discharge without test" and "treat without test" strategies are treated as points on the extremes of an ROC curve—one with 0 sensitivity and 1 specificity, and the other with 1 sensitivity and 0 specificity, respectively, both costing no money.

They compute, for each model, the probability that the test is cost-effective as a function of λ.[18] 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.[19] 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 depend on a decision maker's willingness to pay.

To conclude, Sutton et al. [56] 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" (p. 665). They stress that "even if only one 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".

3.3 The empirical approach - Applications

Sanghera et al. [51] provide a step-by-step guide of the approach proposed in Sutton et al. [56] and use a case study of fetal anemia in which data from a screening test are used in combination with a confirmatory test.[20] 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 underestimate 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.

Jones et al. [27] introduce a statistical method for meta-analyses that accounts for the fact that many studies report sensitivity and specificity at different test thresholds. Their model assumes that a pre-specified or Box–Cox transformation

[18] The so-called cost effectiveness acceptability curve, or CEAC, proposed by Fenwick et al. [18], depicts the probability that a test is optimal as a function of λ.

[19] 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.

[20] Kohn et al. [29] emphasize that the optimal threshold for a diagnostic test varies with the pre-test probability of disease—specifically, the higher the pre-test probability, the lower the optimal threshold should be.

of test results for both diseased and disease-free populations follow a logistic distribution. By estimating the Box–Cox transformation parameter from the data, their approach accommodates a wide range of underlying distributions. They demonstrate the application of their model through two case study meta-analyses, evaluating the accuracy of tests for heart failure and preeclampsia.

Rautenberg et al. [46] 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. [12] systematically assess published model-based economic evaluations in which targeted oncology therapies are evaluated alongside companion diagnostics. They compare the results of economic evaluations that incorporate model parameters for the sensitivity and specificity of the companion diagnostic with those that limit model parameters for the diagnostic to just its cost. Their findings show that the inclusion of sensitivity and specificity in the model parameters leads to significantly different results compared to evaluations that only account for the diagnostic's cost.

Finally, Drakopoulos et al. [13] demonstrate 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, particularly if agents must decide whether to undergo testing. The reasoning behind this is as follows: when a test has low accuracy, it discourages some agents from opting for it. This can be beneficial if the social planner would not prioritize testing these agents in the first place. However, as the test's accuracy improves, these agents may begin to opt for testing, potentially crowding out higher-priority agents (from the social planner's perspective) from testing. This study was inspired by the initial lack of testing during the onset of the COVID-19 pandemic.

KEY TAKEAWAYS

Meta-analyses should consider that the tests studied may differ in the sensitivity and specificity pairs they report, and that these values are not independent of each other. A partial step in the right direction is to assume a constant diagnostic odds ratio, which results in an ROC curve symmetrical around the line where sensitivity equals specificity. A more refined but also more demanding approach introduces a medical decision tree model, which specifies the health consequences of the test's accuracy. Leveraging individual patient data from a single study can lead to more reliable estimates compared to conducting a meta-analysis based on single-point summaries. Several studies that combine both methodological and empirical approaches demonstrate how cost-effectiveness estimates can be biased when these aspects are not properly addressed.

4 Conclusion

This review provides a comprehensive overview of the existing methods for assessing the economic value of diagnostic tests. Health technology assessment plays a crucial role in guiding social decisions about the optimal allocation of limited healthcare resources. When evaluating the value of innovative tests, the approach taken by economists often differs from those employed by other health researchers in several key ways. Economists focus on assessing the value of the information provided by the test. The design and characteristics of the test are central to this evaluation. In terms of accuracy, there is typically a trade-off between false positives and false negatives (or, in medical terminology, between sensitivity and specificity). Economists first optimize this trade-off before measuring the value of the information at the resulting point. It is important to note that the optimal trade-off depends on the consequences of each type of error, as well as the health benefits achieved when the test accurately guides treatment decisions. These outcomes are influenced by society's willingness to pay for health improvements (e.g., the maximum cost per QALY), attitudes toward uncertainty (or ambiguity when precise probabilities are difficult to determine), and the costs associated with both the test and the treatment.

Rather than reiterating the key takeaways from each section, we conclude by highlighting one central result, addressing a significant critique of the approach we have reviewed, and mentioning an especially relevant area for future research.

The development of innovative tests in the realm of personalized medicine (PM) has led to two opposing predictions. On the one hand, if these tests enable the identification of individuals who do not require expensive or ineffective treatments, they may lead to improved health outcomes at lower costs. On the other hand, these tests are often linked to high-cost treatments that would not be approved without the accompanying diagnostic test. In such cases, one would expect to see health improvements, but at significantly higher costs.

The empirical literature provides evidence supporting both effects. Among the studies reviewed in Sect. 2, approximately one fifth to one quarter align with the first hypothesis (lower costs for equivalent or improved health outcomes), while a larger proportion support the second hypothesis, indicating that health improvements are achieved at a cost per QALY that meets societal standards. Additionally, there is limited evidence to suggest that PM tests perform better (in terms of cost per QALY) than non-genetic tests. However, considerable heterogeneity exists in the cost per QALY gained, meaning that some genetic testing procedures may outperform non-genetic ones.

Since the 1970s, the QALY (Quality-Adjusted Life Year) has been recognized as the most rigorous and standardized metric for evaluating health economic outcomes across various healthcare interventions and conditions. However, a growing body of literature highlights several limitations, both methodological and ethical, as well as those arising from specific contexts [41, 45, 54]. One long-standing ethical critique of the QALY is that it may discriminate against elderly individuals and those with disabilities or chronic illnesses. Specifically, extending the lives of people with underlying health conditions results in fewer QALYs than extending the lives of healthier individuals. While such ethical concerns fall outside the purview of economists, they are essential for political decisions regarding social values. Economists, however, can help illuminate the consequences and trade-offs of favoring different metrics.

From a methodological standpoint, the main criticisms of the QALY concern whether the theoretical assumptions necessary for it to be a valid metric hold true in practice, particularly in relation to the methods used for measuring and the sample sources used to value health states. For instance, some studies argue that different populations may evaluate health conditions differently—utility values for healthcare professionals and the general population are likely to vary. Additionally, QALYs have been criticized for not accounting for non-health benefits, particularly societal benefits, such as a faster return to work or improved school performance (see [41]). While QALY has been the primary measure of health benefits in the literature surveyed here, it is far from being the optimal or sole measure that should be considered.

Finally, as highlighted in the introduction, point-of-care tests are becoming increasingly prevalent in medical practice due to their convenience and ability to deliver rapid results, often in decentralized settings. Their unique characteristics—such as speed of results, operational context, integration into clinical workflows, cost-effectiveness, patient acceptability, accessibility, and the regulatory and quality challenges they present—make them distinct from traditional diagnostics and warrant dedicated attention when evaluating their performance and utility. Addressing these aspects comprehensively is beyond the scope of this paper, and we leave this subject for future research.

Authors contributions David Bardey, Philippe De Donder and Vera Zaporozhets have all the same contribution.

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

Data availability No datasets were generated or analysed during the current study.

Code availability Not applicable.

Declarations

Competing interests The authors declare no competing interests.

References

1. Abbott JH, Wilson R, Pryymachenko Y, Sharma S, Pathak A, Chua JYY. "Economic evaluation: a reader's guide to studies of cost-effectiveness. Arch Physiotherapy. 2022;12(1):28.
2. Antoñanzas F, Rodríguez-Ibeas R, Juárez-Castelló C. Pre-approval incentives to promote adoption of personalized medicine: a theoretical approach. Heal Econ Rev. 2019;9(28):2–10.
3. Baker R, Chilton S, Donaldson C, Jones-Lee M, Lancsar E, Mason H, Metcalf H, Pennington M, Wildman J. Searchers vs surveyors in estimating the monetary value of a QALY: resolving a nasty dilemma for NICE. Health Econ Policy Law. 2011;6(4):435–47.
4. Bardey D, De Donder P. Genetic testing with primary prevention and moral hazard. J Health Econ. 2013;32(5):768–779 & 1007–1012.
5. 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): e0146262.
6. Bobinac A, Van Exel NJ, Rutten FF, Brouwer WB. Willingness to pay for a quality-adjusted life-year: the individual perspective. Value Health. 2010;13(8):1046–55.
7. Brent RJ. Cost-benefit analysis versus cost-effectiveness analysis from a societal perspective in healthcare. Int J Environ Res Public Health. 2023;20:4637.
8. Brouwer WB, Koopmanschap MA. On the economic foundations of CEA. Ladies and gentlemen, take your positions! J Health Econ. 2000;19(4):439–59.
9. Chen G, Peirce V. 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. 2020;23(10):1300–6.
10. D'Andrea E, Marzuillo C, Pelone F, De Vito C, Villari P. Genetic testing and economic evaluations: a systematic review of the literature. Epidemiol Prev. 2015;39(4 suppl 1):45–50.
11. Di Ruffano F. L, Harris IM, Zhelev Z, Davenport C, Mallett S, Peters J, Takwoingi Y, Deeks J, Hyde C. Health technology assessment of diagnostic tests: a state of the art review of methods guidance from international organizations. Int J Technol Assess Health Care. 2023. https://doi.org/10.1017/S0266462323000065.
12. Doble B, Tan M, Harris A, Lorgelly P. Modeling companion diagnostics in economic evaluations of targeted oncology therapies: systematic review and methodological checklist. Expert Rev Mol Diagn. 2014;15(2):235–54.
13. Drakopoulos K, Randhawa RS. Why perfect tests may not be worth waiting for: information as a commodity. Manage Sci. 2021;67(11):6678–93.
14. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. 4th edition. Oxford: Oxford University Press; 2015.
15. Eeckhoudt L. Risk and medical decision making (vol. 14). Berlin: Springer; 2002.
16. European Network for Health Technology Assessment. Methods for Health Economics Evaluations. Guideline Based on Current practices in Europe. 2015. https://www.eunethta.eu/wp-content/uploads/2018/01/Therapeutic-medical-devices_Guideline_Final-Nov-2015.pdf.
17. Felder S. Decision thresholds with genetic testing. Eur J Health Econ. 2022;23(6):1071–8.
18. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. Health Econ. 2001;10(8):779–87.
19. Garber AM. Advances in cost-effectiveness analysis of health interventions. In: Handbook of health economics (Vol. 1). Amsterdam: Elsevier; 2000. pp. 181–221.
20. Garber AM, Phelps CE. Economic foundations of cost-effectiveness analysis. J Health Econ. 1997;16(1):1–31.
21. Garrison LP Jr, Towse A. Value-based pricing and reimbursement in personalised healthcare: introduction to the basic health economics. J Person Med. 2017;7(3):10.
22. Grosse SD. Economic analyses of genetic tests in personalized medicine: clinical utility first, then cost utility. Genet Med. 2014;16:225–7.
23. Hatz MHM, Schremser K, Rogowski WH. Is individualized medicine more cost-effective? A systematic review. Pharmacoeconomics. 2014;32(5):443–55.
24. Haute Autorité de Santé. Choices in methods for economic evaluation. A methodological guide. 2012. https://has-sante.fr/upload/docs/application/pdf/2012-10/choices_in_methods_for_economic_evaluation.pdf. Accessed on January 21, 2024.
25. Hirth RA, Chernew ME, Miller E, Fendrick AM, Weissert WG. Willingness to pay for a quality-adjusted life year: in search of a standard. Med Decis Mak. 2000;20(3):332–42.
26. Hunter R, Shearer J. Cost-consequences analysis-an underused method of economic evaluation. National Institute for Health Research; 2014. pp. 4–5.
27. Jones HE, Gatsonsis CA, Trikalinos TA, Welton NJ, Ades AE. Quantifying how diagnostic test accuracy depends on threshold in a meta analysis. Stat Med. 2019;38(24):4789–803.
28. Klibanoff P, Marinacci M, Mukerji S. A smooth model of decision making under ambiguity. Econometrica. 2005;73:1849–92.
29. Kohn MA, Newman TB. 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. Med Decis Mak. 2001;21(6):479–89.
30. Laking G, Lord J, Fischer A. The economics of diagnosis. Health Econ. 2006;15(10):1109–20.
31. Luis AB, Seo MK. Has the development of cancer biomarkers to guide treatment improved health outcomes? Eur J Health Econ. 2021;22:789–810.

32. McNerney R. Diagnostics for developing countries. Diagnostics (Basel). 2015;5(2):200–9.
33. Meltzer DO, Basu A, Sculpher MJ. Theoretical foundations of cost-effectiveness analysis in health and medicine. In: Cost-effectiveness in health and medicine. New York: Oxford University Press; 2016. p. 39–66.
34. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. 2013. https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781.
35. Neumann PJ, Fang CH, Cohen JT. 30 years of pharmaceutical cost–utility analyses: growth, diversity and methodological improvement. Pharmacoeconomics. 2009;27(10):861–72.
36. Neumann PJ, Sanders GD, Russell L, Siegel J, Ganiats T. Cost-effectiveness in health and medicine. Oxford University Press. 2015.
37. Nimdet K, Chaiyakunapruk N, Vichansavakul K, Ngorsuraches S. A systematic review of studies eliciting willingness-to-pay per quality-adjusted life year: does it justify CE threshold? PLoS ONE. 2015;10(4): e0122760.
38. Nimmesgern E, Norstedt I, Draghia-Akli R. Enabling personalized medicine in Europe by the European Commission's funding activities. Pers Med. 2017;14(4):355–65.
39. Oosterhoff M, van der Maas ME, Steuten LM. A systematic review of health economic evaluations of diagnostic biomarkers. Appl Health Econ Health Policy. 2016;14(1):51–65.
40. Pauker SG, Kaussirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980;302(20):1109–17.
41. Pettitt DA, Raza S, Naughton B, Roscoe A, Ramakrishnan A, Ali A, Davies B, Dopson S, Hollander G, Smith JA, Brindley DA. The limitations of QALY: a literature review. J Stem Cell Res Ther. 2016;2016(6):4.
42. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. Med Decis Making. 1988;8(4):279–89.
43. Phillips KA, Ann Sakowski J, Trosman J, Douglas MP, Liang SY, Neumann P. The economic value of personalized medicine tests: what we know and what we need to know. Genet Med. 2014;16:251–7.
44. Pichon-Riviere A, Drummond M, Palacios A, Garcia-Marti S, Augustovski F. Determining the efficiency path to universal health coverage: cost-effectiveness thresholds for 174 countries based on growth in life expectancy and health expenditures. Lancet Glob Health. 2023;11(6):833–42.
45. Rand LZ, Kesselheim AS. Controversy over using quality-adjusted life-years in cost-effectiveness analyses: a systematic literature review. Health Aff. 2021;40(9):1402–10.
46. Rautenberg T, Gerritsen A, Downes M. Health economic decision tree models of diagnostics for dummies: a pictorial primer. Diagnostics. 2020;10(3):158.
47. Redekop W, Mladsi D. The faces of personalized medicine: a framework for understanding its meaning and scope. Value Health. 2013;16(6):S4–9.
48. 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
49. Roberts MS. The next chapter in cost-effectiveness analysis. J Am Med Assoc. 2016;316(10):1049–50.
50. Ryen L, Svensson M. The willingness to pay for a quality adjusted life year: a review of the empirical literature. Health Econ. 2015;24(10):1289–301.
51. Sanghera S, Orlando R, Roberts T. Economic evaluations and diagnostic testing: an illustrative case study approach. Int J Technol Assess Health Care. 2013;29(1):53–60.
52. Sevim D, Felder S. Decision thresholds for medical tests under ambiguity aversion. Front Health Serv. 2022;2: 825315.
53. Siontis KC, Siontis GC, Contopoulos-Ioannidis DG, Ioannidis JP. Diagnostic tests often fail to lead to changes in patient outcomes. J Clin Epidemiol. 2014;67(6):612–21.
54. Schneider P. The QALY is ableist: on the unethical implications of health states worse than dead. Qual Life Res. 2022;31:1545–52.
55. Snowsill T. Modelling the cost-effectiveness of diagnostic tests. Pharmacoeconomics. 2023;41:339–51.
56. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. Med Decis Mak. 2008;28(5):650–67.
57. Téhard B, Detournay B, Borget I, Roze S, De Pouvourville G. Value of a QALY for France: a new approach to propose acceptable reference values. Value Health. 2020;23(8):985–93.
58. Trusheim MR, Berndt ER, Douglas F. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. Nat Rev Drug Discov. 2007;6(4):287–93.
59. United Nations. United Nations the General Assembly. Resolution adopted by the General Assembly on 25 September 2015. 2015.
60. Vallejo-Torres L, García-Lorenzo B, Castilla I, Valcárcel-Nazco C, García-Pérez L, Linertová R, Polentinos-Castro E, Serrano-Aguilar P. On the estimation of the cost-effectiveness threshold: why, what, how? Value Health. 2016;19(5):558–66.
61. Van der Pol S, Garcia PR, Postma MJ, Villar FA, van Asselt ADI. Economic analyses of respiratory tract infection diagnostics: a systematic review. Pharmacoeconomics. 2021;39:1411–27.
62. Vellekoop H, et al. The net benefit of personalized medicine: a systematic literature review and regression analysis. Value in Health. 2022;25(8):1428–38.
63. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. N Engl J Med. 1977;296(13):716–21.
64. World Health Organization. Increasing access to diagnostics through technology transfer and local production. 2011. Available at https://www.who.int/publications/i/item/9789241502375
65. World Health Organization. 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. 2021. Available from: https://iris.who.int/bitstream/handle/10665/339064/9789240019102-eng.pdf?sequence=1
66. Yang Y, Abel L, Buchanan J, Fanshawe T, Shinkins B. Use of decision modelling in economic evaluations of diagnostic tests: an appraisal and review of health technology assessments in the UK. Pharmacoecon Open 2019;3(3):281–91.