Economic Evaluation of Diagnostic Tests in the Absence of a Gold Standard: Addressing Reference Standard Bias

Carlos Felipe Balmaceda Araque, MSc

MPhil

University of York

Department of Health Sciences

April 2025

1 Abstract

Health decision-making must balance clinical benefits, costs, and equity in an environment where diagnostic tests are pivotal yet often lack a definitive reference standard. This thesis explores how imperfections in reference standards bias economic evaluations of diagnostic tests and potentially mislead decision-makers. First, it reviews statistical methods designed to mitigate errors arising from an imperfect reference, highlighting approaches such as latent class analysis and adjustment equations. Next, it examines how these methods have (or have not) been integrated into economic evaluations, underscoring gaps between methodological advances and applied practice.

A conceptual framework is then presented, illustrating how misclassified test accuracy parameters can alter key metrics like incremental cost-effectiveness ratios and net health benefits, resulting in suboptimal decisions. To quantify the magnitude of such bias, simulation studies are conducted across diverse scenarios, varying disease prevalence and the degree of reference standard imperfection. These simulations reveal that modest misclassification in the reference standard can sway conclusions about a test's value, especially in borderline cost-effectiveness contexts.

Finally, the thesis synthesises these findings into pragmatic recommendations for decision-analytic modelling of diagnostics when test accuracy evidence is generated using an imperfect reference standard. Rather than proposing universal adjustment in all health technology assessments, it argues for a proportionate, decision-focused approach: transparent reporting of reference standard limitations, structured sensitivity and scenario analyses, and targeted quantitative adjustments when decision uncertainty is high and credible information on reference standard performance is available. However, the utility of such quantitative adjustments depends critically on whether the assumption of conditional independence between test and reference standard errors can be defended in the specific clinical context; when this assumption is violated, the correction methods themselves may introduce bias. By demonstrating when and how reference standard misclassification can alter economic conclusions, this work aims to support more reliable analyses and resource allocation decisions.

Table of Contents

1 ABSTRACT ..... - 2 -

2 TABLE OF FIGURES ..... - 5 -

3 LIST OF TABLES ..... - 7 -

4 ACKNOWLEDGMENTS ..... - 9 -

5 AUTHOR'S DECLARATION ..... - 10 -

1 DECISION-MAKING IN HEALTHCARE: ADDRESSING CHALLENGES WITH EVIDENCE ..... - 11 -

1.1 Introduction to the Decision-Making Context ..... - 11 -

1.1.1 Health Technology Assessment : A Multidisciplinary Approach ..... - 13 -

1.1.2 Economic Evaluations: Frameworks and Role in HTA Processes ..... - 14 -

1.1.3 Cost-Effectiveness Analysis in Healthcare Decision-Making ..... - 17 -

1.2 Aim of this thesis ..... - 19 -

1.3 Structure of the thesis ..... - 21 -

2 INTRODUCTION TO DIAGNOSTIC TEST RESEARCH ..... - 22 -

2.1 Roles and functions of diagnostic tests ..... - 22 -

2.2 Measures of diagnostic test performance ..... - 25 -

2.2.1 Sensitivity and Specificity ..... - 25 -

2.2.2 Predictive Values ..... - 26 -

2.2.3 Likelihood Ratios ..... - 27 -

2.2.4 Diagnostic Odds Ratio (DOR) ..... - 28 -

2.2.5 Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) ..... - 29 -

2.2.6 Model Calibration and Advanced Performance Measures ..... - 31 -

2.3 Study Design for Evaluating Diagnostic Tests ..... - 33 -

2.4 Test-treatment trials, end-to-end evidence, and linked evidence modelling ..... - 37 -

2.5 Challenges in Diagnostic Test Studies ..... - 38 -

2.5.1 Spectrum Bias and Generalisability ..... - 39 -

2.5.2 Verification Bias ..... - 39 -

2.5.3 Imperfect Reference Standards ..... - 40 -

2.5.4 Observer Variability and Inter-Rater Reliability ..... - 41 -

2.5.5 External Validity and Reproducibility Issues ..... - 41 -

2.6 Reference standards and imperfect reference standards ..... - 43 -

2.7 Discussion ..... - 47 -

3 FROM STATISTICAL METHODS TO ECONOMIC EVALUATIONS: A LITERATURE REVIEW ON ADDRESSING IMPERFECT REFERENCE STANDARD BIAS ..... - 49 -

3.1 THE STATISTICAL CHALLENGE: ADDRESSING IMPERFECT REFERENCE STANDARDS ..... - 51 -
3.1.1 Methods for Imperfect or Missing Reference Standards ..... *- 52 -*
3.2 THE ECONOMIC EVALUATION PERSPECTIVE ..... - 69 -
3.2.1 Imperfect Reference Standards and Bias: Consequences for Cost-Effectiveness ..... *- 70 -*
3.2.2 Integrating Adjusted Accuracy into Economic Evaluations ..... *- 71 -*
3.2.3 Literature on Economic Frameworks and Bias ..... *- 73 -*
3.2.4 Critical Analysis of the Methods-Policy Gap ..... *- 75 -*
3.2.5 Future Directions: Addressing and Reducing the Gap ..... *- 78 -*
3.3 DISCUSSION ..... - 78 -

4 IMPACT OF IMPERFECT REFERENCE STANDARD BIAS ON DECISION-MAKING IN HEALTHCARE: INSIGHTS FOR ECONOMIC EVALUATIONS WITH SINGLE INDEX TEST RELATED TO AN IMPERFECT REFERENCE STANDARD WITH KNOWN ACCURACY ..... - 83 -

4.1 INTRODUCTION ..... - 83 -
4.2 RATIONALE AND NEED FOR THIS FRAMEWORK ..... - 86 -
4.3 UNDERSTANDING THE IMPERFECT REFERENCE STANDARD BIAS: CONCEPTUAL DEFINITIONS AND MECHANISM OF THE BIAS. - 87 -
4.3.1 Implications for economic evaluation ..... *- 88 -*
4.4 CORRECTION METHODS TO ADJUST SINGLE DICHOTOMOUS DIAGNOSTIC TEST WITH AN IMPERFECT REFERENCE STANDARD WITH KNOWN SENSITIVITY AND SPECIFICITY. ..... - 89 -
4.4.1 Limitations of correction methods in economic evaluations ..... *- 92 -*
4.4.2 Applicability of correction methods in cost-effectiveness analysis ..... *- 92 -*
4.5 COST-EFFECTIVENESS ANALYSIS OF DIAGNOSTIC TESTS. ..... - 93 -
4.6 UNALIGNED (MISMATCHED) COST-EFFECTIVENESS MODELS ..... - 94 -
4.6.1 Theoretical foundation ..... *- 100 -*
4.7 THE RELEVANCE OF ADJUSTING METHODS FOR DIAGNOSTIC TESTS IN HEALTHCARE DECISION-MAKING ... - 103 -
4.8 DISCUSSION ..... - 105 -

5 A SIMULATION-BASED EVALUATION OF IMPERFECT REFERENCE STANDARD BIAS IN THE ECONOMIC ASSESSMENT OF DIAGNOSTIC TESTS. ..... - 112 -

5.1 INTRODUCTION ..... - 112 -
5.2 METHODS ..... - 114 -
5.2.1 Selection criteria of correction methods ..... *- 114 -*
5.3 RESULTS ..... - 128 -
5.3.1 Scenarios where conditional independence (CI) assumption holds. ..... *- 128 -*
5.4 DISCUSSION ..... - 144 -

6 DISCUSSION ..... - 157 -

6.1 KEY FINDINGS AND THEORETICAL CONTRIBUTIONS ..... - 157 -

6.2 Findings in the Context of Existing Literature ..... - 159 -
6.3 Implications for Health Policy and HTA Practice ..... - 164 -
6.4 Limitations of the Research ..... - 167 -
6.5 Recommendations for Future Research ..... - 174 -
6.6 Conclusions ..... - 179 -
7 REFERENCES ..... - 182 -
8 APPENDIX ..... - 190 -
1. Staquet et al. (1981) ..... - 190 -
2. Brenner (1996) ..... - 191 -
3. Equations derivate from Brenner (1996) to estimate corrected accuracy. ..... - 192 -
4. Derivation for sensitivity parameter ..... - 192 -
4.1.1. Derivation for specificity parameter ..... - 193 -
8.1 Conditions that lead to illogical results using the Staquet approach. ..... - 194 -
8.2 Conditions that lead to illogical results using the Brenner for conditional dependence approach. ..... - 194 -

2 Table of Figures

Figure 1. Visual representation of likelihood ratios varying the pre- and post-test probability ..... - 28 -
Figure 2. ROC curves for diagnostic tests ..... - 30 -
Figure 3. Example of a calibration plot ..... - 32 -
Figure 4. Flow diagram illustrating the design of a cross-sectional diagnostic accuracy study. ..... - 34 -
Figure 5. Flow diagram outlining the design of a case-control diagnostic accuracy study. ..... - 35 -
Figure 6. Relationship between disease prevalence on predictive values. ..... - 36 -
Figure 7. Decision flowchart for selecting adjustment methods based on case requirements. ..... - 52 -
Figure 8. Stylized decision tree model for cost-effectiveness analysis of a new diagnostic test. The decision model using parameters conditional to the **true-disease status.** - 97 -

Figure 9. Stylized decision tree model for cost-effectiveness analysis of a new diagnostic test. The decision model where parameters are conditional to the **iRS.** .....- 98 -
Figure 10. Example of two unaligned model between probabilities and outcomes. A) unaligned model between the probabilities of having a test result conditioned to the true-disease status and outcomes conditioned to the iRS. B) unaligned model between the probabilities of having a test result conditioned to iRS and outcomes conditioned to the true-disease status. .....- 99 -
Figure 11. Decision model illustrating the role of adjusting method changing from an unaligned model to an aligned model. In red are represented the probabilities from the unaligned model, and in green the adjusted probabilities for the aligned model. ...- 100 -
Figure 12. Decision tree model to quantify the NHB gains of a testing strategy against no testing strategies.....- 123 -
Figure 13. Prevalence vs Test accuracy bias. A) Se and Sp bias for unadjusted estimates. B) Se and Sp bias for corrected estimates using Staquet approach. .....- 129 -
Figure 14. Prevalence vs weighted accuracy bias. A) unadjusted estimates B) corrected estimates using Staquet approach. .....- 130 -
Figure 15. Prevalence vs iNHB bias. A) unadjusted estimates for test 1, B) unadjusted estimates for test 1, C) corrected estimates using Staquet approach for both test, and D) unadjusted estimates for test 1 considering that test 2 is not biased. .....- 132 -
Figure 16. Relationship between the bias of the test accuracy and the bias in the iNHB. - 132 -
Figure 17. True iNHB (Equivalence line in red) vs Unadjusted iNHB (Green dots) and corrected iNHB (Staquet – Blue dots).....- 133 -
Figure 18. Bias in the Value of diagnostic (VoD, showed in the upper figure) and in the value of heterogeneity (VoH, showed in the bottom figure). .....- 134 -
Figure 19. Weighted accuracy bias versus prevalence when CI assumption is relaxed. Figure A) illustrate the estimations for the unadjusted. Figure B) Corrected estimation using Staquet approach. Figure C) Corrected estimation using Brenner approach. The illogical scenarios are showed in grey in plots B) and C). .....- 137 -
Figure 20. Prevalence vs Test accuracy bias. A) Se bias for unadjusted estimates, B) Sp bias for unadjusted estimates. C) Se bias for corrected estimates using Staquet approach, D) Sp bias for corrected estimates using Staquet approach, E) Se bias for

corrected estimates using Brenner CD approach and F) Sp bias for corrected estimates using Brenner CD approach. .....- 138 -
Figure 21. Bias iNHB versus Prevalence. A) Unadjusted iNHB for test 1. B) Unadjusted iNHB for test 2. C) Corrected iNHB using Staquet approach for both test 1 and test 2. D Corrected iNHB using Brenner CD approach for both test 1 and test 2. E) Unadjusted iNHB for test 1 considering that test 2 is not biased.....- 139 -
Figure 22. Relationship between weighted test accuracy bias with iNHB bias. A) For unadjusted estimates. B) Corrected estimates using Staquet. C) Corrected estimates using Brenner CD. .....- 140 -
Figure 23. True iNHB vs Unadjusted iNHB. Figure A) illustrate the estimations for the unadjusted. Figure B) Corrected estimation using Staquet approach. Figure C) Corrected estimation using Brenner approach. In grey is showed the scenarios where correction methods did not reverse the decision. .....- 142 -
Figure 24. Bias in the Value of diagnostic (VoD, showed in the upper figure) and in the value of heterogeneity (VoH, showed in the bottom figure). .....- 143 -

3 List of Tables

Table 1. Key differences between analytical validity and clinical utility .....- 23 -
Table 2. Overview of the Various Functions of Diagnostic Tests.....- 24 -
Table 3. 2x2 contigency table .....- 26 -
Table 4. Summary of Key Diagnostic Performance Metrics.....- 32 -
Table 5. Overview of Two Diagnostic Accuracy Study Designs with Their Descriptions, Strengths, and Limitations .....- 36 -
Table 6. Summary of Key Challenges in Diagnostic Test Studies and Their Implications .- 43 -
Table 7. Summary of the correction methods selected. .....- 91 -
Table 8. 2x2 table to describe the NHB under different circumstances. .....- 118 -
Table 10. Severity modifier and equivalence with both proportional and absolute QALY shortfall.....- 120 -

Table 10. Description of notation for each parameter and estimates involved in the simulation study. .....- 123 -
Table 12. 2x2 table denoting the classification of individuals according to their test results.....- 190 -

4 Acknowledgments

This thesis was conducted from 2021 to 2025 at the Centre for Health Economics (CHE) of the University of York, within the Team for Economic Evaluation and Health Technology Assessment. This work was partially supported by a CHE scholarship. I am also grateful for the additional financial assistance provided over these years.

I extend my heartfelt thanks to my Chilean colleagues and friends, Manuel Espinoza, Nicolás Armijo, Tomás Abbott, and Macarena Vera, who offered their guidance and encouragement despite the distance.

My sincere gratitude goes to my friends and colleagues at the Centre for Health Economics – Adrián Villaseñor, Carlos Rojas, Carlos Chivardi, Iván Ochoa, and Alfredo Palacios – the "Latin-CHE" – who stood by me throughout my time in York, providing friendship, professional advice, and support at every stage of my work.

I would also like to thank my parents, Carlos and Uganda, and my siblings, Rafael and Macarena. A significant part of this thesis has been possible because of you.

Finally, to my beloved wife, Daniela, and our son, Ignacio: without you, this thesis would never have been completed. Your unwavering support, patience, and understanding have sustained me through both the bright moments and the darker times these past years. Thank you for your love, for grounding me, and for always being there.

5 Author's Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references. In addition, any views expressed in this document are exclusive responsibility of the author.

I hereby give my authorization for my thesis, if accepted, to be used for photocopying and inter-library loan. Likewise, I give my consent for the title and the abstract to be made available in all academic dissemination sources, making reference to authorship and copyright.

Chapter 1

1 Decision-Making in Healthcare: Addressing Challenges with Evidence

1.1 Introduction to the Decision-Making Context

Decision-making in healthcare is a multifaceted process that involves balancing competing priorities, including clinical outcomes, cost-effectiveness, equity, and social values, amongst others. Building on this, policymakers, clinicians, and stakeholders are tasked with allocating scarce resources to maximize societal health outcomes. This complexity is further compounded by the dynamic nature of healthcare, characterized by rapid advancements in medical technologies and shifting population health needs (Dolan and Olsen, 2002, Sanders et al., 2016). This raises the question of how to make decisions about the proper allocation of funds.

In contemporary healthcare systems, decision-making extends beyond individual patient care to encompass broader considerations such as public health policies, resource allocation frameworks, and long-term system sustainability. The stakes are high, as decisions often determine who gains access to life-saving interventions and how health outcomes are distributed across populations. This necessitates the use of structured frameworks that integrate diverse forms of evidence to ensure informed and equitable decisions (Drummond et al., 2015).

Healthcare decision-making is uniquely complex due to the specific challenges faced by this field. Among the most pressing is resource scarcity. Even in high-income countries, healthcare budgets are finite, necessitating difficult choices about which treatments and services to prioritize. Rising costs, driven by factors such as aging populations and expensive technological innovations, further strain resources. For example, In the UK, NICE has approved CAR-T therapies such as tisagenlecleucel under the Cancer Drugs Fund, acknowledging their high upfront costs and the need for further data collection to confirm cost-effectiveness. Allocating significant resources to such cutting-edge

therapies necessitates trade-offs, potentially diverting funds from public health programmes or other treatments (National Institute for Health and Care Excellence (NICE), 2020). For instance, allocating funds to cutting-edge therapies may require forgoing investments in broader public health initiatives, creating significant opportunity costs (Drummond et al., 2015). Policymakers must navigate these trade-offs to balance efficiency and equity (Claxton et al., 2002, Sanders et al., 2016).

Evidence gaps add another persistent challenge. Even rigorously designed clinical trials may not provide complete insights into the long-term effectiveness, safety, or real-world applicability of interventions. Emerging technologies, such as novel diagnostic tools or precision medicine approaches, often present compelling theoretical benefits but lack robust post-market evidence. This uncertainty can hinder informed decision-making, leading to delays in adoption or premature investments in unproven technologies (Bossuyt et al., 2003).

Moreover, the multiplicity of stakeholders in healthcare introduces diverse and often conflicting priorities. Patients typically prioritize immediate access to care, focusing on quality and outcomes. Clinicians may emphasize the clinical efficacy and safety of interventions, while health economists and policymakers often prioritize cost-effectiveness and budgetary constraints. Reconciling these differing objectives is challenging but critical to achieving consensus on resource allocation (Dolan and Olsen, 2002).

Rapid advancements in medical technologies, including artificial intelligence, genomics, and telemedicine, promise transformative improvements in care delivery. However, these innovations also introduce complexity into decision-making processes. High initial costs, uncertain long-term benefits, and the need for specialized infrastructure or training can make it difficult to evaluate and implement these technologies effectively. Policymakers must assess not only the immediate value of these innovations but also their broader systemic implications (Sanders et al., 2016, Drummond et al., 2015).

To navigate these complexities, decision-making frameworks must evolve to incorporate multidisciplinary perspectives, integrate evidence from diverse sources, and adapt to emerging challenges. This transition requires robust approaches, such as Health Technology Assessment (HTA), which offers a structured and comprehensive means to

evaluate healthcare interventions within the broader context of system sustainability and equity.

1.1.1 Health Technology Assessment : A Multidisciplinary Approach

HTA is a systematic process that evaluates the medical, social, economic, and ethical implications of healthcare interventions (Drummond et al., 2015). Its primary objective is to provide evidence to inform healthcare decision-making, ensuring that resources are allocated efficiently and equitably. HTA bridges the gap between evidence generation and policy implementation by synthesizing data on clinical efficacy, cost-effectiveness, and societal values.

HTA employs a multidisciplinary approach, integrating insights from clinical medicine, health economics, epidemiology, and ethics. This integration ensures that the assessment captures the diverse dimensions of healthcare interventions, from their direct impact on patient outcomes to their broader implications for public health systems. For example, HTA evaluates not only the clinical effectiveness of a new cancer therapy but also its affordability, potential to reduce health disparities, and feasibility within existing healthcare infrastructures (Claxton et al., 2002).

Uncertainty is a central challenge in healthcare decision-making, and HTA plays a pivotal role in mitigating it. Tools such as probabilistic sensitivity analysis and value of information (VOI) methods are integrated into HTA to assess the potential benefits of reducing uncertainty. These approaches guide policymakers on whether additional research is warranted before adopting a technology, ensuring that decisions are robust and informed by the best available evidence (Claxton et al., 2002).

One innovative application of HTA is the use of Managed Entry Agreements (MEAs). These agreements allow the conditional adoption of new technologies, where reimbursement is linked to the collection of further evidence on real-world performance. MEAs are particularly valuable for high-cost interventions with limited initial evidence, such as precision medicine therapies. By integrating MEAs, HTA facilitates adaptive policy decisions that balance access, evidence development, and financial sustainability (Claxton and Sculpher, 2006).

Additionally, HTA increasingly incorporates Multi-Criteria Decision Analysis (MCDA) to address the diverse and often competing objectives in healthcare. MCDA enables the systematic inclusion of non-economic criteria, such as equity and patient preferences, alongside traditional cost-effectiveness metrics. This approach ensures that HTA evaluations reflect broader societal values and provide decision-makers with a comprehensive understanding of intervention impacts (Phillips and Bana e Costa, 2007). HTA is increasingly recognized as an essential tool for policy-making at both national and international levels. Organizations like the National Institute for Health and Care Excellence (NICE) in the UK and the European Network for Health Technology Assessment (EUnetHTA) have set benchmarks for the integration of HTA into healthcare systems. These organizations emphasize transparency, stakeholder engagement, and the use of standardized methodologies to ensure that HTA findings are both credible and actionable (National Institute for Health and Care Excellence (NICE), 2020).

As healthcare evolves, HTA must adapt to address emerging challenges, such as the rapid pace of innovation in personalized medicine and digital health. Enhancing the integration of societal values and patient preferences into HTA methodologies will be crucial in maintaining its relevance. These adaptations will reinforce the critical role of HTA in shaping evidence-based, equitable, and sustainable healthcare systems. The structured evaluations provided by HTA form the foundation for economic evaluations in healthcare, enabling a deeper exploration of the cost-effectiveness and value of medical interventions.

1.1.2 Economic Evaluations: Frameworks and Role in HTA Processes

Economic evaluation of healthcare is focused on determining the costs and health outcomes of healthcare interventions to inform whether their introduction will lead to improvements in population health (Drummond et al., 2015). In an economic evaluation, it is essential to undertake "the comparative analysis of alternative courses of action" (Drummond et al., 2015), where underlying choices between two or more alternatives are considered. This analytical framework enables policymakers to assess whether a given intervention provides sufficient value relative to other available options.

The theoretical foundations of economic evaluation are deeply rooted in two complementary frameworks: welfarism and extra-welfarism. Welfarism is grounded in classical utility theory, where individual utility serves as the basis for evaluating interventions. This framework emphasizes that individual preferences, typically expressed as willingness to pay, should guide decisions. Cost-Benefit Analysis (CBA) aligns closely with welfarism by monetizing both costs and benefits, enabling comparisons across sectors, including non-health investments. Despite its appeal for cross-sectoral decision-making, welfarism faces significant criticism in healthcare contexts. The monetization of health outcomes, such as life quality improvements, often neglects equity concerns and ethical dimensions, where societal perspectives play a pivotal role (Arrow, 1965, Culyer, 1989).

Extra-welfarism, on the other hand, departs from individual utility maximization to focus on health outcomes as the primary goal of evaluation. This approach is central to healthcare systems aiming to equitably improve population health. Rather than monetary valuations, extra-welfarist methods use health-specific metrics such as life-years gained or Quality-Adjusted Life Years (QALYs). This framework underpins Cost-Effectiveness Analysis (CEA) and Cost-Utility Analysis (CUA), methods designed to prioritize interventions based on their ability to deliver maximum health benefits relative to their costs (Culyer, 2014).

The comparison between these frameworks highlights both synergies and trade-offs. Welfarism offers a broad lens for societal-level resource allocation, making it particularly relevant in governmental budget decisions. However, its heavy reliance on monetary valuation makes it less effective for capturing the intrinsic value of health improvements. Extra-welfarism, by contrast, focuses narrowly on health outcomes, aligning closely with the goals of HTA. Recent advances in economic evaluations have sought to integrate elements of both frameworks, fostering hybrid approaches that balance efficiency with equity and societal preferences.

Economic evaluations in practice employ several methodologies, each tailored to specific decision-making needs. CEA compares interventions by calculating the cost per unit of health outcome gained, such as life-years saved. It is particularly effective for comparing interventions with similar objectives, such as competing cancer treatments. Cost-Utility Analysis (CUA), a refinement of CEA, incorporates QALYs to account for both

the quality and quantity of life. This standardization facilitates comparisons across diseases and interventions, making CUA indispensable in HTA processes (Torrance, 1986, Gold et al., 1996).

Cost-Benefit Analysis (CBA) provides a broader evaluative scope by translating both costs and outcomes into monetary terms. This allows decision-makers to directly compare healthcare investments with non-health alternatives, such as infrastructure or education. However, its reliance on monetary valuation poses challenges in capturing intangible benefits like improved mental health or reduced caregiver burden. Despite these limitations, CBA remains valuable in cross-sectoral resource allocation discussions.

Cost-Minimization Analysis (CMA) is a method reserved for situations where interventions demonstrate equivalent clinical effectiveness. It focuses solely on identifying the least costly option, making it an efficient tool for decisions in areas such as generic drug adoption. Although CMA has declined in prominence due to evolving evaluation standards, it remains relevant for specific applications (Briggs and O'Brien, 2001).

The integration of economic evaluations into HTA processes significantly enhances their decision-making utility. Decision-analytic models, such as Markov models and discrete-event simulations, allow analysts to project long-term costs and outcomes. These models accommodate complexities such as disease progression and treatment adherence, providing a dynamic view of intervention impacts. Probabilistic sensitivity analysis (PSA) further strengthens these models by quantifying uncertainties, ensuring that policymakers can make robust and informed decisions (Briggs, 2006). Additionally, Value of Information (VOI) analysis plays a critical role in identifying areas where further evidence is most valuable, guiding research priorities to reduce uncertainty in future evaluations (Claxton et al., 2002).

Modern economic evaluations increasingly address societal values, incorporating considerations like equity and access into decision-making frameworks. Multi-Criteria Decision Analysis (MCDA) has emerged as a complementary approach, enabling policymakers to integrate non-economic factors, such as ethical considerations and patient preferences, alongside traditional cost-effectiveness metrics. This holistic

perspective ensures that resource allocation decisions align with broader public health goals and societal priorities (Phillips and Bana e Costa, 2007).

In summary, economic evaluations are central to the success of HTA, bridging clinical evidence with policy implementation. By systematically analyzing costs, outcomes, and societal preferences, these evaluations ensure that healthcare interventions are both effective and equitable. As healthcare systems grapple with rising costs and rapid technological advancements, economic evaluations will remain indispensable in achieving sustainable, evidence-based resource allocation.

The next section will explore CEA in Decision-Making, focusing on how this methodology shapes healthcare policies and guides intervention prioritization.

1.1.3 Cost-Effectiveness Analysis in Healthcare Decision-Making

CEA is a foundational methodology in health economic evaluations, developed to assess the efficiency of healthcare interventions. CEA has become indispensable in healthcare decision-making. It provides a structured framework to compare the costs and health outcomes of interventions, helping policymakers prioritise resource allocation. CEA is particularly valuable in contexts of resource scarcity, where efficiency and impact are paramount.

The core metric of CEA is the incremental cost-effectiveness ratio (ICER), which quantifies the additional cost required to achieve one additional unit of health gain, such as life-years saved or disease cases averted. The ICER is calculated as follows:

$$ICER = \frac{C_1 - C_0}{E_1 - E_0}$$

where C_1 and C_0 represent the costs of the new and comparator interventions, respectively, and E_1 and E_0 represent their corresponding health outcomes. The ICER is then compared against a cost-effectiveness threshold, which reflects the maximum amount a society is willing to pay for a unit of health gain. For example, NICE in the UK uses a threshold of £20,000–30,000 per QALY gained as a benchmark for funding decisions (Claxton et al., 2015). However, the application of thresholds varies globally,

with some countries adopting GDP-based thresholds as recommended by the WHO, while others develop context-specific benchmarks (Bertram et al., 2016).

Importantly, in many publicly funded health systems the cost-effectiveness threshold is increasingly understood to represent the opportunity cost of health expenditure rather than a normative willingness-to-pay threshold: the health that is displaced elsewhere in the system when resources are committed to a new intervention. This opportunity-cost interpretation is particularly relevant for diagnostics, where incremental net health benefits may be modest and sensitive to modelling assumptions.

An alternative approach to interpreting CEA results is through the concept of net health benefits (NHB), which simplifies decision-making by expressing outcomes in health units rather than monetary terms. The NHB is calculated as:

$$NHB = \lambda(E_1 - E_0) - (C_1 - C_0)$$

Where λ represents the cost-effectiveness threshold. A positive NHB indicates that the intervention provides greater value than the comparator, while a negative NHB suggests it is not cost-effective. The NHB framework is increasingly used to enhance the transparency and interpretability of CEA results.

CEA's practical utility lies in its ability to rank interventions by value. For instance, analyses of vaccination programmes have demonstrated that they deliver substantial health gains at a fraction of the cost of alternative interventions. In high-income countries like the UK, vaccines often fall well within NICE's cost-effectiveness thresholds, facilitating their rapid adoption. In contrast, low- and middle-income countries face affordability constraints that necessitate adjustments to thresholds or external funding support, as highlighted by WHO guidelines (Bertram et al., 2016).

Another critical aspect of CEA is the time horizon over which costs and outcomes are assessed. Short time horizons may underestimate the cost-effectiveness of interventions with long-term benefits, such as preventive measures or chronic disease management. Decision-analytic models, such as Markov models and discrete-event simulations, address this by projecting long-term outcomes and accounting for complexities like disease progression and treatment adherence. For example, Markov models have been extensively used in smoking cessation programmes to capture

lifetime benefits in reduced healthcare costs and improved health outcomes. These models enhance the robustness of CEA by incorporating the full trajectory of intervention impacts (Briggs, 2006).

Discounting is another key methodological component of CEA, reflecting the reduced value of future costs and benefits. The present value (PV) of a future value (FV) is calculated as:

$$PV = \frac{FV}{(1+r)^t}$$

Where r is the discount rate and t is the time in years. Commonly, discount rates of 3–3.5% are applied in health economic evaluations to standardise the valuation of long-term outcomes.

The integration of QALYs into CEA through Cost-Utility Analysis (CUA) has further advanced its relevance. QALYs combine measures of health quantity (life-years gained) with quality (health-related well-being), enabling comparisons across diseases and interventions. For example, in evaluating cancer therapies, QALYs capture both survival and quality-of-life improvements, ensuring comprehensive assessments of their value (Torrance, 1986, Gold et al., 1996).

As CEA continues to evolve, it remains integral to healthcare policy and planning. Its capacity to identify value-driven interventions while accommodating the constraints of finite resources ensures its relevance in addressing the challenges of modern healthcare systems. Furthermore, the methodology's adaptability to incorporate societal values and long-term evidence strengthens its utility in shaping future healthcare landscapes.

1.2 Aim of this thesis

The main aim of this thesis is:

In assessing the expected value of introducing a new test into routine practice, and where accuracy evidence for the new test is grounded on an imperfect reference standard (iRS), we aim to explore the potential for bias in population health, identify available methods to manage bias and improve our understanding of the usefulness of bias adjustment methods and their potential impact on population health in support of decision making.

This research aims to provide insights to inform decision-making in adopting new tests in the presence of imperfect reference standards and to improve healthcare outcomes ultimately.

Aim 1: Current knowledge of addressing imperfect reference standard bias in economic evaluations.

To review and analyze statistical methods literature that address the issue of lacking a perfect standard for determining the diagnostic accuracy of an index test. Additionally, to identify and summarize the existing literature that discusses, evaluates, or provides guidelines for the economic evaluation of diagnostic tests, particularly focusing on the problem posed by imperfect reference standards.

Aim 2: Constructing a framework for addressing the imperfect reference standard bias on economic evaluations of diagnostic tests when there is a known reference standard.

To determine in what scenarios the imperfect reference standard bias has an impact on decision-making. By employing simulation studies, it aims to investigate when is important to use correction methods for the economic evaluations of diagnostic tests when their accuracy is evaluated against imperfect reference standards. This investigation includes examining the implications of these corrective methods on costs, outcomes, population health, and uncertainty.

This thesis focuses on how imperfect reference standards in diagnostic accuracy research may bias cost-effectiveness analyses of new diagnostic tests, and on exploring ways to identify and address such bias to improve healthcare decision-making. The overarching aim is to examine the scale and implications of this problem, and to propose and assess potential solutions.

To achieve this aim, the thesis is guided by four objectives. The first objective is to review and critically appraise the available evidence on statistical methods designed to handle situations in which the reference standard is imperfect or missing, and to examine how these methods have (or have not) been integrated into the frameworks used for economic evaluations of diagnostic technologies. The second objective is to develop a conceptual and economic framework that demonstrates how errors arising from an

imperfect reference standard may alter cost-effectiveness conclusions. The third objective is to conduct a simulation-based evaluation that quantifies the extent of such bias, compares different correction approaches, and highlights the conditions under which ignoring the bias might significantly affect healthcare decisions. The fourth and final objective is to synthesise these findings into practical recommendations for policy-makers, HTAbodies, and researchers who undertake economic evaluations of new diagnostic tests.

1.3 Structure of the thesis

This thesis is structured in six chapters. Chapter 1 presents the introduction, outlining the rationale, objectives, and scope of the research, as well as the significance of addressing iRS bias in the economic evaluation of diagnostic tests. Chapter 2 provides an introduction to diagnostic test research, offering essential background on the concepts of diagnostic accuracy, reference standards (RS), and the role of diagnostic technologies in health technology assessment. Chapter 3 presents a comprehensive literature review titled "From Statistical Methods to Economic Evaluations: A Literature Review on Addressing Imperfect Reference Standard Bias," which examines the existing methodological approaches used to correct for iRS bias and evaluates how these have been considered in economic evaluations. Chapter 4 introduces the conceptual framework underpinning this thesis, entitled "Impact of Imperfect Reference Standard Bias on Decision-making in Healthcare: Insights for Economic Evaluations with a Single Index Test Related to an Imperfect Reference Standard with Known Accuracy." This chapter establishes the analytical foundation for the empirical investigation by linking diagnostic accuracy bias to decision-making consequences in health economic models. Chapter 5 presents the simulation study, titled "A Simulation-Based Evaluation of Imperfect Reference Standard Bias in the Economic Assessment of Diagnostic Tests", which quantifies the impact of iRS bias on cost-effectiveness outcomes and evaluates the performance of selected correction methods across varying assumptions and data scenarios. Finally, Chapter 6 offers the overall discussion of the thesis, integrating insights from the preceding chapters, reflecting on methodological implications, identifying limitations, and outlining directions for future research.

Chapter 2

2 Introduction to Diagnostic Test Research

Diagnostic tests are fundamental tools in modern medicine, aiding clinicians in detecting, confirming, monitoring, and predicting diseases. They range from simple laboratory tests to complex imaging and molecular diagnostics. The ability to accurately diagnose conditions is crucial for guiding appropriate treatment strategies, improving patient outcomes, and optimising healthcare resources (Bossuyt et al., 2015, Pepe et al., 2001, Zhou et al., 2002).

The evaluation of diagnostic tests is a structured process requiring assessments of their accuracy, utility, and impact on clinical decision-making. Rigorous validation is essential to ensure that tests are not only analytically sound but also provide meaningful clinical insights. As the field of precision medicine evolves, diagnostic testing is increasingly being integrated into personalized healthcare approaches, enabling tailored therapeutic interventions based on individual patient characteristics (Easton et al., 2015, Bossuyt et al., 2003).

This chapter explores the various roles diagnostic tests play in healthcare, key measures of diagnostic performance, methodologies for evaluating test accuracy, study designs used in diagnostic research, and challenges inherent in diagnostic testing. Additionally, the limitations of traditional randomized controlled trials (RCTs) for evaluating diagnostic tests and the implications of lacking a definitive gold standard for certain conditions are discussed.

2.1 Roles and functions of diagnostic tests

Diagnostic tests are fundamental tools in modern medicine, aiding clinicians in detecting, confirming, monitoring, and predicting diseases. They range from simple laboratory tests to complex imaging and molecular diagnostics. The ability to accurately

diagnose conditions is crucial for guiding appropriate treatment strategies, improving patient outcomes, and optimizing healthcare resources

Accurate diagnostic testing is vital in reducing misdiagnoses and ensuring effective healthcare interventions. Poor diagnostic accuracy can lead to inappropriate treatment, increased healthcare costs, and negative patient outcomes (Deeks, 2001, Shaughnessy AF., 2007). It is therefore essential that the evaluation process of diagnostic tests considers both analytical validity and clinical utility, which together determine a test's overall effectiveness (Table 1).

Table 1. Key differences between analytical validity and clinical utility

Aspect	Analytical Validity	Clinical Utility
Definition	The ability of a test to measure what it intends to measure.	The impact of the test result on clinical decision-making and patient outcomes.
Key Components	Precision, accuracy, reproducibility.	Influence on treatment decisions, improvement in health outcomes.
Evaluation Methods	Laboratory-based quality assessments.	Observational studies, randomised controlled trials.

Diagnostic tests can be classified based on their function, ranging from early disease detection to monitoring treatment response.

Screening tests aim to detect disease in asymptomatic individuals at an early stage, allowing for timely intervention. These tests prioritise high sensitivity to minimise false negatives, ensuring that those at risk are identified promptly (Nelson et al., 2016). Examples include mammography for breast cancer screening, colonoscopy for detecting colorectal cancer (Lin et al., 2021), and newborn metabolic screening tests (Wilcken et al., 2003). Early detection through screening improves treatment outcomes and contributes to public health efforts in reducing disease burden.

Diagnostic tests are used to confirm or rule out the presence of a disease in symptomatic individuals or those with positive screening results. These tests balance both sensitivity (Se) and specificity (Sp) to provide an accurate diagnosis (Corman et al., 2020). Techniques such as polymerase chain reaction (PCR) for infectious diseases, histopathological biopsy examination for cancer diagnosis (Eraky et al., 2023), and

advanced imaging modalities like MRI and CT scans for neurological disorders (Eraky et al., 2023) are widely employed in clinical practice.

Prognostic tests provide insight into the likely course of a disease, helping clinicians predict disease progression and tailor treatment accordingly (Crawford et al., 2019). They are particularly important in oncology and cardiovascular diseases, where risk stratification influences therapeutic decisions. Genetic testing for hereditary cancers (Easton et al., 2015), troponin measurements in suspected cardiac events, and tumour markers such as prostate-specific antigen (PSA) in prostate cancer prognosis are key examples of prognostic applications.

Monitoring tests are designed to track disease progression and assess the effectiveness of treatment over time. They play a crucial role in chronic disease management, where ongoing assessment is necessary to adjust therapeutic strategies (Nerenberg et al., 2018). Haemoglobin A1c (HbA1c) testing is integral in diabetes management (Nathan et al., 2005), viral load measurements guide HIV treatment adjustments (Günthard et al., 2016), and continuous blood pressure monitoring is essential for controlling hypertension (Nerenberg et al., 2018). Effective monitoring improves long-term patient outcomes by enabling timely modifications to treatment plans.

Companion diagnostics are essential tools in precision medicine, facilitating personalised treatment approaches by identifying biomarkers that predict therapeutic response (Wolff et al., 2018). These tests ensure that specific treatments are only administered to patients who are most likely to benefit, thus improving efficacy and minimising adverse effects. Examples include HER2 testing for guiding targeted therapy in breast cancer (Sequist et al., 2013), epidermal growth factor receptor (EGFR) mutation analysis for selecting appropriate lung cancer treatments, and pharmacogenomic testing for optimising drug safety and efficacy (Relling and Evans, 2015).

Table 2. Overview of the Various Functions of Diagnostic Tests

Type of Test	Purpose	Examples
Screening	Early disease detection in asymptomatic individuals.	Mammography, colonoscopy, newborn screening.
Diagnostic	Confirmation of disease presence in symptomatic patients.	PCR for infections, biopsy, MRI scans.

Type of Test	Purpose	Examples
Prognostic	Prediction of disease progression and risk assessment.	Genetic testing, cardiac biomarkers.
Monitoring	Tracking disease status and treatment response.	HbA1c, viral load testing, blood pressure monitoring.
Companion Diagnostics	Guiding personalised treatment based on biomarkers.	HER2 testing, EGFR mutation analysis.

The application of these diagnostic tests ensures early detection, accurate disease confirmation, risk stratification, treatment monitoring, and the optimisation of therapeutic interventions. Their proper utilisation is fundamental to achieving improved patient outcomes and the efficient allocation of healthcare resources. Ongoing advancements in diagnostic research, including the integration of artificial intelligence and biomarker discovery, continue to enhance the accuracy and clinical impact of diagnostic tests.

2.2 Measures of diagnostic test performance

A rigorous evaluation of diagnostic tests is critical to ensuring that clinical decisions are founded on reliable and meaningful evidence. The performance of a diagnostic test is determined not only by its ability to correctly classify individuals as diseased or non-diseased but also by how its results influence clinical decision-making. In this section, we describe the key performance metrics—Se, Sp, predictive values, likelihood ratios, diagnostic odds ratio, and receiver operating characteristic (ROC) analysis. Detailed formulas are provided for each measure, and a summary table is included to facilitate comparison and interpretation. In addition, we discuss when and how to use graphical representations to complement numerical data.

2.2.1 Sensitivity and Specificity

Se and Sp are the foundational metrics in diagnostic accuracy studies. Se, also known as the true positive rate, quantifies the proportion of individuals with the disease who are correctly identified by the test. Sp, or the true negative rate, quantifies the proportion of healthy individuals who are correctly classified (Zhou et al., 2002).

Mathematically, Se is defined as:

$$Sensitivity = \frac{TP}{TP + FN}$$

and Sp is defined as:

$$Specificity = \frac{TN}{TN + FP}$$

Where TP is the true positives, TN: true negatives, FP: false positives, and FN: false negatives.

These measures are calculated from the 2×2 contingency table (Table 3) and are independent of disease prevalence. However, because they are prevalence-independent, they may not fully convey the test's performance in a specific clinical context.

Table 3. 2x2 contigency table

	Disease	**No Disease**	
Test Positive	TP	FP	Test positives: TP + FP
Test Negative	FN	TN	Test negatives: FN + TN
	TP + FN	FP + TN	N= TP + FP + FN + TN

TP: True positives, TN: True negatives, FP: False positives, FN: False negatives, N: Total population

2.2.2 Predictive Values

Predictive values, namely the positive predictive value (PPV) and negative predictive value (NPV), provide clinically relevant probabilities by taking disease prevalence into account. The PPV is the probability that a person with a positive test truly has the disease, whereas the NPV is the probability that a person with a negative test is indeed disease-free. Their formulas are:

$$PPV = \frac{TP}{TP + FP}$$

and

$$NPV = \frac{TN}{TN + FN}$$

Unlike Se and Sp, predictive values are influenced by the prevalence of the disease in the population. Hence, in populations with low disease prevalence, even highly accurate tests can yield a relatively low PPV.

2.2.3 Likelihood Ratios

Likelihood ratios (LRs) combine Se and Sp into a single measure that allows clinicians to update the pre-test probability of disease to a post-test probability using Bayes' theorem. There are two primary types:
The Positive Likelihood Ratio (LR+) is given by:

$$LR+= \frac{Sensitivity}{1 - Specificity}$$

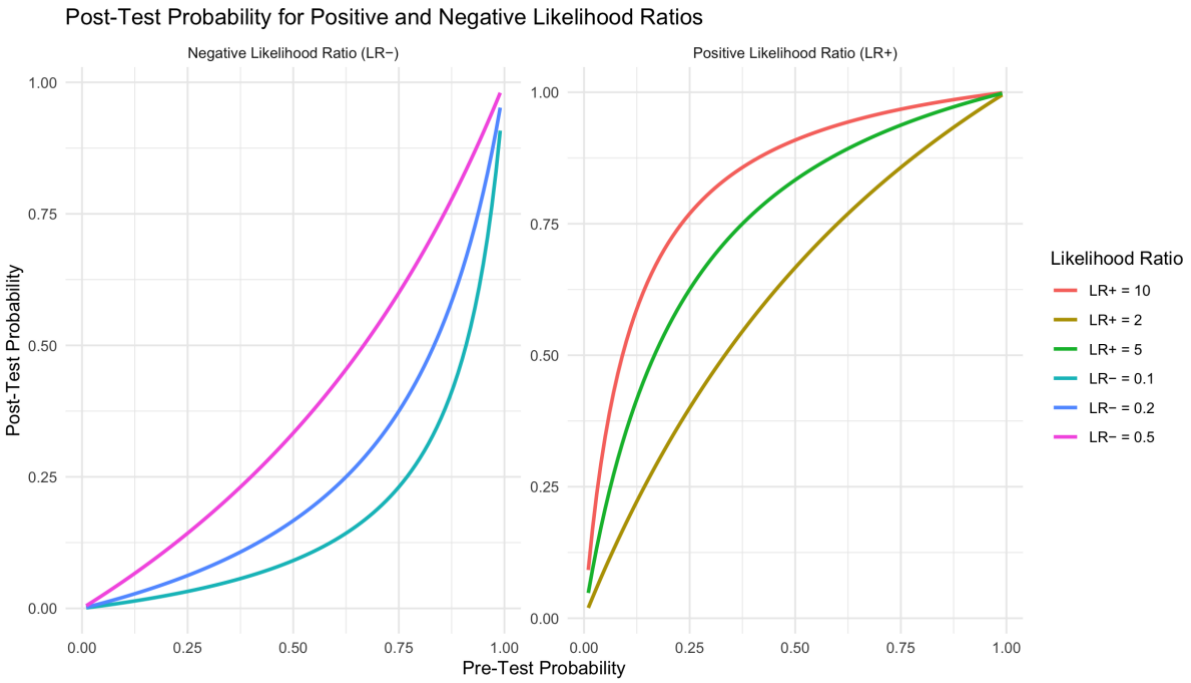
The Negative Likelihood Ratio (LR–) is calculated as:

$$LR-= \frac{1 - Sensitivity}{Specificity}$$

These ratios provide a direct indication of how much a positive or negative test result will alter the odds of having a disease. For instance, a high LR+ (typically above 10) indicates that a positive test result is very effective at ruling in disease, while a low LR– (typically below 0.1) suggests that a negative test result is effective at ruling out disease (Simel et al., 1991).
The Figure 1 illustrates how diagnostic test results modify the probability of disease based on different LR. The left panel shows how a **LR–** lowers the post-test probability, with smaller LR– values (e.g., 0.1) leading to a greater reduction in disease probability. The right panel shows how a **LR+** increases the post-test probability, with larger LR+ values (e.g., 10) resulting in a stronger confirmation of disease presence. These curves

help clinicians understand the real-world impact of test results in refining disease probability.

Figure 1. Visual representation of likelihood ratios varying the pre- and post-test probability



2.2.4 Diagnostic Odds Ratio (DOR)

The Diagnostic Odds Ratio (DOR) is a single summary statistic that combines the information of Se and Sp. It is defined by the formula:

$$DOR = \frac{TP \times TN}{FP \times FN}$$

or equivalently,

$$DOR = \frac{LR+}{LR-}$$

A higher DOR indicates better overall test performance. However, while the DOR provides a useful summary, it lacks the intuitive clinical interpretation of predictive values or likelihood ratios.

2.2.5 Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

The ROC curve is an essential tool for evaluating the performance of diagnostic tests over a range of threshold values. By plotting Se (true positive rate) against 1 – Sp (false positive rate), the ROC curve illustrates the trade-off between true positives and false positives. The Area Under the ROC Curve (AUC) quantifies the overall ability of the test to discriminate between those with and without the disease (Zweig and Campbell, 1993):

- An AUC of 1.0 denotes perfect discrimination.
- An AUC of 0.5 suggests that the test performs no better than chance.

Figure 2 presents the Receiver Operating Characteristic (ROC) curves for two diagnostic tests, illustrating their ability to discriminate between positive and negative cases across varying threshold values. The ROC curve plots Se (true positive rate) against one minus Sp (false positive rate), providing insight into the performance of each test in distinguishing between the two conditions.

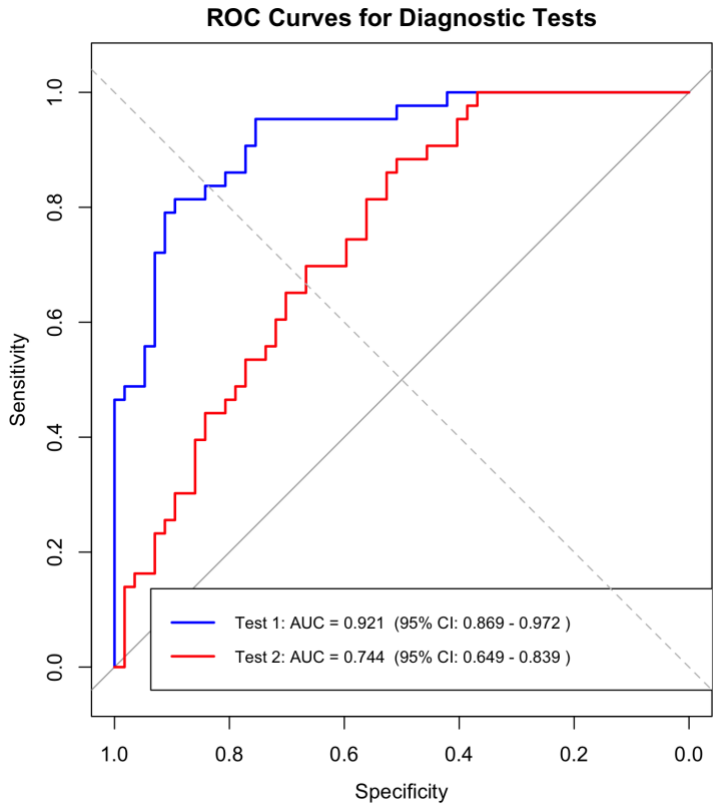
Test 1, represented by the blue curve, demonstrates a strong discriminatory capacity, with an area under the curve (AUC) of 0.921 (95% CI: 0.869–0.972). This value indicates a high level of accuracy, as an AUC close to 1 suggests that the test is highly effective in differentiating between positive and negative cases. In contrast, Test 2, shown in red, exhibits a more moderate diagnostic performance, with an AUC of 0.744 (95% CI: 0.649–0.839). While still performing above the level of a random classifier (AUC = 0.5), the lower AUC value for Test 2 suggests that it is less reliable than Test 1 in correctly classifying cases.

The shape of the ROC curves further supports these observations. Test 1's curve rises steeply towards the upper-left corner of the plot, indicating superior Se and Sp across a range of thresholds. In comparison, Test 2's curve follows a less pronounced trajectory, reflecting a higher rate of misclassification. Additionally, the confidence intervals associated with the AUC estimates provide further insight into the reliability of each test's performance. The narrower confidence interval for Test 1 suggests greater

precision in its AUC estimate, whereas the broader interval for Test 2 reflects greater uncertainty in its discriminatory power.

These findings indicate that Test 1 is the more effective diagnostic tool, offering a significantly better balance between Se and Sp compared to Test 2. The performance differences between the two tests should be taken into account when considering their application in clinical or research settings, particularly where accurate classification is of critical importance.

Figure 2. ROC curves for diagnostic tests



In summary, the ROC curve and AUC are crucial tools for evaluating diagnostic test performance. They provide both visual and quantitative insights into a test's accuracy, helping clinicians and researchers make informed decisions about diagnostic tools. The comparison of Test 1 and Test 2 highlights the importance of considering AUC values and the ROC curve's shape, as well as associated confidence intervals, to understand each test's reliability and precision. Therefore, incorporating ROC analysis in evaluation processes ensures the selection of accurate and reliable diagnostic methods.

2.2.6 Model Calibration and Advanced Performance Measures

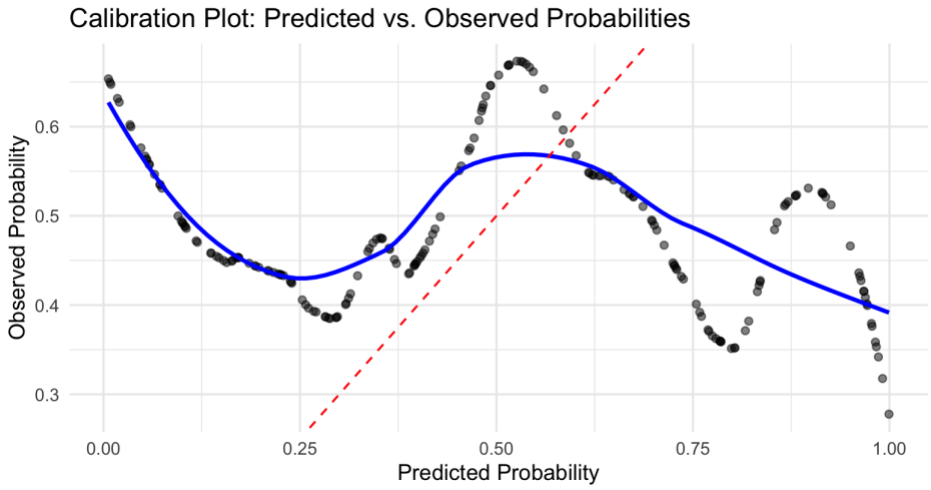
In diagnostic test research, calibration is a critical aspect of model evaluation when tests are incorporated into prognostic or predictive models. While discrimination metrics, such as the area under the receiver operating characteristic curve (AUC), assess a model's ability to differentiate between diseased and non-diseased individuals, calibration evaluates how well the predicted probabilities correspond to actual outcomes. A well-calibrated model ensures that a predicted probability of 70% for disease presence reflects an actual 70% likelihood in real-world observations.

Calibration is often assessed visually using a calibration plot, which compares predicted probabilities against observed proportions. In an ideal model, predictions should align with the diagonal reference line (Figure 2), indicating perfect agreement between expected and actual outcomes. Deviations from this line suggest systematic overestimation or underestimation of disease probability, which may affect clinical decision-making. Statistical tests such as the Hosmer–Lemeshow goodness-of-fit test further quantify calibration quality, identifying models that require recalibration.

Beyond traditional calibration assessment, advanced performance measures, such as net reclassification improvement (NRI) and integrated discrimination improvement (IDI), provide additional insight into the benefit of incorporating a new diagnostic test into an existing predictive framework. These metrics quantify how well a new test improves risk stratification, an essential consideration in personalised medicine, where even marginal refinements in risk estimation can influence therapeutic strategies.

Figure 3 illustrates a calibration plot for a predictive model. The dashed red line represents perfect calibration, while the blue curve depicts the actual relationship between predicted and observed probabilities. In this example, the deviation of the blue curve from the reference line suggests that the model miscalibrates probability estimates in certain ranges, potentially leading to biased clinical interpretations. Addressing these discrepancies through recalibration or model refinement is essential for ensuring reliable diagnostic predictions.

Figure 3. Example of a calibration plot



Finally, the Table 4 summarises all the metrics, presenting their formulas and brief interpretations.

Table 4. Summary of Key Diagnostic Performance Metrics

Metric	Formula	Interpretation	Reference
Sensitivity	$\frac{TP}{TP + FN}$	Proportion of diseased individuals correctly identified.	Zhou, Obuchowski, & McClish (2012)
Specificity	$\frac{TN}{TN + FP}$	Proportion of healthy individuals correctly identified.	Zhou, Obuchowski, & McClish (2012)
Positive Predictive Value (PPV)	$\frac{TP}{TP + FP}$	Probability that a positive test result indicates the presence of disease.	Altman & Bland (1994)
Negative Predictive Value (NPV)	$\frac{TN}{TN + FN}$	Probability that a negative test result indicates the absence of disease.	Altman & Bland (1994)
Positive Likelihood Ratio (LR+)	$\frac{Sensitivity}{1 - Specificity}$	Magnitude of increase in odds of disease with a positive test result.	Simel et al. (1991)

Metric	Formula	Interpretation	Reference
Negative Likelihood Ratio (LR–)	$\frac{1 - Sensitivity}{Specificity}$	Magnitude of decrease in odds of disease with a negative test result.	Simel et al. (1991)
Diagnostic Odds Ratio (DOR)	$\frac{TP \times TN}{FP \times FN}$ or $\frac{LR+}{LR-}$	Overall measure of test performance; higher values indicate better discrimination.	Deeks (2001)

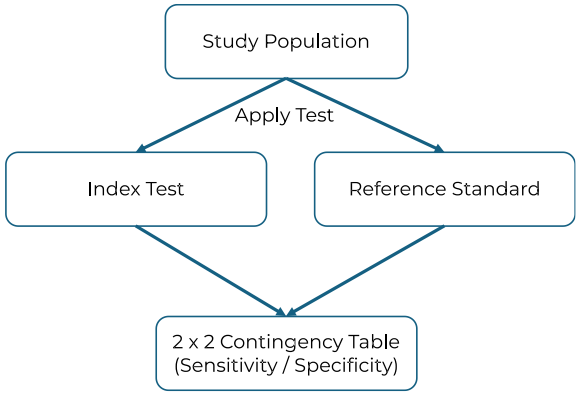
TP = true positives, FP = false positives, TN = true negatives, and FN = false negatives.

2.3 Study Design for Evaluating Diagnostic Tests

Evaluating diagnostic tests requires careful consideration of study design to ensure that the findings are robust, valid, and applicable to clinical practice. In this section, we examine two principal designs: cross-sectional studies (often conducted prospectively) in which the index test and RS are applied to the same participants within a clinically appropriate time window, and case-control studies, which are sometimes used when the target condition is rare. Designs that follow participants over time can be valuable for prognosis or delayed verification, but they are conceptually distinct from standard diagnostic accuracy studies and introduce additional sources of bias (e.g., losses to follow-up and changes in management during follow-up).

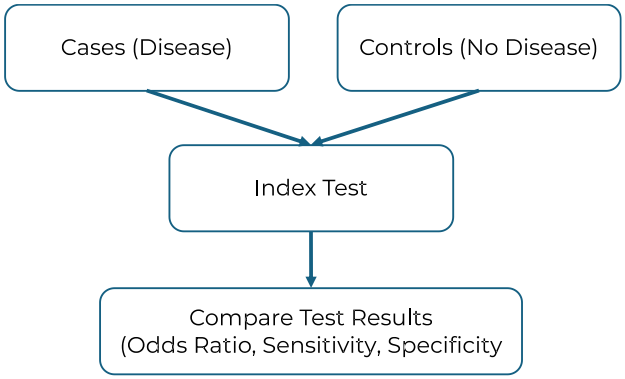
A cross-sectional design is often the initial approach when assessing a new diagnostic test. In this design, a representative sample of the target population is recruited, and every participant undergoes both the index test and the reference standard concurrently. This simultaneous application minimizes the risk of changes in disease status over time and facilitates the immediate computation of performance metrics from a 2x2 contingency table (Figure 4). The data derived in this manner are particularly valuable in the early phases of test evaluation. However, this approach provides only a snapshot of diagnostic performance. It does not capture temporal variations or the progression of disease, which may be relevant in conditions that evolve over time. For instance, while Se and Sp can be directly estimated, the design does not allow an assessment of how these measures might change as the disease advances.

Figure 4. Flow diagram illustrating the design of a cross-sectional diagnostic accuracy study.



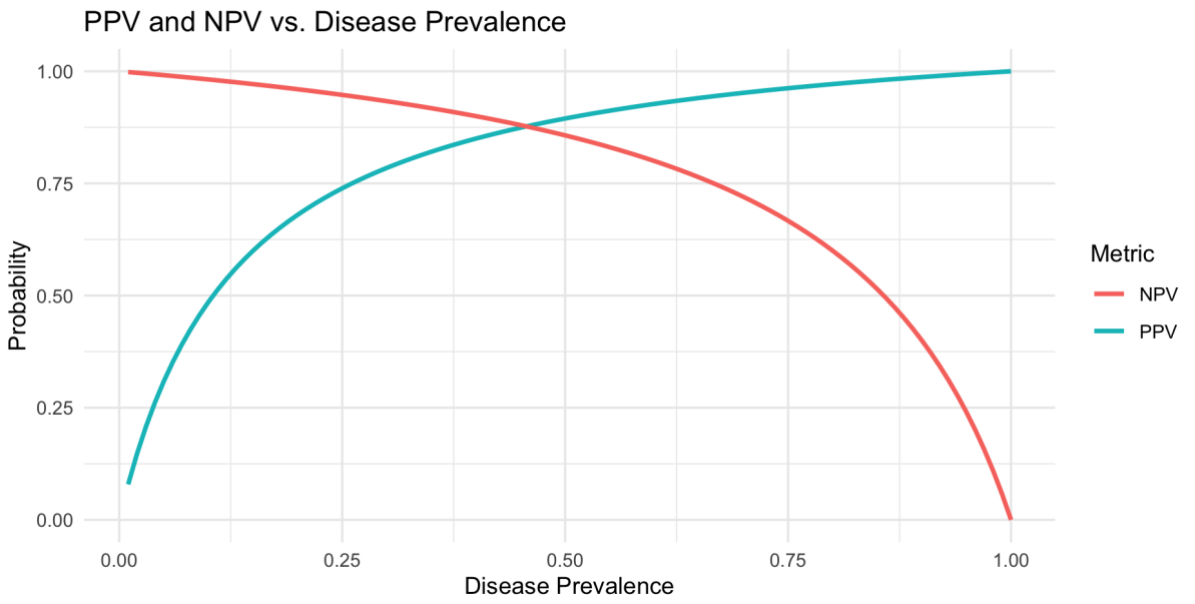
In contrast, case-control studies offer an alternative that is especially suited for conditions with low prevalence. In such studies, individuals are selected retrospectively based on their disease status; those with the disease (cases) and those without (controls) are identified, and the index test is then applied (Figure 5). This approach enables researchers to obtain sufficient numbers of affected individuals, thereby increasing statistical power when the disease is rare. The retrospective nature of the design allows for efficient data collection, yet it also introduces potential biases. Selection bias may occur if the cases and controls are not representative of the full clinical spectrum, while verification bias may arise if the reference standard is not uniformly applied. Care must be taken to ensure that the cases truly reflect the target population and that the control group is free from latent or subclinical forms of the disease. Despite these challenges, case-control studies remain a powerful tool when prospective recruitment is impractical.

Figure 5. Flow diagram outlining the design of a case-control diagnostic accuracy study.



A critical aspect common to all study designs is the influence of disease prevalence on the interpretation of diagnostic performance metrics. While intrinsic measures such as Se and Sp remain unaffected by prevalence, predictive values are highly dependent on the underlying frequency of the disease within the study population. In settings where the disease is rare, even tests with excellent Se and Sp may yield a low positive predictive value. Conversely, the negative predictive value tends to be high when prevalence is low, since most individuals will be disease-free. It is therefore imperative that studies report prevalence-adjusted estimates or include Se analyses that model predictive values over a range of plausible prevalence rates. Graphical representations, such as plots of PPV and NPV against varying prevalence levels, can provide valuable insights into how a diagnostic test may perform in different clinical environments (see example in Figure 6). Such visualisations serve not only as an aid to interpretation but also as a bridge between statistical outcomes and their practical implications in patient care.

Figure 6. Relationship between disease prevalence on predictive values.



NPV: negative predictive value, PPV: positive predictive value

The Table 5 summarises the key characteristics, strengths, and limitations of cross-sectional, and case-control diagnostic accuracy designs.

Table 5. Overview of Two Diagnostic Accuracy Study Designs with Their Descriptions, Strengths, and Limitations

Study Design	Description	Strengths	Limitations
Cross-sectional Studies	The index test and reference standard are applied simultaneously in a specified population.	Reduces temporal variation; straightforward and cost-efficient; offers immediate estimates.	Captures only a specific moment in time, which may not accurately represent the progression of the disease; prone to selection bias if the sample is not representative.
Case-Control Studies	Cases (with disease) and controls (without disease) are selected retrospectively, followed by an index test assessment.	Ideal for rare diseases; needs fewer participants; boosts statistical power for low-prevalence conditions.	Susceptible to selection and verification bias; constrained by the possible non-representativeness of cases and controls; spectrum bias may affect generalizability.

Each design offers a unique perspective on diagnostic performance, and the choice of design should be informed by the specific research question, the nature of the disease, and practical considerations such as available resources and required follow-up duration.

In discussing these designs, it is important to recognise that no single approach is universally superior. Rather, the optimal design depends on the clinical context and on feasibility constraints. For instance, cross-sectional studies are invaluable for preliminary assessments and for tests aimed at immediate diagnosis, whilst case-control studies can be efficient when prevalence is low, albeit at the cost of spectrum and selection biases.

The integration of disease prevalence into the analysis of diagnostic test performance further complicates the design and interpretation of studies. When a test is evaluated in a population with an atypical prevalence—whether higher or lower than that of the general population—the external validity of predictive values may be compromised. Researchers are therefore encouraged to conduct sensitivity analyses and to report performance metrics across a range of prevalence scenarios. Such methodological rigour enhances the clinical utility of the research findings, ensuring that they are relevant and translatable across different patient populations.

2.4 Test-treatment trials, end-to-end evidence, and linked evidence modelling

Diagnostic accuracy is an important intermediate outcome, but it is not the final objective of diagnostic testing. From an HTA perspective, the key question is whether using a test (as part of a diagnostic strategy) improves patient outcomes and represents good value for money, given downstream consequences for treatment decisions, resource use, and health outcomes.

One route to establish this is a test-treatment trial (sometimes described as a diagnostic strategy randomised trial). In such studies, participants are randomised to alternative diagnostic strategies (for example, a new test strategy versus usual care), and subsequent management follows prespecified clinical protocols informed by the test results. The principal endpoints are patient-relevant outcomes and resource use;

diagnostic accuracy may be reported, but it is not necessarily estimable in the classical sense because a definitive reference standard is not always applied to all participants.

Test-treatment trials have an important strength: they provide direct evidence that a diagnostic strategy changes management in a way that improves outcomes. However, they are often difficult to design and execute. Sample sizes may need to be large because only a subset of patients will experience a management change, follow-up may be long, and blinding can be challenging. Contamination between strategies, evolving clinical practice, and rapid test innovation can further limit feasibility.

Because direct trials are uncommon, most diagnostic HTAs rely on evidence linkage and decision-analytic modelling. In a linked-evidence approach, estimates of test accuracy (typically from diagnostic test accuracy studies), evidence on treatment effects, and information on diagnostic pathways are synthesised within a decision model to estimate downstream costs and health outcomes (e.g., QALYs). This is the primary route by which cost-effectiveness of diagnostic technologies is assessed in practice, and it makes explicit how uncertainty and bias in the clinical evidence propagate to decision outcomes.

The work in this thesis sits within this modelling paradigm. It examines how misclassification arising from imperfect reference standards can enter linked-evidence models and distort economic conclusions when model components are conditioned on different disease definitions. The aligned versus unaligned framework presented in later chapters is intended to support internal consistency in model structure, and the simulation study illustrates when targeted adjustments to accuracy inputs may be decision-relevant.

2.5 Challenges in Diagnostic Test Studies

Diagnostic test studies are fraught with methodological complexities that can significantly impact the validity of their findings. These challenges not only influence estimates of test performance—such as Se, Sp, predictive values, and likelihood ratios—but also affect subsequent economic evaluations, where the accuracy of diagnostic tests is a pivotal input. This section provides an in-depth discussion of five principal challenges in diagnostic test studies: spectrum bias and generalisability, verification

bias, imperfect reference standards, observer variability and inter-rater reliability, and external validity and reproducibility issues.

2.5.1 Spectrum Bias and Generalisability

Spectrum bias arises when the study population does not adequately represent the full range of disease severity or the variability present in the target clinical setting. In diagnostic test evaluations, a restricted or non-representative sample may lead to an overestimation or underestimation of a test's accuracy. For instance, studies that recruit predominantly severe cases and clear non-cases may report artificially high Se and Sp, whereas in a real-world scenario—with a more heterogeneous patient population—the test's performance might be less impressive (Ransohoff and Feinstein, 1978). This lack of representativeness limits the generalisability of the study findings, undermining the external validity of the diagnostic evaluation.

The implications for economic evaluation are considerable. Cost-effectiveness analyses rely on robust diagnostic performance estimates to model downstream clinical outcomes and resource utilisation. Overestimation of test accuracy due to spectrum bias may lead to overly optimistic projections of cost savings or health benefits. Conversely, a more conservative estimate in a heterogeneous population can influence the ICER unfavourably, potentially leading decision-makers to underestimate the value of the diagnostic intervention. Thus, ensuring that study populations reflect real-world clinical practice is vital for both accurate diagnostic assessment and subsequent economic modelling.

2.5.2 Verification Bias

Verification bias, also known as work-up bias, occurs when not all study participants undergo the reference standard test, or when the decision to apply the reference standard is influenced by the index test result. This bias can lead to skewed estimates of Se and Sp, as patients with positive index test results are more likely to be verified than those with negative results (Rutjes et al., 2007). Verification bias is particularly

problematic in diagnostic studies where the reference standard is invasive, expensive, or ethically challenging to administer to all participants.

In the context of economic evaluation, verification bias can distort the CEA by misrepresenting the true rates of false positives and false negatives. Such misestimation may lead to inappropriate clinical decisions that either incur unnecessary costs (through overdiagnosis and overtreatment) or result in missed opportunities for early intervention (due to false negatives). Adjusting for verification bias through statistical methods or study design modifications is essential to provide reliable estimates that accurately inform health economic models.

2.5.3 Imperfect Reference Standards

A recurring challenge in diagnostic research is the absence of a perfect gold standard against which to measure the index test. Imperfect reference standards—whether due to inherent limitations in test accuracy or because no single test can capture the true disease state—can lead to biased estimates of diagnostic performance. When the reference standard is flawed, misclassification may occur, resulting in inaccurate Se, Sp, and predictive values (Kohn et al., 2013). In some cases, composite reference standards or latent class analyses are used to mitigate this issue, although these approaches come with their own methodological complexities.

For economic evaluations, the use of an imperfect reference standard introduces uncertainty into the decision model. The cost-effectiveness of a diagnostic test is highly dependent on its ability to correctly classify patients, and errors in the reference standard propagate through the entire analytic model. Sensitivity analyses and Bayesian methods are often employed to account for this uncertainty, thereby ensuring that the economic evaluation reflects a more realistic range of potential outcomes. This uncertainty underscores the need for cautious interpretation of results and transparent reporting of the assumptions underlying the reference standard.

2.5.4 Observer Variability and Inter-Rater Reliability

Observer variability refers to the inconsistencies that may occur when different clinicians interpret the same diagnostic test results. Inter-rater reliability is a measure of the agreement between observers, and low agreement can compromise the reproducibility of test outcomes (Gwet, 2012). Variability can stem from subjective interpretation of images, laboratory results, or clinical findings, particularly when standardised protocols are not strictly followed.

In diagnostic accuracy studies, high observer variability can lead to inconsistent estimates of performance metrics, further complicating the generalisability of the findings. For economic evaluation, observer variability translates into uncertainty in the diagnostic pathway, potentially affecting the projected health outcomes and associated costs. For instance, if the interpretation of a diagnostic imaging test varies widely among radiologists, the downstream clinical decisions—including treatment selection and resource allocation—may be less reliable. Standardisation of test interpretation and rigorous training protocols are essential to minimise these sources of error. Moreover, incorporating measures of inter-rater reliability into the economic model can improve the robustness of the CEA by quantifying the uncertainty associated with diagnostic decision-making.

2.5.5 External Validity and Reproducibility Issues

The external validity of a diagnostic test study refers to the extent to which the findings can be generalised beyond the study setting. Reproducibility is closely linked, referring to the ability to obtain consistent results when the study is repeated under similar conditions. Factors that undermine external validity include differences in patient demographics, variations in disease prevalence, and discrepancies in clinical practice across settings. A study conducted in a highly specialised tertiary care centre may not yield results that are applicable to primary care settings, where patient populations and diagnostic resources differ significantly (Flahault et al., 2005).

From an economic evaluation perspective, the external validity of diagnostic accuracy estimates is critical. Cost-effectiveness analyses are highly sensitive to variations in diagnostic performance; thus, if the test's accuracy is overestimated in a controlled

study environment, the economic model may overstate the benefits of the intervention when applied in routine practice. Reproducibility issues further compound this problem, as inconsistent findings between studies can lead to uncertainty in the economic estimates. Therefore, multi-centre studies and systematic reviews that synthesise data from diverse settings are crucial to enhance both the generalisability and reproducibility of diagnostic test evaluations.

The challenges discussed—spectrum bias, verification bias, imperfect reference standards, observer variability, and issues of external validity and reproducibility—are interrelated and collectively influence the reliability of diagnostic test studies. These methodological issues not only affect the accuracy of diagnostic performance metrics but also have significant implications for economic evaluation. Inaccurate or biased estimates can lead to erroneous conclusions regarding the cost-effectiveness of a diagnostic intervention, ultimately impacting healthcare policy and clinical practice. Researchers must therefore employ rigorous study designs, robust statistical methods, and comprehensive sensitivity analyses to mitigate these challenges. Transparent reporting and adherence to guidelines, such as the STARD (Standards for Reporting Diagnostic Accuracy) initiative, are essential to ensure that the limitations are clearly articulated and that the findings can be interpreted in the appropriate clinical and economic context.

Challenges in diagnostic test studies are multifaceted and pose substantial threats to both internal and external validity. Spectrum bias, verification bias, imperfect reference standards, observer variability, and issues of external validity must be carefully addressed through thoughtful study design and rigorous analytical techniques. These challenges not only influence the estimates of diagnostic accuracy but also propagate uncertainty into subsequent economic evaluations. By recognising and adjusting for these limitations, researchers can ensure that diagnostic test evaluations provide robust and clinically relevant evidence, which is essential for informed decision-making in healthcare (see Table 6).

Table 6. Summary of Key Challenges in Diagnostic Test Studies and Their Implications

Challenge	Description	Impact on Diagnostic Accuracy	Implications for Economic Evaluation
Spectrum Bias	Non-representative sample; limited disease spectrum	Over- or underestimation of Se and Sp	Misleading cost-effectiveness projections due to inaccurate performance estimates
Verification Bias	Selective application of the reference standard based on index test results	Biased estimates of Se and Sp	Inaccurate modelling of downstream outcomes and costs
Imperfect Reference Standards	Use of a non-ideal standard leading to misclassification	Misclassification errors; inaccurate performance measures	Propagation of uncertainty into economic models; requires sensitivity analyses
Observer Variability	Differences in test interpretation among clinicians	Inconsistent and non-reproducible results	Uncertainty in clinical decision-making; potential variability in economic outcomes
External Validity/Reproducibility	Limited generalisability of findings across diverse settings and failure to replicate study results	Results may not reflect real-world performance	Economic evaluations may overestimate benefits if based on non-representative data

2.6 Reference standards and imperfect reference standards

In diagnostic research, the concept of a "gold standard" is central to the evaluation of any new test. A gold standard is defined as a diagnostic test or procedure that is regarded as the best available method for determining the true disease status of an individual. Ideally, it is a method with near-perfect Se and Sp, minimal error rates, and wide clinical

acceptance. In practice, however, the existence of a perfect or near-perfect reference is often more aspirational than real. This section focuses on the multifaceted problems that arise when a gold standard is lacking and examines the profound implications this absence has on the reliability, interpretation, and economic evaluation of diagnostic tests.

A gold standard is meant to represent the "truth" against which new diagnostic tests are measured. In an ideal scenario, this reference test would unambiguously classify individuals as diseased or non-diseased. However, many clinical conditions defy such clear categorisation. For instance, diseases with heterogeneous manifestations—such as autoimmune disorders, early-stage cancers, or emerging infectious diseases—often lack a universally accepted diagnostic criterion. The reference test may itself have inherent limitations such as operator dependency, variability in performance across settings, or an inability to capture the full spectrum of disease severity. Historical examples include the diagnosis of tuberculosis prior to the advent of molecular methods and the challenges in diagnosing Alzheimer's disease where definitive diagnosis may only be possible post-mortem (Bossuyt et al., 2003).

When a true gold standard is unavailable, researchers are forced to rely on what is termed an "imperfect reference standard." Imperfect reference tests, by definition, do not provide flawless diagnostic accuracy. They may be influenced by a range of factors—from technical variability and subjective interpretation to intrinsic limitations in capturing the biological complexity of a disease. For example, in many cancers, histopathological examination remains the reference standard; yet inter-observer variability and differences in tissue sampling can lead to discrepancies in the final diagnosis. In the realm of infectious diseases, culture methods—traditionally considered the gold standard—often have suboptimal Se, particularly when patients have been pre-treated with antibiotics.

The absence of a gold standard has several cascading effects on diagnostic test evaluation. First, it introduces uncertainty into the measurement of key performance metrics such as Se, Sp, predictive values, and likelihood ratios. When the comparator itself is flawed, estimates of these metrics can be biased, sometimes substantially so. For instance, if the reference test misses cases of disease due to low Se, then a new diagnostic test may appear to have a lower false-positive rate than it actually does.

Conversely, an imperfect reference can overestimate the performance of a novel test if the true disease status is misclassified. Such biases not only undermine the internal validity of the study but also limit the external validity or generalisability of the findings.

Decision-analytic modelling represents the predominant approach for economic evaluation of diagnostic tests because tests do not directly improve health but rather inform treatment decisions. Models link test accuracy to patient outcomes through evidence synthesis, combining accuracy data with treatment effectiveness and costs to project long-term consequences (Briggs, 2006). This evidence linkage approach is central to the current thesis, as it provides the framework within which reference standard bias can propagate from accuracy estimates to cost-effectiveness conclusions. The choice of cost-effectiveness threshold—representing society's opportunity cost when healthcare resources are allocated to one intervention rather than another—is particularly consequential for diagnostic tests, which often generate modest incremental benefits where small biases in accuracy could shift conclusions across threshold boundaries (Claxton et al., 2015).

This problem extends into the realm of economic evaluation. Economic models—be they decision trees, Markov models, or cost-effectiveness analyses—rely heavily on accurate diagnostic performance parameters to estimate downstream clinical outcomes, treatment costs, and QALYs. If the baseline data used in these models are derived from studies with an imperfect reference standard, then the resulting economic projections may be misleading. For example, an overestimation of a diagnostic test's accuracy could lead to an inflated expectation of its impact on early disease detection, thereby overestimating potential cost savings and health benefits. In contrast, underestimation of test performance may result in a failure to recognise the value of implementing the new diagnostic approach. This uncertainty poses a significant challenge for policymakers who must decide whether to adopt novel diagnostic technologies based on economic evaluations that are, at their core, built on uncertain diagnostic parameters.The implications of lacking a gold standard are further compounded by the fact that the error introduced by an imperfect reference is often unquantifiable. Unlike random error, which can be statistically modelled and mitigated through large sample sizes, systematic error from an imperfect gold standard can bias results in a consistent and predictable direction. This bias is problematic because it is not always apparent

from the study data itself. Researchers might mistakenly interpret high Se or Sp as indicative of a superior diagnostic test when, in fact, these metrics are artefacts of the flawed reference. As a result, the downstream clinical decisions and economic evaluations that depend on these metrics can be fundamentally compromised.

Moreover, the absence of a gold standard creates significant challenges in the synthesis of diagnostic research. Systematic reviews and meta-analyses aim to aggregate evidence from multiple studies to derive more precise estimates of test performance. However, if the included studies use different, and often imperfect, reference standards, the heterogeneity in diagnostic accuracy estimates can be substantial. This heterogeneity complicates the pooling of data and may necessitate the use of advanced statistical techniques, such as hierarchical modelling or Bayesian meta-analysis, to adjust for variability in the reference standards. In turn, this adds another layer of complexity to both the clinical interpretation and the economic evaluation of diagnostic tests.

The clinical implications are equally critical. Without a definitive gold standard, clinicians face uncertainty in interpreting test results. A patient's diagnosis may be based on a composite of clinical judgement, ancillary tests, and follow-up data rather than a single, definitive test result. This ambiguity can affect treatment decisions, leading to potential over-treatment or under-treatment, both of which carry significant consequences for patient outcomes and healthcare costs. In some instances, the lack of a gold standard may prompt the use of repeat testing or supplementary diagnostic procedures, further increasing the overall cost and complexity of patient care.

Given these multifaceted challenges, limitations in reference standards remain a substantial methodological issues in diagnostic research. It requires researchers to adopt alternative strategies—such as latent class models, composite reference standards, or Bayesian methods—to approximate true disease status. However, while these methods offer potential solutions, they also bring their own assumptions and complexities, and they cannot fully substitute for a definitive gold standard.

In conclusion, the problem of lacking a gold standard in diagnostic test studies is both profound and pervasive. It introduces significant uncertainty into the estimation of diagnostic performance metrics, complicates systematic reviews and meta-analyses, and has far-reaching implications for clinical decision-making and economic evaluation.

Without a reliable reference, the chain of evidence—from diagnostic accuracy to treatment outcomes and economic projections—remains vulnerable to bias and misinterpretation. Recognising and addressing this challenge is thus paramount for advancing the rigour of diagnostic research and ensuring that subsequent economic evaluations are built on a sound and transparent foundation.

2.7 Discussion

The evaluation of diagnostic tests is a fundamental aspect of modern healthcare, ensuring that diagnostic tools are accurate, reliable, and clinically useful. This chapter has outlined key concepts in diagnostic test research, including study designs, performance measures, and challenges such as bias, verification issues, and the absence of a perfect gold standard. These factors collectively influence the interpretation and applicability of diagnostic tests in clinical practice.

One critical issue in diagnostic research is the trade-off between Se and Sp, which can have significant implications for patient management. A highly sensitive test minimises false negatives, making it suitable for conditions were missing a diagnosis could have severe consequences. Conversely, a highly specific test reduces false positives, which is crucial in avoiding unnecessary interventions. The balance between these metrics is often dependent on the clinical context, highlighting the importance of likelihood ratios and predictive values in real-world applications.

Furthermore, while traditional accuracy measures such as ROC curves and AUC values provide insights into test performance, they do not account for the broader impact of diagnostic tests on clinical decision-making and patient outcomes. The concept of test-treatment trials, which integrates diagnostic accuracy with therapeutic implications, is increasingly recognised as an essential approach in evaluating new diagnostic technologies. However, such studies require robust methodologies to ensure that test results translate into meaningful health benefits rather than merely improving classification accuracy.

An equally important consideration is the cost and resource allocation associated with diagnostic testing. The introduction of companion diagnostics and personalised medicine has led to the development of highly specific tests that can tailor treatment to

individual patients. However, these innovations also raise concerns regarding their economic feasibility, particularly in resource-limited settings. The increasing complexity of diagnostic strategies necessitates a shift from purely technical evaluations towards a more holistic approach that includes economic and health system considerations.

While diagnostic accuracy is a critical determinant of a test's clinical utility, it does not provide a complete picture of its real-world value. The implementation of a diagnostic test carries not only clinical but also economic implications, influencing healthcare costs, resource allocation, and overall system efficiency.

In this chapter, the foundation of diagnostic test research is explored, with attention given to key concepts such as Se, Sp, study designs, and common forms of bias. By illustrating both the theoretical underpinnings and practical considerations in diagnostic accuracy assessment, this chapter addresses part of the first objective of the thesis, namely to build an informed understanding of diagnostic metrics and the methodological challenges that arise. This background sets the stage for the subsequent review of imperfect reference standard methods and economic evaluations in Chapter 3. The next chapter explores the fundamentals of CEA, a key methodology in health economics that evaluates whether a diagnostic test provides sufficient clinical benefit relative to its cost. Understanding CEA in the context of diagnostic testing allows for a more comprehensive assessment, ensuring that innovations in diagnostics are not only scientifically valid but also economically sustainable within healthcare systems

Chapter 3

3 From Statistical Methods to Economic Evaluations: A Literature Review on Addressing Imperfect Reference Standard Bias

Accurate diagnostic tests are pivotal to modern healthcare, guiding treatment decisions, resource allocation, and public health strategies. Economic evaluations of diagnostics, such as cost-effectiveness analyses (CEAs), rely heavily on estimates of test performance—typically Se and Sp—to quantify the clinical and economic value of adopting new technologies. However, these estimates are inherently contingent on the quality of the reference standard used to ascertain true disease status. When the reference standard is imperfect—misclassifying patients as false positives or false negatives—the resulting inaccuracies in Se and Sp propagate through economic models, distorting cost-effectiveness ratios and undermining the validity of healthcare decisions. This challenge, termed imperfect reference standard bias, represents a critical methodological gap in the evaluation of diagnostic technologies, with far-reaching implications for patient outcomes and resource efficiency.

The consequences of such bias are not merely theoretical. Overestimating a test's Se due to a flawed reference standard may lead to premature adoption of a technology that misses true cases, delaying treatment and exacerbating disease burden. Conversely, underestimating Sp could result in unnecessary follow-up testing or treatment, squandering limited healthcare resources. For instance, when latent tuberculosis infection is evaluated against an imperfect reference (e.g., tuberculin skin testing), biased accuracy estimates may mislead models assessing the cost-effectiveness of screening programs, potentially diverting funds from more impactful interventions. These scenarios underscore the urgency of addressing imperfect reference standard bias to ensure economic evaluations reflect true diagnostic performance.

Over the past two decades, statistical methodologies have emerged to mitigate this bias. Approaches such as verification bias correction, latent class models, and algebraic adjustments aim to disentangle the true accuracy of an index test from the imperfections of its reference. For example, Bayesian latent class models leverage multiple imperfect tests to infer unobserved disease status, while algebraic corrections adjust observed Se and Sp using known error rates of the reference standard. Despite their potential, these methods remain underutilized in health technology assessments (HTAs). Many economic evaluations continue to treat reference standards as infallible, perpetuating bias in cost-effectiveness conclusions. This disconnect between methodological advances and applied practice highlights a pressing need to synthesize existing knowledge, evaluate the strengths and limitations of correction methods, and clarify their role in supporting robust decision-making.

Objective and Scope of This Chapter

This chapter addresses Objective 1 of the thesis by systematically reviewing statistical methods designed to address imperfect reference standard bias and analyzing their integration into economic evaluations of diagnostic tests. First, it categorizes correction techniques by their applicability to three scenarios: (1) missing or partial verification of disease status, (2) universally applied but imperfect reference standards, and (3) contexts with no gold standard. Second, it evaluates the economic implications of these methods through case studies and methodological frameworks, illustrating how bias-adjusted accuracy estimates alter cost-effectiveness outcomes. Finally, it identifies persistent gaps in practice—such as the limited adoption of latent class models in HTAs—and discusses their implications for policymakers.

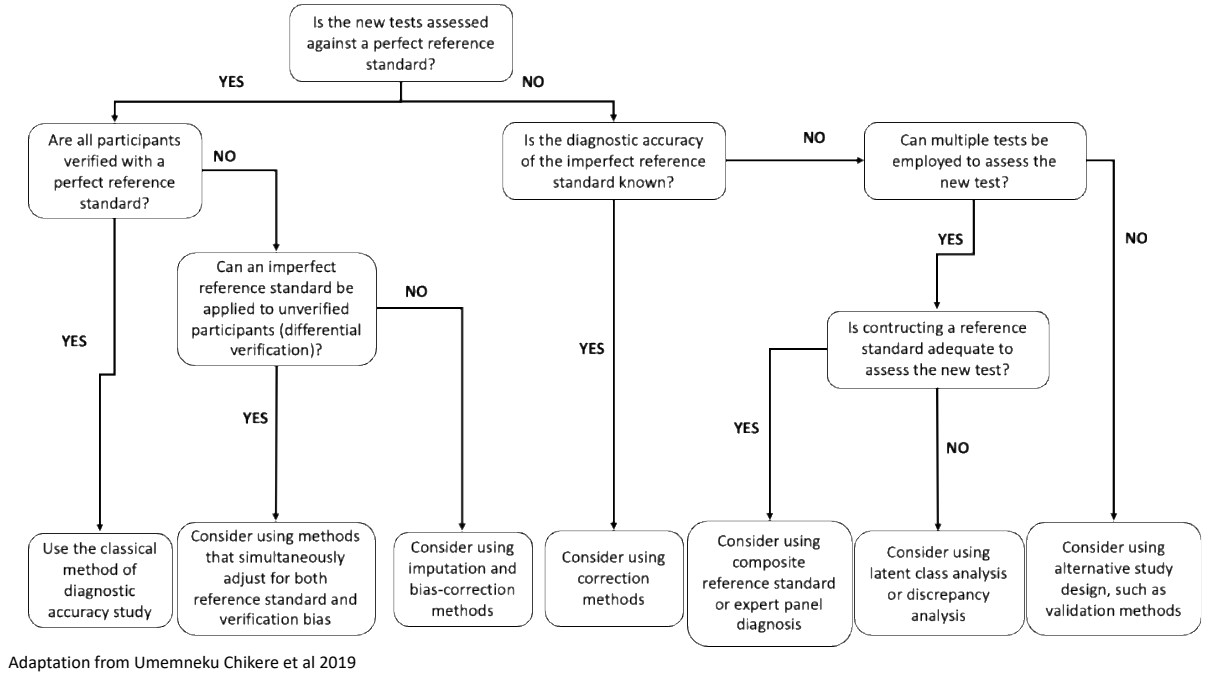
By consolidating current knowledge, this review not only clarifies the trade-offs inherent in different correction methods but also sets the stage for the development of a practical framework (Objective 2) to guide analysts in mitigating bias. In doing so, it advances the broader aim of this thesis: to ensure that decisions about diagnostic technologies are grounded in evidence that reflects both clinical and economic realities, ultimately improving healthcare outcomes and resource allocation.

3.1 The Statistical Challenge: Addressing Imperfect Reference Standards

Diagnostic accuracy studies ideally compare an index test's results to a reference standard (or "gold standard") that perfectly reflects the true disease status. In practice, many reference standards are imperfect – they may misclassify some patients – or may be unavailable for some participants. Using an imperfect or partial reference can bias estimates of Se, Sp, and other measures of test performance (Rutjes et al., 2007). This problem, often called imperfect reference standard bias (or gold standard misclassification bias), has prompted a variety of statistical solutions. Several systematic reviews have catalogued these methods, notably Rutjes et al. (2007), Reitsma et al. (2009), and Umemneku Chikere et al. (2019). Broadly, the methods can be grouped into four categories: (1) methods that impute or adjust for missing reference data (to address verification bias when not all patients receive the reference standard), (2) methods that correct for imperfections in a known reference standard, (3) methods that construct a new reference standard from multiple sources (including expert panel consensus or latent class modeling), and (4) methods that abandon the classical accuracy paradigm and instead validate the test against other clinical information or outcomes (Rutjes et al., 2007). Each approach rests on different statistical frameworks and assumptions. Figure 7 provides a high-level decision flowchart for selecting an appropriate method depending on the scenario.

The following sections, I review in depth the published evidence describing several statistical methods to adjust by the imperfect reference standard bias or to deal with scenarios where a reference standard is not available or it does not exist. These article were identified from systematic reviews using terms such as "adjusting methods", "diagnostic test", "imperfect reference standard bias", "absence of reference standard", "absence of gold standard", amongst others. In this review I explain their statistical underpinnings and suitable conditions, examine evidence from applications (including accuracy/bias comparisons and real-world use), discuss their limitations, and consider implications for economic evaluations of diagnostics.

Figure 7. Decision flowchart for selecting adjustment methods based on case requirements.



Adaptation from Umemneku Chikere et al 2019

3.1.1 Methods for Imperfect or Missing Reference Standards

The Figure 7 outlines method selection based on whether a perfect gold standard exists and is applied to all patients. If a definitive gold standard exists and is used for every patient, classical diagnostic accuracy analysis (simple Se/Sp calculation) is appropriate. If not, the chart considers several key questions: Are some participants missing the reference verification? Is the accuracy of the available (imperfect) reference known? Can multiple imperfect tests be used, or can an alternative reference be applied for those without the gold standard? Depending on these conditions, different statistical solutions are recommended – for example, imputation or bias-correction methods to handle partial verification, algebraic correction methods if the reference's error rates are known, or latent class and composite reference standard approaches if no single adequate reference standard exists. This framework (originally developed by Reitsma et al. (2009) and updated by Umemneku Chikere et al. (2019)) illustrates the breadth of strategies required to mitigate imperfect reference standard bias in diagnostic studies.

3.1.1.1 Missing/Partial Verification

3.1.1.1.1 Imputation and Verification-Bias Correction Methods.

Verification bias occurs in diagnostic accuracy studies when only a subset of patients undergo verification by the reference standard, often those testing positive on the initial (index) test or meeting specific clinical criteria (Kohn et al., 2013). This selective verification can significantly distort estimates of diagnostic accuracy—typically inflating Se and artificially reducing Sp. To mitigate this bias, a range of statistical techniques have been developed, collectively referred to as imputation and verification-bias correction methods.

Early methodological approaches, such as those proposed by Begg and Greenes (1983) and Gray et al. (1984), identified verification bias as a type of selection bias and employed maximum likelihood or Bayesian models to account for non-randomly missing data. These approaches typically require assumptions about the missingness mechanism—specifically, whether missing verification data are ignorable (Missing at Random, MAR) or non-random (MNAR). Under the MAR assumption, logistic regression or multiple imputation methods can effectively model and impute missing disease status based on observable patient characteristics and test outcomes, resulting in unbiased Se and Sp estimates (Umemneku Chikere et al., 2019).

More sophisticated methods address scenarios of non-ignorable missingness (MNAR), where verification is selectively applied, often guided by clinical judgment or specific patient characteristics. Techniques such as selection modeling jointly address the verification process and test outcomes, offering more accurate adjustments but requiring more complex modeling and substantial assumptions or the availability of instrumental variables (Hui and Zhou, 1998, Alonzo, 2014). Bayesian approaches have also gained prominence, leveraging prior knowledge alongside observed data to correct for verification bias. These methods have demonstrated robust performance in multiple clinical contexts, notably oncology and cardiology, where ethical or practical constraints limit universal verification. For example, Bayesian modeling has effectively adjusted Se estimates for cancer screening tests, providing more realistic appraisals aligned with subsequent clinical follow-up outcomes (Begg and Greenes, 1983, Hui and Zhou, 1998).

However, these verification-bias correction methods rely heavily on critical assumptions regarding the nature of missing data, especially the MAR assumption (missing at random). When the MAR assumption holds true, maximum likelihood corrections and Bayesian approaches typically eliminate verification bias effectively (Umemneku Chikere et al., 2019). Conversely, if the MNAR scenario is inaccurately modeled, residual bias may remain. Therefore, explicit modeling of the verification process and the use of instrumental variables or sensitivity analyses are essential for robust bias correction (Umemneku Chikere et al., 2019). Additionally, verification-bias corrections depend significantly on adequate sample sizes for both verified and unverified cases. Insufficient data can increase uncertainty in imputed values, potentially diminishing the reliability of corrected accuracy estimates. Modern computational tools (e.g., Bayesian Markov Chain Monte Carlo (MCMC) methods) have made complex corrections manageable, although they require careful convergence assessment and advanced statistical expertise.

Practical considerations also include the interpretability and acceptability of corrected estimates by clinical stakeholders. Transparent communication is crucial, particularly regarding assumptions about the verification process. Despite these caveats, verification-bias correction methods have consistently been shown to improve diagnostic accuracy estimates and thus enhance the reliability of subsequent economic evaluations.

In summary, imputation and verification-bias correction methods significantly improve accuracy estimates in diagnostic test evaluations when verification bias is present. Their effective application, however, depends critically on the appropriate characterization of the verification mechanism and availability of sufficient data. Analysts must carefully evaluate these assumptions, perform sensitivity analyses, and clearly communicate uncertainties to ensure the credibility and usefulness of adjusted accuracy findings in economic evaluations.

3.1.1.2 Differential Verification (Multiple Reference Standards).

Differential verification bias occurs when distinct reference standards are applied to different patient subgroups, leading to inconsistent diagnostic accuracy estimates

(Kohn et al., 2013). This form of bias is particularly prevalent in studies where a definitive gold standard is infeasible for all participants, necessitating the use of alternative—and often less reliable—reference tests. Such differential misclassification complicates the interpretation of Se and Sp, as error rates vary across verification pathways.

To address this, statistical methodologies have been developed to jointly model the accuracy of the index test and reference standards. Bayesian latent class models (LCMs) have emerged as a robust solution, enabling probabilistic estimation of true disease status without assuming a single infallible reference (de Groot et al., 2011). By integrating prior distributions with observed test outcomes, LCMs simultaneously estimate the Se and Sp of both the index test and multiple reference standards, adjusting for discrepancies in verification protocols (Umemneku Chikere et al., 2019). These models are explored in detail in subsequent sections.

Alternative approaches include hierarchical Bayesian models, which extend LCMs by incorporating random effects to account for variability in test performance across subgroups or clinical settings (Lu et al., 2018). Such models relax the restrictive conditional independence (CI) assumption (i.e., that tests are independent given disease status), thereby reducing bias when tests share common sources of error (Gardner et al., 2000). Multiple imputation techniques offer another corrective strategy, simulating plausible disease statuses for unverified patients based on covariates and test results, though they require robust auxiliary data to ensure validity (Umemneku Chikere et al., 2019).

Empirical applications underscore the utility of these methods. For instance, in Alzheimer's disease research, Bayesian LCMs have reconciled discrepancies between post-mortem histopathology (high Sp but limited availability) and clinical-imaging criteria (lower Sp). By jointly analysing these verification pathways, studies have derived adjusted Se estimates for cerebrospinal fluid biomarkers that align more closely with longitudinal clinical outcomes (Umemneku Chikere et al., 2019). Similarly, hierarchical models have improved the accuracy of tuberculosis diagnostics in paediatric populations, where reference standards like culture and Xpert MTB/RIF exhibit variable performance (Schumacher et al., 2016).

However, these methods present significant challenges. First, the validity of Bayesian LCMs depends heavily on the appropriateness of prior distributions and model

assumptions. Sensitivity analyses are critical to assess robustness, as divergent priors or dependency structures can yield markedly different accuracy estimates (Umemneku Chikere et al., 2019). Second, computational complexity and the need for specialised statistical expertise may hinder widespread adoption, particularly in resource-limited settings.

In conclusion, differential verification bias poses a substantial threat to the validity of diagnostic accuracy studies employing multiple reference standards. Bayesian LCMs, hierarchical models, and multiple imputation techniques provide principled frameworks to adjust for such bias, enhancing the reliability of Se and Sp estimates. Nevertheless, meticulous model specification, sensitivity analyses, and transparent reporting of assumptions remain imperative to ensure credible results for clinical and economic decision-making.

3.1.1.3 Absence of a Known Perfect Reference Standard

3.1.1.3.1 Reference Standard Misclassification Correction Methods.

When all patients are verified by a reference test that is imperfect, a different set of methods can be used to adjust for the reference standard's known limitations. These methods treat the situation as one of misclassification in a binary classifier (the reference) and apply algebraic corrections to the index test's apparent Se/Sp. In essence, one can derive formulas relating the observed 2×2 table (index test vs. imperfect reference) to the true disease status if the reference test's sensitivity (Se_{iRS}) and specificity (Sp_{iRS}) are known or can be estimated. For example, if the reference is only 90% sensitive, the index test will appear to miss some cases simply because the reference missed them.

Methods by Gart and Buck (1966) and later by Brenner (1996) formalized this adjustment: one "unscrambles" the contingency table assuming the reference misclassifies disease at a known rate. Several correction methods were identified in the literature review by Umemneku Chikere et al. (2019) – at least seven distinct approaches across five articles (Gart and Buck, 1966, Brenner, 1996, Staquet et al., 1981, Emerson et al., 2018, Trikalinos and Balion, 2012), and recently Sherwin (2022). Most yield point estimates for the index test's true accuracy.

Notably, Emerson et al. (2018) introduced a method that does not give a single corrected value but rather computes bounds (an upper and lower possible value) for Se and Sp. This bounding approach acknowledges uncertainty in the reference's imperfection or in potential dependency between index and reference errors: instead of assuming a precise correction, it reports the range within which the true performance lies, given the imperfect reference.

Conditions for applying these correction methods include having reliable, externally obtained estimates of the reference standard's Se and Sp (for example, from prior studies or meta-analyses of the reference test) (Rutjes et al., 2007). If such information is available, one can plug those values into the correction formulas. These methods are most appropriate when the reference test is well-studied (so its imperfection is quantifiable) and when the index test is being evaluated in the same context. For instance, if culture is used as the reference for a new tuberculosis test, and culture's Se (~80% in that context) and Sp (~99%) are known, one can adjust the new test's observed sensitivity upward to account for the ~20% of true cases missed by culture (Rutjes et al., 2007).

Algebraic correction methods have been applied in fields like epidemiology and occupational health, often termed "adjusting for an alloyed gold standard" (Brenner, 1996). They are straightforward to apply and computationally trivial, which historically made them popular. In practice, when the reference's error rates are accurately known, these corrections can substantially improve the accuracy estimates. For example, in HIV diagnostics, early evaluations of rapid tests used composite reference standards; by applying a misclassification correction (assuming the composite missed certain acute cases), researchers obtained Se estimates that better matched long-term clinical follow-up (Rutjes et al., 2007).

Compared to unadjusted analysis, the corrected values tend to be closer to what one would get if a true gold standard were available (as evidenced in simulation studies and case studies) (Umemneku Chikere et al., 2021). However, the empirical success of these methods hinges on correct input values. In some real-world studies, using a slightly wrong assumption about reference Se has led to over-correction, producing implausible estimates (e.g., Se >100% or negative false rates) (Umemneku Chikere et al., 2021).

Emerson's bounding method has been demonstrated on biomarker studies: rather than picking a single "corrected" point, it provides a high-low range which successfully bracketed the truth in reported examples (Emerson et al., 2018). Users of the bounds approach have noted that while it does not give a tidy answer, it usefully conveys the uncertainty due to the imperfect standard.

Direct comparisons of different correction formulas (Gart vs. Brenner vs. others) show they algebraically agree under similar assumptions, so they tend to yield the same result when applied to a given dataset (Umemneku Chikere et al., 2021, Staquet et al., 1981, Gart and Buck, 1966, Brenner, 1996). Thus, the differences are more in presentation (bounds vs. point estimates) and in how they handle assumptions like dependency.

Thus, correction methods can be very accurate when their prerequisites are met, but their performance is entirely dependent on the quality of prior knowledge about the reference test. The primary limitation of these methods is the requirement for known reference standard characteristics. Often, we do not actually know the Se of the reference test with certainty—we may only suspect it's imperfect. Using a guessed value can introduce new bias. For instance, if the reference is less sensitive than assumed, the correction will under-adjust (leaving residual bias), whereas if the reference is better than assumed, one might over-adjust. Moreover, many correction formulas assume independence of errors between the index test and reference (Brenner, 1996, Umemneku Chikere et al., 2019, Staquet et al., 1981, Gart and Buck, 1966, Sherwin, 2022). This means they presume that the events of both tests missing the disease in the same patient are uncorrelated. In reality, that might not hold: for example, if both the index and reference test are less likely to detect milder disease, their errors will be correlated. Ignoring this (i.e., using a simple independent model) could underestimate the index test's sensitivity.

Methods like Emerson's bounding approach try to address this by not committing to a single set of assumptions, but traditional methods do not account for correlation (Emerson et al., 2018). Additionally, implementing these corrections in practice requires a consensus on the reference standard's performance; in some cases, experts disagree on how imperfect the reference is (e.g., different studies give different Se estimates). From a practical standpoint, these methods are easy to calculate but can be

conceptually challenging to explain to clinicians—the idea that the reported sensitivity "assumes a reference sensitivity of X%" needs careful communication.

Despite these caveats, when a reference standard's performance is well characterized (e.g., certain lab tests or long-standing diagnostics), correction methods provide a quick and useful bias adjustment. They are less useful when the reference standard is novel or poorly studied (since one has no reliable values to input), or when multiple reference tests are used without a clear "base" test to adjust for.

Albert (2009) vs. Direct Correction Methods

In contrast to the algebraic or bounding approaches described above, and different with what Umemneku Chikere et al. (2019) reported, Albert (2009) presents a latent class analysis framework for estimating test performance when no perfect reference is available. Rather than directly correcting a 2×2 table, this approach models the true disease status as an unobserved latent variable, simultaneously estimating each diagnostic test's Se and Sp. While it addresses many of the same concerns about imperfect references, this model-based methodology is not typically viewed as a direct misclassification correction—instead, it belongs to the broader family of latent class or Bayesian methods.

In conclusion, when a reference standard's Se and Sp can be reliably estimated, algebraic correction techniques can significantly improve the accuracy of the index test estimates. These methods, however, require careful attention to the assumptions made—particularly about the reference standard's performance and the independence of errors between tests. In scenarios where such assumptions are difficult to validate or where the reference standard's performance is unknown, broader latent class models (like Albert (2009)) may be more appropriate, though they follow a distinct methodology. Ultimately, choosing among these approaches depends on the availability and reliability of external reference data, the plausibility of assumptions about error independence, and the level of uncertainty acceptable in decision-making contexts.

3.1.1.4 No Perfect Reference Standard

3.1.1.4.1 Latent Class Models and Bayesian Approaches (Multiple Imperfect Tests).

The challenge of evaluating diagnostic tests in the absence of a perfect reference standard is a common and longstanding issue in diagnostic research. When no single test can be trusted as the truth, researchers may have to rely on multiple imperfect tests (including the index test) and aim to infer each test's accuracy parameters—Se and Sp—without a definitive gold standard. Latent class analysis (LCA) offers a principled statistical framework for this scenario (Umemneku Chikere et al., 2019). The essential idea is to regard the true disease status as latent, treating all observed test results merely as fallible indicators of an unobservable construct.

A classical formulation of LCA posits a binary latent variable indicating disease presence or absence. The observed data consist of multiple test outcomes—often modelled as Bernoulli variables conditional on the latent disease class (Hui and Walter, 1980). Under basic assumptions such as CI, the model can be identified with two or more tests applied to two or more populations with distinct disease prevalences (Collins and Huynh, 2014). This yields estimates for each test's Se and Sp, as well as disease prevalence. The logic is that each test measures the same underlying condition, but all are subject to errors.

However, the classical latent class approach (often termed the Hui–Walter paradigm) is limited if key assumptions, notably CI, are violated. Tests may share correlated errors—for instance, if two molecular assays fail in patients with low bacterial load (Vacek, 1985). Models must then allow for dependence parameters to capture these correlations explicitly (Gardner et al., 2000). Additionally, many LCA formulations assume that the test performance (i.e., Se and Sp) remains constant across different subpopulations, which may be unrealistic for highly heterogeneous populations.

Researchers have tackled these limitations with a variety of extended and more flexible models. Some adapt the basic latent class framework by adding covariance terms or random effects for correlated tests (Collins and Huynh, 2014). Others incorporate covariates—such as age or disease stage—to account for variability in test accuracy across subgroups. When only two or three tests are available, such expansions can lead to non-identifiability, since the number of parameters often exceeds the effective degrees of freedom (Albert, 2009). In practice, analysts may resort to either collecting

data from additional distinct populations or making further assumptions (for example, that one test is conditionally independent of others).

Bayesian formulations provide further avenues for handling model complexity and uncertainty. Rather than seeking a single best point estimate, a Bayesian model yields posterior distributions for sensitivities, specificities, and disease prevalence. This allows the analyst to incorporate prior knowledge—whether from previous studies or expert opinion (Dendukuri and Joseph, 2001, Menten et al., 2008)—and to explore uncertainty with Markov Chain Monte Carlo (MCMC) methods. In principle, Bayesian LCA can reduce sensitivity to questionable assumptions, especially if well-chosen priors constrain parameter space (Collins and Huynh, 2014). However, strong priors risk imposing external beliefs on the data, and one must carefully check how priors influence posterior inferences.

CI states that, given the latent disease status, the test results are mutually independent. This assumption greatly simplifies LCA but can be unrealistic. Many tests share correlated biases—particularly those based on similar physiological markers—or rely on similar measurement instruments. Violations of CI can distort accuracy estimates (Umemneku Chikere et al., 2019). Several extended LCA models relax CI by introducing additional parameters that capture correlations between tests—often described as covariance or random effects approaches (Gardner et al., 2000). Yet each added parameter typically requires more data or more assumptions to preserve identifiability.

Systematic reviews highlight the usefulness of LCA in veterinary diagnostics, microbiology, and other fields where multiple imperfect tests are the norm (Collins and Huynh, 2014). For instance, Baughman et al. (2008) compared a composite reference standard with a latent class approach in pertussis diagnosis, discovering that the composite reference standard misclassified some cases. LCA revised the estimated sensitivities upward for certain PCR-based tests. Similarly, Schumacher et al. (2016) demonstrated that culture—a common reference for paediatric tuberculosis—might miss clinically important subsets of TB patients, inflating the apparent Sp of other tests when culture is assumed perfect.

Albert (2009) shows how LCA can incorporate external estimates or hierarchical structures, emphasising that, unlike algebraic correction methods, these models do not treat one test's error rates as known. Instead, they jointly estimate disease prevalence

and each test's accuracy, thereby disentangling misclassifications from all tests simultaneously. This can reduce biases from relying on a single imperfect reference but hinges on explicit assumptions about dependence structures, priors, or the distribution of latent variables.

LCA-based results can be highly sensitive to model specification, sample size, and the number of tests. Adding more tests, each with distinct error patterns, can improve identifiability and precision but increases computational burdens. Models that allow test accuracy to vary across subpopulations (for example, different clinical sites) or across the disease spectrum can introduce further parameters that are difficult to estimate reliably (Collins and Huynh, 2014). Bayesian implementations help by encoding assumptions in priors, but model non-identifiability remains a problem if data are insufficiently rich.

Despite these concerns, latent class models and Bayesian expansions remain invaluable for managing the no-gold-standard scenario. They permit the estimation of test accuracies without assuming that any single test is ground truth. When multiple imperfect tests are available, LCA—supplemented by either frequentist or Bayesian methods—can offer more realistic estimates than naive approaches that treat one test as perfect. The key is careful modelling, clarity about assumptions (notably CI), and, ideally, robust sensitivity analyses to assess how results shift under alternate dependency or prior structures (Umemneku Chikere et al., 2019).

Despite its theoretical appeal, latent class analysis has seen limited uptake in diagnostic HTA due to several practical challenges. The method's critical assumption of conditional independence between tests—that test results are independent given true disease status—is frequently violated when tests share common error sources or are methodologically correlated, potentially biasing accuracy estimates (Dendukuri and Joseph, 2001). Model identifiability issues arise when attempting to estimate more parameters than observable data patterns support, requiring subjective prior distributions that can influence results (Albert and Dodd, 2004). Additionally, LCA requires specialized statistical software and Bayesian expertise not universally available in HTA contexts, while data requirements (multiple tests on same patients, substantial sample sizes) often exceed what published accuracy studies provide. These implementation barriers help explain why simpler algebraic adjustment methods or

careful study selection may be more pragmatic alternatives for many diagnostic evaluations.

In summary, latent class analysis has become a go-to strategy in diagnostic evaluation when a perfect reference standard is lacking. Building on early formulations (Hui and Walter, 1980), modern LCAs handle a variety of complexities, from correlated errors to hierarchical priors. Nonetheless, the reliability of these methods depends on the plausibility of the assumptions and the availability of data that can support them. In contexts where test performance cannot be sensibly assumed constant, or where test outcomes are clearly interdependent, LCA must be extended or combined with additional design elements (such as multiple populations or more tests). By integrating the insights from Albert (2009) and Collins and Huynh (2014), one can better appreciate how LCA provides a principled alternative to direct misclassification correction methods, albeit at the cost of increased model complexity and the need for strong interpretive caution.

3.1.1.4.2 Composite Reference Standards and Expert Panel Diagnosis.

An alternative approach to dealing with the absence of a perfect reference standard is the use of composite reference standards (CRS) or consensus (expert panel) diagnoses. Instead of relying on a single test to define true disease status, multiple imperfect tests are combined via logical rules such as AND or OR. If a patient tests positive on any of the constituent tests, an OR-rule composite labels them as diseased, whereas an AND-rule composite requires all tests to be positive. Variations, such as the k-of-n rule, attempt to balance Se and Sp (Dendukuri et al., 2018). In paediatric tuberculosis, for instance, researchers sometimes label a patient a “true TB case” if they have clinical signs plus a positive X-ray or lab test, effectively mixing AND and OR criteria.

Composite references standard can increase the Se (or Sp) relative to any single test by designating fewer (or stricter) criteria for disease classification. Where multiple imperfect tests each address different shortcomings, combining them might produce a stronger proxy for true disease status. Consequently, when evaluating a novel test, the composite reference standard (rather than a single flawed test) serves as the benchmark. These methods have been widely applied in areas ranging from cardiology—

where ECG, biomarker levels, and clinical symptoms may define myocardial infarction—to infectious diseases such as endocarditis (Duke's criteria) or paediatric tuberculosis (Dendukuri et al., 2018, Umemneku Chikere et al., 2019). Another non-algorithmic approach, known as expert panel or consensus diagnosis, assembles a group of clinicians to review all the available evidence, effectively a human-driven composite reference standard (Umemneku Chikere et al., 2019). By pooling multiple data sources, panels aim to approximate true disease status as accurately as possible.

Nevertheless, composite reference standards introduce their own complexities. Since these rules merely combine multiple imperfect tests, they do not guarantee a correct classification. An OR-rule may yield high Se but over-call disease (labelling some healthy individuals as "diseased"), leading to an underestimated Sp for the index test (Naaktgeboren et al., 2016). Conversely, an AND-rule typically misses mild or borderline cases, artificially lowering an index test's apparent sensitivity if the composite fails to count individuals who only test positive on one component. Investigators must be mindful of such dynamics, as composite reference standards can inadvertently exaggerate or mask the true performance of the new diagnostic.

Potential Biases:

- **Incorporation Bias**: If the index test is incorporated into the composite, the reference outcome is no longer independent. This artificially inflates agreement, often resulting in overestimated Se or Sp (Dendukuri et al., 2018). Good practice excludes the index test from the composite rule.
- **Rule-Selection Bias**: The choice of AND vs. OR, or the specific weighting in k-of-n, can have large effects on measured accuracy. A rule that skews toward high Se inevitably lowers Sp, and vice versa. Researchers may inadvertently introduce bias by selecting one rule over another.
- **Context Specificity**: A composite that works well in a tertiary-care environment might not translate to a different clinical setting with fewer resources or different patient characteristics. Reproducibility and external validity become issues if the underlying tests or their operating thresholds differ.
- **Subjectivity in Expert Panels**: While comprehensive, consensus panels can be inconsistent if the mix of experts, knowledge of the index test results, or weighting

of individual tests is not standardised. Inter-panel variation can be substantial (Bertens et al., 2013).

Black et al. (2002) combined DNA probe and NAAT tests as a new “improved reference,” thus capturing infections that would be missed by culture alone. With this more sensitive composite, the apparent sensitivity of other tests rose since fewer genuine positives were missed. Another example was provided by Baughman et al. (2008), who showed that a composite based on clinical criteria plus lab tests misclassified some individuals, leading LCA to produce higher Se estimates for certain PCR assays.

These examples highlight that composite reference standards can improve upon a single flawed gold standard by merging different tests. Yet, as demonstrated by simulation and real-data analyses, composites may distort an index test’s measured accuracy, depending on how their criteria inflate or suppress positivity rates (Naaktgeboren et al., 2016). For disease states with nuanced manifestations—like systemic lupus—expert panels can approximate truth more comprehensively than single tests, but they remain subjective and prone to bias if not carefully blinded.

There are some limitation regarding with the use of these methods. Composite reference standards often do not quantify the uncertainty in their own classification—they are treated as “final” truth, ignoring that each constituent test has partial Se/Sp. Overlooking this can lead to overconfidence in derived accuracy estimates (Dendukuri et al., 2018). Furthermore, if the rule is overly permissive (OR rule), many healthy subjects test positive. If too strict (AND rule), true cases are missed. Methods like dual composite reference standards (dCRS) attempt to bracket accuracy by comparing the index test against two extremes—one highly sensitive, one highly specific—yielding performance bounds (Tang et al., 2018). Still, these remain more practical heuristics than robust statistical corrections.

Composite and consensus reference standards offer a practical compromise when multiple imperfect tests are available and no single reliable gold standard exists. By combining sources of evidence, they can capture more true positives (or negatives) than individual tests, sometimes reducing misclassification. However, their own limitations—including potential incorporation bias, rule-driven distortions of Se and Sp, and

subjectivity—mandate careful interpretation. Many authors emphasise validating composites or complementing them with additional analyses (e.g., latent class analysis or sensitivity analyses) to gauge how biases might influence final estimates (Umemneku Chikere et al., 2019). While composites remain a popular solution in real-world diagnostic evaluation, their results should not be accepted uncritically: the rule logic, context, and potential bias sources all require scrutiny to ensure that the apparent improvement in accuracy genuinely reflects a closer approximation of the truth.

In summary, composite and consensus reference standards can be pragmatic solutions in the absence of a perfect gold standard, but they are far from panaceas. They reduce some biases but introduce others, and they rely on crucial assumptions about test complementarity and completeness. Investigators should interpret composite-based estimates with an awareness of these limitations and, where feasible, compare or combine them with more formal modelling approaches such as latent class analysis to gauge the robustness of conclusions.

3.1.1.4.3 Other Indirect and Hybrid Methods.

Beyond the principal categories of managing imperfect reference standards, researchers have devised additional techniques that do not yield classical Se and Sp estimates but can still offer valuable evidence on a test's potential utility. Discrepant resolution (discordance analysis) is one such approach, wherein the index test is first compared with an initial reference test. Any discordant results (index test positive, reference negative or vice versa) are then subjected to a third "resolver" test assumed to represent the truth for those specific cases (Umemneku Chikere et al., 2019). Historically, this procedure was adopted to resolve disagreements, often by employing an expert panel or more invasive verification. However, it is recognised to introduce bias, because only discordant samples receive further verification, potentially inflating the new test's apparent Se or Sp (e.g., if the resolver test is prone to confirming positives in borderline cases). Although some modifications, such as verifying a random sample of concordant results, have been proposed to reduce bias, discrepant resolution remains problematic and is now discouraged by many experts (Umemneku Chikere et al., 2019, Hadgu, 1996).

Another indirect method involves evaluating inter-test agreement, rather than accuracy against an established gold standard. Here, one uses statistics such as Cohen's kappa to gauge how closely the new test's results match those of an existing test (Umemneku Chikere et al., 2019, Zaki et al., 2012). Although high agreement could suggest that both tests measure the same underlying signal, it does not guarantee correctness—two tests might agree but both miss mild cases. Agreement studies can thus be used as preliminary or supportive evidence but do not constitute a robust measure of accuracy by themselves.

Researchers have also investigated the test positivity rate in different populations, comparing, for instance, the proportion of positive results among symptomatic patients versus healthy controls. A much higher positivity rate among symptomatic individuals may imply that the test correlates with disease status, although differences in prevalence or population characteristics can confound this inference. As a result, positivity rates are typically used only in exploratory contexts and must be interpreted cautiously (Umemneku Chikere et al., 2019).

In the absence of a reliable reference, some investigators resort to alternative study designs. One is the case-control study, which preselects diseased cases (perhaps determined via follow-up or clinical consensus) and disease-free controls, allowing estimation of the test's discriminatory capacity without per-individual verification (Kohn et al., 2013). However, case-control designs risk spectrum bias, since the diseased group may be more severe, and controls disproportionately healthy, artificially inflating accuracy measures. Another possibility is follow-up (delayed) verification, whereby one verifies disease status over time, observing who develops the condition or shows progression (Rutjes et al., 2007). This approach, while more realistic for progressive diseases, can be logistically demanding and subject to intervening clinical actions (e.g., treatments) that affect disease manifestation. Analytical validity studies constitute a final indirect approach, focusing on whether the test detects the correct analyte or target under controlled conditions, rather than clinical disease. This strategy, frequently used in laboratory medicine or assay development, can confirm that the test is technically measuring what it purports to, but it does not establish clinical Se or Sp in real-world contexts.

All these strategies—discrepant resolution, agreement indices, positivity rates, and specific study designs—offer partial, indirect insights into test performance and typically provide weaker evidence than more rigorous methods. Their main utility is often exploratory or supplementary, bridging gaps until more definitive evaluations are feasible (Umemneku Chikere et al., 2019). Discrepant analysis, in particular, has been shown repeatedly to produce inflated accuracy estimates, so its use has markedly declined. Agreement measures like kappa fail to indicate which test is correct; positivity rate comparisons do not quantify Se or Sp. Meanwhile, case-control designs can yield overoptimistic accuracy estimates because of atypical distributions of diseased and healthy individuals, and delayed verification remains prone to confounding by treatment or dropout. Despite these caveats, indirect and hybrid approaches can be informative when integrated carefully into a broader evidence synthesis—for instance, initially testing an assay's basic detection capability in vitro, confirming it in a carefully chosen case-control sample for a preliminary signal of clinical validity, and subsequently performing delayed follow-up to gauge real-world performance. Yet, as emphasised by (Umemneku Chikere et al., 2019), these methods should not replace more robust bias-correction strategies or comprehensive economic evaluations. Instead, they serve as interim or complementary solutions, highlighting areas warranting additional research and informing whether large-scale or more rigorous designs are needed.

Addressing imperfect reference standard bias requires a toolkit of statistical methods and a careful study design. The literature offers an array of solutions – from bias-correction formulas and missing data approaches to sophisticated latent class models and pragmatic outcome-based validations – each with their own uses, strengths, and pitfalls. Empirical studies and systematic reviews show that no single method is universally best; rather, the choice depends on context (e.g. whether an alternative reference or multiple tests are available, whether assumptions like CI are tenable, and how results will be used in practice) (Rutjes et al., 2007). Often, a combination of methods can provide complementary insights. For example, one might report primary accuracy estimates from a composite reference standard, but also perform a latent class analysis as a sensitivity check, and perhaps adjust for any verification bias if not all patients were verified. Transparent reporting of assumptions and uncertainties is

essential, as emphasized by modern guidelines and authors like Reitsma et al. (2009) and Dendukuri et al. (2018).

For decision-makers – whether clinicians interpreting a test or policy-makers weighing cost-effectiveness – understanding these methods is crucial. By recognizing how each method tries to correct or circumvent reference standard imperfections, and being aware of the remaining biases, one can make more informed judgments about a diagnostic test's true value. In economic evaluations, explicitly integrating these methodological insights (for instance, using bias-adjusted accuracy estimates or conducting uncertainty analyses around them) can prevent suboptimal decisions that might arise from flawed data. The field continues to evolve, with ongoing research into hybrid models, improved composite definitions (like the dual composite approach (Tang et al., 2018)), and the interplay between diagnostic accuracy and clinical outcomes. Ultimately, the goal is a robust evaluation of diagnostic tests that accounts for the real-world fact that the "gold standard" is often far from pure – and to do so in a way that supports sound clinical and economic decision-making.

3.2 The Economic Evaluation Perspective

The integration of diagnostic tests into health systems demands careful attention not only to their clinical value but also to their economic impact. Economic evaluations—commonly in the form of cost-effectiveness analyses—seek to establish whether a new diagnostic adds sufficient patient benefit (for instance, measured in QALYs) relative to any increased costs. In the context of diagnostics, however, a major complication arises from the measurement of test accuracy: many studies assume a perfect "gold standard" against which the new test is judged. When that benchmark is itself imperfect, the Se and Sp estimates for the index test may be biased, leading to skewed outcomes in the resulting cost-effectiveness model.

Researchers and HTA bodies have increasingly recognised this problem, as it has the potential to distort decisions about adopting or rejecting a test. While methodological advances exist for adjusting or mitigating this bias (Trikalinos and Balion, 2012, Xie et al., 2017, Xie et al., 2022c, Soares et al., 2018), their uptake in real-world decision-making remains uneven. The following sections review the key literature on economic

evaluations in the context of imperfect reference standards, exploring the consequences of bias, possible strategies for correction, and the current guidance from HTA agencies. The discussion references multiple frameworks (Severens and van der Wilt, 1999, Sutton et al., 2008, Laking et al., 2006, Soares et al., 2018, Snowsill, 2023, van der Pol et al., 2021) and case examples (Xie et al., 2017, Johnson et al., 2019, Yang et al., 2019, Novielli et al., 2010), culminating in a critical assessment of persistent gaps in both methodology and practice.

3.2.1 Imperfect Reference Standards and Bias: Consequences for Cost-Effectiveness

The presence of an imperfect reference standard introduces fundamental uncertainty in defining true disease status, which is the linchpin of measured test accuracy. When a reference test fails to identify all diseased individuals (lacking perfect Se) or mistakenly classifies healthy individuals as diseased (lacking perfect Sp), the observed performance of the index test against that reference deviates from its true performance in detecting or ruling out the disease (Trikalinos and Balion, 2012). This misestimation may be small in some contexts but can be significant for diseases with incomplete reference processes (for instance, in microbial culture, where prior antibiotic use can suppress growth, or in imaging, where the chosen reference is also subject to reading errors).

Such distortion affects cost-effectiveness in multiple ways. First, if the index test's Se is underestimated, the economic model may conclude that fewer true cases are detected at an earlier stage, thus diminishing the perceived health gains and cost offsets attributable to the new diagnostic. The analysis might then appear to show an unfavourable incremental cost-effectiveness ratio (ICER). Conversely, overestimating Se can lead to an overly optimistic appraisal, allowing a test with limited real benefit to appear cost-effective. Similar logic applies to Sp, whose misestimation alters the number of false positives, with downstream consequences for unnecessary treatments and associated costs. Because most diagnostic economic models rely on these parameters to simulate patient outcomes and resource use, any error can propagate significantly, particularly when the cost or health implications of false negatives or false positives are considerable (Severens and van der Wilt, 1999, Laking et al., 2006).

Case examples confirm that these biases are not theoretical curiosities but have practical impact. Xie et al. (2017) documented how a flawed reference for pneumonia diagnosis—namely, culture—underestimated the true performance of a new urinary antigen test. Had the economic evaluation proceeded with naïve Se and Sp figures, the test might have been deemed less cost-effective than it truly was. Their adjustment using a Bayesian latent class meta-analysis yielded more accurate input parameters and consequently a more favourable incremental cost per correctly diagnosed case. Likewise, (Johnson et al., 2019) explored a latent tuberculosis screening test scenario in which the reference was partial clinical follow-up, leading to missed subclinical infections. By acknowledging that references can fail to capture all true TB cases, the authors demonstrated that strategies using a "supposedly moderate-accuracy" screening tool were better than expected once the reference's limitations were factored in. These examples illustrate how missed disease (or mislabelled healthy individuals) in the reference can tip an evaluation's conclusion about whether or not a diagnostic is cost-effective.

3.2.2 Integrating Adjusted Accuracy into Economic Evaluations

When an imperfect reference is acknowledged, several strategies can be used to refine the accuracy inputs in an economic model, thereby producing a more valid representation of cost and outcome implications. While detailed statistical foundations have been laid out in preceding sections, a summary here provides context for their economic significance.

3.2.2.1 Direct Statistical Adjustment

Some authors employ relatively direct methods if the reference standard's own Se and Sp can be reliably estimated. Algebraic correction formulas (Gart and Buck, 1966, Staquet et al., 1981, Brenner, 1996, Sherwin, 2022) "unmix" the observed 2×2 table of test results by applying known misclassification rates of the reference. If, for instance, the reference is believed to have 90% Se and 98% Sp, one can adjust the index test's apparent performance to reflect a true disease distribution. This correction is

straightforward when independence assumptions hold and the reference's performance is well characterised by external data. However, it becomes less reliable or even unworkable if the errors in the reference are correlated with errors in the index test, or if the reference standard's accuracy is itself unknown or disputed (Trikalinos and Balion, 2012).

3.2.2.2 Latent Class and Hierarchical Models

Where more extensive data exist (such as multiple tests all applied to the same cohort), analysts can turn to latent class analysis or hierarchical meta-analysis to estimate each test's accuracy without presupposing a perfect reference (Collins and Huynh, 2014, Xie et al., 2017). These models treat the true disease status as unobserved, using the patterns of agreement and disagreement among tests to infer their respective sensitivities and specificities. Once derived, these less-biased accuracy estimates feed into cost-effectiveness models. Although conceptually appealing and supported by methodological exemplars (Sutton et al., 2008, Novielli et al., 2013b), they demand higher-level statistical expertise and patient-level data for multiple tests, limiting their routine application in HTA contexts pressed for time and resources.

3.2.2.3 Scenario Analyses and Sensitivity Testing

HTA bodies often advocate scenario analyses to explore how robust cost-effectiveness findings are to different assumptions about reference standard imperfection (Soares et al., 2018, Yang et al., 2019, van der Pol et al., 2021). For instance, one scenario might assume the reference standard is truly 100% sensitive, another might assume it fails to catch 5% of cases, while a further scenario ascribes a plausible distribution of reference test errors. By recalculating cost-effectiveness outcomes in each scenario, one can see how crucial the reference standard's imperfection is to the final decision. Probabilistic sensitivity analyses can also incorporate distributions reflecting uncertainty in the reference's accuracy. Although these approaches can be less statistically rigorous than full latent class models, they are more feasible and transparent to policy committees, enabling them to judge whether the new test's value remains stable or collapses under an alternative assumption.

3.2.3 Literature on Economic Frameworks and Bias

3.2.3.1 Early Developments and Foundations

A foundational review by Severens and van der Wilt (1999) surveyed diagnostic economic evaluations from the mid-1990s, identifying general methodological shortcomings and emphasising that many studies failed to mention key elements such as analytical perspective or the need for sensitivity analysis. Their critique set the stage for subsequent frameworks that began to incorporate the unique nature of diagnostic tests, including their dependency on an accurate reference.

Laking et al. (2006) took this further by proposing a conceptual piece titled "The economics of diagnosis," which pointed out that diagnostic tests fundamentally partition patients into disease-positive and disease-negative streams, each incurring distinct costs and outcomes. They showed how cost-effectiveness depends intimately on test performance (Se, Sp), which in turn is typically measured against a gold standard. Although Laking et al. (2006) did not specifically address imperfect references, their recognition that "the dependence on a gold standard can be illusory" foreshadowed the need to handle inaccuracy in the reference itself.

Novielli et al. (2010) provided a systematic appraisal of UK HTA reports that used diagnostic models, concluding that many real-world analyses either did not synthesise accuracy evidence properly or relied on simplistic assumptions. They found that the majority of reports used independent meta-analytic estimates of Se and Sp with no correlation and no recognition of reference standard limitations. The authors argued that ignoring such nuances can lead to suboptimal decisions, a point reinforced by later simulation and case studies (Novielli et al., 2013a).

3.2.3.2 Evolving Frameworks

By the late 2000s, Sutton et al. (2008) introduced a unified Bayesian approach to combine diagnostic accuracy meta-analysis with decision-analytic modelling, promoting consistent propagation of uncertainty. This approach allowed for correlated parameters and different data sources. While their example did not explicitly incorporate

reference standard error, it set a platform to do so if reference standard metrics were themselves uncertain.

More recently, Soares et al. (2018) proposed an analytical framework for valuing diagnostic tests, devoting specific attention to the scenario of no perfect reference. They suggested that analysts might (1) use latent variable methods, (2) evaluate the test by patient outcomes directly, or (3) model the reference standard's own inaccuracies if partial data exist. Soares et al. (2018) acknowledged that these expansions add complexity but contended that ignoring the problem is a bigger risk to decision quality. Snowsill (2023) then highlighted that few consensus guidelines exist for systematically adjusting for imperfect references in cost-effectiveness models, meaning modellers often improvised by applying scenario analyses or partial corrections.

van der Pol et al. (2021) offered a structured approach to health-economic evaluations of diagnostic technologies, emphasising the importance of describing pathways and referencing all relevant downstream consequences. Their report indicates that an imperfect reference standard is a known issue in economic evaluations, although their main focus is on transparency and comprehensiveness in reporting how test accuracy was ascertained. They urge explicit mention of any "alloyed gold standard," which remains an area of known methodological tension.

3.2.3.3 Guidance from HTA Agencies

Studies such as Yang et al. (2019) and Snowsill (2023) underscore that official agencies, including NICE in the UK and CADTH in Canada, have begun incorporating dedicated diagnostic programmes within their appraisal processes. These agencies are aware that biases from imperfect references can undermine cost-effectiveness estimates. NICE, for instance, instructs Evidence Review Groups to carefully appraise the accuracy data, checking if the supposed gold standard truly warrants that title. If doubts arise, NICE suggests that model assumptions about accuracy be tested thoroughly (NICE, Diagnostics Assessment Programme manual) (National Institute for Health and Care Excellence, 2022a, National Institute for Health and Care Excellence, 2022b). CADTH similarly acknowledges that if no accepted gold standard exists, analysts must either justify their chosen reference or incorporate plausible errors in that reference (CADTH,

2019). While neither agency demands a single uniform approach, both strongly encourage scenario analyses, thorough risk-of-bias assessments, and transparency about any uncertain reference standard performance.

The NICE Diagnostics Guidance for various specific technologies often features explicit references to the possibility that the reference standard might have missed cases or introduced differential verification. In such instances, committees might request further research or consider a broader range of accuracy inputs. CADTH's instructions for companion diagnostics highlight that if the reference standard itself is subject to misclassification, sensitivity analyses must reflect that (CADTH, 2019). This combination of mandated transparency and recommended scenario exploration underscores that agencies are cognisant of reference standard bias, even if they stop short of obliging the use of advanced correction models in every appraisal.

3.2.4 Critical Analysis of the Methods-Policy Gap

Despite the existence of methods that address reference standard bias, an evident gap persists in real HTA submissions and published economic evaluations. Analysts frequently use accuracy estimates from the published literature without re-evaluating whether the reference standard was imperfect. They may rely on a single 2×2 table or meta-analysis that itself treated the reference as infallible. This situation arises from multiple practical barriers.

One barrier is data availability: adjusting for reference standard bias often requires additional data on that reference's Se and Sp or on multiple tests used concurrently. Many published studies do not present such information. Another barrier is time constraints: re-analysing raw data or fitting advanced Bayesian latent class models can exceed the resources or time allocated for a typical HTA submission. A further obstacle is the complexity of interpretation: committees, which often include non-statisticians, can find simple scenario analyses more transparent than a fully fledged latent class re-estimation.

As a result, while the theoretical impetus to correct for bias is strong (Sutton et al., 2008, Xie et al., 2017), the real practice is often to note the imperfection, present a scenario or two that might account for it, and then proceed with the main base-case analysis

assuming minimal or moderate imperfection (Novielli et al., 2010, Yang et al., 2019). The consequence is a somewhat conservative approach: agencies do recognise that the reference might not be perfect, but they do not always invest in complex re-analyses. This leaves open a possibility that cost-effectiveness conclusions are not fully reflecting the test's true performance, especially if the bias is large or systematically pushes accuracy in one direction.

Another phenomenon is that committees might weigh other forms of evidence, such as patient outcomes or real-world observational data, to temper the risk of bias. If a test is found, in observational contexts, to reduce hospital admissions or improve patient satisfaction, that might offset concerns about whether the reported sensitivity was slightly under or overestimated. These intangible or qualitative factors can sway the final recommendation in ways that do not appear purely data-driven. Hence, the gap between theory and practice is not simply a matter of ignorance; it is shaped by pragmatic considerations of evidence and the multi-factorial nature of HTA decisions.

3.2.4.1 Synthesis of Methods and Relevance to Economic Evaluation

A closer examination of the correction techniques described above reveals important strengths and weaknesses that are often underestimated in the economic evaluation of diagnostic tests. For instance, Staquet et al. (1981) propose a relatively straightforward formulaic adjustment for Se and Sp when the reference standard's accuracy is known, which can be implemented with minimal data requirements. However, this approach depends heavily on the assumption that the reference standard's misclassifications occur uniformly across all patients, which is not always the case in more complex clinical contexts. Brenner's method, by contrast, accommodates conditional dependence (CD), thereby recognising that the iRS can induce correlated errors in test classifications. Although Brenner's approach is more flexible conceptually, it requires more detailed data inputs, such as estimates of covariance terms. Latent class models take a different route, dispensing with a single definitive reference altogether and instead inferring 'true' disease status probabilistically from multiple indicators. While this can be a powerful strategy when several tests are available, it raises further concerns about

model convergence, prior assumptions, and the complexity of interpreting posterior probabilities in a policy-making environment.

Real-world HTAprocesses illuminate how difficult it can be to apply these methods in practice. Although guidance from entities such as the National Institute for Health and Care Excellence (National Institute for Health and Care Excellence, 2022a, National Institute for Health and Care Excellence, 2022b) highlights the need for rigorous diagnostic evidence, numerous HTA submissions still rely on data generated via an imperfect or unverified reference standard. In one recent NICE appraisal of an imaging-based tool for diagnosing a rare neurological disorder, for example, the absence of a universally accepted confirmatory test meant that the underlying cost-effectiveness model depended on suspiciously high accuracy estimates derived from a reference standard of uncertain validity. The appraisal committee explicitly acknowledged the potential for misclassification bias to inflate the perceived value of this diagnostic innovation, but did not issue detailed guidance on how to correct or quantify that bias within the model. This illustrates a broader systemic gap: while some methodological correction approaches exist (Staquet, Brenner, latent class analyses), their integration into economic modelling remains sporadic, and there is no standard requirement to evaluate how iRSs might alter cost-effectiveness results.

In identifying these methodological limitations and practical examples, this chapter directly contributes to the first objective of this thesis, namely to review and critically appraise the literature on dealing with iRSs in both statistical and economic frameworks. By highlighting the existing gaps between proposed correction approaches and real-world HTA practice, the discussion further clarifies why new or more systematic methods are needed. This analysis sets the stage for the conceptual framework introduced in Chapter 4, which aims to offer a structured way of incorporating bias corrections into decision models, and for the simulation study in Chapter 5, which provides a direct empirical demonstration of how misclassification can drive unwarranted variance in estimated cost-effectiveness outcomes.

3.2.5 Future Directions: Addressing and Reducing the Gap

Scholars and HTA specialists continue to advocate for bridging the methodology-policy gap. One proposed strategy is to incorporate reference standard bias as a standard dimension of risk of bias in systematic reviews that feed into HTA (Trikalinos and Balion, 2012). In that event, meta-analyses of diagnostic performance would come with an indication of how uncertain the estimates might be, based on the chosen reference. This uncertainty could then be integrated into the economic model via probabilistic sensitivity analysis. Another approach, championed by Xie et al. (2022c), is to publish user-friendly scripts or toolkits for latent class meta-analysis, enabling HTA analysts to re-estimate accuracy fairly quickly if multiple test data are available. Over time, as these tools become more accessible and computationally efficient, their usage might become more commonplace.

Furthermore, future primary research can be designed to mitigate these biases at source. For instance, investigators might incorporate composite or multi-stage references that catch more true cases, or verify a subset of negative cases thoroughly, ensuring that the final accuracy data is less prone to missing disease. Soares et al. (2018) point out that robust trial and observational designs, combined with thorough follow-up, can yield near-gold standard evidence that reduces the need for complicated post-hoc corrections. Coupled with broader changes in reporting—such as the new QUADAS iteration that explicitly addresses partial verification and differential reference standards—the field may gradually shift towards better data that obviates or minimises the risk of major reference standard bias.

3.3 Discussion

In critically examining the range of methods available for addressing iRS bias—covering verification bias corrections, composite and consensus references, latent class models, and algebraic adjustments—this chapter confirms that flawed benchmarks in diagnostic accuracy studies pose significant methodological challenges with far-reaching implications for economic evaluations. Misclassifications introduced by iRSs jeopardise the validity of measured Se and Sp, creating potential distortions in cost-effectiveness results. This concern is neither hypothetical nor trivial; the examples from Xie et al. (2017)

on pneumonia and Johnson et al. (2019) on latent tuberculosis provide persuasive empirical evidence that seemingly small misalignments in reference accuracy can shift economic conclusions in substantive ways. Recent frameworks (Sutton et al., 2008, Soares et al., 2018) and HTA guidance (National Institute for Health and Care Excellence, 2022a, National Institute for Health and Care Excellence, 2022b, CADTH, 2019) acknowledge these issues, advising analysts to adopt scenario analyses or more sophisticated modelling approaches where feasible.

Despite notable methodological advances, routine practice often lags behind theory. Economic evaluations in real-world HTA processes frequently rely on simple acceptance of Se/specificity estimates as though the reference standard were perfect, with only limited scenario testing for potential errors. High computational demands, lack of raw data, time constraints, and the need for transparent reporting collectively dampen the widespread application of advanced methods such as Bayesian latent class analyses. Many assessments, as a result, settle for partial corrections or scenario-based exploration rather than undertaking thorough re-estimations of test accuracy. This shortfall underscores a tension between methodological rigour and pragmatic feasibility in health technology assessments.

Furthermore, while latent class models have received considerable academic attention, highlighting their capacity to estimate a test's true performance in the absence of any definitive reference, economic evaluations seldom incorporate these fully. Their uptake is constrained by the availability of multiple test data, complex assumptions on CI, and the need for advanced statistical expertise to implement and interpret results. Composite or consensus reference standards similarly provide a partial solution, but they introduce other biases—especially if the index test inadvertently influences the composite outcome or if the OR/AND rule artificially inflates or deflates measured accuracy. The literature thus portrays a field in transition: essential awareness of the ramifications of iRSs coexists with the partial adoption of statistical remedies.

In addition to these methodological issues, there are practical and organisational barriers that help explain the slow uptake of latent class analysis in routine HTA. Latent class models can be difficult to specify transparently, their assumptions (especially CI and identifiability) are not straightforward to validate from the available evidence base, and outputs such as posterior class probabilities can be harder to communicate to non-

technical stakeholders. Given typical HTA timelines and requirements for auditable modelling, analysts may legitimately prefer simpler approaches (e.g., comparative accuracy evidence, careful study selection, and scenario analyses) unless the decision is highly sensitive to the assumed disease status definition and the data support more complex modelling.

Beyond latent class methods, simpler adjustment formulas (Gart and Buck, 1966, Brenner, 1996, Staquet et al., 1981, Sherwin, 2022) offer a more direct way to correct observed Se and Sp if estimates of the reference's own accuracy can be gleaned from external data. Yet these formulas remain largely absent in cost-effectiveness studies, possibly because few economic modellers are familiar with them, or because reference accuracy is unknown, or because correlation in errors between the index and reference violates the independence assumptions needed for straightforward algebraic corrections. There is thus a particular vacuum in the economic literature around the application of direct correction methods—such as those by Staquet et al. (1981)—despite their theoretical simplicity. While the required input data (the reference test's known or assumed misclassification rates) may be available, analysts rarely incorporate them into the final model.

The consistent theme emerging from this review is the mismatch between methodological potential and real-world usage. On one hand, the academic discourse, as shaped by works like Albert (2009), Collins and Huynh (2014), and Umemneku Chikere et al. (2019), showcases robust techniques that can recast biased accuracy estimates into more realistic parameters for cost-effectiveness modelling. On the other, much of routine HTA remains bound by constraints of time, data scarcity, and pragmatic interpretability. The immediate implication is that decision-makers should interpret economic evaluations of diagnostics with caution if the underpinning accuracy data stems from potentially flawed references. Those evaluations that fail to engage in any form of bias correction or thorough scenario testing risk under- or overestimating a technology's value.

From this vantage point, the chapter underscores opportunities for future work. First, systematically incorporating direct correction approaches (e.g., Staquet or Sherwin formulations) into economic models could be a pragmatic middle ground for contexts in which advanced latent class analyses exceed resource limitations. Formalising these

approaches—complete with uncertainty distributions around reference standard Se/Sp—could allow modellers to approximate corrected diagnostic parameters with fewer assumptions than full latent class frameworks. Second, structured reporting guidelines, building on tools such as QUADAS-2 (Whiting et al., 2011) or the NICE Diagnostics Guidance manual (National Institute for Health and Care Excellence, 2022b), might require authors to detail whether reference standard inaccuracies were considered and, if so, how they affected model inputs. Greater transparency and standardisation could then prompt more consistent application of bias adjustments.

Finally, as the thesis moves forward to explore the use of correction methods such as Staquet et al. (1981) in economic evaluations, the path is clear. While latent class analysis has become a widely discussed approach, the direct algebraic corrections remain underrepresented in HTA. Investigating how these simpler algorithms perform in real-world cost-effectiveness scenarios—and whether they can feasibly align with the data constraints of typical diagnostic evaluations—will be a central focus. By testing these correction methods within a simulation or case study framework, subsequent chapters aim to provide a more practical blueprint for analysts seeking to mitigate iRS bias without overly complex modelling.

In conclusion, iRS bias represents a foundational concern in the accuracy assessments that underpin diagnostic economic evaluations. Despite a growing body of literature demonstrating both the nature of the bias and the means to address it, implementation gaps remain. This review, by mapping the statistical solutions, economic implications, and observed practices, sets the stage for the next chapters, where the role of algebraic correction methods (like Staquet's) in economic models will be systematically investigated. The aim is to determine whether these methods can achieve a balance between rigour and feasibility, thereby helping to close the theory-practice divide and improve the reliability of cost-effectiveness evidence for diagnostic technologies.

This chapter extends the discussion of diagnostic test evaluation by focusing on the specialised problem of iRSs. It surveys the statistical literature on handling scenarios in which the definitive disease status is uncertain, and it also examines how (and whether) these approaches have been incorporated into health economic analyses. In doing so, this chapter fulfils a key part of the first objective, which is to review and critically

appraise existing methods. It also lays the groundwork for the conceptual framework presented in Chapter 4 by highlighting the research gap that the framework aims to fill.

Chapter 4

4 Impact of Imperfect Reference Standard Bias on Decision-making in Healthcare: Insights for economic evaluations with single index test related to an imperfect reference standard with known accuracy

4.1 Introduction

The evaluation of diagnostic tests plays a crucial role in healthcare decision-making, influencing both clinical management and resource allocation. The reliability of these evaluations is predicated upon the accuracy of diagnostic tests, which is typically assessed by comparing the test in question to a reference standard. Ideally, a diagnostic test's Se and Sp are estimated against a gold standard (GS), which represents the true disease status without error. However, in many clinical settings, a perfect GS is unavailable, and an iRS must be used instead. This limitation introduces a form of systematic error known as iRS bias, which distorts test accuracy estimates and, consequently, has significant implications for health economic evaluations (Begg, 1987, Kohn et al., 2013, Schmidt and Factor, 2013).

The impact of iRS bias on diagnostic accuracy is well-documented, particularly in relation to the dependencies between test errors. When the errors of a new diagnostic test and the reference standard are conditionally independent, the bias tends to underestimate Se and Sp (Kohn et al., 2013). Conversely, when the errors are correlated, the bias may either inflate or deflate Se and Sp estimates depending on the nature and direction of the correlation (Vacek, 1985, Gardner et al., 2000). Positive dependencies, such as those observed in serological tests with cross-reactivity, can artificially elevate accuracy estimates, whereas negative dependencies, though less common, may lead to underestimation (Rifkin, 1995, Lu et al., 2018). Understanding and adjusting for these

dependencies is critical to ensuring robust diagnostic evaluations and subsequent economic assessments.

Decisions on whether to adopt a diagnostic test hinge on accurately translating its Se and Sp into estimates of cost, resource use, and patient outcomes. When the reference standard contains misclassifications, even small inaccuracies can magnify as they propagate through an economic model, potentially reversing a test's apparent cost-effectiveness. Without addressing this bias, decision-makers risk endorsing diagnostic interventions on skewed data, particularly in contexts where small changes in Se or Sp drive substantial changes in patient management. This chapter thus details the conceptual foundation for how reference standard errors can misalign a model's parameters with true disease status, underscoring the critical importance of correcting—or at least quantifying—such biases in any robust health technology assessment.

Despite the recognised effects of iRS bias in diagnostic accuracy studies, its influence on CEA remains an underexplored area in health economics. CEA is a key methodological tool in HTA, guiding the allocation of healthcare resources by comparing costs and health outcomes across competing interventions (Briggs, 2006, Soares et al., 2018). Within this framework, diagnostic test accuracy is a fundamental parameter that affects disease prevalence estimates, treatment decisions, and ultimately, the calculation of incremental cost-effectiveness ratios (ICERs). If diagnostic accuracy is systematically biased due to an iRS, the entire economic evaluation framework may be compromised, leading to suboptimal or even harmful decision-making (Denwood et al., 2024).

A central challenge in incorporating diagnostic accuracy into economic evaluations is ensuring alignment between test performance measures and the reference standard used in modelling health outcomes. Mismatches occur when diagnostic evaluations rely on case definitions derived from an iRS, leading to discrepancies between estimated and true disease probabilities. This misalignment can distort cost-effectiveness comparisons and obscure the relative value of diagnostic interventions (Janes et al., 2011, Ferrante di Ruffano et al., 2012b). To address this issue, some researchers advocate for adjustment methods that account for iRS bias within economic models

(Snowsill, 2023). However, the development and systematic evaluation of such approaches remain limited.

Several statistical correction methods have been proposed to adjust Se and Sp estimates for iRS bias when the accuracy of the reference standard is known. These include proportion-based equations (Gart and Buck, 1966, Brenner, 1996) and patient-count data approaches (Staquet et al., 1981). While these methods have been applied in diagnostic accuracy studies, their integration into economic evaluations remains largely unexplored. Umemneku Chikere et al. (2021) conducted one of the few studies assessing correction methods, demonstrating that while Staquet's method performs well under CI assumptions, it is less robust in scenarios with correlated test errors. However, their analysis was limited to diagnostic accuracy outcomes, without extending to economic implications.

The primary aim of this chapter is to develop a systematic framework for integrating correction methods into health economic evaluations. This framework extends previous work on diagnostic accuracy adjustments (e.g., Umemneku Chikere et al. (2021)) by considering their applicability within CEA and exploring their impact on decision-making processes in HTA. Specifically, this chapter seeks to:

- Examine the theoretical foundations of correction methods for adjusting Se and Sp in the presence of an iRS.
- Develop a structured framework for integrating these adjustments into economic evaluations of diagnostic tests.
- Assess the implications of correction methods on cost-effectiveness outcomes, with a focus on decision consistency and robustness in HTA.
- Identify key challenges and limitations in applying these methods within real-world healthcare decision-making.

By addressing these objectives, this chapter contributes to the growing body of literature on improving the methodological rigour of economic evaluations involving diagnostic tests. The proposed framework aims to facilitate more accurate and reliable decision-making in healthcare by mitigating the biases introduced by iRSs. In doing so, it aligns with broader efforts to enhance the transparency and validity of HTA methodologies

(Soares et al., 2018, Snowsill, 2023). The following sections will first provide an overview of existing correction methods before presenting the proposed framework and discussing its theoretical and practical implications.

4.2 Rationale and need for this framework

Economic evaluations of diagnostic tests must be grounded in accurate estimates of test performance to ensure reliable decision-making in healthcare. However, current approaches often fail to adequately address the systematic bias introduced by iRS. This oversight leads to misestimated diagnostic accuracy, ultimately affecting disease prevalence calculations, cost-effectiveness estimates, and healthcare resource allocation.

While several statistical correction methods have been developed to adjust for iRS bias (e.g., Gart and Buck (1966), Staquet et al. (1981), Brenner (1996)) , their application in economic evaluations remains underutilised and inconsistent. The absence of a structured approach results in variable methodological choices across studies, making it difficult to compare findings and derive generalisable policy recommendations (Umemneku Chikere et al., 2021). Moreover, CEA frameworks often assume that diagnostic test accuracy is either measured against a perfect RS or that any imperfections do not substantially impact economic outcomes. These assumptions are problematic in real-world settings, where diagnostic errors can propagate through decision models, leading to misinformed policy choices (Denwood et al., 2024, Snowsill, 2023). The lack of methodological consensus on handling iRS bias in economic evaluations is a critical gap that must be addressed.

This chapter proposes a structured framework to integrate correction methods systematically into economic evaluations. This framework aims to:

- Enhance methodological consistency across HTA studies by providing a standardised approach to incorporating diagnostic accuracy adjustments.
- Improve the validity of cost-effectiveness estimates by ensuring diagnostic test accuracy is systematically corrected for iRS bias.

- Support better healthcare decision-making by providing policymakers with robust, bias-adjusted economic evaluations that reflect real-world diagnostic performance.

By filling this methodological gap, the proposed framework will ensure that economic evaluations more accurately reflect the true value of diagnostic interventions. This is particularly crucial for healthcare systems that rely on iRSs, where failing to adjust for bias can lead to inefficient resource allocation and suboptimal patient outcomes.

4.3 Understanding the imperfect reference standard bias: conceptual definitions and mechanism of the bias.

iRS bias occurs when a diagnostic test is evaluated against a reference that does not perfectly reflect the true disease status. Unlike a GS, an iRS introduces systematic misclassification errors that affect Se and Sp estimates (Begg, 1987, Kohn et al., 2013). The extent of this bias depends on the accuracy of the iRS itself and whether the misclassification errors are conditionally independent (CI) or dependent (CD) in relation to the test being evaluated (Gardner et al., 2000, Vacek, 1985).

The direction of this bias depends on whether the test errors are assumed to be independent. If the errors are independent but conditional on the true disease status, the bias will underestimate Se and Sp. If the errors are correlated, the Se and Sp will increase if the covariances are positive. (Kohn et al., 2013). For instance, two serological tests detecting antibodies for an infectious disease may have positively correlated errors if both tests exhibit cross-reactivity to a similar pathogen, leading to simultaneous false positives (Gardner et al., 2000, Kohn et al., 2013). In addition, if the errors are inversely correlated (negative dependencies), the bias will underestimate the Se and Sp (Vacek, 1985, Rifkin, 1995). Negatively correlated errors are uncommon but can occur under specific circumstances. For instance, as blood flows passively through the mitral valve to fill the heart's main pumping chamber before contraction, the rate at which the inflow velocity decreases is closely linked to the size of the valve area. This decline in velocity is routinely measured using Doppler ultrasound. Typically, a smaller valve area

corresponds to a slower rate of velocity decline. However, if the heart muscle becomes thickened and stiff, the inflow velocity may decrease more rapidly despite a narrowed valve area. This is because the stiffened chamber walls create a sharply rising resistance to blood flow as the chamber fills, potentially leading to an overestimation of the valve area.

Since thickened heart muscle is also associated with mitral valve calcification, errors in measuring the mitral valve area using this method can exhibit negative correlations with other methods. In this case, the thickened muscle acts as a hidden confounding factor linking the errors in these diagnostic approaches (Rifkin, 1995). Another example might occur when two tests rely on complementary mechanisms; for example, a molecular test (e.g., PCR) and a culture-based test for bacterial detection might exhibit negatively correlated errors, where one method is prone to false positives due to detecting non-viable bacteria while the other has false negatives due to difficulties growing certain strains (Vacek, 1985, Lu et al., 2018).

Correlation or dependencies of test errors might exist when tests share similar methodologies or biological processes (Gardner et al., 2000, Lu et al., 2018, Vacek, 1985). However, determine the direction and magnitude of dependencies can be challenging if it is not estimated as part of the analysis of diagnostic studies. Gardner et al. (2000) explores the relationship between conditional covariance and kappa values, as this last one depend on the prevalence, Se/Sp and their covariance. The lack of relationship between both measures does not allow the use of kappa as a proxy of conditional covariance.

Neglecting whether CI can be assumed can lead to misdiagnosis or inappropriate treatment decisions (Collins and Huynh, 2014, Gardner et al., 2000, Vacek, 1985, Rifkin, 1995).

4.3.1 Implications for economic evaluation

The presence of iRS bias can substantially impact economic evaluations by distorting key parameters that CEAs. Some major implications include:

- Misestimation of Disease Prevalence: Systematic misclassification leads to incorrect prevalence estimates, affecting baseline risk calculations used in decision models.
- Unreliable ICERs: If diagnostic accuracy is biased, the estimated costs and health benefits associated with different interventions may be miscalculated, leading to flawed ICER estimates.
- Suboptimal Resource Allocation: When economic evaluations fail to account for iRS bias, healthcare systems may allocate funding to interventions that appear cost-effective due to biased test performance estimates rather than their true effectiveness.
- Impact on Treatment Decisions: Biased diagnostic accuracy affects downstream clinical decisions, leading to inappropriate treatment recommendations and potentially lower overall health system efficiency.
- Variability in HTA Outcomes: Without standardised approaches for adjusting test accuracy, different studies may reach inconsistent conclusions regarding the cost-effectiveness of the same diagnostic strategy.

Given these consequences, it is critical to develop systematic approaches to integrating correction methods into economic evaluations to improve the reliability of healthcare decision-making.

4.4 Correction methods to adjust single dichotomous diagnostic test with an imperfect reference standard with known sensitivity and specificity.

The presence of an iRS in diagnostic evaluations necessitates the use of statistical correction methods to mitigate bias in Se and Sp estimates.

Correction methods are algebraic derivations developed to deal with the iRS bias and designed to be used on dichotomous diagnostic test related to an iRS with known accuracy. For these methods to be feasible two conditions must be met (i) the Se and Sp of the iRS need to be known and assuming it is known with certainty, and (ii) assumes that the index test is conditionally independent of the iRS.

Numerous correction methods have been developed to adjust for misclassification errors when the iRS has known accuracy. These methods broadly fall into two categories: (i) proportion-based methods and (ii) patient-count data methods. While the literature includes a wide range of correction techniques, this chapter focuses on methods that are widely applied in HTA and CEA, computationally feasible within economic models, and transparent in their application.

Although multiple correction methods have been proposed in the literature, this chapter focuses on those that are widely used in economic evaluations and can be feasibly incorporated into cost-effectiveness models. Many alternative correction methods, such as those proposed by Gart and Buck (1966), Staquet et al. (1981), Brenner (1996), Sherwin (2022), Emerson et al. (2018), and Trikalinos and Balion (2012), provide valuable theoretical contributions but are either computationally intensive or lack direct integration into economic models. Sherwin's algebraic correction, based on Youden's Index, offers an alternative approach to adjusting for iRS bias but has not yet been empirically validated through simulation studies. Similarly, Emerson et al.'s framework focuses on defining performance bounds rather than providing direct corrections applicable in economic evaluations. Trikalinos and Balion's recommendations emphasise concordance statistics rather than direct Se/Sp correction, making their integration into cost-effectiveness models more complex. Moreover, the authors have not presented the equations in their article.

The correction methods examined in this chapter are selected based on their widespread application in diagnostic accuracy research and their adaptability within decision-analytic modelling frameworks. Specifically, these methods include proportion-based and patient-count data correction techniques, such as those developed by Gart and Buck (1966), Staquet et al. (1981), Brenner (1996).

Gart and Buck (1966) was the first on provide equations to correct the apparent Se and Sp using a proportion-based approach. Then, Staquet et al. (1981) developed equations following a patient-count approach. The derivation for both Gart and Buck and Staquet equations were not reported by the authors. Nevertheless, Sherwin (2022) due to lack of evidence about whether these equations were correct, he derivate his own equations and then demonstrated its equivalence with Gart and Buck and Staquet approaches. Brenner (1996) adopt an alternative perspective to derivate equations in order to

estimate the apparent Se and Sp from the true Se and Sp and its respective alloyed gold standard from a case-control study. The aim was studying the effect of the misclassification bias of the alloyed gold standard. The proposed equations were obtaining for scenarios assuming CI and for positive CD, reporting its respective derivation. We will refer as Brenner CI with the equations designed when CI is assumed, and Brenner CD to the equations with positive CD structures. Trikalinos and Balion (2012) mentioned that it is possible to correct accuracy due to iRS bias, but no equations nor derivations was reported.

To date, Umemneku Chikere et al. (2021) is the only study has assessed the performance of any of these correction methods. Despite the authors discussed three methods – Gart and Buck, Staquet, and Brener – They decided to assess Staquet and Brenner, because of they considered to Gart and Buck approach equivalent to Staquet. The evaluation was made using simulation studies and empirical data analysis across three clinical datasets. The analysis was performed assuming that test errors are conditional independence and conditional to the disease status but also considering dependencies structures. The simulations reveal that Staquet's method outperforms Brenner CI approach when CI is assumed, but falters with extreme disease prevalence or CD. Despite the efforts to assess the methods on three scenarios and varying the prevalence on one of them, these analyses provide limited information that represent the different situations that a CEA would face. In addition, Umemneku Chikere et al. (2021) utilise Brenner's equations without made any reorder, as this equations aims to estimate the apparent Se (Sp) and not the corrected accuracy.

The Table 7 below summarises the key correction methods for iRS bias, highlighting their theoretical foundation, assumptions, and applicability.

Table 7. Summary of the correction methods selected.

Method	Category	Key Assumptions	Advantages	Limitations
Gart & Buck (1966)	Proportion-based	Assumes CI between test and iRS	Simple algebraic correction	Inaccurate if test and iRS errors are correlated
Staquet et al. (1981)	Patient-count	Requires count data for false positives/negatives	More precise in well-structured datasets	Limited applicability when patient-

Method	Category	Key Assumptions	Advantages	Limitations
				level data is unavailable
Brenner (1996) CI	Proportion-based	Assumes CI but allows flexibility in parameter estimation	Useful for case-control studies	Cannot accommodate conditional dependencies
Brenner (1996) CD	Proportion-based	Allows for positive CD	More adaptable to real-world data	Increased model complexity

4.4.1 Limitations of correction methods in economic evaluations

While correction methods help mitigate iRS bias in diagnostic accuracy, their integration into economic evaluations presents several challenges. A fundamental limitation is the requirement for precise estimates of the Se and Sp of the iRS, which may be difficult to obtain in real-world settings. Incorrect assumptions about the accuracy of the RS can lead to biased adjustments, potentially distorting rather than improving cost-effectiveness analyses.

Moreover, many correction methods assume CI between the diagnostic test and the iRS. When this assumption does not hold, the adjusted estimates may still be inaccurate. This limitation is particularly problematic in economic evaluations where misclassification errors influence disease prevalence estimates, treatment decisions, and ultimately, ICERs. Given the sensitivity of cost-effectiveness models to small changes in diagnostic accuracy, unverified assumptions can substantially alter reimbursement and policy decisions.

Additionally, patient-count-based correction methods often necessitate individual-level data, which may not always be available in economic evaluations, further restricting their applicability.

4.4.2 Applicability of correction methods in cost-effectiveness analysis

The use of correction methods in economic evaluations must be carefully considered in relation to their impact on decision-making. Correction techniques should be applied when the presence of an iRS is likely to introduce significant bias in diagnostic test

performance, particularly when misclassification errors have demonstrable consequences for treatment selection and long-term patient outcomes. However, their use should be justified based on the availability of accurate RS performance estimates and the feasibility of incorporating corrections within existing modelling structures.

In practical terms, corrected Se and Sp values should be integrated into disease prevalence estimates within decision-analytic models. Additionally, sensitivity analyses should be conducted using multiple correction methods to assess their influence on ICER variability.

It is also crucial to ensure transparency in the application of correction methods within HTA reports. Any adjustments made to diagnostic accuracy estimates should be clearly documented, along with the underlying assumptions and their potential impact on cost-effectiveness results. Failing to properly account for iRS bias may lead to misleading economic evaluations and suboptimal resource allocation. However, applying correction methods without adequately testing their assumptions can introduce additional bias rather than mitigate it. Economic evaluations should, therefore, consider both the necessity and the feasibility of applying correction techniques within their respective analytical frameworks.

4.5 Cost-effectiveness analysis of diagnostic tests.

In the economic evaluation of diagnostic tests, integrating critical parameters within the context of an iRS is essential. Comprehensive information is the cornerstone of health economic assessments, often requiring the synthesis of data from multiple sources (Briggs, 2006). Key parameters in the CEA of diagnostic tests include: (i) the probability of disease presence ($p(D^+)$, related to the true disease status), (ii) the probability of a specific test result given the disease status ($i.e, Se\ and\ Sp$), and (iii) the net benefits of treatment options contingent on the actual disease status ($NB(Treat|D), NB(No\ Treat|D)$).

Ideally, test accuracy is measured against a perfect RS that inform us the true disease status:

$$Se_{True} = P(T^+|D^+), \quad Sp_{True} = P(T^-|D^-)$$

Ensuring unbiased estimates of disease probability ($p(D^+)$) and ($NB(Treat|D)$, $NB(No\ Treat|D)$) – leading to accurate cost and health outcome evaluations. Alternatively, when evaluations rely on parameters conditional on an iRS rather than the true disease status (e.g., Se and Sp, pre-test probabilities, and payoffs), a consistent alignment of these parameters ensures unbiased results (Soares et al., 2018).

Thus, the parameters condition on an iRS are give by:

$$P(Pre-Test) = p(iRS^+)$$

$$Se_{app} = P(T^+|iRS^+), \quad Sp_{app} = P(T^-|iRS^-),$$

$$NB(Treat|iRS^+), \qquad NB(Treat|iRS^-)),$$

$$NB(No\ Treat|iRS^+), \qquad NB(No\ Treat|iRS^-))$$

4.6 Unaligned (mismatched) cost-effectiveness models

Denwood et al. (2024) highlighted that "mismatches" occur when diagnostic evaluations rely on case definitions derived from a reference test that does not align with the actual target condition, such as the true disease status. These mismatches introduce bias into Se and Sp estimates, particularly when the RS fails to accurately represent the disease state. Building on this foundation, we extend the concept to CEA by introducing the "aligned model," which ensures that all parameters—diagnostic accuracy, disease probabilities, and economic outcomes—are consistently harmonized under a common reference framework, whether derived from a perfect or iRS.

A key challenge arises when linking test results to broader comparative analyses, as complete information for all comparators is rarely available. For example, not all diagnostic strategies are evaluated using the same RS, disease prevalence is often uncertain, and strong assumptions may be required. Moreover, the lack of long-term outcome data, even when available, often prevents their integration into comparative evaluations across clinical settings (Soares et al., 2018, Janes et al., 2011). Practical constraints, such as limited time, budget, sample size, and ethical considerations, further complicate these analyses (Hot et al., 2021, Rick et al., 2020, Varse et al., 2019).

Consequently, the limited availability of test-treatment studies often obscures the relationship between diagnostic test results and patient outcomes (Ferrante di Ruffano et al., 2012a).

The validity of CEA results depends on the consistency of parameter conditions across the model. An "aligned model" ensures that all parameters—test accuracy, disease probabilities, and treatment outcomes—are conditioned on the same RS, whether it is a perfect RS (true disease status) or an iRS. This alignment is crucial because it avoids inconsistencies that can distort the comparison of diagnostic strategies.

In an “aligned” model, both clinical and economic components are expressed relative to the true disease state or the iRs. This allows each branch of the decision tree (or state transition probabilities in a Markov model) to reflect the actual outcomes. By contrast, in an “unaligned” model, key probabilities such as false-positive or false-negative rates and pay-offs are anchored to different disease states, resulting in a mismatch between how test accuracy, pre-test probabilities or prevalnce, and outcomes are defined. As Chapter 3 demonstrated, neglecting the difference between these two approaches can obscure the real magnitude of a test’s benefits or costs, especially when the RS systematically under- or over-diagnoses certain patient subgroups.

In an aligned model:

Internal Consistency Maintains Validity: When all parameters are conditioned on the iRS, the systematic biases introduced by the iRS are integrated consistently into both the accuracy estimates and the resulting outcomes (e.g., misclassified individuals are assigned treatments based on their incorrect classification). As a result, the relative cost-effectiveness comparisons between strategies remain valid, even though the absolute accuracy of the test may be biased.

Bias Cancelation: The misclassifications from the iRS affect both the estimated test accuracy and the downstream outcomes. This alignment ensures that any systematic bias cancels out when comparing strategies, as long as the same RS is used throughout the model.

In contrast, unaligned models, where parameters are conditioned on different statuses (e.g., test accuracy on the iRS and outcomes on the true disease status), introduce

discrepancies that lead to biased results. These mismatches create inconsistencies in the probability calculations, distorting the cost-effectiveness evaluations.

For example, consider a CEA comparing three diagnostic tests where no gold standard exists, and outcomes are synthesized relative to a shared iRS. Even if the iRS introduces misclassification, the health consequences of the classified individuals reflect this bias. Thus, the relative cost-effectiveness results remain unbiased because all parameters are aligned.

“Aligned models” address this issue by ensuring that all parameters are conditioned on the same status. In such models, cost-effectiveness results remain unbiased. Figure 8 illustrate an aligned model where all parameters are conditional on the true disease status. Another aligned scenario occurs when all parameters are conditional on an iRS (Figure 9).

In these cases, biased accuracy does not equate to biased cost-effectiveness results. For example, in a CEA comparing three diagnostic tests without a gold standard, if the outcomes are synthesized for both accuracy and health consequences relative to a shared iRS, the misclassification introduced by the iRS is consistently accounted for, ensuring unbiased CE results despite biased accuracy.

Figure 8. Stylized decision tree model for cost-effectiveness analysis of a new diagnostic test. The decision model using parameters conditional to the ***true-disease status.***

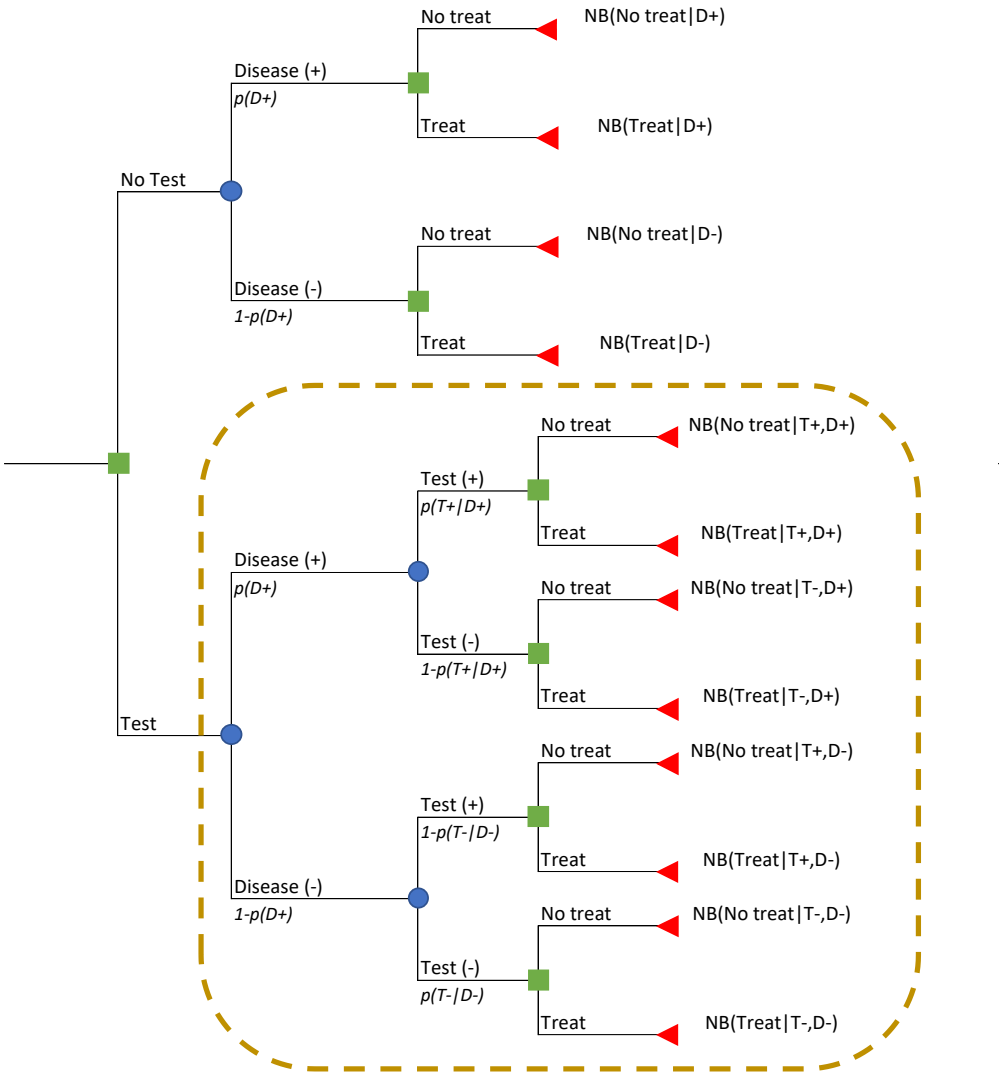
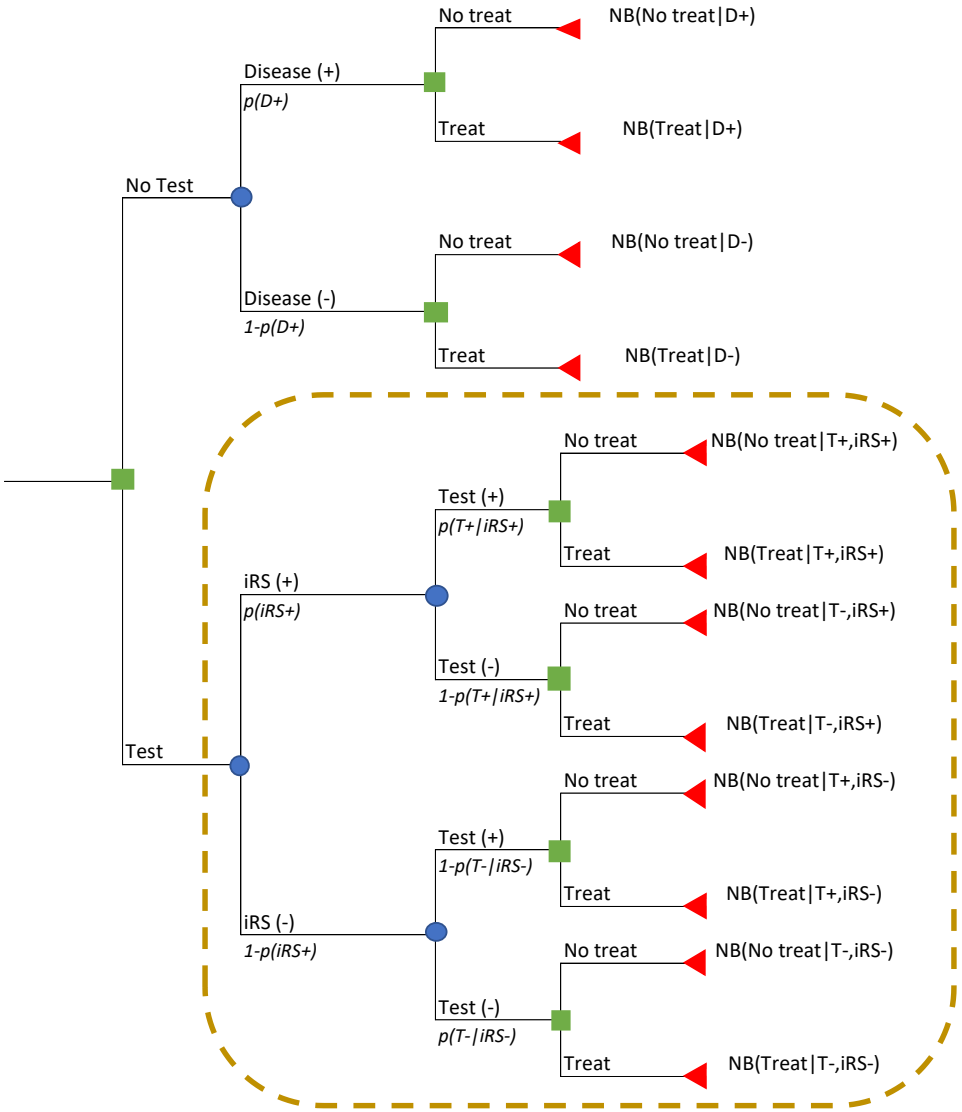


*Figure 9. Stylized decision tree model for cost-effectiveness analysis of a new diagnostic test. The decision model where parameters are conditional to the **iRS.***



In contrast, unaligned models, where parameters are conditioned on different statuses, produce biased CE results. Figure 10 illustrates two such scenarios: (a) test results based on the true disease status and outcomes based on an iRS, and (b) test results based on an iRS and outcomes based on the true disease status. These mismatches distort the probabilities and outcomes, resulting in biased evaluations. Adjusting methods, such as those shown in

Figure 11, correct test accuracy to align with the true disease status, resolving these inconsistencies. Adjusted probabilities, denoted with an asterisk (e.g., p(T|D)*), ensure accurate CEA results by harmonizing the parameters across the model.

Figure 10. Example of two unaligned model between probabilities and outcomes. A) unaligned model between the probabilities of having a test result conditioned to the true-disease status and outcomes conditioned to the iRS. B) unaligned model between the probabilities of having a test result conditioned to iRS and outcomes conditioned to the true-disease status.

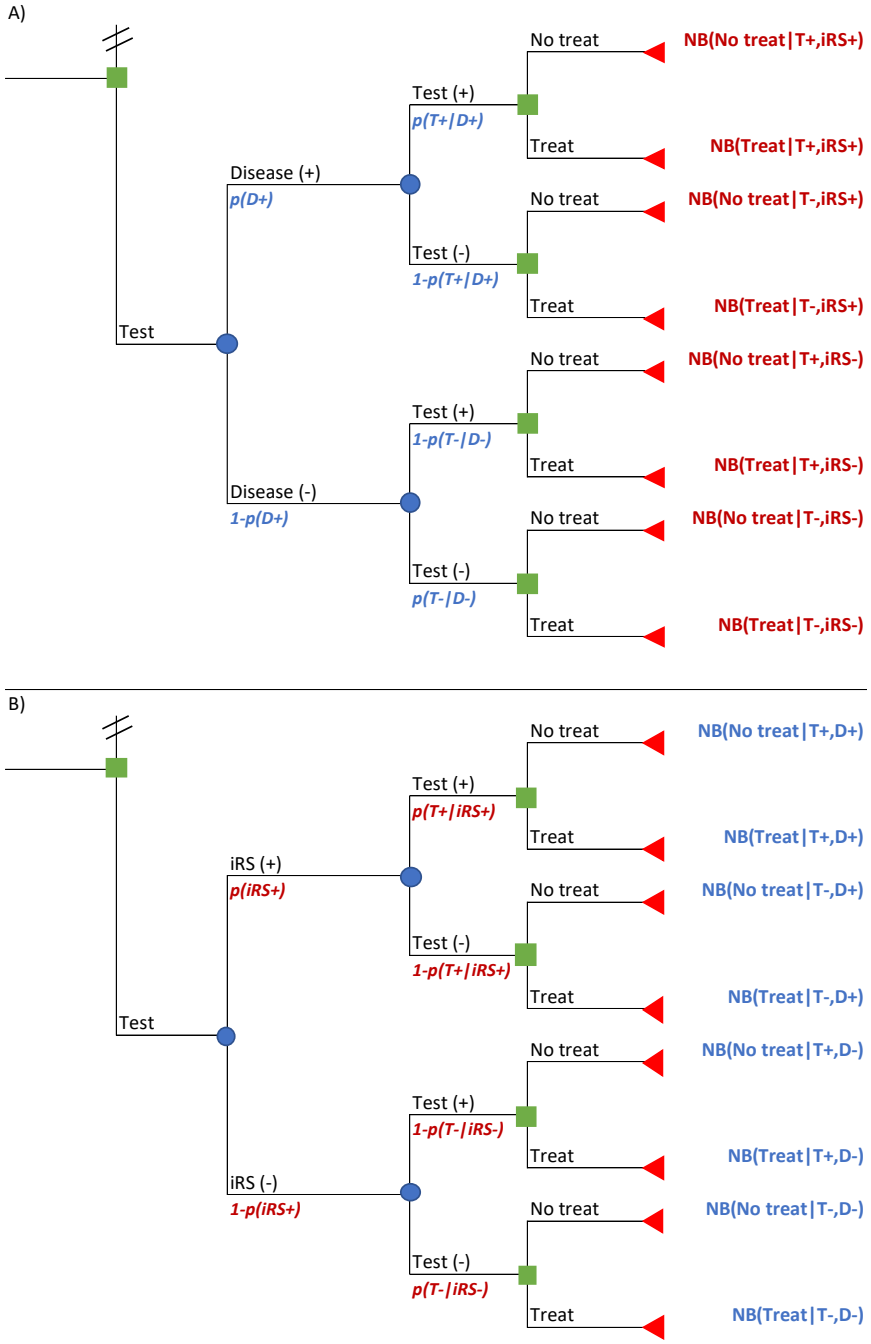
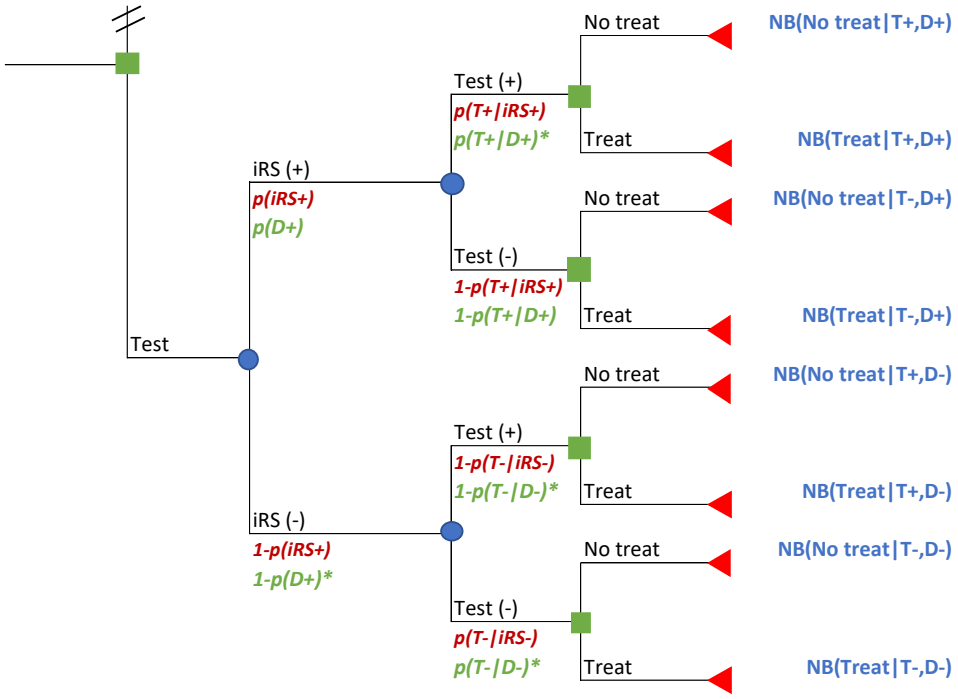


Figure 11. Decision model illustrating the role of adjusting method changing from an unaligned model to an aligned model. In red are represented the probabilities from the unaligned model, and in green the adjusted probabilities for the aligned model.



The concept of alignment in CEA of diagnostic tests is critical to ensuring unbiased results. By aligning all parameters—whether with a perfect RS or an iRS—models can account for inherent biases and produce valid comparisons of diagnostic strategies. Misalignment, however, introduces discrepancies that distort evaluations, necessitating the use of adjustment methods to correct for bias.

4.6.1 Theoretical foundation

Consider a simplified decision model for a diagnostic test that guides a treatment decision. When the diagnostic test is used to decide whether to treat, the expected NHB – when all parameters are aligned with the true disease status, is given by:

$$E[NHB] = p(D^+) \cdot [Se \cdot NB(Treat|D^+) + (1 - Se) \cdot NB(No\ Treat|D^+)] + \left(1 - p(D^+)\right) \cdot [(1 - Sp) \cdot NB(Treat|D^-) + Sp \cdot NB(No\ Treat|D^-)]$$

In this aligned model, both the diagnostic accuracy parameters and the outcome measures are defined with respect the true disease status *D*.

When a diagnostic test is evaluated against an iRS, we obtain "apparent" Se and Sp. These estimates are biased because the iRS does not perfectly represent the true disease status.

If one uses these apparent values directly in the decision model whilst the outcomes remain conditional on the true disease state *D*, the expected NHB becomes:

$$
\begin{aligned}
&E[NHB]_{misaligned} \\
&\quad = p(D^+) \cdot \left[Se_{app} \cdot NB(Treat|D^+) + \left(1 - Se_{app}\right) \cdot NB(No\ Treat|D^+)\right] \\
&\quad + \left(1 - p(D^+)\right) \\
&\quad \cdot \left[\left(1 - Sp_{app}\right) \cdot NB(Treat|D^-) + Sp_{app} \cdot NB(No\ Treat|D^-)\right]
\end{aligned}
$$

The bias in the NHB due to misalignment can then expressed as:

$$Bias = E[NHB]_{misaligned} - E[NHB]$$

If we assume that the difference between the true and apparent test parameters is $\Delta Se = Se_{app} - Se$ and $\Delta Sp = Sp_{app} - Sp$, and if we denote the benefit differences for patients with and without disease as:

$$
\begin{aligned}
\Delta NB^+ &= NB(Treat|D^+) - NB(No\ Treat|D^+), \\
\Delta NB^- &= NB(No\ Treat|D^-) - NB(Treat|D^-),
\end{aligned}
$$

Then, a first-order approximation of the bias is:

$$Bias \approx p(D^+) \cdot \Delta Se \cdot \Delta NB^+ + (1 - p(D^+)) \cdot \Delta Sp \cdot \Delta NB^-$$

Thus, the bias in NHB depends on the difference between the apparent and true Se and Sp, the prevalence, and the magnitude of the difference in net benefits between the treatment options.

In practical terms, if ΔSe and/or ΔSp are nonzero – that is, if the diagnostic accuracy is mismeasured- the resulting CEA will be biased.

4.6.1.1 Aligned model with parameters conditional on the imperfect reference standard

If the iRS is used, one can achieve an aligned model if all parameters are consistently defined with respect to the same references. For instance, if both diagnostic accuracy and outcomes are conditioned on the iRS, then the expected NHB is computed as:

$$
\begin{aligned}
E[NHB]_{iRS-aligned} \\
&= p(iRS^+) \\
&\cdot \left[Se_{app} \cdot NB(Treat|iRS^+) + \left(1 - Se_{app}\right) \cdot NB(No\ Treat|iRS^+)\right] \\
&+ \left(1 - p(iRS^+)\right) \\
&\cdot \left[\left(1 - Sp_{app}\right) \cdot NB(Treat|iRS^-) + Sp_{app} \cdot NB(No\ Treat|iRS^-)\right]
\end{aligned}
$$

If all the model inputs – both the accuracy measures and the outcome data – are consistently derived relative to the iRS, then even though the absolute values might be biased relative to the true disease state, the comparative cost-effectiveness results (e.g., incremental NHB between competing strategies) may remain unbiased because the systematic bias cancels out. This is what we refer to as an “aligned model”.

4.6.1.2 Correction methods: Bridging the gap

To convert and unaligned model (where diagnostic accuracy is measured on the iRS but outcomes on the true disease state) into an aligned one, correction methods are applied. As outlined in earlier sections, the true Se and Sp can be recovered from the apparent ones.

Once the corrected (aligned) parameters Se^* and Sp^* are obtained, they are used in the NHB calculation:

$$E[NHB]_{corrected}$$
$$= p(D^+) \cdot [Se^* \cdot NB(Treat|D^+) + (1 - Se^*) \cdot NB(No\ Treat|D^+)]$$
$$+ \left(1 - p(D^+)\right) \cdot [(1 - Sp^*) \cdot NB(Treat|D^-) + Sp^* \cdot NB(No\ Treat|D^-)]$$

Because these parameters are now aligned with the true disease state, the resulting CEA (i.e., the NHB) is unbiased relative to the true underlying clinical scenario.

4.7 The relevance of adjusting methods for diagnostic tests in healthcare decision-making

In the domain of diagnostic test accuracy, the influence of adjusting methods to mitigate bias from an iRS is substantial. These methods are integral not only for statistical correction but also play a critical role in guiding CEA. The impact of bias adjustment on healthcare decisions—whether it shifts the decision towards adopting or rejecting a diagnostic intervention—can vary significantly based on the analytical context.

For instance, in a scenario where the prevalence of the disease is low, a high false-positive rate associated with an iRS may lead to overestimation of the test's benefits due to the underestimation of the misclassified individuals, thereby influencing the decision to adopt it. Conversely, in a high-prevalence setting, the same false-positive rate may have a lesser impact on the decision-making process, as the test may still be considered beneficial due to a higher number of true positives. Although the bias-adjustment methods could lead to more nuanced decision-making in both contexts, the actual change in adoption or rejection of the test might be more pronounced in one setting over the other.

In situations where the bias-adjusted outcomes do not alter the adoption or rejection decision, the value of these methods lies in enhancing the precision of the data, supporting more accurate healthcare prioritization, and offering a clearer estimation of the health gains at the population level. Despite their importance, in the realm of CEA, the bias introduced by an iRS is often overlooked due to the absence of a widely accepted framework for addressing this issue. However, there is a growing body of literature where

some authors have put forth diverse perspectives, enriching the debate and slowly shaping the dialogue around this complex topic.

Soares et al. (2018) highlight that in the absence of a perfect RS, outcomes are intrinsically linked to test results, as the true disease status cannot be definitively known. They argue that this interdependence necessitates an analytical approach that thoroughly accounts for the entwined nature of test results and imperfections of the available RS. They recommend avoiding assumptions of independence between test results and outcomes, advising against using test accuracy data unless these dependencies are explicitly addressed. Their conservative outlook emphasizes the need for more precise tests and a rigorous analytical framework in cases where a perfect RS is unavailable.

Snowsill (2023) explores the same challenge but proposes an alternative modeling approach. While acknowledging Soares et al.'s perspective, Snowsill argues for linking outcomes directly to test results through latent disease state modeling. He critiques reliance solely on observed test results without considering the underlying latent disease state, emphasizing the importance of incorporating prognostic errors and systematically addressing biases to reflect true patient risks. This latent-state modeling approach ensures that biases and inaccuracies in test performance are captured, enabling more robust economic evaluations. His method aligns well with randomized test-treatment studies but requires assumptions about the latent disease state in the absence of direct data.

Thus, while Soares et al. (2018) advocate for cautious use of test accuracy data and focus on refining analytical frameworks for cases without a RS, Snowsill (2023) provides a complementary perspective by recommending latent disease state modeling as a way to integrate and correct for bias systematically. Together, these contributions highlight the necessity of addressing iRS bias to improve the reliability of healthcare decision-making.

4.8 Discussion

Chapter 4 addressed a critical gap in HTA by exploring how bias from iRSs can distort the CEA of diagnostic tests. This issue is of more than theoretical concern: cost-effectiveness models rely on accurate diagnostic input data, and if those data are biased by a faulty RS, the entire evaluation can be compromised. In the context of HTA, where decisions guide resource allocation, such bias risks leading to suboptimal or even harmful recommendations. In short, iRS bias matters because a diagnostic strategy evaluated against a flawed "gold" standard may misinform policy decisions, undermining the goal of maximising health benefits with limited resources.

Aligned vs. Unaligned Models – A Conceptual Distinction: The chapter's principal theoretical contribution is the introduction of an aligned versus unaligned model framework to formally describe how iRS bias enters cost-effectiveness modelling. In an aligned model, all parameters in the analysis – test accuracy, disease risk, and patient outcomes – are conditioned on the same definition of disease status. This could mean everything is defined against the true disease state (if a perfect reference or definitive clinical outcome data exist), or everything is defined relative to the same iRS. By ensuring internal consistency in this way, aligned models avoid the inconsistencies that otherwise distort comparisons between interventions. In practical terms, when all inputs share the same reference framework, any systematic misclassification affects both the estimated test performance and the projected health outcomes in tandem. Such bias may skew absolute outcome measures, but it cancels out in relative comparisons across strategies. As a result, within an aligned model the incremental cost-effectiveness of competing tests remains unbiased by iRS error – all strategies are essentially evaluated on a level playing field despite the imperfect yardstick.

In contrast, an unaligned model is one in which different parts of the evaluation are conditioned on different disease definitions (for example, test accuracy might be calculated against an iRS, while health outcomes are modelled on true disease presence, or vice versa). Chapter 4 demonstrated that such mismatches introduce probability inconsistencies that propagate bias through the model. When the test performance metrics do not align with the disease state underpinning outcome calculations, the cost and effect estimates become distorted. The model is effectively "mixing" incompatible information – akin to measuring speed in miles per hour but

distance in kilometres. As Denwood et al. (2024) put it, diagnostic evaluations that rely on a case definition from an imperfect test are prone to "mismatches" between observed and true disease status. Our aligned/unaligned framework extends this principle into the economic domain: it explains why a model that naïvely combines an iRS-derived accuracy with true disease-based outcomes (an unaligned approach) will yield biased cost-effectiveness results. Conversely, maintaining a single, coherent RS throughout the analysis (an aligned approach) prevents those mismatches and the ensuing bias. In summary, bias in test accuracy does not automatically equate to bias in cost-effectiveness results – it depends on how the model is structured. The aligned vs. unaligned distinction makes explicit the condition for unbiased evaluation: consistent alignment of all parameters with the same underlying disease definition.

This framework is a novel contribution that helps fill the methodological gap highlighted in recent literature. Denwood et al. (2024) provide a conceptual backdrop by warning that "all that glitters is not gold" – a supposedly gold-standard test that is imperfect will misrepresent true disease and thereby mislead interpretations of a new test's performance. Chapter 4 builds on this insight, demonstrating how such misrepresentation can further mislead economic evaluations if not properly handled. The aligned model concept ensures that the evaluation remains internally valid even when dealing with an imperfect reference, a point which aligns with broader efforts in HTA to improve methodological rigour in diagnostic test assessment.

Soares et al. (2018) notably stressed the challenges of valuing diagnostics when no perfect RS exists; they argued that analysts must not simply plug in accuracy data without accounting for the dependencies and uncertainties introduced by an iRS. They highlighted the importance of linking test results directly to patient outcomes within a decision modelling framework as a manner to address these challenges. Snowsill (2023) echoed these concerns and went further to suggest latent disease-state modelling, effectively tying patient outcomes to the unobserved true disease status rather than to observed (imperfect) test results. He cautioned that relying only on the observed test positives and negatives can be misleading if the RS is flawed, and advocated for models that incorporate the reference's imperfections explicitly. The framework presented in this chapter complements these perspectives by providing a unifying theoretical

structure for handling iRS bias. It offers a clear criterion for model integrity (alignment of parameters) that underpins approaches like those of Soares et al. (2018) and Snowsill (2023). In doing so, it answers recent calls for more transparent and principled methods to integrate diagnostic uncertainty into economic evaluations. Put simply, this chapter's contribution is to make plain why and when an iRS will bias a cost-effectiveness result, thereby equipping analysts with insight into how to correct or avoid such bias. This is an important theoretical advance with direct practical utility, given that previous reviews have found a lack of consensus or guidance on this issue.

The aligned vs. unaligned model framework provides a heuristic tool for practitioners conducting cost-effectiveness models of diagnostic tests. It emphasises that when dealing with test accuracy data derived from a study with an imperfect reference, one must either keep the entire analysis on that same evidential basis or adjust the data to approximate the true disease basis. In practice, if a perfect reference is unavailable, an "aligned-on-iRS" approach would use the imperfect reference consistently for defining who is "diseased" in both the accuracy estimates and in projecting health outcomes. This means the model's outcomes (such as life-years gained or cases detected) are interpreted in terms of the iRS's disease definition. The benefit of this approach is internal consistency – any misclassification by the RS affects treatment decisions and outcomes in the model just as it did in the clinical reality, so comparisons between strategies are fair. Indeed, when all strategies are assessed with the same flawed yardstick, relative cost-effectiveness rankings remain valid. However, the downside is that the analysis is then inherently evaluating a biased notion of disease; a more sensitive new test might appear to offer little benefit if the model "thinks" those additional true cases are false positives. Therefore, analysts often prefer to realign the model to the true disease state – essentially, to correct the accuracy inputs for RS errors so that outcomes can be modelled for actual disease. Our framework underscores that such realignment is not just an add-on but a necessity for an unbiased assessment if one wants to inform correct decision-making. Techniques to do this include using published Se/Sp of the reference test to adjust observed accuracy (e.g. with correction formulas) or employing latent class models that treat true disease status as an unknown variable to be inferred. The former correspond to what Soares et al. (2018) describe - modelling the reference's inaccuracies or using latent variables – while the latter represents an alternative

approach grounded in latent variable modelling, and what Snowsill (2023) recommends (linking outcomes to a latent true state). The aligned/unaligned framework provides a guiding principle behind all these methods: ensure that the decision tree in your economic model is not inadvertently mixing truths and quasi-truths. In practical terms, this could be implemented as a structural check when building decision-analytic models, and it could inform HTA guidelines. Notably, agencies like NICE and CADTH are increasingly aware of the need to incorporate diagnostic accuracy issues in their appraisals (Snowsill, 2023). The framework from this chapter offers a conceptual tool to improve such appraisals, helping modellers decide when a simple scenario analysis on test accuracy might suffice and when a more sophisticated bias correction is warranted. Ultimately, applying this framework can enhance the credibility and fairness of cost-effectiveness comparisons between diagnostic strategies in the presence of iRS.

While the aligned/unaligned model framework is theoretically robust, it rests on some simplifying assumptions that deserve critical reflection. Firstly, the chapter's analysis treats the Se and Sp of the reference test as if they were known quantities. In reality, we rarely know these parameters with certainty; they are themselves estimates subject to sampling error and context dependence. Assuming a fixed 90% Se for an iRS, for example, ignores the uncertainty or variability in that estimate. This is important because if the true sensitivity of the reference is lower or higher than assumed, our adjusted "true" accuracy for the index test would be off, potentially reintroducing bias. In practice, analysts should conduct sensitivity analyses on the assumed RS performance (as suggested by Soares et al. (2018)) or use Bayesian approaches to propagate uncertainty in the reference's accuracy through to the outcomes. For our conceptual framework, we held these values constant to clarify the structural impact of misalignment, but one should acknowledge that parameter uncertainty in reference test accuracy is an additional challenge in real applications. Secondly, our framework assumes that any errors in the RS are random (non-differential) misclassifications that apply uniformly. We did not explicitly model situations where the index test and reference might share correlated errors or where RS errors might depend on patient subgroups. If, for instance, an imperfect reference systematically misses the most severe disease cases (perhaps because those cases inhibit the test's detection), and a new test also struggles in those

cases, the biases might be more complex than a simple misalignment. Correlated errors can either exacerbate or mitigate bias in unpredictable ways. While the aligned model concept still holds (you must consistently handle the definitions), adjusting for bias in such scenarios may require more complex statistical modelling than the straightforward methods considered in Chapter 4. This is a limitation of the current framework: it assumes we can characterize the iRS bias with a couple of numbers (Se and Sp) and correct for it, which may be an oversimplification when dependencies exist. Another important consideration is that an analysis fully aligned on an iRS, by design, evaluates outcomes on the basis of that iRS's view of the world – a view that is imperfect. This ensures internal consistency but means the analysis might under-estimate the true health benefits of a better test. In an iRS-aligned model, a patient whom the reference misses is treated as if they truly do not have the disease, so the model might assign them no treatment and no adverse outcome when in reality that patient would suffer morbidity without treatment. A more sensitive test that correctly identifies such patients will, in the real world, improve outcomes, but an iRS-aligned model could fail to fully capture that benefit (it would regard those extra positives as false alarms). Thus, alignment on an iRS, while unbiased in a relative sense, does not equate to correctness in an absolute sense. The practical implication is that decision-makers should interpret iRS-aligned analyses with caution. Such an analysis can tell us which strategy is better given the evidence at hand, but it might not reveal the true magnitude of benefit. This limitation reinforces why adjustment toward the true disease state is ultimately desirable: whenever feasible, one should attempt to translate an iRS-based evaluation into one that reflects actual disease outcomes, either by collecting better reference data or by using the kind of modelling approaches noted above. In summary, the framework's assumption of full alignment as a solution is valid for avoiding internal bias, but it does not overcome the fundamental loss of information caused by not having a perfect reference. That gap can only be filled by further data or strong assumptions – a point to which we remain alert. Despite these caveats, the aligned vs unaligned model framework stands as a useful theoretical advance. It offers clarity about the conditions under which iRS bias will and will not affect cost-effectiveness conclusions, guiding both analysts and methodological researchers. By pinpointing "misalignment" as the culprit for bias, it provides a target for future methodological development (for example, devising new ways to adjust models so that

alignment is achieved). This contribution is timely, as recent literature has criticised the paucity of formal guidance in diagnostic test economic modelling (Snowsill, 2023, Soares et al., 2018) and called for improved frameworks. Chapter 4's framework begins to answer that call. Looking ahead, these theoretical insights beg the question of their practical significance. How large are the errors we risk by ignoring iRS bias? And how effective are the proposed solutions under realistic conditions of uncertainty? To address these questions, Chapter 5 will undertake a simulation-based exploration of iRS bias in cost-effectiveness models. By simulating diagnostic scenarios under controlled conditions (varying the RS, test characteristics, disease prevalence, etc.), we will assess the performance of aligned models versus unaligned ones, and evaluate the impact of applying correction methods. This empirical analysis will illustrate the magnitude of bias in outcomes like cost per QALY when models are mis-specified, and will test whether aligning the model (or using latent true disease modelling) indeed yields results closer to the truth. In essence, the next chapter will validate and stress-test the aligned/unaligned framework. This not only serves to demonstrate the practical utility of the framework but also helps refine it: the simulations may reveal, for example, in which situations a fully iRS-aligned model is sufficient and in which situations an adjustment is critical.

This chapter has established the theoretical basis for understanding the impact of iRS bias on cost-effectiveness evaluations. The analysis demonstrates that iRS bias influences disease prevalence estimates, cost allocations, and incremental cost-effectiveness ratios in systematic and predictable ways. Furthermore, correction methods offer potential solutions, but their theoretical properties vary depending on underlying statistical assumptions and economic model structures. However, while these theoretical relationships highlight the importance of accounting for iRS bias, empirical evidence is necessary to assess the practical magnitude and real-world implications of these distortions.

Building on the methodological insights gained in Chapter 3, this chapter introduces a conceptual framework for understanding the impact of iRSs on cost-effectiveness estimates. The framework clarifies the pathways through which errors in diagnostic classification translate into distorted cost and outcome projections, thereby addressing the second thesis objective. Establishing this theoretical basis is essential for interpreting the results of the simulation study in Chapter 5.

In light of these considerations, this chapter has shown why iRSs pose a critical threat to accurate cost-effectiveness inferences. The next chapter operationalises the aligned and unaligned approaches in a controlled simulation, varying both disease prevalence and RS accuracy to illustrate the potential scale of bias. By explicitly testing these scenarios, we can determine the conditions under which an uncorrected reliance on imperfect reference data may significantly mislead health technology appraisal.

Chapter 5

5 A Simulation-Based Evaluation of Imperfect Reference Standard Bias in the Economic Assessment of Diagnostic Tests.

5.1 Introduction

The evaluation of diagnostic test performance within HTA has often been constrained by assumptions of methodological perfection—an idealised framework where disease status is unequivocally determined, and diagnostic tests are measured against an infallible gold standard. Yet, this paradigm rarely reflects the complexity of real-world clinical practice. Instead, the landscape of diagnostic assessment is fraught with uncertainties, particularly when a definitive gold standard does not exist or is infeasible to implement. In such cases, decision-makers must rely on iRS, the very nature of which introduces systematic errors into economic models and distorts the cost-effectiveness landscape.

A striking challenge emerges when the bias inherent to an iRS is not merely a theoretical inconvenience but a structural flaw that propagates through cost-effectiveness models, altering clinical decision-making and resource allocation. The issue is further complicated when diagnostic accuracy assessments are conditioned on assumptions that do not hold in practice, such as the assumption of CI in latent class models (LCMs). As recent studies (Okkaoglu et al., 2024) have demonstrated, departures from this assumption can lead to misleading estimates of test Se and Sp, ultimately affecting the downstream interpretation of economic evaluations. This is particularly problematic in high-stakes medical conditions, such as cancer screening, infectious disease surveillance, and antimicrobial resistance management, where diagnostic uncertainty translates directly into morbidity, mortality, and financial burden on healthcare systems.

Simulation studies offer a powerful lens through which these distortions can be systematically interrogated. Unlike traditional statistical adjustments, simulations allow

for the controlled manipulation of parameters to examine the magnitude and direction of biases introduced by iRS-driven misclassification errors. For instance, in tuberculosis screening, where multiple imperfect tests are often used, simulation studies have demonstrated how reliance on an iRS can lead to underestimation of true disease prevalence, ultimately influencing health policy decisions (Menten et al., 2008). Similar patterns emerge in oncology, where reliance on biomarker-driven diagnostics without a perfect RS has been shown to shift treatment recommendations significantly (Dendukuri and Joseph, 2001).

Recent advances in Bayesian hierarchical models and machine learning techniques have expanded the possibilities for refining diagnostic accuracy assessments, yet these methods remain underutilised in HTA frameworks. Bayesian models, for example, allow for the explicit modelling of uncertainty and dependency structures, offering a probabilistic approach to account for iRS-induced bias (Vacek, 1985). Despite their potential, these models are rarely integrated into cost-effectiveness analyses, largely due to computational complexity and a lack of standardised methodologies for their application in economic evaluation.

Moreover, evidence of economic evaluations incorporating simulation in diagnostic tests remains scarce. Existing studies primarily focus on simulating issues from clinical trials, such as sample size and design, to evaluate their impact on cost-effectiveness results (Xie et al., 2022a, Xie et al., 2022b, Xie et al., 2022d). To the best of my knowledge, no study has evaluated the application of correction methods in the CEA of diagnostic tests using a single test and an iRS with known accuracy.

This chapter does not seek to reiterate the well-established principles of diagnostic test evaluation but develops a structured simulation framework that quantifies the effects of iRS bias on health economic evaluations. Specifically, it will:

1. Evaluate the performance of existing correction methods (e.g., Staquet and Brenner approaches) for adjusting iRS-induced misclassification errors in diagnostic accuracy estimates.
2. Implement a stylised decision-tree model that reflects diagnostic processes, allowing for comparison of economic models where test performance is

conditioned on an iRS (aligned) versus those assuming an idealised gold standard (unaligned).

3. Quantify economic evaluation metrics, including Incremental Net Health Benefit (iNHB), Value of Diagnostic Information (VoD), and Value of Heterogeneity (VoH), under different assumptions of test dependency structures.
4. Examine the influence of CD and independence assumptions on bias propagation in correction methods and CEA outcomes.

By integrating these components, this chapter provides a comprehensive framework for understanding how iRS bias affects economic evaluations, ensuring that decision models reflect the true uncertainty surrounding diagnostic accuracy. In doing so, this work seeks to improve the reliability of cost-effectiveness assessments in settings where a definitive gold standard is unavailable, ultimately leading to more informed and resilient healthcare decision-making.

5.2 Methods

To achieve the objectives of this chapter, simulation studies will be conducted using prespecified “true” values for the accuracy parameters of the testing strategy, including Se and Sp, as well as for the net health benefit (NHB). These true values will serve as a benchmark to calculate both the initial biased estimates and the subsequent corrected results.

5.2.1 Selection criteria of correction methods

To determine which correction method would be evaluated in this chapter, three selection criteria were defined.

1. The method must utilise accuracy information of an iRS that relates to the true disease status.
2. It must allow analysis with a single index test and a single iRS.
3. It must be validated through simulation studies to establish its robustness and credibility.

Based on these criteria, Staquet and Brenner were identified as the most suitable. In scenarios where test errors are CI and conditional on the true disease status, the Staquet method was selected because it outperformed Brenner under CI conditions. However, in cases where dependencies between test errors were considered, Brenner's CD method was chosen as it is specifically designed for scenarios with positive CD.

Despite being developed for CI scenarios, Staquet often produced more accurate results but occasionally led to illogical estimates, such as Se or Sp values exceeding 1 or falling below 0. Therefore, both Staquet and Brenner CD were assessed in situations where test errors were dependent, providing a comprehensive evaluation across different dependency structures.

To ensure the clarity and reproducibility, we adopted the ADEMP framework - Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures - proposed by Morris et al. (2019).

5.2.1.1 Data-generating mechanism

This section provides a detailed explanation of the data-generating mechanism (DGM) employed in this study to simulate the CEA of diagnostic tests. Unlike a standard DGM, which generates datasets representative of clinical trials, this approach uses randomly generated point-estimate values to populate a stylised decision tree model (Morris et al., 2019). This strategy is particularly suited to modelling diagnostic and treatment scenarios under uncertainty, focusing on economic and health outcomes.

5.2.1.1.1 Diagnostic accuracy parameters

The $p(D^+)$ (prevalence or true disease state), the $p(T^+|D^+)$ (sensitivity), and the $p(T^-|D^-)$ (specificity) were modelled using a uniform distribution where its minimum and maximum parameters were 0 and 1, respectively. To model the biased or unadjusted accuracy parameters of a new test, it is necessary to simulate the probabilities of having a positive or negative test result of the new test and the iRS as well - $p(T, iRS)$- following the equations proposed by Vacek (1985):

$$
\begin{aligned}
p(T^+, iRS^+) &= \Big(\big(p(D^+) p(T^+|D^+) p(T^+|iRS^+) + cov_{1,1|1} \big) \\
&\quad + \Big(\big(1 - p(D^+)\big)\big(1 - p(T^-|D^-)\big)\big(1 - p(T^-|iRS^-)\big) + cov_{0,0|0} \Big) \Big) \\
p(T^+, iRS^-) &= \Big(\big(p(D^+) p(T^+|D^+)\big(1 - p(T^+|iRS^+)\big) - cov_{1,1|1} \big) \\
&\quad + \Big(\big(1 - p(D^+)\big)\big(1 - p(T^-|D^-)\big) p(T^-|iRS^-) - cov_{0,0|0} \Big) \Big) \\
p(T^-, iRS^+) &= \Big(\big(p(D^+)\big(1 - p(T^+|D^+)\big) p(T^+|iRS^+) - cov_{1,1|1} \big) \\
&\quad + \Big(\big(1 - p(D^+)\big) p(T^-|D^-)\big(1 - p(T^-|iRS^-)\big) - cov_{0,0|0} \Big) \Big) \\
p(T^-, iRS^-) &= \Big(\big(p(D^+)\big(1 - p(T^+|D^+)\big)\big(1 - p(T^+|iRS^+)\big) + cov_{1,1|1} \big) \\
&\quad + \Big(\big(1 - p(D^+)\big) p(T^-|D^-) p(T^-|iRS^-) + cov_{0,0|0} \Big) \Big)
\end{aligned}
\tag{1}
$$

Where the $cov_{1,1|1}$ is the covariance term for a positive test result and positive iRS given the presence of the disease and $cov_{0,0|0}$ is the covariance term for a negative test result and negative iRS given the absence of the disease. When the CI between tests is assumed, the $cov_{1,1|1}$ and $cov_{0,0|0}$ will be zero.

Therefore, the biased Se and Sp are estimated as:

$$
p(T^+|iRS^+) = \frac{p(T^+, iRS^+)}{\big(p(T^+, iRS^+) + p(T^+, iRS^-)\big)} \tag{2}
$$

$$
p(T^-|iRS^-) = \frac{p(T^-, iRS^-)}{\big(p(T^-, iRS^+) + p(T^-, iRS^-)\big)} \tag{3}
$$

When the assumption of CI is relaxed, the covariance terms are modelled using a uniform distribution ranging from the minimum to the maximum covariance terms

estimable. The minimum and maximum values for the covariances were estimated using the equations stated by

Covariance terms for the disease group

$$max(-(1-p(T^+|iRS^+))(1-p(iRS^+|D^+)), -p(T^+|iRS^+)p(iRS^+|D^+)) \leq cov_{1,1|1} \leq \min(p(T^+|iRS^+)(1-p(T^+|iRS^+)), (p(iRS^+|D^+)(1-p(T^+|iRS^+))) \quad (4)$$

Covariance terms for the non-disease group

$$max(-(1-p(T^-|iRS^-))(1-p(iRS^-|D^-)), -p(T^-|iRS^-)p(iRS^-|D^-)) \leq cov_{0,0|0} \leq \min(p(T^-|iRS^-)(1-p(T^-|iRS^-)), (p(iRS^-|D^-)(1-p(T^-|iRS^-))) \quad (5)$$

5.2.1.2 Modelling a cost-effectiveness analysis of diagnostic tests

In the CEA of diagnostic tests, three key parameters are essential for modelling the consequences of misclassification when treating subgroups: (i) the prevalence, p(D+), (ii) the performance of the test, p(T|D), and (iii) the associated payoffs. If the new test is not perfect (i.e., its accuracy is less than one), some individuals will be misclassified, leading to a series of downstream consequences. A similar effect is observed in cases involving an iRS, where the true number of misclassified individuals is not known with certainty. Consequently, the expected opportunity loss depends on several factors, including the misclassification rate—characterised by false positives (FP) and false negatives (FN)—the prevalence, and the net health benefit (NHB) gains of the testing strategy.

To define the function for expected NHB gains, it is necessary to consider not only the diagnostic test itself but also additional factors that influence decision-making. These include the treatment effect of the intervention used following diagnosis and the patient's risk profile.

In a population-level simulation of NHB, we will account for both true positive and true negative disease statuses, as well as whether individuals received treatment. The

corresponding payoffs will be estimated for each scenario. Specifically, ***"QALY$_0$"*** represents the expected QALY of an individual without the disease, ***"Cost$_{max}$"*** refers to the maximum cost that an individual could incur as a consecuence of having a disease, "***a"*** denotes the severity of the condition, "***b"*** represents the treatment effect, "***c"*** reflects the toxicity associated with treatment, and "***t"*** denotes the penalty related to the cost of the test and any consequences arising from its application. Table 8 defines the NHB functions for disease status, whether treated or untreated.

Table 8. 2x2 table to describe the NHB under different circumstances.

	D$^+$	**D$^-$**
Trt$^+$	$[(QALY_0 \times (1 - a \times b)) - \frac{Cost_{max} \times a \times b}{\lambda}] - c - t$	$QALY_0 - c - t$
Trt$^-$	$[(QALY_0 \times (1 - a)) - \frac{Cost_{max} \times a}{\lambda}] - t$	$QALY_0 - t$

D$^+$: Disease positive, individuals have the disease; D$^-$: Disease negative, individuals have not the disease; Trt$^+$: Treatment positive, individuals are receiving treatment; Trt$^-$: Treatment negative, individuals are not receiving any treatment, QALY$_0$: baseline QALYs. a: Burden of disease, b: Treatment effect, c: Toxicity, t: Test penalization.

QALY$_0$ and Cost$_{max}$ parameter: Baseline Net Health Benefit

The baseline NHB, meaning the health state and costs of healthy individuals. A NHB is defined as a function of health outcomes (QALY) and costs, using a standard cost-effectiveness threshold (λ) - £20,000-£30,000/QALY for England (National Institute for Health and Care Excellence, 2022b):

$$NHB = QALY - \frac{Cost}{\lambda} \quad (6)$$

By assuming that health outcomes of a healthy individual throughout their lifetime is measured in terms of QALE (Quality-Adjusted Life Expectancy), the baseline NHB definition is:

$$NHB_0 = QALE_{healthy} - \frac{Cost\ of\ being\ healthy}{\lambda} \quad (7)$$

Furthermore, if being healthy has no costs for the health system, the NHB = $QALE_{Healthy}$. In this chapter we will refer indistinctly $QALE_{Healthy}$, $QALY_0$, and $QALY_{Healthy}$.

McNamara et al. (2023) have published the QALE norms for the English population, indicating that a new-born has a QALE of 24.9 when assuming a 3.5% discount rate and an equivalence between male and female populations.

Thus, the simulation will use values from 0 to 24.9, representing different time horizons.

In terms of costs, we have defined that a maximum cost ($Cost_{max}$) as the highest value that an individual with the disease could incur owing to a disease. Reviewing the expected total costs of the independent economic evaluations of the diagnostic assessment reports (DARs) published for the National Institute for Health and Care Excellence (NICE) (2024), the maximum cost was reported in the DG45 (National Institute for Health and Care Excellence (NICE), 2022) report with a total cost of ~£210,000, whilst the minimum was ~£50 reported by the DG38 report (National Institute for Health and Care Excellence (NICE), 2019). The average expected total cost was ~£12,000. Therefore, I decided to establish the maximum costs in £250,000.

a parameter – Burden of disease

The $\boldsymbol{a}$ refers to the shortfall that adjusts the $QALY_0$ and $Cost_{max}$ by severity, considering the losses in health outcomes and the additional costs for disease management.

A practical manner to correct by severity in HTA is using proportional shortfalls (PS) or absolute shortfalls (AS) (Skedgel et al., 2022). Both PS and AS are estimated as follows:

$$PS = \frac{Disease\ related\ QALY\ loss\ (AS)}{Remaining\ QALY\ expectation\ in\ absence\ of\ disease} \quad (8)$$

Where

$$Absolute\ shortfall\ (AS) = Remaining\ QALY\ expectation\ in\ absence\ of\ disease - Remaining\ QALY\ expectation\ with\ disease$$

The National Institute for Health and Care Excellence (2022b) has mentioned in its manual QALY weightings as a severity modifier based on absolute and proportional shortfall. This severity modifier is an adjustment factor applied to account for the burden of severe disease, allowing treatments for more sever conditions to receive a higher cost-effectiveness threshold, thereby improving their likelihood of being recommended for the NHS funding. In addition, the manual has reported an equivalence table where:

Table 9. Severity modifier and equivalence with both proportional and absolute QALY shortfall

QALY weight	Proportional QALY shortfall	Absolute QALY shortfall
1	Less than 0.85	Less than 12
x1.2	0.85 to 0.95	12 to 18
X1.7	At least 0.95	At least 18

Source: National Institute for Health and Care Excellence (2022b)

Let assumes that *a* = 0 represent no disease (perfect health) – *PS* = 0, and *a* = 1 represent maximal severity (the worst possible disease state) – PS = 1. The QALY of an individual with the disease is:

$$QALY(a) = QALY_0 \times (1 - a) \tag{9}$$

Where QALY(*a*) refers to the QALY of an individual with the disease condition as function of severity.

Alternatively, the costs related of an individual completely health (*a* = 0) will be equal zero, and increasing as ***"a"*** grows (more severe disease). Using a linear definition the cost of an unhealthy individual is:

$$Cost(a) = \begin{cases} 0, & S = 0 \\ Cost_{max} \times a, & 0 < a < 1 \\ Cost_{max}, & a = 1 \end{cases} \tag{10}$$

b parameter – Treatment Effect

The $\boldsymbol{b}$ denotes the treatment effect (TE) of being treated in reducing health outcomes losses and other disease-treatment costs. TE ranges from 0 to 1, where 0 means the shortfall is completely eliminated (i.e., curative treatments), 1 means no TE, and values above 1 indicate the treatment is less beneficial than control. For simplicity, we will constrain $\boldsymbol{b}$ between 0 and 1, assuming a treatment cannot be worse than doing nothing.

c parameter - Toxicity

The parameter $\boldsymbol{c}$ represents the QALY losses associated with toxicity, adverse events (AEs), or any additional health or cost consequences arising from treatment. Although $\boldsymbol{c}$ can, in principle, range from zero to infinity, a treatment will only be valuable if it provides a favourable risk–benefit ratio for patients with the disease. Mathematically, if

$$c > \left[\left(\left(QALY_0 \times (1 - a \times b)\right) - \frac{Cost_{max} \times a \times b}{\lambda}\right) - \left(\left(QALY_0 \times (1 - a)\right) - \frac{Cost_{max} \times a}{\lambda}\right)\right] \quad (11)$$

then the net benefit gain from treating a diseased patient is outweighed by toxicity, rendering the treatment never cost-effective. Here, $\left[[(QALY_0 \times (1 - a \times b)) - \frac{Cost_{max} \times a \times b}{\lambda}] - [(QALY_0 \times (1 - a)) - \frac{Cost_{max} \times a}{\lambda}]\right]$ is the maximum NHB gain for treating an individual who actually has the disease (ignoring toxicity). Therefore, to ensure a treatment can remain cost-effective in at least some scenarios, we define $\boldsymbol{c}$ as a fraction of this maximum gain, by sampling a random proportion $\psi \in [0,1]$ and setting:

$$c = \psi\left[\left(\left(QALY_0 \times (1 - a \times b)\right) - \frac{Cost_{max} \times a \times b}{\lambda}\right) - \left(\left(QALY_0 \times (1 - a)\right) - \frac{Cost_{max} \times a}{\lambda}\right)\right] \quad (12)$$

t parameter – Test

The parameter $\boldsymbol{t}$ denotes the NHB penalty for administering the diagnostic test, including its direct costs and any health-related inconveniences. To avoid scenarios where testing is more harmful than treatment itself, we impose a second logical constraint. If

$$t > \left[\left[\left(\left(QALY_0 \times (1 - a \times b)\right) - \frac{Cost_{max} \times a \times b}{\lambda}\right) - \left(\left(QALY_0 \times (1 - a)\right) - \frac{Cost_{max} \times a}{\lambda}\right)\right] - c\right] \quad (13)$$

then even patients with the disease would not find testing + treatment worthwhile. The term $\left[[(QALY_0 \times (1 - a \times b)) - \frac{Cost_{max} \times a \times b}{\lambda}] - [(QALY_0 \times (1 - a)) - \frac{Cost_{max} \times a}{\lambda}]\right] - c$ represents the remaining NHB of treating a diseased individual after subtracting toxicity. Hence, we define $\boldsymbol{t}$ as another random fraction $\phi \in [0,1]$ of that remaining:

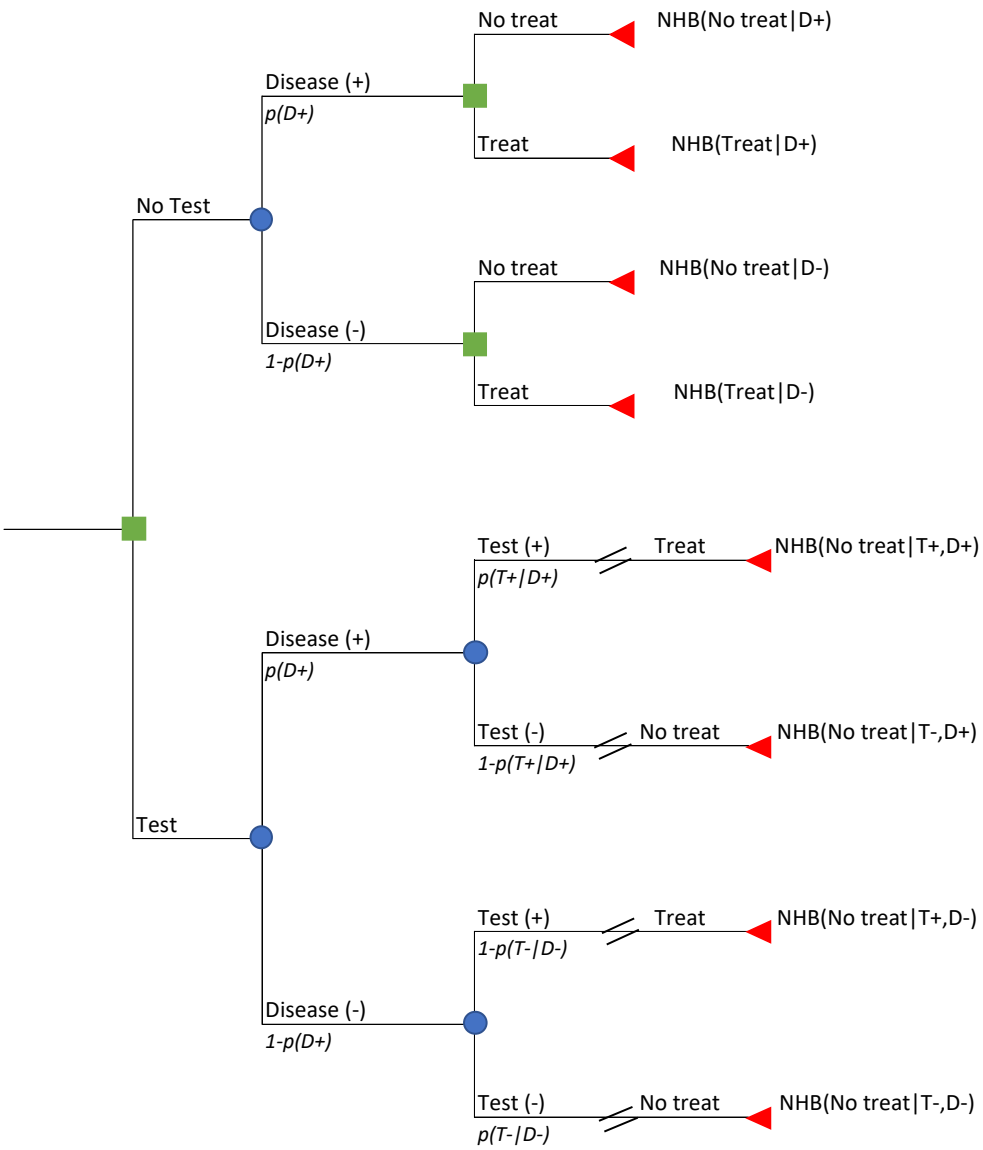
$$t = \phi\left[\left[\left(\left(QALY_0 \times (1 - a \times b)\right) - \frac{Cost_{max} \times a \times b}{\lambda}\right) - \left(\left(QALY_0 \times (1 - a)\right) - \frac{Cost_{max} \times a}{\lambda}\right)\right] - c\right] \quad (14)$$

5.2.1.2.1 Decision model

A decision tree model was created to compare a testing strategy with two non-testing strategies: (i) no test, no treatment (nt_nt), and (ii) no test, treat all (nt_t).

Figure 12 shows the decision model's structure and parameters.

Figure 12. Decision tree model to quantify the NHB gains of a testing strategy against no testing strategies.



The decision model is parametrised using the variable listed in Table 10.

Table 10. Description of notation for each parameter and estimates involved in the simulation study.

Notation	Description
$p(T^+\|D^+)$	Probability of having a positive test result given the true disease status (True Se)
$p(T^-\|D^-)$	Probability of having a negative test result given the true disease status (True Sp)
$p(iRS^+\|D^+)$	Probability of having a positive iRS result given the true disease status (iRS-Se)

Notation	Description
$p(\overline{iRS^-}\|D^-)$	Probability of having a negative iRS result given the true disease status (iRS-Sp)
$p(D^+)$	Probability of having a disease (Prevalence)
N_{pt}	Total number of patients in each hypothetical trial
$QALY_0$	QALY of a healthy individual
$Cost_{max}$	Maximum Cost
a	Burden of disease
b	Treatment effect
c	Toxicity
ψ	Proportion of the NHB gains by treating diseased individuals (Toxicity)
t	NHB losses by administering test
ω	Proportion that defines the size of the t parameter
$\widehat{cov}_{1,1\|1}$	The covariance estimates term for the disease group
$\widehat{cov}_{0,0\|0}$	The covariance estimates term for the non-disease group
$p(\widehat{T^+\|\imath RS^+})$	The estimate of the probability of having a positive test result given the iRS is positive (Biased or Unadjusted Se)
$p(\widehat{T^-\|\imath RS^-})$	The estimate of the probability of having a negative test result given the iRS is negative (Biased or Unadjusted Sp)

5.2.1.3 Estimands

The simulation will focus on four main estimands:

- Diagnostic accuracy parameters (Se and Sp)
- CE results in terms of NHB and iNHB for each strategy.
- The expected value of diagnostic information (VoD), defined as the difference in NHB between the chosen testing strategy and the best non-testing strategy.
- The expected value of heterogeneity (VoH), also referred as the value of a perfect test, which represent the difference in NHB between a hypothetical test with perfect Se and Sp – equal to 1, and the current testing strategy.

Each of these estimands will be calculated under three scenarios: using the *true* test performance, an *adjusted* (i.e., bias-corrected) performance, and the *unadjusted* (i.e., biased) performance.

5.2.1.4 Methods

Two algebraic-correction methods—Staquet et al. (1981) and Brenner (1996) —were selected for analysis based on the criteria outlined earlier. The Staquet method was exclusively evaluated under the assumption of CI, owing to its established performance in such scenarios (Staquet et al. (1981)). For conditions involving dependent test errors between diagnostic tests, both the Staquet and Brenner correction methods (CD) were applied, albeit with modifications to address methodological assumptions.

First, Brenner's CD equations, originally formulated for scenarios with positively correlated tests, were extended in our analysis to accommodate both positive and negative correlations. Second, while Staquet's method was not designed for CD between test errors, its inclusion served two purposes: (i) to evaluate its bias-correction efficacy in these conditions, given its superior performance relative to Brenner CD in prior studies (Umemneku Chikere et al., 2021); and (ii) to address concerns raised by Umemneku Chikere et al. (2021), who reported implausible Se and Sp estimates (exceeding the [0,1] range) under extremely high or low disease prevalence. These findings prompted our investigation into the robustness of Staquet's method across diverse simulated conditions, with a view to informing its broader applicability.

For Staquet et al. (1981), corrected sensitivity (Se'_T) and specificity Sp'_T were derived using the following equations:

$$Se'_T = \frac{Tp\ Sp_{iRS} - pn}{N(Sp_{iRS} - 1) + (pp + np)} \tag{15}$$

$$Sp'_T = \frac{Tn\ Se_{iRS} - np}{N\ Se_{iRS} - (pp + np)} \tag{16}$$

Here, the apostrophe (') denotes corrected estimates. Tp and Tn represent the total counts of positive and negative test results, respectively. The terms pp, pn, np, nn correspond to the counts of individuals with specific combinations of test and iRS outcomes, where *p* indicates a positive result and *n* a negative result.

For Brenner (1996) method, corrected estimates under positive C) were calculated by rearranging the original equations. Brenner's framework derives apparent sensitivity (Se_{ap}^{dep}) and specificity (Sp_{ap}^{dep}) from true accuracy parameters and iRS values. The corrected estimates were obtained as follows:

$$Se_{ap}^{dep} = \frac{P\ Se_T\ + (1-P)(1-Sp_{iRS})}{P\ Se_{iRS} + (1-P)(1-Sp_{iRS})} \tag{17}$$

$$Se_T^{dep'} = \frac{Se_{ap}^{dep}\big(Se_{iRS}p(D^+) + (p(D^+)-1)(Sp_{iRS}-1)\big) - (Sp_{iRS}-1)(p(D^+)-1)}{p(D^+)} \tag{18}$$

$$Sp_{ap}^{dep} = \frac{P\ (1-Se_{iRS}) + (1-P)\ Sp_T}{P\ (1-Se_{iRS}) + (1-P)\ Sp_{iRS}} \tag{19}$$

$$Sp_T^{dep'} = \frac{Sp_{ap}^{dep}\big(Sp_{iRS}(p(D^+)-1) + p(D^+)(Se_{iRS}-1)\big) - (Se_{iRS}-1)p(D^+)}{p(D^+)-1} \tag{20}$$

In these equations:

The apostrophe (') denotes corrected estimates. The superscript ***dep*** indicates an assumption of positive CD**.** $\boldsymbol{p(D^+)}$ represents disease prevalence, and ***P*** refers to the proportion of positive

5.2.1.5 Performance Measure

The correction methods' performance was measured by the bias of Se and Sp, the weighted average of se/sp (Test error), prevalence, and their impact on iNHB. The bias was calculated as follows:

$$Bias(Se) = Se_{bias/adjusted} - Se_{true}$$
$$Bias(Sp) = Sp_{bias/adjusted} - Sp_{true}$$

$$Test\ error = p(D^+)(1 - Se) + (1 - p(D^+))(1 - \text{Sp}) \quad (21)$$

$$Bias(test\ error) = Test\ error_{biased/adjusted} - Test\ error_{true} \quad (22)$$

$$Bias(iNHB) = iNHB_{biased/adjusted} - iNHB_{true} \quad (23)$$

5.2.1.6 Coding and execution of scenarios

The simulation study was implemented using R statistical software. To systematically explore the parameter space governing CEA of diagnostic tests, Latin Hypercube Sampling (LHS) was employed as the primary sampling strategy. LHS offers distinct advantages over conventional Monte Carlo methods for this study's aims. Unlike Monte Carlo simulations, which randomly sample parameters independently, LHS stratifies each parameter's range into equally probable intervals, ensuring efficient coverage of high-dimensional parameter spaces while minimising clustering or gaps in sampled values (McKay et al., 1979). This property is critical for evaluating scenarios where interactions between parameters—such as disease prevalence, test accuracy, and covariance terms—drive decision-relevant biases in CEA results. By generating 10,000 scenarios via LHS, the study captures a broad spectrum of plausible conditions, including rare but consequential combinations (e.g., extremely low prevalence with high test sensitivity), which are often under sampled in Monte Carlo approaches.

Parameters were varied according to their theoretical or empirical ranges. Probabilistic parameters (e.g., Se, Sp, prevalence) were sampled uniformly between 0 and 1, while QALYs of healthy individual ($QALY_0$) was drawn uniformly from 0 to 24.9 QALYs, and $Cost_{max}$ between £0 and £250,000. For scenarios relaxing the CI assumption, covariance terms were constrained within their permissible bounds (Equations 4–5) using rejection sampling to maintain biological plausibility.

This approach aligns with the ADEMP framework's emphasis on structured, transparent data-generating mechanisms (Morris et al., 2019), enabling robust evaluation of correction methods across diverse simulated conditions while optimising computational efficiency.

5.3 Results

Evidence suggests that the impact of prevalence on Se and Sp biases is noteworthy, with the magnitude of the bias increasing at both lower and higher levels of prevalence (p(D+)) (Umemneku Chikere et al., 2021). In CEA of diagnostic tests, bias predominantly arises from the weighted sum of Se, Sp, and P(D+), as articulated in equation 21. This implies that deviations in CEA outcomes are more significantly affected by the composite influence of Se, Sp, and P(D+) rather than by the independent accuracy of these diagnostic parameters. To address this, we expanded the Umemneku Chikere et al. (2021) analysis to include the weighted test accuracy alongside p(D+), aiming to provide a more comprehensive understanding of how prevalence influences CE outcomes.

In this document, we refer to "bias accuracy" denoting to the bias in Se and Sp separately. Conversely, "bias test" refers to the bias in weighted parameter.

5.3.1 Scenarios where conditional independence (CI) assumption holds.

A total of 10,000 scenarios were simulated. In 5.26% of these, bias introduced by the iRS actually changed the final decision regarding which test strategy was optimal. In the remainder, the decision did not differ despite observable bias.

5.3.1.1 Test accuracy bias.

Figure 13 and Figure 14 visualise the bias in Se and Sp against disease p(D+). Overall, unadjusted results exhibit sizeable dispersion, revealing underestimation or overestimation of Se and Sp—especially pronounced at low or high prevalence. By contrast, applying the Staquet correction yielded nearly unbiased estimates across all prevalence levels.

The plots in Figure 13 offer insight into how the prevalence of a condition influences bias in diagnostic test accuracy. The X-axis denotes the condition's prevalence within a population, and the Y-axis measures the bias for Se and Sp in unadjusted estimates. The plot shows that as p(D+) increases, bias in unadjusted Se decreases whereas bias in Sp tends to increase. Using Staquet's method (blue dots), the bias falls to zero across all

prevalence levels, confirming its ability to offset prevalence-related misclassification under CI.

Figure 13. Prevalence vs Test accuracy bias. A) Se and Sp bias for unadjusted estimates. B) Se and Sp bias for corrected estimates using Staquet approach.

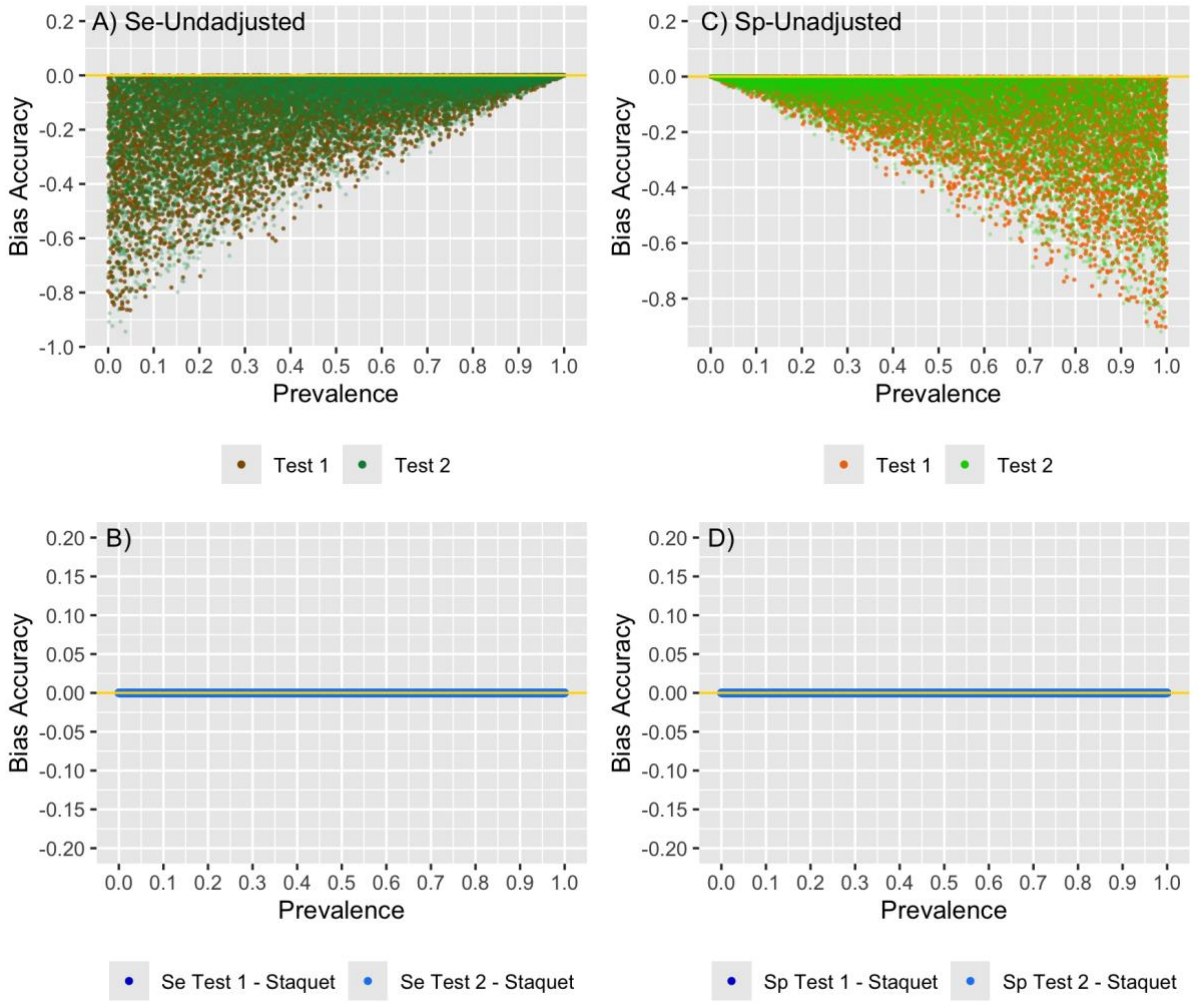
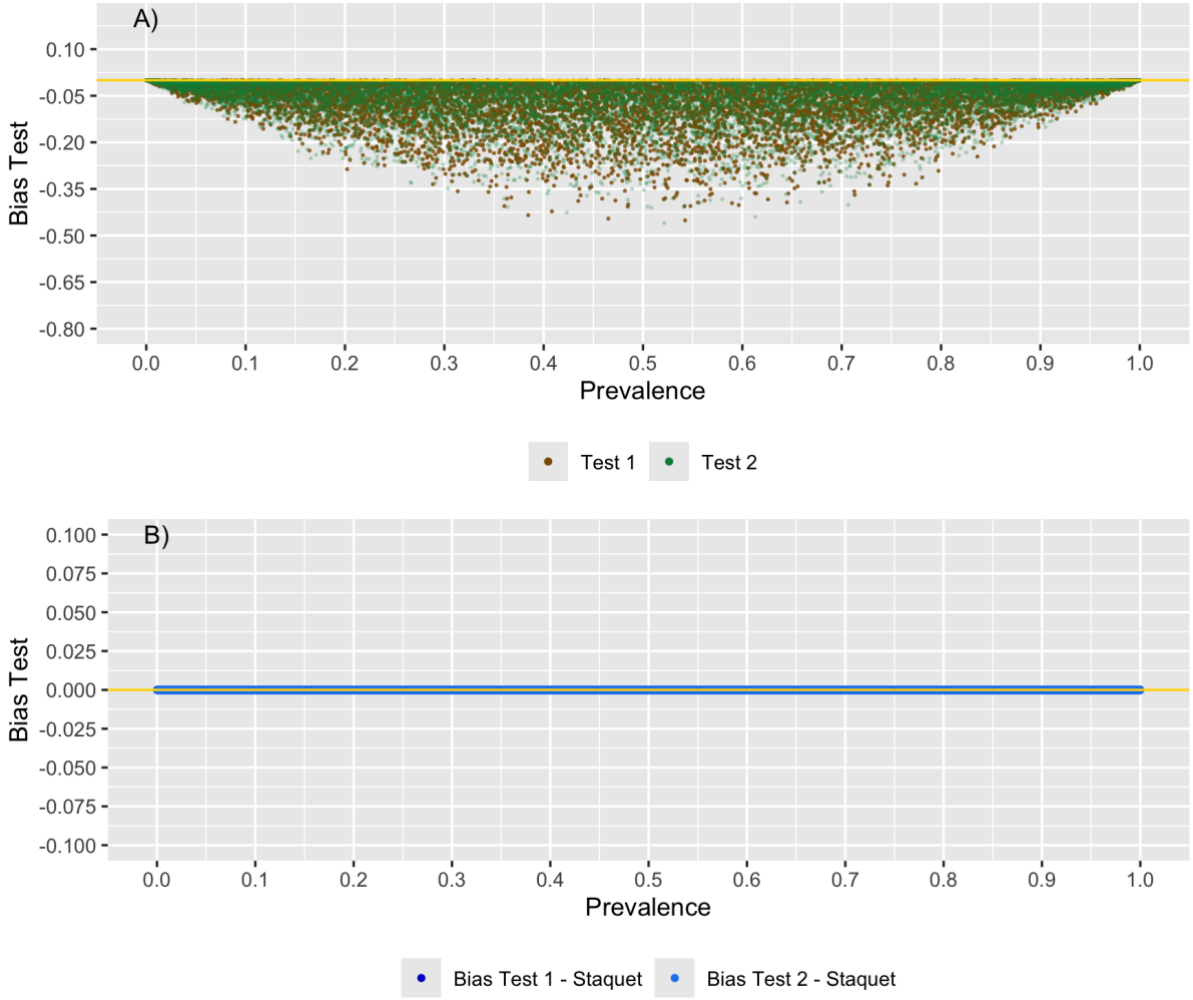


Figure 14 depicts the variation in weighted test accuracy across different p(D+) levels for both unadjusted and corrected accuracies. The X-axis represents the prevalence of a condition within a population, while the Y-axis measures the bias in weighted accuracy for unadjusted estimates, indicated by brown (Test 1) and green dots (Test 2) (Figure 14a). In contrast to the patterns observed in Figure 13, the bias decreases as p(D+) deviates from middle values and becomes more pronounced at moderate prevalence levels. This suggests that the bias is most significant when p(D+) is at intermediate levels. Conversely, corrected accuracy utilizing the Staquet approach, as illustrated in Figure

14b, achieves unbiased estimates across all prevalence levels, demonstrating the efficacy of this correction method in eliminating the impact of prevalence on test accuracy.

Figure 14. Prevalence vs weighted accuracy bias. A) unadjusted estimates B) corrected estimates using Staquet approach.



5.3.1.2 Bias in cost-effectiveness analysis owing to imperfect reference standard bias.

The iNHB bias across different prevalence levels exhibits a slightly different pattern to what is shown in Figure 14, albeit with greater variability. Figure 15 illustrates the bias in

the iNHB for Test 1 and Test 2 (y-axis) across prevalence (x-axis). Figure 15a and Figure 15b show that most of the simulations underestimate the true value of the iNHB, similar to what is observed in Figure 14. However, there are a few scenarios where the iRS bias overestimates the results. This overestimation is attributed to both tests being biased; therefore, in the incremental analysis between the testing strategies, as the NHB of the comparator is underestimated, the iNHB tends to be larger. Figure 15d illustrates the effect when only one test is biased, observing only underestimated results, as is the case in Figure 14. Additionally, corrections with the Staquet approach produce unbiased estimates for both tests, as observed in Figure 15c.

Furthermore, the increased variability is due to the magnitude of the bias being influenced by factors beyond just the accuracy parameters. There is a complex non-linear relationship between the bias in test accuracy and the iNHB bias, where minor inaccuracies in diagnostic testing can result in disproportionately large errors in iNHB estimates.

The complex and non-linear relationship between the biases (test accuracy and iNHB) is illustrated in Figure 16, where larger magnitudes of bias in test accuracy may produce small magnitudes of bias in terms of the iNHB. Figure 16a shows that the Staquet approach fully corrects the bias as aforementioned.

Figure 15. Prevalence vs iNHB bias. A) unadjusted estimates for test 1, B) unadjusted estimates for test 1, C) corrected estimates using Staquet approach for both test, and D) unadjusted estimates for test 1 considering that test 2 is not biased.

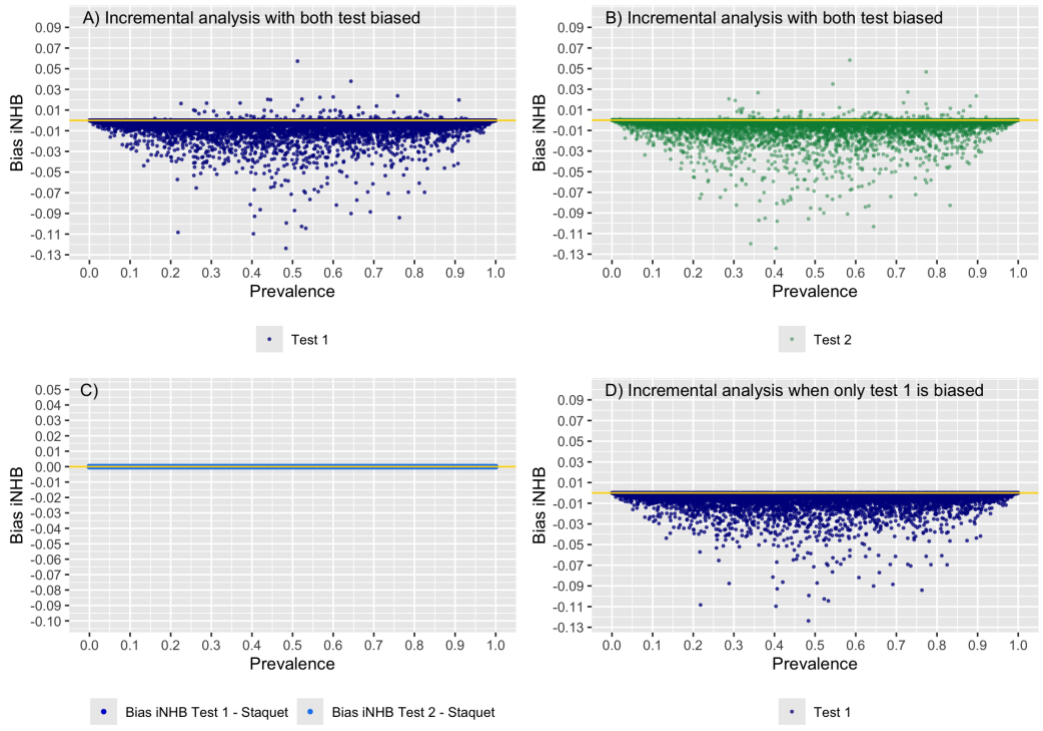


Figure 16. Relationship between the bias of the test accuracy and the bias in the iNHB.

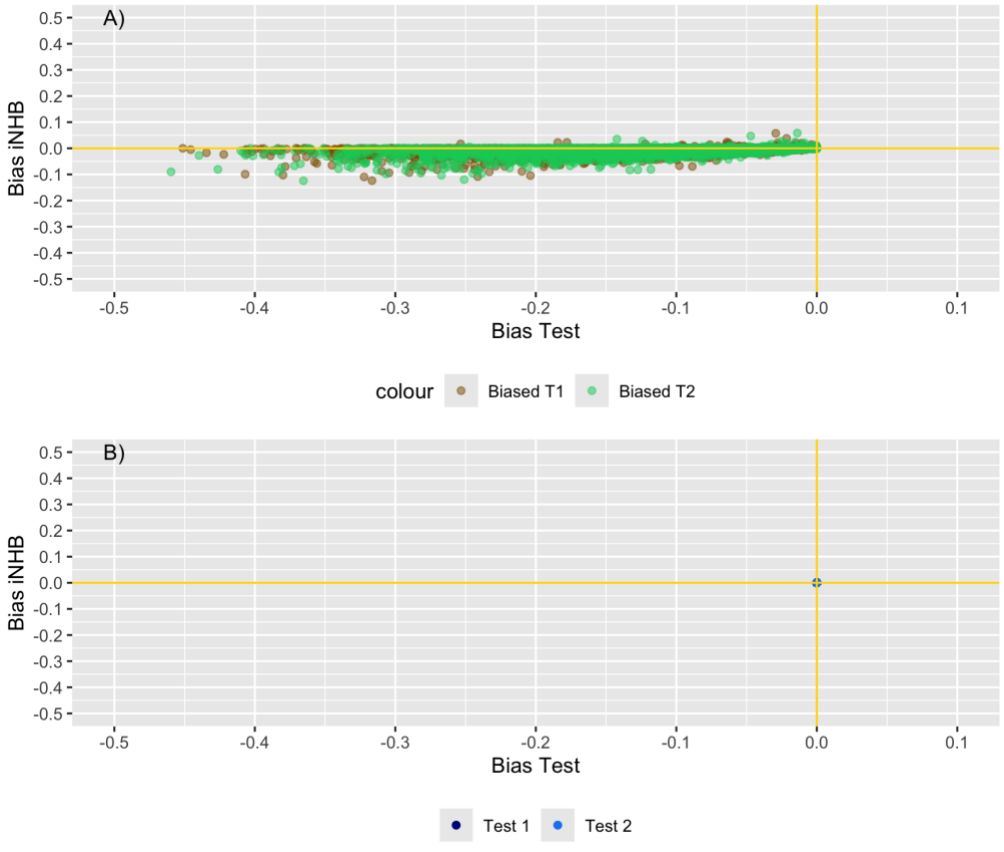
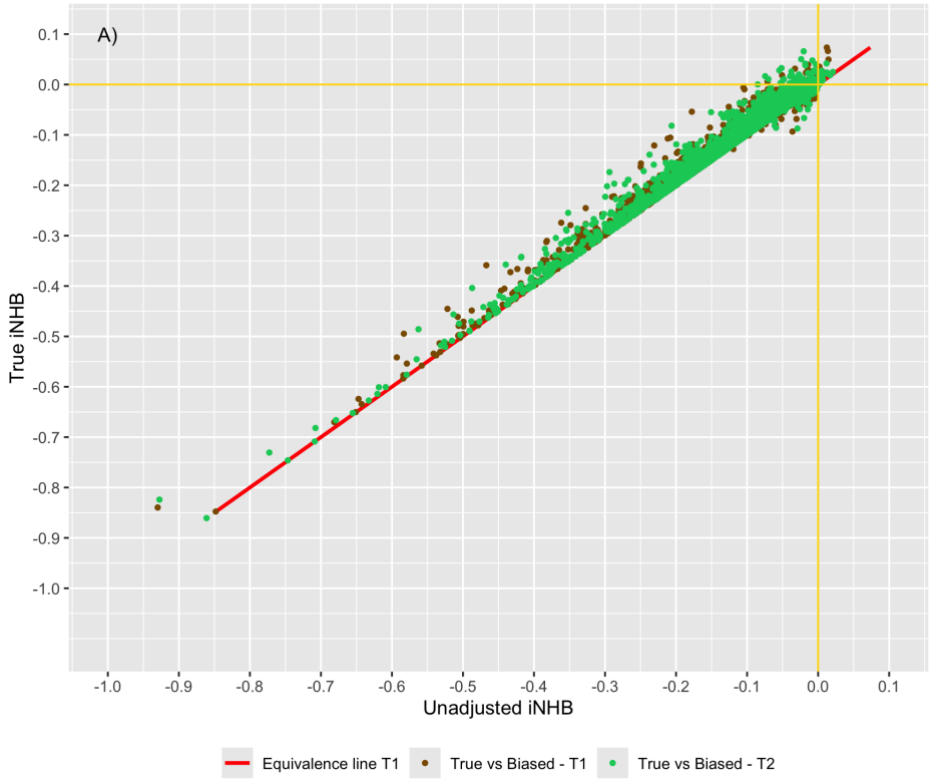


Figure 17 illustrates the relationship between unadjusted iNHB on the horizontal axis and true iNHB on the vertical axis, marked by a line of equivalence that serves as a benchmark for unbiased estimation. This equivalence line, running diagonally across the plot, represents the ideal scenario where unadjusted iNHB equals the true iNHB, indicating no bias. The brown and green dots, representing unadjusted iNHB, show a noticeable deviation from the equivalence line, highlighting the presence of bias in the initial estimates. This bias is variable and, as previously demonstrated, does not display a linear relationship with increasing unadjusted iNHB values.

Figure 17. True iNHB (Equivalence line in red) vs Unadjusted iNHB (Green dots) and corrected iNHB (Staquet – Blue dots).



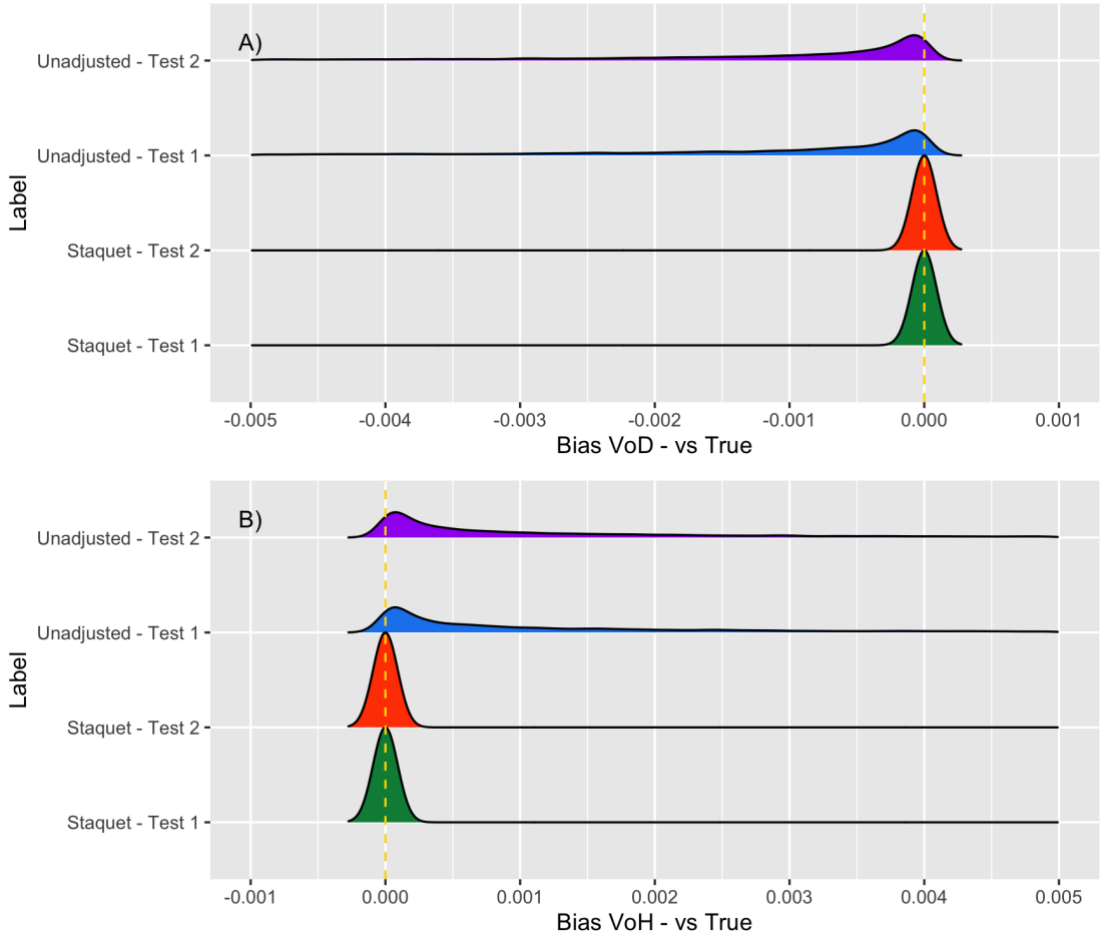
Another way to assess the iRS bias is by measuring its impact in terms of the Value of Diagnostic test (VoD) and Value of Heterogeneity (VoH). The VoD refers to the value that a diagnostic test has when compared against a non-testing strategy (e.g., the difference between the NHB produced by Test 1 against the best non-testing strategy (no test/no treat); $NHB_{T1} - NHB_{(no\ test\ -\ no\ treat)}$). Alternatively, the VoH, or the value of the perfect test,

refers to the additional net benefits that a testing strategy could produce as a result of correctly identifying all cases.

In Figure 18, we illustrate the biased VoD and VoH by comparing unadjusted and corrected estimates with the true values of those metrics. These results are summarised through a density plot to condense all the simulated scenarios. Figure 18a shows that unadjusted estimations underestimate the VoD. However, this underestimation is no greater than 0.005 NHB, and a large density of scenarios underestimate the VoD by no more than 0.001 NHB. Furthermore, the corrected results provide unbiased estimates, as observed in previous findings.

Figure 18b shows that the VoH is overestimated when estimated by calculating the perfect test against the biased – unadjusted – testing strategy. This overestimation occurs because the NHB informed by the biased test is lower than the true value. Once again, the correction methods provide a full adjustment of the estimators.

Figure 18. Bias in the Value of diagnostic (VoD, showed in the upper figure) and in the value of heterogeneity (VoH, showed in the bottom figure).



5.3.1.3 Scenarios considering dependence between tests errors.

When simulations were conducted, taking into account the dependence of test errors, only 529 (5.29%) scenarios resulted in a decision change due to iRS bias for either Test 1 or Test 2. Similarly to the scenarios with CI, the remaining simulations produced biased estimates where the decision was not reversed when correction methods were applied, as the unadjusted values remained on the same side of the decision as the true estimates.

5.3.1.4 Test accuracy bias.

Brenner's method, designed for scenarios with positive covariance, has shown promise for adaptation to negative covariance scenarios, though its accuracy warrants further examination. Umemneku Chikere et al. (2021) compared the performance of Brenner's covariance method and the Staquet approach, the latter being utilised despite its lack of support for dependence structures. They discovered that the Staquet approach yielded illogical estimates (Se or Sp greater than one or less than zero) when the prevalence was below 0.2 or above 0.85. Within the prevalence range of 0.2 to 0.85, illogical results from the Staquet method were unlikely, and the likelihood decreased further as the sample size increased —to less than 0.5% of the sample (sample size ~1,000 individuals). Conversely, Brenner's method did not produce illogical outcomes. However, in contrast to Umemneku Chikere et al. (2021)), our simulations using Brenner's covariance method resulted in illogical estimates in 58.39% of cases. It is important to note that our Brenner covariance equations were reordered, as Umemneku Chikere et al. (2021) used equations to estimate apparent Se and Sp and not to correct them. Meanwhile, applying the Staquet method, which typically ignores correlations, led to illogical outcomes in 59.74% of our simulations, necessitating the exclusion of these results. It is important to note that Umemneku Chikere et al. (2021) based their findings on constant Se and Sp values (0.8 and 0.7, respectively) and iRS Se and Sp (both 0.9), while varying the prevalence from zero to one. Our findings, derived from simulations with uniformly distributed parameter variations, account for the discrepancies observed when

compared to Umemneku Chikere et al. (2021). The conditions that lead to illogical results for both Brenner covariance and Staquet are detailed in the appendix.

The bias observed in unadjusted estimates of Se and Sp, as plotted against prevalence, exhibits a behaviour similar to that seen in situations involving CI, where the accuracy parameters are underestimated (see Figure 19). However, in cases of overestimation, the bias shows the opposite trend, increasing at extreme prevalence values.

Figure 19 presents a comparative analysis of the impact of different adjustment methods on test accuracy bias, as related to disease prevalence. The x-axis represents disease prevalence, while the y-axis shows the weighted accuracy bias. In Figure 19a, the broad distribution of brown and green dots illustrates significant bias that the adjustment methods fail to correct. Moving to Figure 19b, the distribution becomes more concentrated with blue dots, potentially indicating a reduction in bias through the application of the Staquet method. However, this method's applicability is limited, as it sometimes produces illogical results, and therefore, what we think could be corrected is, in fact, scenarios that we are not observing due to illogical results. Figure 19c, marked by blue dots, reveals the Brenner covariance method's tendency to slightly reduce the weighted test accuracy bias at prevalence below 0.5 and to overestimate it above this threshold, showcasing the nuanced performance of these methods in bias correction. Surprisingly, the Staquet method achieves a bias reduction in 28.1% of the logical scenarios examined for test 1 (n=4,030), and 26.9% for test 2. Alternatively, the Brenner covariance method does so in 100% of cases for test 1, and 77.7% for test 2, highlighting the varying efficacy of these approaches in managing test accuracy bias.

Figure 19. Weighted accuracy bias versus prevalence when CI assumption is relaxed. Figure A) illustrate the estimations for the unadjusted. Figure B) Corrected estimation using Staquet approach. Figure C) Corrected estimation using Brenner approach. The illogical scenarios are showed in grey in plots B) and C).

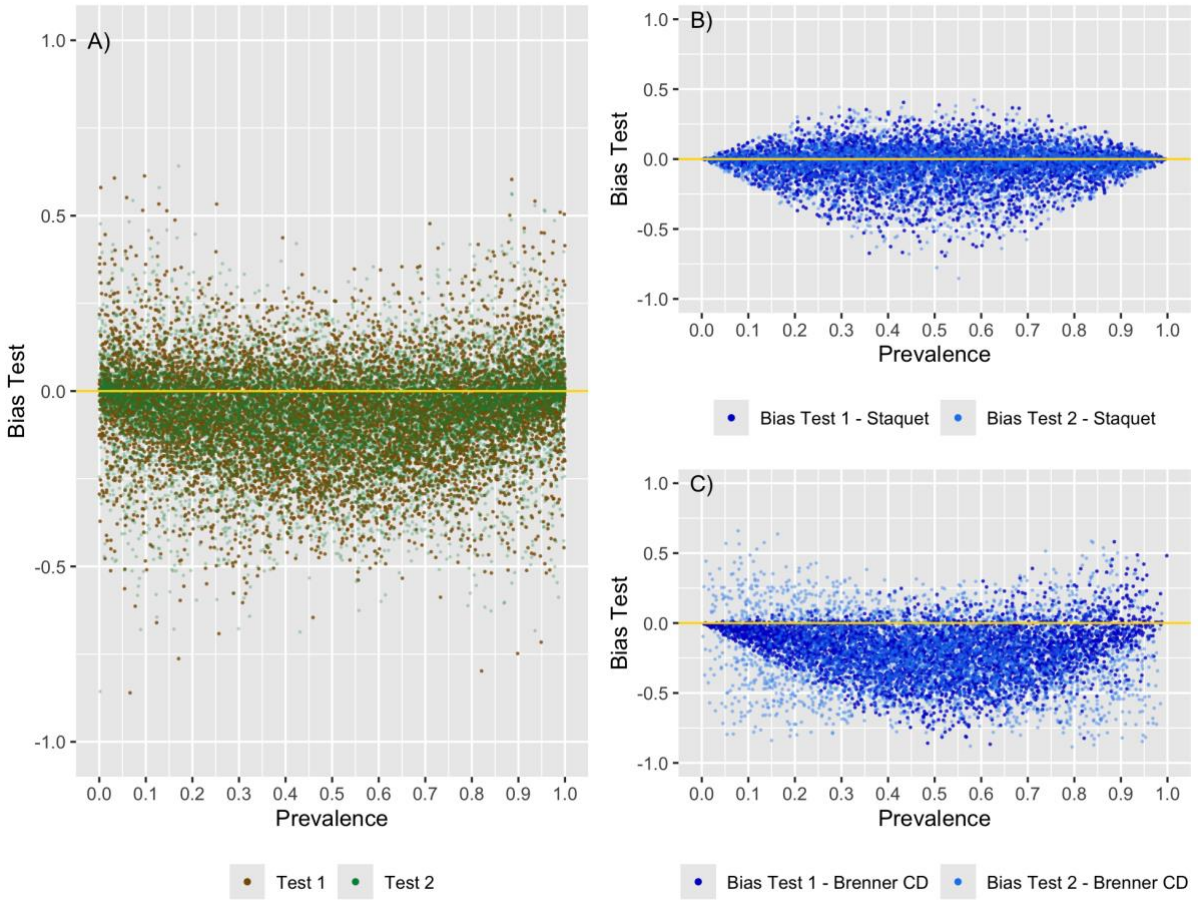
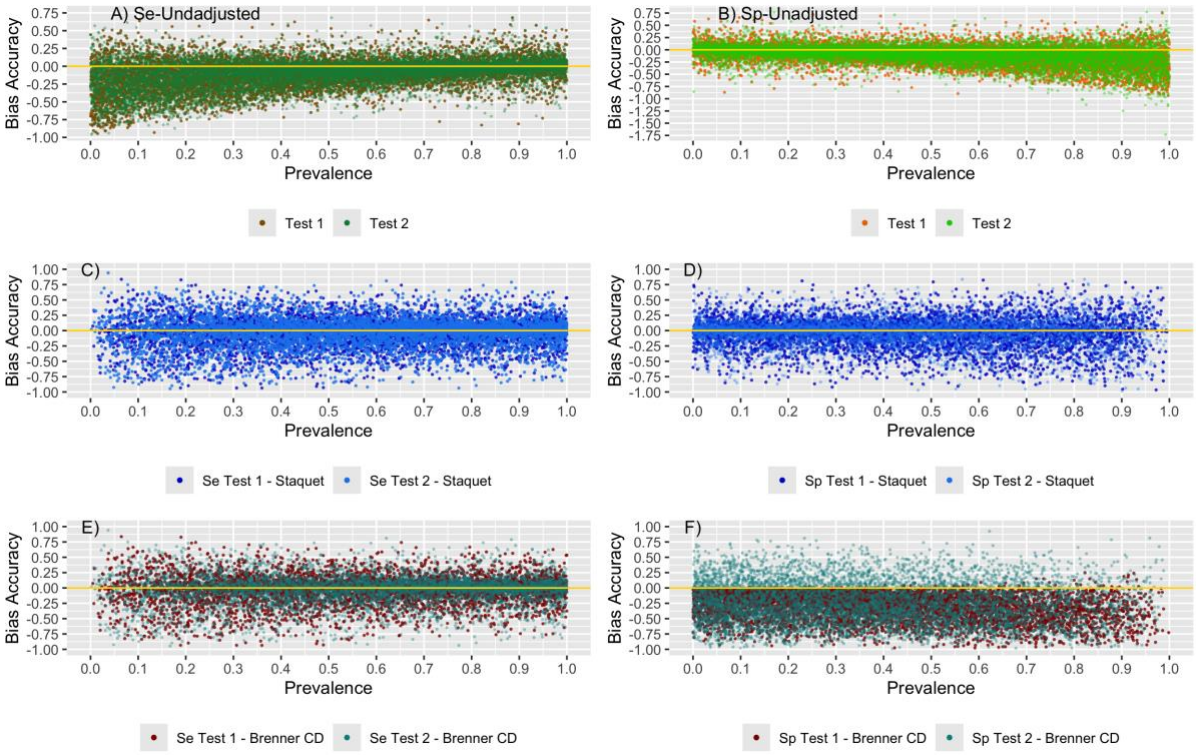


Figure 20 illustrates the bias for both Se and Sp (y-axis) independently across the prevalence (x-axis) for test 1 and test 2. Figure 20a and Figure 20b depict the bias for the unadjusted accuracy parameters, showing similar behaviour to the CI scenarios when results are underestimated. In these cases, the Se bias increases as prevalence decreases, and the Sp bias increases at higher prevalence values. However, when the parameters are overestimated, no clear pattern is observed, and the magnitude of the bias is not influenced by prevalence.

Although correction methods tend to reduce bias in certain situations, we found that in many scenarios where the bias was low — high prevalence with Se and low prevalence with Sp — using correction methods often increased the bias, though it did decrease it in some cases (see Figure 20c-f). Therefore, both correction methods are ineffective at correcting the Se and Sp when the assumption of CD is relaxed.

Figure 20. Prevalence vs Test accuracy bias. A) Se bias for unadjusted estimates, B) Sp bias for unadjusted estimates. C) Se bias for corrected estimates using Staquet approach, D) Sp bias for corrected estimates using Staquet approach, E) Se bias for corrected estimates using Brenner CD approach and F) Sp bias for corrected estimates using Brenner CD approach.



5.3.1.5 Bias in cost-effectiveness analysis owing to imperfect reference standard bias.

Figure 21 plots corrected iNHB bias on the y-axis against disease prevalence on the x-axis, revealing that bias in iNHB does not show significant differences across the prevalence, illustrating high variability across the scenarios.

As previously described, correction methods could potentially increase the bias, especially at central values of prevalence (approximately 0.5). Moreover, the bias reduction observed at extreme prevalence values may be explained by the existence of uncorrected scenarios that result in illogical estimates when correction methods are applied.

Furthermore, the effectiveness of the correction methods can be gauged by the instances in which they led to a reversal of decision-making. The Staquet approach resulted in a decision reversal in 0.48% of the cases, whereas the Brenner CD method did so in only 0.30% of the cases. Interestingly, most of these reversals tended to occur when the other correction method did not reverse the decision. For instance, when a scenario was reversed using the Staquet approach, this was not the case when the Brenner CD method was applied, and vice versa.

Complementing these results, Figure 21 shows the complex and non-linear relationship between biases, where the iNHB bias is on the x-axis, and the weighted test accuracy bias is on the y-axis. Additionally, Figure 21b and Figure 21c shows the few scenarios where the decision was reversed by using the correction methods.

Figure 21. Bias iNHB versus Prevalence. A) Unadjusted iNHB for test 1. B) Unadjusted iNHB for test 2. C) Corrected iNHB using Staquet approach for both test 1 and test 2. D Corrected iNHB using Brenner CD approach for both test 1 and test 2. E) Unadjusted iNHB for test 1 considering that test 2 is not biased.

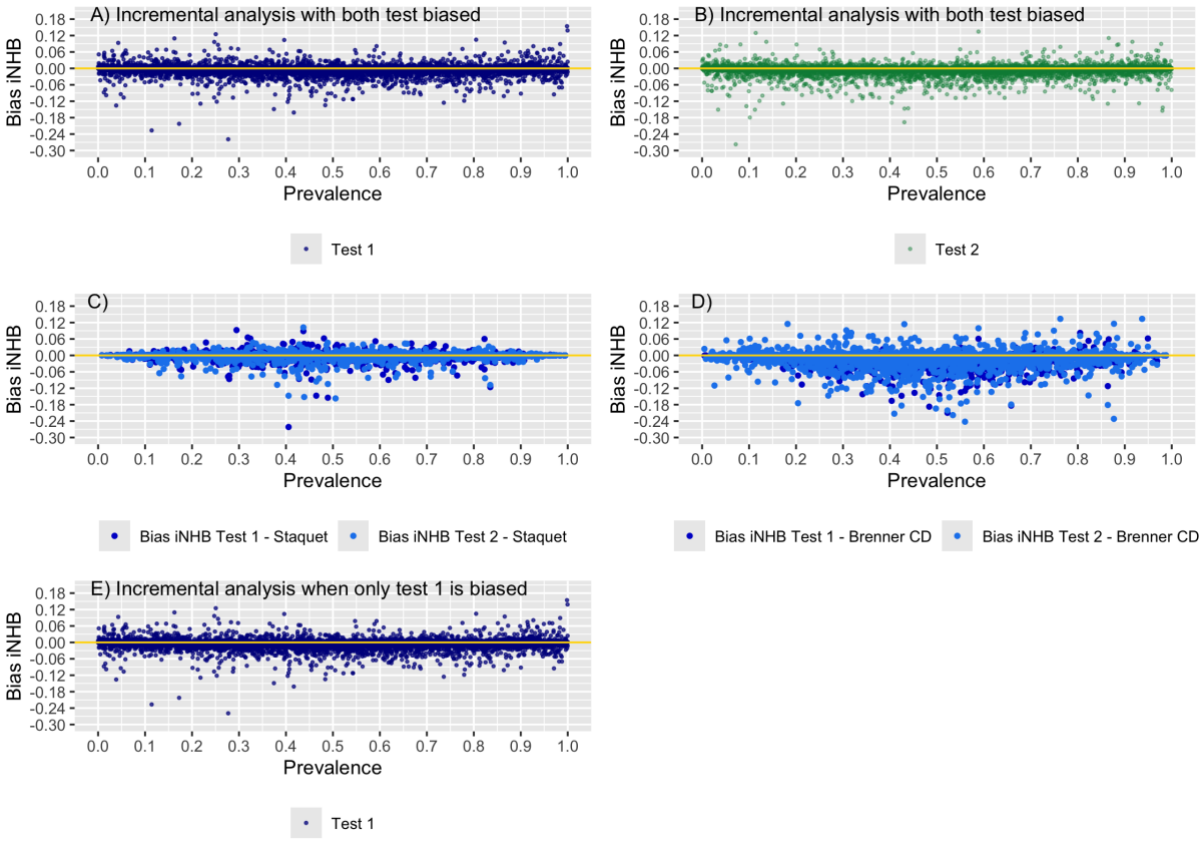


Figure 22. Relationship between weighted test accuracy bias with iNHB bias. A) For unadjusted estimates. B) Corrected estimates using Staquet. C) Corrected estimates using Brenner CD.

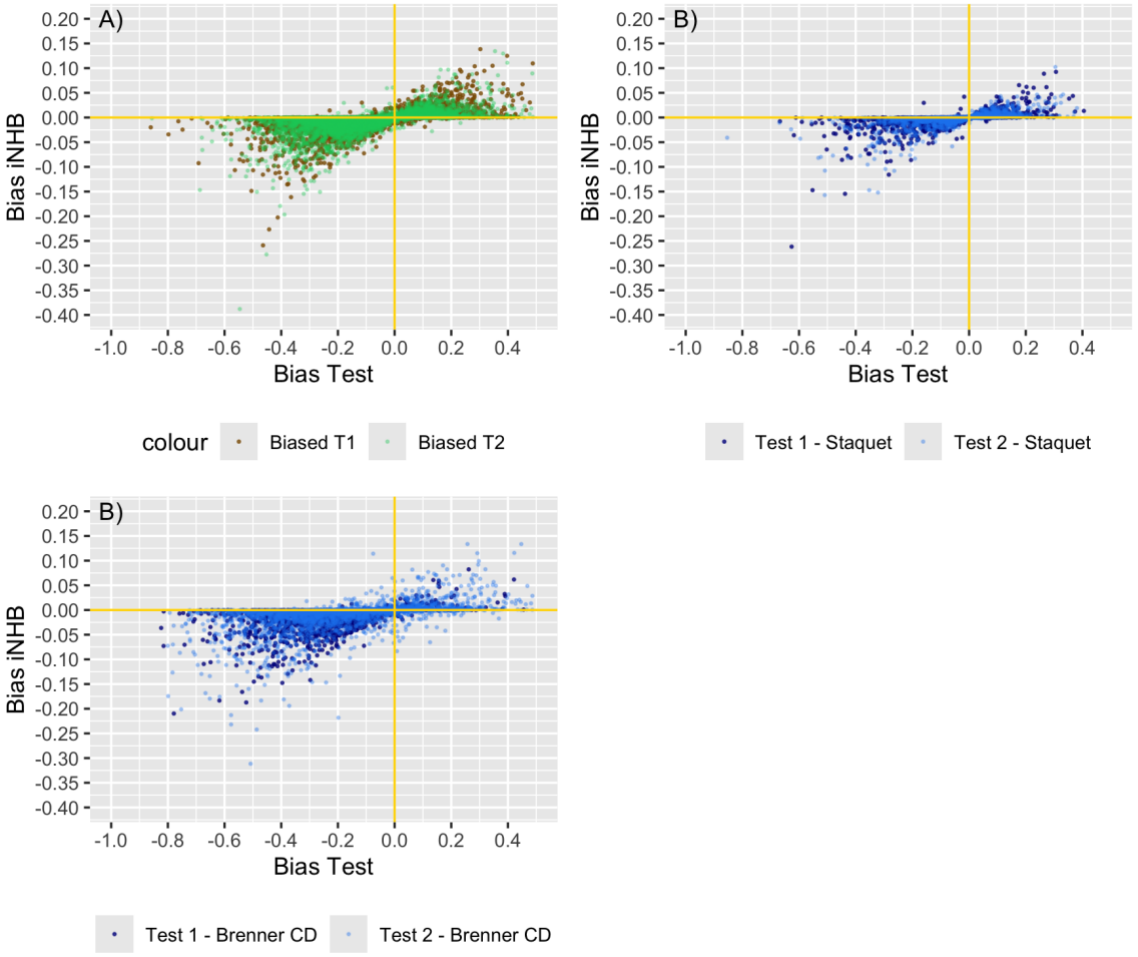


Figure 23 provides a comparative visualisation of the relationship between the true incremental net health benefit (iNHB) and its unadjusted or corrected counterparts, presented across separate subplots. Central to these comparisons is the equivalence line (depicted in red), representing the point at which the true iNHB aligns precisely with the estimated iNHB. This equivalence line thus serves as a reference for unbiased estimation.

Figure 23a, attention is focused on the unadjusted iNHB values, plotted on the x-axis. The distribution of brown and green data points—representing the "True vs Biased" comparisons for test 1 and test, respectively—is notably dispersed. The considerable scatter observed relative to the equivalence line underscores significant discrepancies

between the true and unadjusted iNHB values, manifesting as both overestimations and underestimations.

Figure 23b illustrates iNHB values corrected using the Staquet approach, excluding illogical outcomes. Blue points signify scenarios where corrections led to reversed decisions, whereas grey points indicate cases without such reversals. The corrected cases, particularly those marked by blue points, exhibit improved proximity to the equivalence line, suggesting enhanced alignment of the iNHB estimates with their true values. Nevertheless, despite this improvement, the overall performance of the Staquet method in terms of bias reduction remains limited.

Figure 23c further examines corrected iNHB values, this time employing the Brenner CD method, again excluding illogical scenarios. The blue points indicating reversed decisions exhibit tighter clustering around the equivalence line compared to the Staquet approach. This closer clustering indicates relatively improved performance of the Brenner method. Notably, the nature of bias correction differs between the two approaches: the Staquet method typically corrects underestimated unadjusted estimates, whereas the Brenner CD method predominantly addresses overestimated scenarios.

Through this nuanced comparison, Figure 23elucidates the differential impacts of the Staquet and Brenner CD correction methods on the accuracy of iNHB estimation. It highlights each method's relative effectiveness in aligning corrected iNHB values with their true counterparts, underscoring the critical importance of method selection based on the specific nature of bias present in unadjusted estimates.

Overall, the analysis demonstrates the significance of selecting an appropriate method for estimating iNHB to ensure accuracy. Although certain correction methods produce illogical estimates, the adjustment of logical cases notably reduces bias and influences decision-making. Specifically, decision reversals tend to occur when correcting negative iNHB values with the Staquet approach and positive values with the Brenner CD method. Nevertheless, even corrected estimates are not entirely free from bias, contrasting with scenarios involving complete information. This finding highlights the necessity for careful selection of estimation methods to achieve reliable outcomes.

Figure 23. True iNHB vs Unadjusted iNHB. Figure A) illustrate the estimations for the unadjusted. Figure B) Corrected estimation using Staquet approach. Figure C) Corrected estimation using Brenner approach. In grey is showed the scenarios where correction methods did not reverse the decision.

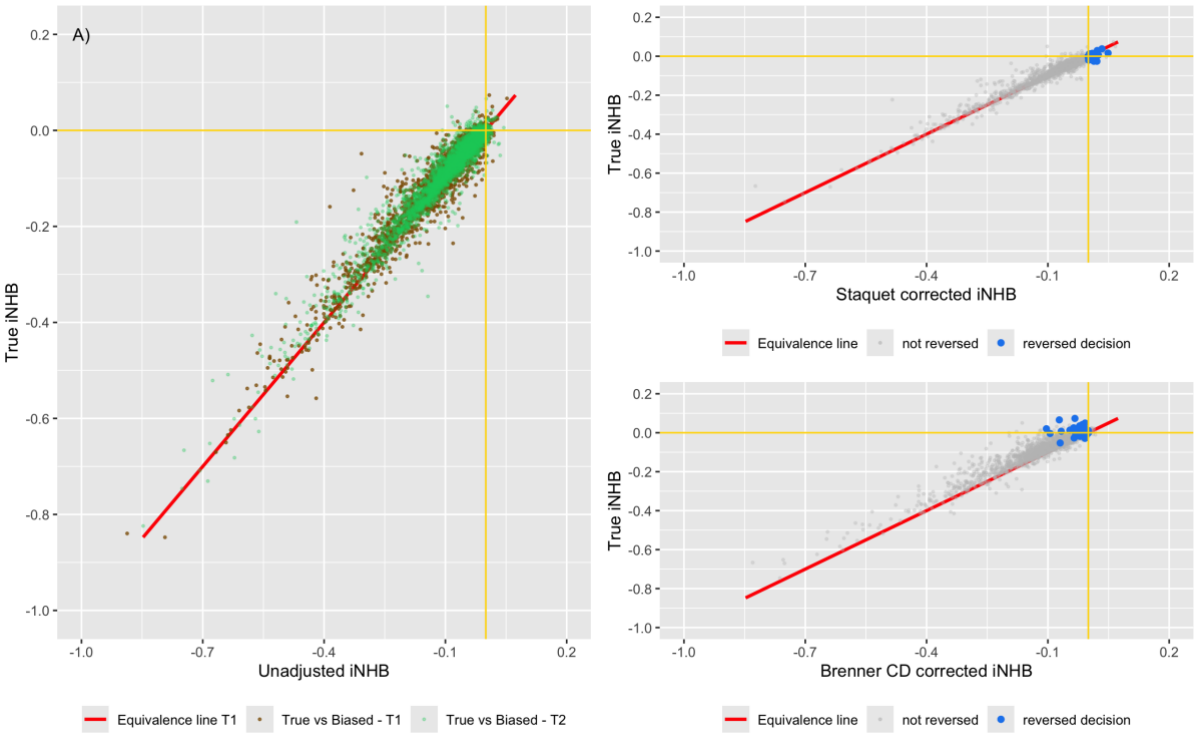


Figure 24 illustrates the magnitude of bias for the VoD and VoH through density plots, summarising the distribution of bias across simulated scenarios.

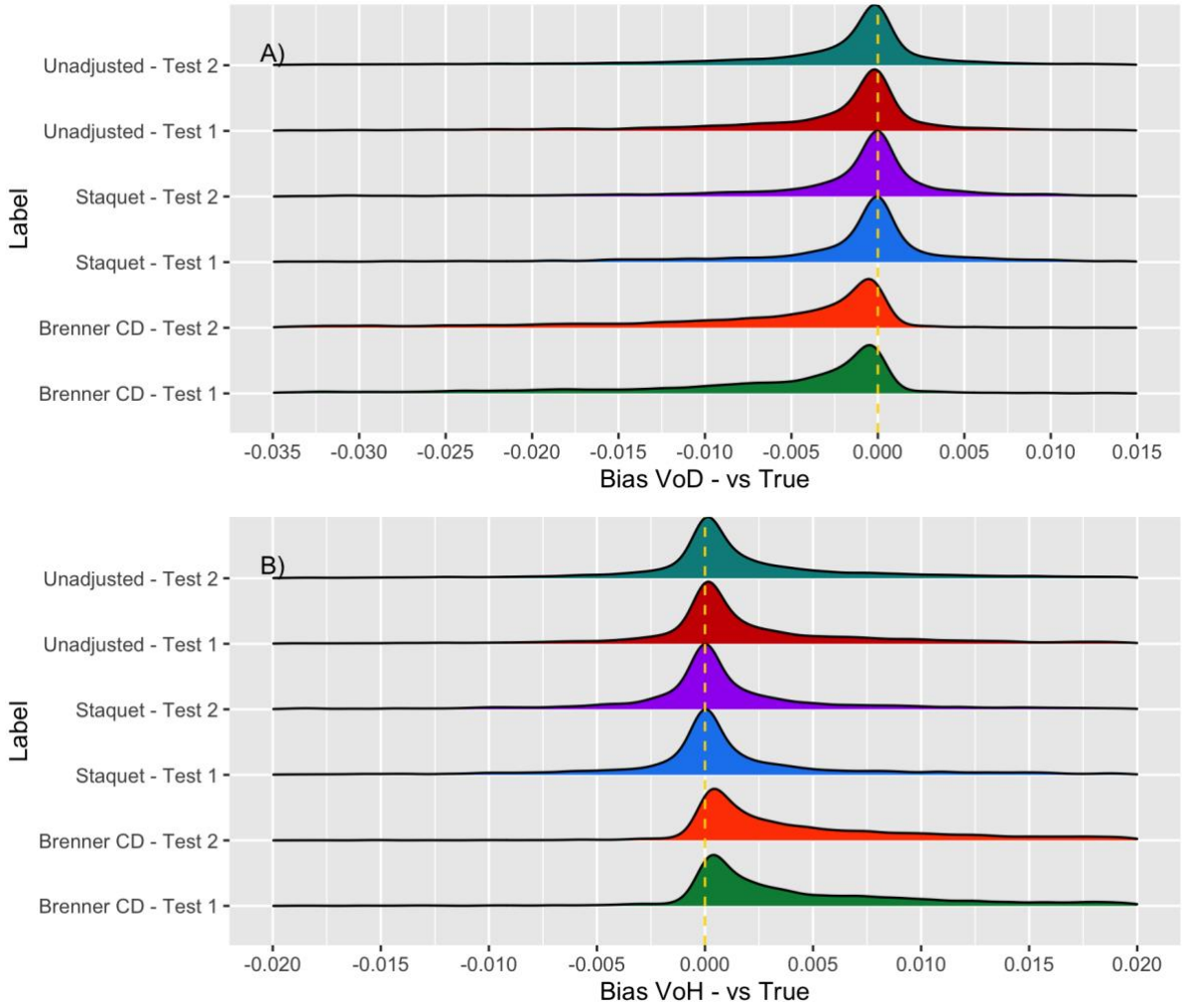
In Figure 24a, the analysis of bias in the VoD reveals both underestimation and overestimation in the unadjusted results for both testing strategies, consistent with earlier findings. Evaluating the results from correction methods, it becomes apparent that their efficacy is limited, particularly when considering CD scenarios. The Staquet correction approach yields a bias distribution closely mirroring that of the unadjusted results, indicating minimal improvement. Conversely, the Brenner CD approach demonstrates a tendency to reduce overestimated results; however, this correction comes at the expense of increasing the frequency of underestimated cases. This shift is evidenced by the right-skewed distribution, contrasting with the more symmetrical distribution observed in the unadjusted results.

Figure 24b summarises the bias analysis for VoH. The resulting distribution pattern closely parallels that of the VoD analysis. Unadjusted VoH estimates display an approximately symmetrical distribution around the zero line, signifying a balance

between underestimated and overestimated cases. Similar to VoD, the Staquet-corrected VoH estimates show minimal deviation from the unadjusted results. However, application of the Brenner CD approach notably shifts previously overestimated scenarios into underestimated scenarios, resulting in a left-skewed distribution.

Overall, Figure 24 highlights the relative limitations of both correction methods, with the Brenner CD approach notably altering the bias distribution at the risk of shifting the type of bias rather than achieving comprehensive bias correction.

Figure 24. Bias in the Value of diagnostic (VoD, showed in the upper figure) and in the value of heterogeneity (VoH, showed in the bottom figure).



5.4 Discussion

This thesis set out two primary objectives, and the simulation results directly inform both. The first objective was to review and summarise methods for adjusting diagnostic accuracy when an iRS is used, and to examine how such adjustments have been (or could be) integrated into economic evaluations. In Chapter 3, a comprehensive literature review identified various approaches, ranging from simple count-based corrections (Staquet et al., 1981) to proportion-based adjustments (Brenner, 1996) and more complex latent class models (Dendukuri and Joseph, 2001, Umemneku Chikere et al., 2019). This review highlighted that, despite a rich methodological literature on handling iRS bias in diagnostic accuracy studies, integration of these methods into health economic models remains limited. Few published cost-effectiveness analyses of diagnostic tests explicitly adjust for bias introduced by an iRS; most assume the reported Se and Sp are unbiased, effectively treating the iRS as a gold standard. This gap set the stage for the second objective: using simulation studies to investigate how iRS bias might propagate to population-level decision-making, costs, outcomes, and uncertainty in an HTA context.

The simulation study was designed to fulfil the second objective and did so by constructing a stylised decision-tree model for a diagnostic strategy under various scenarios of test accuracy and disease prevalence. Crucially, the simulation introduced controlled bias in test Se and Sp by assuming an iRS with known error rates. By comparing scenarios with no correction (naïve use of biased accuracy) versus those with bias correction methods applied, the analysis directly addressed how failing to adjust for iRS bias can affect cost-effectiveness results. The results demonstrate that the objectives were met: the simulation confirmed that iRS bias can meaningfully skew economic outcomes, and it evaluated the extent to which existing correction methods (identified under Objective 1) can mitigate this. This represents, to the best of our knowledge, a novel synthesis of diagnostic accuracy methodology and economic evaluation. Previous simulation work in health economics has primarily focused on issues like parameter uncertainty and trial design (Bryant et al., 2009, Xie et al., 2022a, Xie et al., 2022b, Xie et al., 2022d), with no prior study examining the explicit impact of RS misclassification on CEA. By addressing this gap, the thesis achieves both of its aims:

it not only catalogues the methodological tools available for iRS adjustment but also provides quantitative insight into their importance (or lack thereof) for decision-making through a rigorous simulation approach.

The findings carry important implications for HTA, particularly in decision-making contexts like those of the National Institute for Health and Care Excellence (NICE) in the UK. In HTA, the adoption of a new diagnostic test often hinges on evidence of improved outcomes and cost-effectiveness compared to current practice. If the diagnostic accuracy evidence is biased by an iRS, there is a risk that HTA decisions could be misguided**.** Our simulations indicate that when test Se or Sp is misestimated due to iRS bias, the resulting cost-effectiveness metrics (such as incremental cost-effectiveness ratios or net health benefits) can be distorted. For instance, a test that truly has high Se might appear less effective if the RS misses some disease cases (leading to an underestimate of the test's true positives), potentially causing the test to be undervalued in the model. Conversely, a test's benefits might be overestimated if the RS's imperfections inflate the test's apparent performance. In a threshold-based decision setting, even a modest bias in estimated QALYs or costs can tip the recommendation on whether a diagnostic is approved or not.

In practical terms, this means NICE and similar agencies should be cautious when appraising diagnostics that lack a definitive gold standard. The NICE diagnostics assessment process (National Institute for Health and Care Excellence, 2022b) does acknowledge that an "iRS" can lead to bias in accuracy estimates, but currently the approach to handling this is often limited to qualitative considerations or sensitivity analyses. Our results support more formal adjustments: for example, incorporating bias-corrected Se and Sp in base-case or scenario analyses. In recent NICE diagnostic guidance (e.g. rapid tests for streptococcal infections (National Institute for Health and Care Excellence (NICE), 2019); and the PredictSURE IBD test for Crohn's disease (National Institute for Health and Care Excellence (NICE), 2022)), the evidence was based on comparative accuracy studies where the RS may not have been perfectly sensitive or specific. Had our bias adjustment framework been applied in such cases, the conclusions about cost-effectiveness might differ – either reinforcing confidence in a truly effective test or highlighting uncertainty for a test whose benefits were overstated due to RS errors. This is especially pertinent in high-impact clinical areas (e.g. cancer

screening or infectious disease diagnostics) where diagnostic uncertainty translates directly into patient outcomes and healthcare costs (Dendukuri and Joseph, 2001, Menten et al., 2008). Decision-makers should consider requiring analysts to assess the robustness of a diagnostic's economic value to potential iRS bias, much as one would test other structural assumptions in a model.

Another implication is in how uncertainty is represented in HTA. Typically, probabilistic sensitivity analysis in cost-effectiveness models will draw samples for Se and Sp from reported distributions (often derived from meta-analyses). If those underlying estimates are biased by an iRS, the entire distribution is shifted away from the truth. Our study shows that this systematic bias will not be captured by confidence intervals or probabilistic analyses unless the model explicitly adjusts for it. Thus, HTA agencies might be overconfident in a decision if they see a narrow cost-effectiveness acceptability curve that in reality is centred on biased inputs. For example, we found that the VoD – the net health benefit gain from using the test strategy – was systematically underestimated when RS bias was ignored, although the magnitude was typically small (on the order of a few thousandths of an NHB unit in our base-case simulations). In a borderline decision, even a small underestimation of benefit could reduce the probability that a cost-effective test is adopted when it should be. Conversely, measures like the value of a perfect test (essentially the upper bound of what perfect information could achieve) were overestimated under iRS bias in our results, which could mislead stakeholders about the potential upside of future improvements or research. In summary, the HTA implication is that bias due to an iRS is a hidden source of error that can propagate through an economic evaluation. Accounting for it leads to more accurate estimates of cost-effectiveness and more appropriate uncertainty ranges, thereby supporting sound decision-making in organisations like NICE.

Staquet's method is a count-based algebraic adjustment that takes the 2x2 classification results of the index test compared to the iRS and uses prior knowledge of the iRS's own Se and Sp to produce corrected estimates of the index test's performance (Staquet et al., 1981). In this thesis, Staquet was the primary correction technique when we assumed CI between the index test and RS errors. Under CI, test errors are not

systematically correlated—i.e. whether the iRS misclassifies a given patient is not directly related to whether the index test does so.

In the CI scenarios of our simulation, Staquet generally performed well in retrieving near-unbiased estimates of Se and Sp (Umemneku Chikere et al., 2021). Across a wide range of prevalence values, we observed that naive (unadjusted) accuracy could be severely distorted by iRS bias (especially at very low or high prevalence), whereas Staquet's correction moved the estimated Se and Sp closer to the "true" values. In turn, when these corrected values were fed into the decision model, the incremental net benefit and overall cost-effectiveness findings aligned much more closely with the scenario in which the test was measured against a perfect RS. The utility of the Staquet method thys depends critically on whether the CI assumption can be defended in the specific clinical and methodological context. If the assumption of CI is valid (or close to valid), the Staquet approach can successfully mitigate iRS bias and reduce the risk that an incorrect adoption/rejection decision might be made. However, when test and RS errors are correlated-a realistic scenario in many clinical contexts where tests rely on similar biological processes or measurement principles-Staquet's performance deteriorates markedly, as demonstrated in our CD simulations.

Nevertheless, we observed two limitations for Staquet. First, it depends on having external or literature-based estimates of how imperfect the RS is. If the accuracy of that RS is itself highly uncertain (or estimated from a small validation sample), Staquet's corrected values may carry additional uncertainty that we did not fully propagate here. Second, and more critically, Staquet's algebraic corrections can produce illogical estimates (e.g. Se <0 or >1) in extreme conditions—particularly under very low or very high prevalence, small sample sizes, or when the core assumption of CI is violated. Consistent with reports by Umemneku Chikere et al. (2021), these "impossible" estimates primarily appeared when prevalence was outside the ~0.2–0.8 range or when correlation between the index test and iRS existed (which directly violates CI). Although such scenarios were relatively rare in our simulations, they underscore that the Staquet method lacks an internal mechanism to constrain Se and Sp to [0, 1] under all sampling conditions. Analysts must therefore interpret Staquet-corrected values with caution, especially in studies characterised by extreme prevalence or where there is a known correlation between test errors.

Where test and iRS errors were explicitly correlated, we turned to Brenner's approach for CD (Brenner, 1996), referred to here as Brenner CD. This variant extends Brenner's original "alloyed gold standard" framework to account for the possibility that the iRS's classification error may be positively or negatively correlated with the index test's error (e.g. if both tests are based on similar biochemical principles, or if patient/clinical factors influence both). Brenner CD works by incorporating covariance terms into the correction formula, effectively adjusting for the partial overlap in misclassifications between the index test and iRS.

In principle, Brenner CD should provide more appropriate estimates under correlated errors than would a method (like Staquet) that strictly assumes independence. However, in our simulations, Brenner CD often struggled in practice. Although it avoided the "illogical" estimates sometimes seen under Staquet, we observed instances where Brenner CD led to substantial residual bias—particularly an underestimation of the true Se and Sp. In other words, the method tended to produce corrected estimates that erred on the conservative side, potentially "penalising" the index test's performance. In cost-effectiveness terms, this could lead to undervaluing a test that truly offers significant health benefits. The reasons for this underestimation may lie in Brenner's original derivation, which was oriented toward correcting measures of association (e.g. exposure–disease odds ratios) and not always calibrated for direct estimation of Se/Sp. Moreover, in our simulations, the covariance structure and parameter ranges were intentionally wide to represent many real-world possibilities. Under certain parameter combinations (e.g. strong positive correlation plus moderate prevalence), we found that Brenner CD did not fully eliminate iRS bias and sometimes introduced new distortions in the final estimates.

A further complication is that Brenner CD can produce a high rate of "illogical" outcomes if the underlying assumptions about the sign and magnitude of correlation are not carefully constrained. In practice, we imposed mathematically valid bounds on the covariance (following Vacek (1985)) and used rejection sampling to ensure only biologically plausible parameter sets were included. Even so, many simulated scenarios yielded borderline or out-of-bounds corrections. This highlights that while Brenner CD is explicitly designed for CD, real-world complexities (e.g. varying degrees of positive or negative correlation, or small sample sizes) may still hamper its reliability.

Taken together, the results indicate that neither Staquet nor Brenner CD is a panacea once correlation in errors is introduced. Staquet ignores correlation altogether and therefore is not theoretically correct in a dependent scenario, whereas Brenner CD attempts to handle dependence but may lead to high variance or systematic underestimation of accuracy if the covariance structure is not well estimated. These issues mirror the concerns expressed by Okkaoglu et al. (2024) and others, who emphasise that violating independence can markedly distort estimates from simpler correction approaches. In more complex or data-rich contexts, Bayesian latent class models (Dendukuri and Joseph, 2001, Menten et al., 2008) might prove superior because they can explicitly model correlations between multiple imperfect tests under a single unified framework. Such models, however, demand more advanced statistical expertise and often large datasets to ensure robust parameter estimation.

From the perspective of CI, our simulations show that Staquet is an attractive option: it can markedly reduce bias introduced by an iRS, thereby stabilising cost-effectiveness outputs. Yet when the assumption of independence is even partly violated, Staquet no longer guarantees reliable estimates. Brenner CD is intended for correlated test errors but can still yield problematic outcomes when correlation is strong or the data are sparse. Consequently, both methods must be carefully matched to the nature of the testing scenario and the available data.

In practical diagnostic HTA, analysts would ideally (1) confirm or refute CI using clinical/technical knowledge of the tests, (2) consider the magnitude and direction of any correlation in errors, and (3) choose (or develop) a correction approach appropriate for those conditions. If strong correlation is suspected, one might either attempt a more sophisticated method (e.g. Bayesian modelling) or conduct sensitivity analyses around the level of correlation to gauge how much it alters the final cost-effectiveness decision. Although the simulations underscore that simpler algebraic methods can improve on naive estimates, they also highlight the risk of misapplication. Ultimately, the analyst should remain vigilant about how assumptions—particularly about test error dependence—will propagate through to clinical and economic outcomes.

A broader methodological takeaway from this work is the demonstrated value of simulation modelling in understanding how biases propagate through decision models. By using a simulation-based approach, we were able to systematically manipulate the

presence and degree of RS imperfection and observe the consequences for diagnostic accuracy measures and economic outcomes. This approach provided several advantages. First, it allowed us to establish a "ground truth" for each scenario – we knew the true Se, Sp, costs, and outcomes because we specified them – and then to compare the observed outcomes under an iRS to that truth. Such a comparison is generally impossible with real-world data, where the true disease status is by definition unknown when the reference stadanrd is imperfect. Simulation thus gave us a unique lens to quantify bias directly (Morris et al., 2019). For example, we could explicitly measure how many QALYs were lost or misallocated due to RS misclassification, which is not feasible to detect in a live HTA without a perfect reference stadnard.

Second, the simulation was able to cover a broad range of scenarios and parameter values, far beyond what any single empirical study could examine. We employed Latin Hypercube Sampling to efficiently explore the multidimensional parameter space (McKay et al., 1979), running 10,000 simulated scenarios that varied disease prevalence, test characteristics, and costs. This breadth is important because the impact of iRS bias is context-dependent. By sampling across low, medium, and high prevalence settings, and across different cost/QALY trade-offs, we could identify patterns – for instance, bias in sensitivity estimation was most pronounced at the extremes of prevalence (very low or very high), which in turn had the largest effect on outcomes in those extreme cases. We also captured rare but important combinations (such as a very inaccurate RS paired with a very accurate new test) that might not yet be documented in literature but are plausible in practice. This comprehensive exploration adds robustness to our conclusions: we can be confident that the observations (e.g. Staquet typically corrects bias well; Brenner tends to under-correct; extreme prevalence poses challenges) are not artifacts of one particular set of assumptions, but hold across many plausible worlds. In contrast, a non-simulation approach (like a purely analytical derivation) might have required simplifying assumptions or could only address a narrower question (Morris et al., 2019). Our simulation effectively serves as a "stress test" of the decision model under bias, revealing its behaviour under both typical and edge-case conditions.

Furthermore, simulation modelling proved valuable for comparing methods head-to-head in a controlled environment. We were able to apply no correction vs. Staquet vs. Brenner to the same underlying scenarios to see how each method performed in parallel.

This is a powerful way to assess methodological performance (Umemneku Chikere et al., 2021, Morris et al., 2019). The side-by-side comparison under identical conditions enhances the validity of our recommendations on which correction method to prefer. It also allowed us to discover interactions – for instance, how the performance of Staquet's correction is influenced by prevalence, or how Brenner's method behaves when the test is better vs. worse than the RS. These nuanced insights are a direct product of the simulation approach.

Finally, the use of simulation has a didactic value for the HTA community. It illustrates in concrete terms how a bias in a diagnostic study can ripple through to influence cost per QALY. Often, these concepts are conveyed only abstractly in guidelines (National Institute for Health and Care Excellence (2022b) mentions the possibility of iRS bias but provides little quantitative guidance). By contrast, our simulation puts numbers on the issue – for example, showing that under certain conditions an unadjusted analysis might misestimate net health benefit by a small fraction (e.g. 0.005 QALYs), which might seem negligible, but in other cases the error could be larger and even change the optimal decision. This kind of insight is only possible through modelling. As such, this thesis advocates for greater use of simulation modelling as a tool to understand and communicate methodological uncertainties in HTA. Just as sensitivity analyses and scenario analyses are standard for exploring uncertainty in economic evaluations, simulation studies can play a complementary role in exploring structural uncertainties or biases that are not easily captured by closed-form analytical solutions.

Beyond the specific results, this research makes several broader methodological contributions. First, it demonstrates a framework for integrating diagnostic accuracy correction methods into economic evaluation. Previous applications of bias correction were largely confined to the clinical epidemiology domain (e.g. adjusting Se/Sp in diagnostic accuracy studies). We have shown how to carry those adjustments through to the endpoints that matter in HTA – costs and QALYs – thereby bridging a gap between diagnostic test evaluation and CEA. This is a step toward more holistic diagnostic HTA models that simultaneously consider clinical validity and economic value. Importantly, the framework is generic: although we used a stylised decision tree, the same approach could be applied to more complex models (such as Markov models or microsimulations) and to different diseases. By laying out this framework and identifying where biases enter

and how they can be corrected, the thesis provides a template for future researchers facing similar issues.

Second, the research offers insights that can inform HTA methodology guidelines. The findings suggest that when dealing with a diagnostic test evaluated against an iRS, analysts should at minimum conduct a sensitivity analysis to assess the impact of potential misclassification bias (as also recommended conceptually by National Institute for Health and Care Excellence (2022b)). Our results go further to indicate that a quantitative bias adjustment (like Staquet's) should be considered, especially if there is evidence or suspicion that the RS is significantly imperfect. By quantifying how much error in cost-effectiveness results can arise (or be fixed) through such adjustments, this work provides an evidence base to support those recommendations. It contributes to the broader movement in HTA of improving decision models by incorporating better characterisations of uncertainty and bias (Skedgel et al., 2022, Umemneku Chikere et al., 2021). In essence, this thesis adds to the methodological toolkit available for diagnostic assessments and encourages its adoption in practice.

Third, this work highlights the conditions under which bias might materially affect decision outcomes. Not every bias will change a policy decision, and resources for analysis are finite. Our simulation indicates that in many scenarios, the cost-effectiveness rankings of interventions (e.g. test vs no test) remained the same with or without bias correction – in other words, the bias, while present, did not flip the decision. Typically this was the case when the differences in outcomes were large enough that a small accuracy bias was negligible, or when both the new test and the comparator were similarly affected by the same bias (cancelling out in incremental terms). However, we also identified scenarios where the bias did have a decision-altering impact, such as when a new test's advantage over standard care was marginal. A small underestimation of the new test's effectiveness (due to an insensitive RS) could make it appear not cost-effective when in truth it was (a false negative decision for adoption), and vice versa. These findings can help decision-makers and analysts to prioritise which evaluations need advanced adjustments: for example, tests that on initial analysis are borderline or for which the RS is known to be problematic should receive more careful bias scrutiny. This kind of prioritisation is a methodological contribution to the practice of evidence assessment.

Finally, the thesis contributes to the academic literature by comparing and contrasting two specific correction methods in the context of economic evaluation – something not previously done. While Umemneku Chikere et al. (2021) compared multiple bias correction methods on statistical grounds (accuracy of Se/Sp estimates), our work extends this comparison to the economic consequences of using those methods. The conclusion that in scenarios assuming CI, Staquet's method outperforms in preserving cost-effectiveness results is an original contribution that can guide both future research and immediate practice.

Notwithstanding its contributions, this research has several limitations that should be acknowledged, many of which point to directions for future research. First, the simulation scenarios were necessarily simplified. We modelled a single test and single (imperfect) RS in a decision-tree structure. Real-world diagnostic pathways can be more complex – for example, involving multiple tests (in sequence or combination), or tests that provide continuous results rather than dichotomous outcomes. Extending the bias adjustment framework to such complex scenarios is a logical next step. Future research could examine, for instance, how iRS bias affects a sequence of tests (like an initial screening test followed by a confirmatory test) or how to adjust for iRS in situations of composite reference standards (where multiple imperfect measures are used to adjudicate disease status). These expansions would increase the applicability of the findings to a wider array of HTA problems.

Second, the issue of CD between test and RS errors, which we touched upon, deserves deeper investigation. Our work confirmed that when the test and the iRS are not independent, simple correction methods break down. However, our exploration of dependence was limited in scope. Future studies could systematically vary the degree of correlation between test and RS misclassification errors to map out the impact on both accuracy estimates and economic outcomes. Additionally, developing or testing alternative methods that can handle dependent errors would be valuable. One avenue is to use Bayesian latent class models that include a covariance parameter (Dendukuri and Joseph, 2001), integrating those models with cost-effectiveness calculations. This would be computationally intensive but feasible with modern Bayesian software, and it would enable a more holistic propagation of uncertainty, where the inference about true test performance (accounting for dependency) is directly fed into the economic model. Such

an approach could also naturally incorporate uncertainty about the RS's performance, something we handled in a deterministic way.

Third, the practicality of implementing bias corrections in HTA needs further exploration. While the Staquet method is conceptually straightforward, analysts may face challenges in obtaining the required inputs. Research could focus on how to elicit or estimate the accuracy of an iRS. For example, could one use expert elicitation to quantify the Se and Sp of a commonly used but imperfect diagnostic benchmark? Or could one design studies that measure patient outcomes under different RS assumptions to back-calculate RS accuracy? Moreover, user-friendly tools or guidelines could be developed for HTA practitioners to apply these corrections. Currently, the use of such methods is rare in HTA, possibly due to lack of awareness or perceived complexity. Demonstrating their use on real case studies would help. A future case study could take a published NICE diagnostic model and re-analyse it with bias correction to see how results change; this would concretely illustrate the value (or limitations) of the approach in practice and might accelerate adoption.

Fourth, linking bias adjustments to value of information analysis could be an important research direction. HTA decisions often consider whether further research is worthwhile. If a diagnostic's cost-effectiveness is highly uncertain partly because of iRS bias, then one could compute the Expected Value of Perfect Information (EVPI) specifically for the RS accuracy. This "value of a better reference standard" might justify investing in a study that uses a more definitive RS (for instance, long-term follow-up of patients to ascertain true outcomes, or the development of a new gold standard). Our thesis did not explicitly calculate such measures, but it paved the way by quantifying how much bias in the RS can move the economic results. Future work could build on this to determine when eliminating RS bias would appreciably change the decision (and thus be worth the effort and cost).

Finally, while our simulation was extensive, there is scope to improve the realism of simulation inputs. We drew many parameters from uniform distributions to be agnostic about what scenario might occur. A possible extension is to base simulations on specific disease areas using evidence-based priors – for example, simulate around the range of Se/Sp reported for a particular disease's tests, and cost/QALY inputs from relevant cost-effectiveness analyses. This would produce findings directly translatable to that context

(e.g., “in tuberculosis testing, failing to adjust for iRS bias of X% could mislead decisions by Y amount of QALYs”). It would also allow incorporating correlations between parameters (beyond the diagnostic accuracy realm) to see how combined uncertainties play out. Incorporating patient-level heterogeneity (for instance, through microsimulation) is another future path, examining if iRS bias might disproportionately affect certain subgroups or distributional outcomes – a consideration that could link to HTA discussions on health equity.

In conclusion, this simulation study has illuminated the impact of iRS on the economic evaluation of diagnostic tests and examined how methodological corrections can address the resultant bias. The thesis objectives were addressed by, first, synthesising the literature on available correction methods and, second, demonstrating through simulation how these methods influence cost-effectiveness outcomes. The results underscore that diagnostic accuracy is not just a clinical or technical issue, but one that can alter health-economic conclusions. Even though in many scenarios the quantitative impact on cost-effectiveness was modest, we have identified plausible conditions where ignoring iRS bias would lead to suboptimal or incorrect decisions. The encouraging news is that we are not without tools to tackle this bias: under the assumption of CI between test and RS erros, simple adjustments like the Staquet method can substantially align the analysis with the true state of the world, thereby supporting better-informed HTA decisions. However, we also caution that these tools have substantial limits, especially when the CI assumption is violated-which is a realistic scenario in many clinical contexts where test and RS errors are collerated through shared biological mechanism or measurement dependencies.

From a broader perspective, this work contributes to improving the rigour and credibility of HTA involving diagnostics. As healthcare moves increasingly toward precision medicine, biomarkers and complex diagnostics, the absence of clear gold standards will become more common, not less. Thus, the issues addressed in this thesis are likely to grow in relevance. By proactively investigating bias correction in economic evaluation, this research adds to the foundation for future guidelines and best practices. It prompts HTA practitioners to ask, “How sure are we about the true accuracy of this test, and what does that mean for our decision?” – a crucial question when lives and resources are at stake. In the end, the goal of HTA is to make the best possible use of evidence for the

benefit of patients and society. Recognising and adjusting for iRS bias is one more step towards that goal, ensuring that when we recommend a diagnostic technology, we do so on the basis of evidence that truly reflects reality.

This chapter presents the simulation study designed to explore the degree and consequences of iRS bias under a range of assumptions, including different disease prevalence levels and conditional dependencies. By comparing the effects of selected bias correction methods on measures of cost-effectiveness, the analysis directly addresses the third thesis objective. The findings here ultimately inform the recommendations discussed in Chapter 6, particularly regarding the practical implications for policymakers.

Chapter 6

6 Discussion

6.1 Key Findings and Theoretical Contributions

This thesis set out to explore how an iRS can bias the economic evaluation of diagnostic tests, and to develop a framework to understand and mitigate this bias. The research introduces the concept of "aligned vs. unaligned" cost-effectiveness models as a way to characterise and resolve discrepancies that arise when diagnostic accuracy is measured against an imperfect benchmark. Through a simulation study, we tested this framework and quantified the extent to which iRS bias can distort cost-effectiveness conclusions. The key findings and contributions of the thesis can be summarised as follows:

Conceptual Framework – Aligned vs. Unaligned Models: We defined an aligned model as one in which all probabilities and outcomes in the decision-analytic model are conditioned on the same reference standard (whether perfect or imperfect), whereas an unaligned model has parameters conditioned on different standards (e.g. test characteristics on an iRS but outcomes on true disease status). This framework is a novel theoretical contribution that builds on ideas of "mismatched" disease models noted in the literature (Denwood et al., 2024). In essence, an aligned model maintains internal consistency and avoids self-contradictory assumptions. We demonstrated that when a model is fully aligned, the relative cost-effectiveness of diagnostic strategies remains unbiased by reference standard imperfections, because any systematic error affects all arms equally. By contrast, unaligned models introduce bias – the misclassification from the iRS skews the estimated costs and health outcomes unevenly, leading to distorted ICERs or net benefit results.

Impact of iRS Bias on Cost-Effectiveness: Using simulation experiments, we showed that iRS bias can indeed alter cost-effectiveness outcomes, although the magnitude varies. In many simulated scenarios, the bias in incremental outcomes (e.g. cost per QALY or net health benefit) was modest, meaning it did not change which diagnostic strategy was preferred. Typically, if a new test was clearly superior or clearly inferior, a small misestimation of Se or Sp due to an iRS did not flip the result. However, in borderline

cases – when the new test's advantage over the comparator was marginal – iRS bias sometimes led to different decisions. For example, if the reference standard missed some true cases (imperfect Se), the new test's benefit could be underestimated, potentially making it look not cost-effective when in truth it would be (a false-negative decision for adoption). Conversely, an overly optimistic reference standard could inflate a test's perceived value, risking a false-positive adoption decision. These findings highlight that while iRS bias will not always be decision-altering, there are plausible conditions where ignoring it would lead to suboptimal or incorrect health policy decisions.

Validation of Correction Methods in an Economic Context: We extended methods originally devised for correcting diagnostic accuracy measures and evaluated them in terms of their effect on cost-effectiveness results. Two correction approaches were examined in depth: the classical Staquet method (Staquet et al., 1981) which adjusts observed Se and Sp given estimates of the reference standard's accuracy, and Brenner's method (Brenner, 1996) which provides an adjustment accounting for potential CD between index test and reference standard errors. While these methods have been discussed in epidemiological literature, their performance in the health economic context had not been systematically compared before. Our simulations showed that Staquet's algebraic correction effectively preserved cost-effectiveness results under the assumption of CI between test and reference standard – in other words, when reference standard misclassification was random with respect to the index test, this simple adjustment realigned the model and produced unbiased incremental outcomes. Brenner's method, designed to handle CD, showed mixed results: it partially corrected biases in scenarios with correlated errors, but did not eliminate them entirely and even introduced slight distortions in certain cases. Notably, we found that if the test and iRS errors were strongly positively correlated (violating the CI assumption), neither correction fully restored unbiased results, underscoring the challenges of dependent misclassifications. This comparative analysis of correction methods' impact on cost-effectiveness is an original contribution of the thesis, moving beyond previous studies that evaluated such methods only in terms of statistical accuracy (Umemneku Chikere et al., 2021).

Bridging Diagnostic Accuracy and Economic Outcomes: A central contribution of this research is methodological integration. We demonstrated how to carry adjustments for diagnostic accuracy through to the endpoints that matter in HTA – costs and QALYs. Previous corrections for iRS bias were largely confined to the domain of diagnostic accuracy studies (e.g. adjusting Se/Sp in a clinical study analysis). We showed, both conceptually and via simulation, that these adjustments can be embedded within a decision-analytic model to adjust outcome estimates. This effectively bridges a gap between diagnostic test evaluation and economic evaluation. By doing so, the thesis moves the field toward more holistic diagnostic HTA models that simultaneously account for clinical validity (correct diagnosis of disease) and economic value (costs and outcomes of subsequent treatment decisions). Importantly, the framework and simulation were kept disease-agnostic and model-agnostic – we illustrated the principles using a stylised decision tree model, but the aligned vs. unaligned logic and correction approach can, in principle, be applied to more complex modelling structures (Markov models, microsimulations) and any disease area where the reference standard is imperfect. This generalisability means our framework acts as a template that future researcher can adapt to their specific diagnostic context.

In summary, the thesis confirms that iRSs pose a real and non-trivial threat to the validity of cost-effectiveness analyses of diagnostics. It offers a clear framework (aligned vs unaligned models) for understanding when and why such bias arises, and it provides evidence on how much this bias can matter in practice. Additionally, it adds to the toolkit of solutions by evaluating correction methods within an economic analysis context, thereby contributing practical guidance on managing iRS bias in HTA.

6.2 Findings in the Context of Existing Literature

The challenges illuminated by this thesis resonate strongly with prior observations in the literature on diagnostic tests and HTA. Historically, methodologists have cautioned that all that glitters is not gold in diagnostic evaluation – in other words, that what we treat as the disease “truth” in studies can be a flawed construct, leading to downstream errors (Denwood et al., 2024). Denwood et al. (2024) recently introduced an interpretive framework for diagnostic test evaluation, emphasising how the case definition of disease

(often implicitly set by the reference standard test) can misalign with the actual target condition. They demonstrated with clinical examples that if a reference standard test defines disease imperfectly, "mismatches" occur between test results and true disease status, introducing bias into Se and Sp estimates. Our aligned vs. unaligned model concept builds directly on this insight. In effect, what we call an unaligned model in CEA is the downstream consequence of the kind of mismatch Denwood et al. (2024) describe in diagnostic accuracy studies. By extending this concept to the economic realm, we have shown that a mis-specified case definition (through an iRS) can propagate bias all the way to cost-per-QALY calculations. This theoretical extension answers a call in the literature for greater clarity on how to integrate diagnostic uncertainty into decision models (Soares et al., 2018, Snowsill, 2023). To position our work among existing frameworks: establishing the value of diagnostics in HTA has been an area of growing attention over the past decade. Soares et al. (2018) provided an analytical framework for considering the value of diagnostic and prognostic tests, aligning their evaluation with the principles used for therapeutic interventions. A key message from Soares et al. (2018) was that the value of a diagnostic test cannot be determined by accuracy alone – it must be inferred through the impact on patient outcomes and decisions, often requiring a linked evidence approach. Our findings reinforce and add nuance to this perspective. We show that not only must one link test accuracy to outcomes, but one must also be mindful of the quality of the accuracy data itself. If the accuracy is measured against an iRS, one could be systematically under- or overestimating the test's true clinical utility. In other words, our work underscores that diagnostic value assessment needs to account for uncertainty in what constitutes "truth". By quantifying how misclassification errors in a reference standard test affect outcomes, we contribute evidence to support recommendations (like those of Soares et al. (2018)) that analysts conduct rigorous sensitivity analyses around diagnostic parameters when direct evidence of patient impact is lacking. Another important context for our study is the literature on methods to evaluate diagnostic tests when a gold standard is not available. A variety of algebraic approaches have been proposed over the years, ranging from patient-count approaches to proportion-based approaches (Reitsma et al., 2009, Rutjes et al., 2007, Umemneku Chikere et al., 2019, Sherwin, 2022). These methods aim to infer the "true" test performance in the absence of a perfect standard or to correct the estimates obtained

with an imperfect standard. However, as reviews by researchers like Umemneku Chikere et al. (2019) have shown, each approach comes with assumptions and complexities. This thesis does not create new statistical methods for diagnostic accuracy. Instead, it takes existing algebraic corrections—such as those proposed by Staquet and Brenner—and integrates them directly into the economic evaluation framework. This integration addresses a notable gap in the literature. Whilst Sutton et al. (2008) emphasized the importance of linking synthesized test accuracy data with decision models, and Soares et al. (2018) pointed to a broader methodological disconnect between diagnostic test evaluation and HTA, our work operationalizes this bridge by explicitly adjusting for iRS-induced bias in the model's input parameters. We demonstrate how even simple algebraic corrections can substantially affect cost-effectiveness outcomes, offering a tractable method to improve the credibility of HTA for diagnostic technologies.

Our simulation results also contextualise previous case studies and examples. Several authors have documented instances where ignoring reference standard imperfections could mislead decision-makers. Xie et al. (2017), for example, evaluated a new pneumonia diagnostic and showed that using traditional meta-analytic methods (which assume a perfect reference standard like culture) underestimated the test's accuracy. By employing a Bayesian latent class meta-analysis (which does not assume a perfect gold standard), they obtained higher sensitivity estimates, which in turn led to a more favourable cost-effectiveness profile for the test. Their work provided proof that accounting for iRS bias could change an ICER sufficiently to alter conclusions. Our findings generalise this insight: we similarly found that when reference standard bias was present, a naive model (uncorrected, unaligned) could yield a less favourable impression of a test's value compared to a model that adjusts for the bias. The difference in outcomes was typically small in absolute terms – in our base-case simulations, the net health benefit of a test strategy might be underestimated by only a few hundredths of a QALY – but this gap could become critical in scenarios where decisions hinge on fine margins (such as a technology that is borderline cost-effective). Thus, our results are in line with case-specific analyses like Xie's, while offering a broader theoretical explanation (alignment) for why those differences arise. Another recently published study by Johnson et al. (2019) highlights the impact of misclassification on cost-effectiveness by replacing the conventional (imperfect) diagnostic pathway with a more

accurate test. Rather than statistically correcting for reference standard bias, they assume near-perfect accuracy from a new genetic screening approach (massively parallel sequencing). This effectively sidesteps the iRS problem altogether, demonstrating how removing misclassification can improve both economic and clinical outcomes. However, because they do not explicitly model or correct for a reference standard's imperfections, their method differs from the direct adjustments we examine in this thesis. Our work also intersects with literature on composite reference standards and verification biases, which are related issues arising in diagnostic studies. Dendukuri et al. (2018) argued that using composite reference standards (combining multiple imperfect tests) can introduce new forms of bias, often giving a false sense of security about having a "better" reference standard. They highlight that composite reference standards might inflate accuracy if not carefully implemented. In our context, composite reference standards were not explicitly modelled, but the aligned/unaligned framework still applies; one could consider a composite reference standard as another iRS, and the need to align the economic model to whichever standard is used remains critical. In fact, if an HTA uses composite reference standard data for one test and single-test reference data for another comparator, this would be a classic unalignment scenario likely to produce biased comparisons. Although we did not simulate this specific nuance, our framework provides a conceptual warning that mixing reference definitions in a single model (composite vs single) without adjustment would be problematic. This observation supports calls in the literature (Dendukuri et al., 2018) for careful reporting and handling of composite reference standards. It also complements methodological recommendations like those by Trikalinos and Balion (2012), who suggested that analysts report how uncertain diagnostic performance is, given the lack of a gold standard, possibly using measures like concordance or bounds rather than a single estimate. Our results echo the spirit of those recommendations by showing that explicitly accounting for bias in the reference standard—and exploring different assumptions for its misclassification— is essential for credible results. Importantly, our thesis responds to critiques about the lack of formal guidance in diagnostic economic modelling. Snowsill (2023) remarked that despite the increasing use of decision models for diagnostic technologies, there is still a paucity of structured guidance compared to therapeutic areas. The aligned vs unaligned model framework that we propose can be

seen as a step toward such guidance. It gives a clear criterion that analysts can check: are all my inputs and outcomes conditioned on the same definition of disease, or am I inadvertently mixing "truth" and "reference" conditions? This criterion is straightforward to apply and can flag potential bias early in model development. The framework also guides the choice of solution – if an unalignment is detected, one can either restructure the model to be aligned (e.g. express everything relative to the iRS, acknowledging its imperfections) or apply an adjustment to bridge from the iRS to truth. The fact that our simulation validates this framework by showing that aligned models yield unbiased incremental results (even when the reference standard is imperfect) lends weight to its usefulness. It essentially provides evidence to back a rule-of-thumb for practice: keeping your model aligned avoids internal inconsistencies. This is a timely contribution, as noted, because both Snowsill (2023) and earlier authors have called for improved frameworks to handle exactly such issues. We also connect with broader principles in decision-analytic modelling. The idea that biases in intermediate parameters can affect outcomes is not unique to diagnostics – it parallels concerns in other areas like using surrogate endpoints for therapies. In CEA of drugs, for instance, if one uses a surrogate outcome (e.g. tumour response) instead of the true outcome (survival), one must either validate the surrogate or adjust for its imperfection. Similarly, in diagnostics, test accuracy is a "surrogate" for patient outcome impact. Our work can be seen as part of the broader trend of ensuring model inputs truly reflect reality or are adjusted when they do not. Frameworks by Severens and van der Wilt (1999) and Laking et al. (2006) laid out how diagnostic models should incorporate the downstream consequences of false positives and false negatives. We add to that by demonstrating that the frequency of those false positives/negatives itself might be mis-estimated if based on flawed reference standard data. In other words, we enrich the classical understanding (Severens and van der Wilt, 1999) that "any error can propagate significantly, particularly when the consequences of misdiagnosis are considerable". We provide the additional insight that those errors can be systematic (biased) rather than just random and must be corrected to avoid systematically skewed results. Finally, our findings are consistent with and lend support to the arguments for more rigorous uncertainty analysis in HTA. There is increasing acknowledgement that uncertainty is multidimensional – not just statistical uncertainty in estimates, but structural uncertainty (e.g. what if our reference

standard is wrong?). Recent guidance (e.g. van der Pol et al. (2021)) and panels have encouraged analysts to be explicit about such uncertainties. By quantifying the impact of reference standard misclassification on cost-effectiveness, our study gives practical impetus to those recommendations. We can now say: if you suspect your reference standard is imperfect, here is how large the error in outcomes could be, and here are methods to adjust for it. This merges theoretical recommendations with empirical evidence, thereby enriching the literature's discourse on best practices for diagnostic HTA.

6.3 Implications for Health Policy and HTA Practice

The implications of this research should be interpreted proportionality considering both its findings and its limitations. The work demonstrates a mechanism by which iRS can bias linked-evidence economic models, and it shows that such bias can materially affect conclusions in some borderline cost-effectiveness scenarios. At the same time, the simulations indicate that the absolute magnitude of bias is often modest and frequently does not alter the preferred diagnostic strategy. On that basis, it would be premature to advocate quantitative corrections for RS misclassification as a routine procedural requirement in HTA.

Several contextual factors temper the scope for immediate policy change. First, with advancing methodological understanding over recent decades, the concept of a definitive "gold standard" is no longer seen as achievable in most diagnostic contexts (Reitsma et al., 2009, Rutjes et al., 2007), and HTA processes have increasingly adapted to working with imperfect and heterogeneous evidence. Second, the value of a diagnostic strategy may arise through pathway effects—such as earlier access to effective treatment, reduced use of invasive procedures, or improved treatment targeting—and does not always require an improvement in classical accuracy metrics (Ferrante di Ruffano et al., 2012b, Soares et al., 2018). Third, heterogeneity across diagnostic accuracy studies, combined with well-established sources of bias such as case-control study design and spectrum effects, may exert more influence on model results than RS misclassification in many practical applications (Rutjes et al., 2007). Empirical evidence consistently identifies case-control designs as generating the largest bias effects (Lijmer

et al., 1999), logically suggesting that study design selection should receive priority attention over post-hoc correction methods.

Moreover, comparative accuracy studies—which directly compare index tests against a common imperfect reference—increasingly represent an alternative approach that mitigates RS effects without requiring correction methods. Such studies are inherently consistent with the principle of alignment emphasised in the conceptual framework presented in Chapter 4. Where comparative evidence exists, selecting the least biased accuracy studies or prioritising head-to-head comparisons may be a more tractable route to reducing decision risk than applying formal algebraic corrections. This is particularly salient given that the availability of credible data to parameterise reference standard performance remains a major practical barrier. While point estimates from validation studies or expert elicitation could theoretically be explored, the reliability of expert judgement on test accuracy—a technical property not readily amenable to clinical intuition—is questionable.

A proportionate practical implication for agencies such as the NICE is therefore not "always correct" but "always appraise and test materiality". At minimum, diagnostic HTAs should explicitly document the RS underpinning accuracy estimates, consider the risk that the reference standard misclassifies disease status, and explore plausible impacts through structured sensitivity or scenario analyses. NICE already instructs Evidence Review Groups to consider RS quality and conduct sensitivity analyses when uncertainty exists (National Institute for Health and Care Excellence, 2022b). The present work reinforces the importance of those instructions by quantifying potential consequences, but does not mandate a step-change in process.

Targeted quantitative adjustment using methods such as those evaluated in Chapter 5 can be viewed as an optional analytical extension when several conditions are satisfied: (i) the appraisal is genuinely borderline relative to a decision threshold, (ii) credible information exists to parameterise RS performance with appropriate uncertainty propagation, (iii) the CI assumption between test and RS errors is tenable in the specific clinical context, and (iv) the expected decision consequences justify the additional analytical effort and specialist statistical expertise required. Importantly, the correction methods evaluated in this thesis perform poorly under CD (Chapter 5), limiting their applicability in scenarios where test and RS errors are correlated—a realistic situation in

many clinical contexts where tests rely on similar biological processes or measurement principles.

Even when these conditions are met, replication of the approach across real HTA case studies and piloting within actual appraisal workflows would be necessary before any standardised correction procedure could be recommended. Such piloting should examine not only technical feasibility but also the time burden on analysts, interpretability for decision committees, and robustness to alternative parameterisations of RS performance. The present thesis provides proof of concept for the approach but does not constitute the evidence base required to justify systemic changes to HTA processes.

It is also important to recognise that cost-effectiveness is not as critical a determinant of HTA decisions for diagnostic tests as it is for pharmaceuticals. Diagnostic appraisals often place substantial weight on clinical utility, pathway feasibility, acceptability to patients and clinicians, and the strength of evidence linking test results to downstream outcomes (Ferrante di Ruffano et al., 2012b, Bossuyt et al., 2003). This may reflect recognition of the greater intrinsic uncertainty involved in modelling the cost-effectiveness of tests compared to treatments, given the longer and more complex causal chain linking test performance to patient outcomes. Within this broader decision context, bias assessment in accuracy parameters should complement—not dominate—other evaluative considerations.

The framework's conceptual contribution—distinguishing aligned from unaligned models—may prove more immediately useful than the specific correction algorithms. Ensuring that all model parameters (test accuracy, disease prevalence, treatment outcomes) are conditioned on consistent disease definitions is a structural validity check analogous to ensuring consistent patient populations across data sources. This principle could be incorporated into existing quality assessment frameworks without requiring formal bias adjustment. For instance, HTA reviewers could systematically ask whether downstream outcomes in the model are linked to the same disease definitions as the test accuracy inputs, flagging misalignment as a potential source of internal inconsistency.

For international HTA bodies and global health contexts, the findings may have particular resonance in settings where iRS are ubiquitous—for example, in neglected tropical

diseases or resource-constrained environments where definitive diagnostic verification is infeasible (Dendukuri et al., 2018). In such contexts, transparent acknowledgement of RS limitations and structured exploration of their implications through scenario analysis may support more equitable evidence assessment. However, the same data and capacity constraints that make imperfect references common in these settings also limit the feasibility of formal correction methods, reinforcing the case for a proportionate rather than universal approach.

In summary, the implications for practice are measured. HTA agencies should enhance transparency around RS quality, systematically explore the materiality of potential misclassification through scenario analyses, and consider the principle of model alignment when validating decision-analytic structures. Quantitative correction methods represent a specialist tool that may be valuable in carefully selected borderline cases where the necessary conditions are satisfied, rather than a routine procedural step applicable to all diagnostic evaluations. The long-term impact of this work lies not in immediate procedural change, but in raising awareness of an underappreciated source of bias, providing a structured framework for thinking about its implications, and offering methods that future research can refine, validate, and selectively deploy where most needed.

6.4 Limitations of the Research

Like any modelling and simulation study, this research has limitations that must be acknowledged. These limitations not only point towards the need for further work, which we outline in the next section, but also explain why the proportionate approach to policy implications advocated in Section 6.3 is appropriate. Here we critically reflect on the main constraints and assumptions of our thesis:

(i) Simplified Decision Model Structure: We deliberately employed a simplified decision tree model focusing on a single index test versus a comparator (which could be “no test” or an alternative test) with a single iRS. This abstraction was useful for isolating the effect of iRS bias, but it does not capture the full complexity of real-world diagnostic pathways. In practice, diagnostic strategies can involve multiple tests in sequence or combination,

as well as downstream treatment decisions that form intricate pathways. In practice, diagnostic strategies can involve multiple tests in sequence or combination, as well as downstream treatment decisions that form intricate pathways. For example, a patient might undergo an initial screening test, followed by a confirmatory test if positive, and then treatment – each stage could have its own RS issues. Our analysis did not evaluate how bias might accumulate or interact in such multi-test algorithms. Similarly, we treated test outcomes as dichotomous (positive/negative); many modern diagnostics yield continuous or ordinal results (e.g. biomarker levels, imaging scores). The decision thresholds and subsequent outcomes for such tests were beyond our scope. Consequently, the generalisability of our quantitative findings (such as the percentage bias in net benefit) to more complex scenarios is uncertain. The aligned vs. unaligned concept should still hold in complex models, but the specific impacts and the ease of applying corrections could differ. Extending the framework to richer models is necessary before one can confidently apply it to, say, a screening program with multiple stages or a combination of tests (we suggest this as future research).

(ii) Assumption of Known RS Characteristics: Our use of algebraic correction methods (Staquet and Brenner approaches) assumes that we have or can obtain estimates of the iRS's own Se and Sp. In the simulation, we treated those as known inputs (drawing from distributions). In reality, the accuracy of the RS is often unknown or very uncertain – after all, if it were well-known, one might argue the reference standard isn't so "imperfect" or at least the problem is half solved. This presents a circular challenge: to correct bias, you need to know how biased the RS is. This limitation is a principal reason why routine application of quantitative correction methods in HTA would be premature, as noted in Section 6.3 In practice, such information might come from external studies (e.g. a validation of the RS against a third test or long-term follow-up). In our analysis, we did not fully integrate uncertainty in RS accuracy into the economic model's probabilistic sensitivity analysis; we mostly examined deterministic scenarios where we varied prevalence and assumed certain fixed levels of RS misclassification. This is a limitation because it downplays

the impact of uncertainty in the bias correction itself. An incorrect assumption about the RS's sensitivity could lead to over-correction – for instance, if we overestimate how imperfect the RS is, the Staquet method might overshoot and yield implausible adjusted values (such as a negative false rate or >100% Se, a phenomenon reported by Umemneku Chikere et al. (2021) in real-case applications). We did observe in some simulated instances that using a mis-specified correction (thinking the RS was worse than it is, or vice versa) led to nonsensical or illogical results (e.g. negative incremental QALYs or reversal of a clearly dominant strategy). This highlights that our correction methods require accurate inputs about the RS, which are not always available. The data requirements for reliable correction thus constitute a major practical barrier to implementation, supporting the recommendation in Section 6.3 that quantitative adjustment should be reserved for cases where credible parameterisation is feasible. Our study did not resolve how to obtain those inputs reliably – we assumed them. While we propose future work on eliciting or estimating these parameters (see below), the reliability of expert judgement on technical accuracy measures remains questionable.

(iii) Limited Exploration of Conditional Dependence: We acknowledged and partially explored the issue of CD between the index test and RS errors – that is, the situation where the likelihood of the index test being correct or wrong is correlated with the reference standard test being correct or wrong. We included a version of Brenner's method aiming to adjust for CD and ran simulations with some degree of correlation. However, our exploration was not exhaustive. We did not, for example, vary the correlation coefficient systematically across a full range or consider different structural forms of dependence (such as one test missing mild disease and the other missing severe disease, etc.). The dependence model in our simulation was relatively simple. Real diagnostic tests may have complex dependencies: for instance, two tests might both rely on the presence of a certain antibody, or one test might be more likely to be positive when the other is, beyond the disease status (e.g. spectrum effects or patient factors causing correlated results).

Critically, our findings showed that the simple corrections perform poorly when CD is introduced – they did not provide a robust solution for the CD case, aside from suggesting that latent class models could handle it. This limitation is a second principal reason for the proportionate policy stance in Section 6.3. For a test and RS that share a common flaw (say both miss the disease in the same subset of patients), our aligned model concept alone is not enough to eliminate bias – even an aligned model would consistently misclassify that subset across all strategies, potentially leading to a wrong conclusion about absolute benefit (though relative comparison might still cancel out some bias). Since CD is realistic in many clinical context where tests rely on similar biological processes or measurement principles, the limited applicability of correction methods under such conditions constrains their utility for routine HTA application. We also did not incorporate multiple iRS (differential verification) in the main analysis; that could be seen as a form of dependence or heterogeneity we left out.

(iv) Simulation Parameterisation and Scope: By necessity, our simulation used a finite set of scenarios and parameter ranges. We tried to be comprehensive by varying disease prevalence, test accuracy levels, and reference standard accuracy levels across wide ranges (often assuming uniform distributions to be agnostic). This approach ensured we were not tailoring to a specific disease, but it also means the results are somewhat diffuse – we did not calibrate scenarios to specific diseases like cancer or HIV or COVID-19 where the stakes and typical accuracies are known. Therefore, the relevance of our quantitative results to any given real-world case might be limited. For example, a decision-maker in breast cancer screening might ask: "Your simulation found biases of a few hundredth of an NHB; is that the case for mammography with imperfect biopsy RS?" Our generic approach cannot answer that directly without plugging in mammography-specific data.

The lack of disease-specific case studies where we apply our framework to actual published HTA models is a significant limitation. It would have strengthened the thesis to demonstrate the bias correction on an existing NICE diagnostic model post hoc, providing concrete evidence of real-world

impact This absence of empirical validation in actual HTA contexts is another reason why recommending immediate procedural changes would be premature-replication and piloting across real case studies are necessary prerequisites, as noted in Section 6.3. We consciously focused on theoretical and simulation evidence to keep the scope manageable, but this means the thesis stops short of a full real-world validation.

Additionally, our simulations largely assumed binary decision outcomes (test leads to treatment or not) and immediate payoff in QALYs/costs. We did not simulate long-term feedback loops (e.g. the consequences of a false negative might be mitigated if the patient returns later with symptoms). In reality, healthcare systems have some ability to catch errors (a false negative today might be diagnosed in a subsequent visit). Our model was a one-shot decision model. This could overstate the impact of a diagnostic accuracy error on outcomes in some conditions or understate it in others. In short, the external validity of our numerical findings is limited by these design choices.

(v) Focus on Bias in Accuracy, Not in Other Inputs: We singled out RS bias as the distortion of interest. However, diagnostic economic models can have many other uncertainties and biases – for example, spectrum bias (the test may perform differently in practice than in the study sample), treatment effect modifiers (if patients diagnosed earlier do better only if certain conditions hold), or even costing issues. Our discussion touched on how iRS bias might interact with other uncertainties, but our simulation did not incorporate, say, variability in treatment effectiveness or cost variations. This isolation of a single bias source, while methodologically appropriate for proof of concept, means we cannot assess the relative importance of RS misclassification compared to other sources of model uncertainty. As noted in Section 6.3, heterogeneity across diagnostic accuracy studies and well-established biased such as case-control study design may be more influential in many applications.

One could argue that in a full decision model, the effect of iRS bias might be overshadowed or confounded by these other factors. We treated the model in isolation to measure one effect at a time, which is a strength for internal

validity but a limitation for seeing the big picture. Decision-makers must consider all uncertainties jointly. We have provided insight into one piece of that puzzle, potentially improving the handling of diagnostic accuracy inputs, but the overall decision uncertainty will also include many other elements. We did not formally compute joint uncertainty metrics (like the probability of cost-effectiveness) with and without bias – doing so would require integrating our bias model into a probabilistic sensitivity analysis fully, which was outside our scope. This limits our claims somewhat: we can say "ignoring iRS can bias results in this direction," but we cannot always say how it compares to, for instance, sampling uncertainty in trial data or generalisation error.

(vi) Implementation Practicalities: While we explored correction methods and demonstrated their potential value, we did not deeply engage with the practical difficulty of using these methods in real HTA workflows. Many HTA analysts use readily available software (like spreadsheet models or standard decision tree software) and may not be equipped to implement custom bias-correction code. The feasibility from capacity standpoint-requiring specialist statistical expertise, additional analytical time, and potentially custom programming-is a practical constraint on implementation that reinforces the case for selective rather than routine application. As a limitation, we did not provide, for example, a user-friendly tool or algorithm that agencies could immediately pick up. Our work stops at demonstrating value and suggesting adoption, but there is a gap between an academic demonstration and broad implementation.

(vii) Data availability is another practical limitation: we assumed one can get estimates of RS accuracy from literature or experts. In some cases, literally no data exist on how good the RS is. We did not solve that data gap, just highlighted it. Thus, while we champion adjusting for iRS bias, we must admit that our proposed solutions might be challenging to execute without further support or research, making it a limitation in terms of immediate impact on practice.

(viii) Disease-Agnostic Approach – Strength and Weakness: By design, our thesis remained disease-agnostic. This is a strength in that it applies widely, but a

weakness in that it might not fully convince stakeholders in any one field. Specialists might find the discussion somewhat detached from practical clinical nuance. For example, in oncology diagnostics, there might be specific considerations (like how RS age over time with new biomarkers) that we did not address. Similarly, in infectious diseases, there are pathogen dynamics and prevalence fluctuations that were not modelled. Therefore, while we maintain the principles are general, the lack of disease-specific instantiation could limit the perceived applicability. This was a necessary trade-off to keep the thesis focused on methodology, but it is worth noting as a limitation – especially since some effects (like the relationship of prevalence to predictive value and thereby to cost-effectiveness) may behave differently in diseases with certain prevalence ranges. We did vary prevalence, but real diseases cluster at particular prevalence levels that might emphasize certain portions of our results more than others.

Despite these limitations, we took care to ensure that our conclusions are qualitatively robust. We conducted extensive sensitivity checks within the simulation (e.g. testing extreme values, verifying that aligned models produced expected unbiased results as a sanity check, confirming that when RS was made perfect our methods did nothing unusual, etc.). Importantly, the limitations underscore not only areas where caution is needed in interpretation and where further research is required, but also why the measured policy stance adopted in Section 6.3 is appropriate. The data requirements, methodological constraints under CD, absence of real-world validation, and practical implementation barriers collectively indicate that this work provides proof of concept and a framework for thinking about RS bias, rather than a ready-to-implement procedural change for HTA agencies. They also pave the way for the next steps in research – many of which we anticipated and discuss below. In the end, none of these limitations undermine the fundamental conceptual message of the thesis: that iRS can bias CEA and that this bias can be understood and partially corrected. However, they do qualify how far one can take our quantitative estimates and provide a clear mandate that further work-including replication, piloting, and disease-specific validation- is required before these findings could be translated into operational changes in HTA practice.

6.5 Recommendations for Future Research

Building on the insights and limitations discussed, there are several promising avenues for future research. These range from methodological advancements to applied studies and even policy-oriented investigations. We propose the following directions as particularly worthwhile:

Extension to Complex Diagnostic Pathways: Future research should extend the aligned vs. unaligned framework to more complex diagnostic strategies, including multiple-test combinations and algorithms. This could involve simulation or analytical modelling of scenarios where, for example, an initial screening test is followed by a confirmatory test, or where two different tests are used in tandem (parallel testing). How does iRS bias play out in a sequence of tests? If the initial test has an iRS and the confirmatory test has its own RS, what is the net effect on cost-effectiveness outcomes? There may be unique "alignment" challenges in such pathways – for instance, ensuring that both tests' performance metrics and downstream outcomes are measured on the same disease definitions. Additionally, composite reference standards (using multiple sources to determine disease status) should be examined. Research could simulate cases of AND/OR rule composite reference standards and see how to integrate those into economic models without bias. The goal would be to test whether our framework applies to a wider array of HTA problems, reflecting real-world diagnostic processes. This might also involve developing decision models beyond simple trees, such as state-transition models for diseases where diagnostics influence long-term trajectories (e.g. Markov models for chronic illness management with ongoing testing).

Advanced Methods for Dependent Errors: Given the limitations we faced with CD, an important line of inquiry is the development and testing of bias-adjustment methods that can handle dependent misclassification errors. One promising avenue is to integrate Bayesian LCMs directly with economic modelling. For example, one could use a Bayesian approach where the true disease status of each simulated patient is a latent variable informed by both the index test and reference standard test results, along with a specified covariance structure (Dendukuri and Joseph, 2001). This latent true status would then feed into the cost and outcome calculation. Such a model would inherently account for any correlation between test and reference standard errors via the covariance parameter. Modern Bayesian computational tools (e.g. JAGS or Stan) could

be used to perform joint estimation of diagnostic accuracy and cost-effectiveness outcomes. This would be computationally intensive, but feasible – especially with increasing computing power. A study could systematically vary the correlation between test and reference standard errors and use a Bayesian LCM approach to assess whether it recovers unbiased estimates of incremental cost-effectiveness and under what conditions such approaches outperform simpler methods. This would inform whether more complex (and computationally heavy) approaches are worth the effort compared to simpler adjustments. It would also allow incorporation of uncertainty in reference standard accuracy in a probabilistic way, rather than our deterministic adjustment. The result could be a framework where instead of plugging in a single corrected sensitivity, one propagates a distribution of possible true sensitivities into the model, yielding a posterior distribution for the ICER that fully reflects RSndard uncertainty. Additionally, exploring other statistical approaches like simulation-extrapolation (SIMEX) for measurement error correction (Sevilimedu and Yu, 2022), or employing instrumental variables within decision models (if some external instrument affects true disease but not the reference standard misclassification), could be innovative ways to handle dependency. The overarching aim is to evaluate whether bias correction can be made more robust in situations where the simple independence assumption doesn't hold, and to understand the trade-offs between methodological complexity and practical utility.

Elicitation and Estimation of Reference Standard Accuracy: As noted, one bottleneck in applying bias corrections is knowing the imperfection of the reference standard. Future research could address this through methods to derive or elicit the accuracy of iRSs. One approach is expert elicitation: engaging clinicians or diagnosticians in a formal elicitation exercise to quantify their beliefs about the reference standard's Se and Sp. For example, if pathologists commonly miss a certain subtype of disease on biopsy, experts might estimate how often that occurs. However, the reliability of expert judgement on technical accuracy properties-which are not readily amenable to clinical intuition- remains an open question requiring empirical investigation. These elicited distributions could then be used in bias-adjusted analyses if shown to be sufficiently reliable. Research is needed to refine elicitation protocols specific to diagnostic accuracy (perhaps building on techniques used for eliciting treatment effects or epidemiological parameters) and to validate such approaches against empirical benchmarks where available. Another

approach is meta-analytic evidence: often multiple studies use different RS; comparing their reported accuracy could give clues about RS quality. One could perform a meta-regression on reported sensitivities of a test across studies with different reference standards to back-calculate an implied reference standard Se. Additionally, designing validation studies is key. Future applied research could pair up with ongoing diagnostic trials to incorporate a subset of patients who receive a more definitive work-up (the closest to a gold standard) and use that to benchmark the RS. Methods like capture-recapture (as in Ferrante di Ruffano et al. (2012b)) could be further explored for estimating the completeness of case detection by a reference standard test. Overall, a research program that provides practical guidance or even software to help analysts obtain estimates of RS accuracy (be it through literature or expert judgment) would be necessary precondition for broader consideration of the methods explored in this thesis. Development of HTA Tools and Case Studies: To bridge the gap between theory and practice, future work should focus on demonstrating these bias-adjustment methods in real HTA case studies and developing user-friendly tools or templates. One idea is to take published NICE diagnostic models (many of which are publicly available in academic papers or NICE documentation) and re-analyse them introducing iRS bias adjustments. For instance, a case study could look at a NICE evaluation of a diagnostic for which we suspect the reference standard was imperfect (like the strep throat rapid test evaluation, or a cancer diagnostic using an imperfect biopsy). By applying our framework – perhaps first running the original model (unaligned) and then an aligned or corrected version – we can assess the practical difference it makes and whether correction materially affects conclusions. This would serve as empirical validation necessary before any procedural recommendation could be made. If it turns out that in past evaluations the conclusions would have changed with bias correction, that would provide evidence to support considering more rigorous assessment of RS quality. Conversely, if it shows minimal impact, that helps delineate when such complexity can be safely skipped. Alongside case studies, creating guidelines or checklists for analysts could be valuable if empirical work demonstrates sufficient impact to justify the additional burden. We envision something like: Step 1: Identify if your RS is imperfect; Step 2: if yes, decide whether to use scenario analysis, algebraic correction, or Bayesian modelling based on data availability; Step 3: ensure alignment in model structure; Step 4: present results with and

without correction. Formalising this process in an HTA methodological guide or a tutorial paper could be considered once the approach has been piloted and validated. Additionally, software implementation would be useful – for example, an R package or an add-on to existing decision modelling software that performs Staquet's adjustment automatically given user inputs. Some researchers (Xie et al., 2022c) have started providing scripts for latent class meta-analysis. Expanding such toolkits to cover on-the-fly bias correction within cost-effectiveness models would help practitioners if and when the evidence base supports selective application. Investment in capacity building – workshops or training for HTA analysts – could accompany this, ensuring that the next generation of health economists is aware of iRS and the considerations involved in assessing their impact.

Value of Information and Research Prioritisation: We touched on the idea of computing the Expected Value of Perfect Information (EVPI) for the reference standard's accuracy – essentially, the value of eliminating uncertainty caused by iRS bias. Future research can formalise this concept. Specifically, studies could integrate bias modelling into a VOI analysis framework to determine when it is most worthwhile to collect better RS data. This involves calculating the change in net benefit uncertainty when moving from current imperfect knowledge to hypothetical perfect knowledge of true disease status. By doing so across different scenarios, one could identify thresholds (e.g. if reference standard Se is below X or disease prevalence is Y, EVPI is high). Such analysis would inform policy: for which diagnostics should scarce research funds be allocated to conduct, say, a definitive study with long-term follow-up or a new more accurate reference standard test? Moreover, the concept of "value of a better reference standard" or the "value of a perfect test" could be extended to design optimal studies. If we know the EVPI is high, what kind of study would provide that information? Perhaps a large cohort with dual testing, or development of a novel RS (like an autopsy study in infectious disease, or a composite including genomic data). Methodologically, this means incorporating our bias parameters into the probabilistic model and performing standard VOI computations (which is computationally feasible if one can run a simulation model repeatedly, as we did). One interesting extension is to consider partial EVPI for different parameters – e.g. value of eliminating uncertainty in RS sensitivity vs RS specificity vs sample size etc. This could highlight which aspect of an iRS is most problematic. Our thesis laid the foundation

by quantifying how bias in RS affects outcomes; the next step is to attach an explicit value to reducing that bias.

Addressing Bias in Systematic Reviews feeding HTA: On a related note, there is scope for collaboration between the systematic review community (particularly those doing meta-analyses of diagnostic accuracy) and the economic modelling community. Trikalinos and Balion (2012), as mentioned, suggested that the results of diagnostic accuracy reviews should indicate the uncertainty due to lack of gold standard. Future research could pilot ways to express this in HTA submissions. For example, a Cochrane diagnostic test accuracy review could present not just a summary Se but also an adjustment for potential reference standard bias (perhaps by incorporating studies with an alternative RS). How can that information be ported into cost-effectiveness models seamlessly? Research efforts might include developing reporting standards – akin to QUADAS – that specifically flag RS issues and provide quantitative bias estimates. There is already movement in updating QUADAS to consider these issues. Our alignment concept could inform such reporting: e.g. labelling studies or analyses as aligned or not. A future project could work on a framework where any new diagnostic entering HTA first goes through a bias assessment pipeline (covering RS and other biases), and that output is used directly in model parameterisation. This might require interdisciplinary teams and consensus building and would need piloting before any systematic implementation could be recommended.

Exploring Patient and Societal Perspectives: Another extension is to investigate how the existence of reference standard uncertainty affects patients and society's acceptance of diagnostic technologies. While more of a social science angle, one could study if stakeholders (patients, clinicians) are aware of these uncertainties and how they weigh them. For example, if clinicians know a RS is imperfect, do they trust a new test's results, or do they behave differently (perhaps ordering additional tests)? This could affect real-world cost-effectiveness if physicians compensate for perceived test unreliability. Incorporating such behaviour into future models (maybe through adaptive decision-making or second opinion testing) would provide a more holistic view. It also intersects with ethics and communication – should patients be informed that a diagnosis might be wrong because the standard isn't perfect, and would that change their preferences or quality of life (which feeds into QALYs)? These questions are speculative but worth

exploring as diagnostics become more patient-centric (e.g. patient-chosen genomic testing).

Integration with Equity and Subgroup Analysis: We mentioned that iRS bias could disproportionately affect certain subgroups. Future research could simulate or analyse distributional consequences of RS bias. For instance, if a RS is less accurate in women than men (perhaps due to physiological differences), then an unadjusted analysis might undervalue a diagnostic for women relative to men. Correcting bias in such a case is not just a general accuracy issue but an equity concern. Studying specific examples (say, cardiovascular diagnostics that historically had lower accuracy in women because reference definitions were male-centric) and seeing how bias-adjusted modelling might alter conclusions for subpopulations would be an important contribution. This dovetails with current interests in HTA on health equality and fairness. Our framework could be expanded to explore whether alignment considerations differ within subgroups – or to identify subgroup misalignment.

In summary, there is a rich research agenda to further develop and apply the ideas from this thesis. We particularly emphasise the need for practical translational work – taking these methods into the wild of actual HTAs and refining them based on real-world complexity. At the same time, methodological innovation (especially using Bayesian and evidence synthesis techniques) will strengthen the theoretical underpinnings and allow tackling cases (like dependent errors) that we simplified. By pursuing these future directions, the field can move closer to a state where iRS bias is routinely accounted for in economic evaluations, much as sampling uncertainty is today.

6.6 Conclusions

In conclusion, this thesis has examined the implications of iRSs on the economic evaluation of diagnostic tests and investigated methods to address the resulting bias. We have confirmed that when the “gold standard” is tarnished, the ripples extend to cost-effectiveness conclusions – diagnostic accuracy is not just a technical detail of test evaluation, but a determinant of health-economic outcomes. Encouragingly, our work shows that we are not powerless against this bias: by understanding the concept of alignment and exploring correction techniques, we can contribute to improving the

validity of diagnostic cost-effectiveness models. The objectives laid out at the start of this research have been met. First, we synthesised the literature on how to handle situations with no true gold standard, spanning both statistical and decision-analytic domains. We identified strengths and limitations of approaches such as verification bias adjustments, latent class modelling, and simple analytic corrections. This review highlighted a crucial gap – while many methods existed to estimate or correct test accuracy, their application in cost-effectiveness studies was rare and not well understood. Second, we introduced a structured framework (aligned vs. unaligned models) to clarify how misclassification errors enter decision models, addressing the theoretical link between diagnostic and economic realms. Third, through simulation, we demonstrated how these misclassifications can distort outcomes like the iNHB and how applying bias corrections influences those outcomes. Importantly, we found that in many realistic scenarios the quantitative impact of iRS bias on cost-effectiveness metrics was modest, which is reassuring – routine decisions might not be wildly off-track. However, we also identified clear cases where ignoring iRS bias would lead to incorrect decisions (e.g. missing the chance to adopt a truly cost-effective test). This answers the practical "so what" – it matters in specific, non-negligible situations. Finally, we translated these findings into recommendations for HTA practice and future work, thereby fulfilling the broader aim of the thesis: not only to diagnose the problem, but to suggest cures and preventive measures for the HTA community. From a broader perspective, our contributions lie in improving the rigour and credibility of diagnostic HTA. As healthcare continues to advance, we see an explosion of new diagnostics, whether they be genomics, imaging AI tools, biomarkers, or point-of-care tests. Paradoxically, the better and more innovative these tests become, often the harder it is to evaluate them because we lack a clear gold standard – consider genetic tests for complex disorders or diagnostic algorithms based on machine learning. The absence of perfect reference standards will likely become more common, not less, in the future of medicine (Denwood et al., 2024, Snowsill, 2023). Therefore, the issues tackled in this thesis are only going to grow in importance. By proactively investigating methods to correct reference standard bias, we add to the foundation on which future guidelines and best practices can be built. In particular, we hope that the concept of ensuring alignment in models and checking for reference standard bias will be internalised by analysts. It

prompts a critical question that every evaluator of a diagnostic should ask: “How sure are we about the true accuracy of this test, and what does that mean for our decision?” Too often in the past, that question was glossed over by assuming the reference standard was truth. Ultimately, the goal of HTA is to make the best possible use of evidence for the benefit of patients and society. This thesis contributes to that goal by shedding light on a subtle but meaningful source of bias that can creep into our evidence base. Recognising and adjusting for iRS bias is one more step toward ensuring that our evaluations of diagnostic technologies truly reflect reality. It helps guard against the twin risks of unwarranted enthusiasm for a test that is not as good as it seems and missed opportunities where a beneficial test is undervalued due to misleading evidence. Through further development and validation of the approaches explored in this work – potentially including selective bias analyses in borderline cost-effectiveness cases and scenario analyses to bound uncertainty – decision-makers may eventually have greater confidence that when they recommend a diagnostic test for use in the healthcare system, they do so on solid ground. In closing, we return to the motivation behind this research: the recognition that evidence is imperfect, yet decisions must be made. The work presented here does not make evidence perfect – but it does provide a clearer lens through which to view imperfect evidence. Using that clearer lens, HTA practitioners can better discern the true value of diagnostics and make choices that improve health outcomes and use resources wisely. It is our hope that this general discussion, and the thesis as a whole, will inform future assessments and inspire ongoing efforts to refine the evaluation of diagnostic tests in the absence of a gold standard.

7 References

ALBERT, P. S. 2009. Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. *Statistics in Medicine,* 28**,** 780-797.

ALBERT, P. S. & DODD, L. E. 2004. A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error without a Gold Standard. *Biometrics,* 60**,** 427-435.

ALONZO, T. A. 2014. Verification Bias: Impact and Methods for Correction when Assessing Accuracy of Diagnostic Tests. *REVSTAT-Statistical Journal,* 12**,** 67–83-67–83.

ARROW, K. J. 1965. Uncertainty and the Welfare Economics of Medical Care: Reply (The Implications of Transaction Costs and Adjustment Lags). *The American Economic Review,* 55**,** 154-158.

BAUGHMAN, A. L., BISGARD, K. M., CORTESE, M. M., THOMPSON, W. W., SANDEN, G. N. & STREBEL, P. M. 2008. Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. *Clin Vaccine Immunol,* 15**,** 106-14.

BEGG, C. B. 1987. Biases in the assessment of diagnostic tests. *Statistics in Medicine,* 6**,** 411-423.

BEGG, C. B. & GREENES, R. A. 1983. Assessment of Diagnostic Tests When Disease Verification is Subject to Selection Bias. *Biometrics,* 39**,** 207-215.

BERTENS, L. C. M., BROEKHUIZEN, B. D. L., NAAKTGEBOREN, C. A., RUTTEN, F. H., HOES, A. W., VAN MOURIK, Y., MOONS, K. G. M. & REITSMA, J. B. 2013. Use of Expert Panels to Define the Reference Standard in Diagnostic Research: A Systematic Review of Published Methods and Reporting. *PLOS Medicine,* 10**,** e1001531.

BERTRAM, M. Y., LAUER, J. A., DE JONCHEERE, K., EDEJER, T., HUTUBESSY, R., KIENY, M.-P. & HILL, S. R. 2016. Cost–effectiveness thresholds: pros and cons. *Bulletin of the World Health Organization,* 94**,** 925-930.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L., LIJMER, J. G., MOHER, D., RENNIE, D., DE VET, H. C., KRESSEL, H. Y., RIFAI, N., GOLUB, R. M., ALTMAN, D. G., HOOFT, L., KOREVAAR, D. A. & COHEN, J. F. 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Bmj,* 351**,** h5527.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L. M., LIJMER, J. G., MOHER, D., RENNIE, D. & DE VET, H. C. 2003. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Bmj,* 326**,** 41-4.

BRENNER, H. 1996. Correcting for exposure misclassification using an alloyed gold standard. *Epidemiology,* 7**,** 406-10.

BRIGGS, A. 2006. *Decision modelling for health economic evaluation*, Oxford University Press.

BRIGGS, A. H. & O'BRIEN, B. J. 2001. The death of cost-minimization analysis? *Health Econ,* 10**,** 179-84.

BRYANT, J., COOPER, K., PICOT, J., CLEGG, A., RODERICK, P., ROSENBERG, W. & PATCH, C. 2009. Diagnostic strategies using DNA testing for hereditary haemochromatosis in at-risk populations: a systematic review and economic evaluation. *Health Technol Assess,* 13**,** iii, ix-xi, 1-126.

CADTH 2019. Guidelines for the Economic Evaluation of Health Technologies: Canada

APPENDIX — SPECIFIC GUIDANCE FOR TREATMENTS
WITH COMPANION DIAGNOSTICS. 4th Edition ed.
CLAXTON, K., MARTIN, S., SOARES, M., RICE, N., SPACKMAN, E., HINDE, S., DEVLIN, N., SMITH, P. C. & SCULPHER, M. 2015. Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. *Health Technol Assess,* 19**,** 1-503, v-vi.
CLAXTON, K., SCULPHER, M. & DRUMMOND, M. 2002. A rational framework for decision making by the National Institute For Clinical Excellence (NICE). *Lancet,* 360**,** 711-5.
CLAXTON, K. P. & SCULPHER, M. J. 2006. Using Value of Information Analysis to Prioritise Health Research. *PharmacoEconomics,* 24**,** 1055-1068.
COLLINS, J. & HUYNH, M. 2014. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Stat Med,* 33**,** 4141-69.
CORMAN, V. M., LANDT, O., KAISER, M., MOLENKAMP, R., MEIJER, A., CHU, D. K., BLEICKER, T., BRüNINK, S., SCHNEIDER, J., SCHMIDT, M. L., MULDERS, D. G., HAAGMANS, B. L., VAN DER VEER, B., VAN DEN BRINK, S., WIJSMAN, L., GODERSKI, G., ROMETTE, J. L., ELLIS, J., ZAMBON, M., PEIRIS, M., GOOSSENS, H., REUSKEN, C., KOOPMANS, M. P. & DROSTEN, C. 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill,* 25.
CRAWFORD, E. D., HEIDENREICH, A., LAWRENTSCHUK, N., TOMBAL, B., POMPEO, A. C. L., MENDOZA-VALDES, A., MILLER, K., DEBRUYNE, F. M. J. & KLOTZ, L. 2019. Androgen-targeted therapy in men with prostate cancer: evolving practice and future considerations. *Prostate Cancer Prostatic Dis,* 22**,** 24-38.
CULYER, A. J. 1989. The Normative Economics of Health Care Finance and Provision. *Oxford Review of Economic Policy,* 5**,** 34-58.
CULYER, A. J. 2014. *The Dictionary of Health Economics*, Edward Elgar Publishing.
DE GROOT, J. A., DENDUKURI, N., JANSSEN, K. J., REITSMA, J. B., BOSSUYT, P. M. & MOONS, K. G. 2011. Adjusting for differential-verification bias in diagnostic-accuracy studies: a Bayesian approach. *Epidemiology,* 22**,** 234-41.
DEEKS, J. J. 2001. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *Bmj,* 323**,** 157-62.
DENDUKURI, N. & JOSEPH, L. 2001. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics,* 57**,** 158-67.
DENDUKURI, N., SCHILLER, I., DE GROOT, J., LIBMAN, M., MOONS, K., REITSMA, J. & VAN SMEDEN, M. 2018. Concerns about composite reference standards in diagnostic research. *BMJ: British Medical Journal,* 360.
DENWOOD, M., NIELSEN, S. S., OLSEN, A., JONES, H. E., COFFENG, L. E., LANDFRIED, G., NIELSEN, M. K., LEVECKE, B., THAMSBORG, S. M., EUSEBI, P., MELETIS, E., KOSTOULAS, P., HARTNACK, S., ERKOSAR, B. & TOFT, N. 2024. All that glitters is not gold: An interpretive framework for diagnostic test evaluation using Ascaris lumbricoides as a conceptual example. *PLOS Neglected Tropical Diseases,* 18**,** e0012481.
DOLAN, P. & OLSEN, J. A. 2002. *Distributing Health Care: Economic and Ethical Issues*, Oxford University Press.
DRUMMOND, M. F. M., SCULPHER, M. J., TORRANCE, G. W., O'BRIEN, B. J. & STODDART, G. L. 2015. *Methods for the Economic Evaluation of Health Care Programmes*, Oxford University Press.
EASTON, D. F., PHAROAH, P. D., ANTONIOU, A. C., TISCHKOWITZ, M., TAVTIGIAN, S. V., NATHANSON, K. L., DEVILEE, P., MEINDL, A., COUCH, F. J., SOUTHEY,

M., GOLDGAR, D. E., EVANS, D. G., CHENEVIX-TRENCH, G., RAHMAN, N., ROBSON, M., DOMCHEK, S. M. & FOULKES, W. D. 2015. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med,* 372**,** 2243-57.

EMERSON, S. C., WAIKAR, S. S., FUENTES, C., BONVENTRE, J. V. & BETENSKY, R. A. 2018. Biomarker validation with an imperfect reference: Issues and bounds. *Statistical Methods in Medical Research,* 27**,** 2933-2945.

ERAKY, A. M., BECK, R. T., TREFFY, R. W., AARONSON, D. M. & HEDAYAT, H. 2023. Role of Advanced MR Imaging in Diagnosis of Neurological Malignancies: Current Status and Future Perspective. *J Integr Neurosci,* 22**,** 73.

FERRANTE DI RUFFANO, L., DAVENPORT, C., EISINGA, A., HYDE, C. & DEEKS, J. J. 2012a. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *Journal of Clinical Epidemiology,* 65**,** 282-287.

FERRANTE DI RUFFANO, L., HYDE, C. J., MCCAFFERY, K. J., BOSSUYT, P. M. M. & DEEKS, J. J. 2012b. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ,* 344**,** e686.

FLAHAULT, A., CADILHAC, M. & THOMAS, G. 2005. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol,* 58**,** 859-62.

GARDNER, I. A., STRYHN, H., LIND, P. & COLLINS, M. T. 2000. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Preventive Veterinary Medicine,* 45**,** 107-122.

GART, J. J. & BUCK, A. A. 1966. Comparison of a screening test and a reference test in epidemiologic studies: A probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology,* 83**,** 593-602.

GOLD, M. R., SIEGEL, J. E., RUSSELL, L. B. & WEINSTEIN, M. C. 1996. *Cost-Effectiveness in Health and Medicine*.

GRAY, R., BEGG, C. B. & GREENES, R. A. 1984. Construction of Receiver Operating Characteristic Curves when Disease Verification Is Subject to Selection Bias. *Medical Decision Making,* 4**,** 151-164.

GüNTHARD, H. F., SAAG, M. S., BENSON, C. A., DEL RIO, C., ERON, J. J., GALLANT, J. E., HOY, J. F., MUGAVERO, M. J., SAX, P. E., THOMPSON, M. A., GANDHI, R. T., LANDOVITZ, R. J., SMITH, D. M., JACOBSEN, D. M. & VOLBERDING, P. A. 2016. Antiretroviral Drugs for Treatment and Prevention of HIV Infection in Adults: 2016 Recommendations of the International Antiviral Society-USA Panel. *Jama,* 316**,** 191-210.

GWET, K. 2012. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*.

HADGU, A. 1996. The discrepancy in discrepant analysis. *The Lancet,* 348**,** 592-593.

HOT, A., BOSSUYT, P. M., GERKE, O., WAHL, S., VACH, W. & ZAPF, A. 2021. Randomized test-treatment studies with an outlook on adaptive designs. *BMC Medical Research Methodology,* 21**,** 110.

HUI, S. L. & WALTER, S. D. 1980. Estimating the Error Rates of Diagnostic Tests. *Biometrics,* 36**,** 167-171.

HUI, S. L. & ZHOU, X. H. 1998. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res,* 7**,** 354-70.

JANES, H., PEPE, M. S., BOSSUYT, P. M. & BARLOW, W. E. 2011. Measuring the performance of markers for guiding treatment decisions. *Ann Intern Med,* 154**,** 253-9.

JOHNSON, S. R., CARTER, H. E., LEO, P., HOLLINGWORTH, S. A., DAVIS, E. A., JONES, T. W., CONWELL, L. S., HARRIS, M., BROWN, M. A., GRAVES, N. & DUNCAN, E. L. 2019. Cost-effectiveness Analysis of Routine Screening Using

Massively Parallel Sequencing for Maturity-Onset Diabetes of the Young in a Pediatric Diabetes Cohort: Reduced Health System Costs and Improved Patient Quality of Life. *Diabetes Care,* 42**,** 69-76.
KOHN, M. A., CARPENTER, C. R. & NEWMAN, T. B. 2013. Understanding the direction of bias in studies of diagnostic test accuracy. *Acad Emerg Med,* 20**,** 1194-206.
LAKING, G., LORD, J. & FISCHER, A. 2006. The economics of diagnosis. *Health Economics,* 15**,** 1109-1120.
LIJMER, J. G., MOL, B. W., HEISTERKAMP, S., BONSEL, G. J., PRINS, M. H., VAN DER MEULEN, J. H. & BOSSUYT, P. M. 1999. Empirical evidence of design-related bias in studies of diagnostic tests. *Jama,* 282**,** 1061-6.
LIN, J. S., PERDUE, L. A., HENRIKSON, N. B., BEAN, S. I. & BLASI, P. R. 2021. Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *Jama,* 325**,** 1978-1998.
LU, D., ZHOU, C., TANG, L., TAN, M., YUAN, A. & CHAN, L. 2018. Evaluating accuracy of diagnostic tests without conditional independence assumption. *Statistics in Medicine,* 37**,** 2809-2821.
MCKAY, M. D., BECKMAN, R. J. & CONOVER, W. J. 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics,* 21**,** 239-245.
MCNAMARA, S., SCHNEIDER, P. P., LOVE-KOH, J., DORAN, T. & GUTACKER, N. 2023. Quality-Adjusted Life Expectancy Norms for the English Population. *Value in Health,* 26**,** 163-169.
MENTEN, J., BOELAERT, M. & LESAFFRE, E. 2008. Bayesian latent class models with conditionally dependent diagnostic tests: a case study. *Stat Med,* 27**,** 4469-88.
MORRIS, T. P., WHITE, I. R. & CROWTHER, M. J. 2019. Using simulation studies to evaluate statistical methods. *Statistics in Medicine,* 38**,** 2074-2102.
NAAKTGEBOREN, C. A., OCHODO, E. A., VAN ENST, W. A., DE GROOT, J. A. H., HOOFT, L., LEEFLANG, M. M. G., BOSSUYT, P. M., MOONS, K. G. M. & REITSMA, J. B. 2016. Assessing variability in results in systematic reviews of diagnostic studies. *BMC Medical Research Methodology,* 16**,** 6.
NATHAN, D. M., CLEARY, P. A., BACKLUND, J. Y., GENUTH, S. M., LACHIN, J. M., ORCHARD, T. J., RASKIN, P. & ZINMAN, B. 2005. Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. *N Engl J Med,* 353**,** 2643-53.
NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2022a. *Diagnostic Assessment Programme* [Online]. Available: https://www.nice.org.uk/About/What-we-do/Our-Programmes/NICE-guidance/NICE-diagnostics-guidance [Accessed].
NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2022b. *NICE health technology evaluations: the manual* [Online]. 31 January 2022. Available: https://www.nice.org.uk/process/pmg36/chapter/introduction-to-health-technology-evaluation [Accessed June 2022].
NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE) 2019. Rapid tests for group A streptococcal infections in people with a sore throat (DG38).
NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE) 2020. Tisagenlecleucel for treating relapsed or refractory diffuse large B-cell lymphoma after 2 or more systemic therapies. Technology appraisal guidance [TA567].
NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE) 2022. PredictSURE IBD and IBDX to guide treatment of Crohn's disease (DG45).
NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE). 2024. *Diagnostics Assessment Programme* [Online]. Available:

https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-diagnostics-guidance#get-a-technology-assessed [Accessed April 2024].

NELSON, H. D., CANTOR, A., HUMPHREY, L., FU, R., PAPPAS, M., DAEGES, M. & GRIFFIN, J. 2016. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. *Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation.* Rockville (MD): Agency for Healthcare Research and Quality (US).

NERENBERG, K. A., ZARNKE, K. B., LEUNG, A. A., DASGUPTA, K., BUTALIA, S., MCBRIEN, K., HARRIS, K. C., NAKHLA, M., CLOUTIER, L., GELFER, M., LAMARRE-CLICHE, M., MILOT, A., BOLLI, P., TREMBLAY, G., MCLEAN, D., PADWAL, R. S., TRAN, K. C., GROVER, S., RABKIN, S. W., MOE, G. W., HOWLETT, J. G., LINDSAY, P., HILL, M. D., SHARMA, M., FIELD, T., WEIN, T. H., SHOAMANESH, A., DRESSER, G. K., HAMET, P., HERMAN, R. J., BURGESS, E., GRYN, S. E., GRéGOIRE, J. C., LEWANCZUK, R., POIRIER, L., CAMPBELL, T. S., FELDMAN, R. D., LAVOIE, K. L., TSUYUKI, R. T., HONOS, G., PREBTANI, A. P. H., KLINE, G., SCHIFFRIN, E. L., DON-WAUCHOPE, A., TOBE, S. W., GILBERT, R. E., LEITER, L. A., JONES, C., WOO, V., HEGELE, R. A., SELBY, P., PIPE, A., MCFARLANE, P. A., OH, P., GUPTA, M., BACON, S. L., KACZOROWSKI, J., TRUDEAU, L., CAMPBELL, N. R. C., HIREMATH, S., ROERECKE, M., ARCAND, J., RUZICKA, M., PRASAD, G. V. R., VALLéE, M., EDWARDS, C., SIVAPALAN, P., PENNER, S. B., FOURNIER, A., BENOIT, G., FEBER, J., DIONNE, J., MAGEE, L. A., LOGAN, A. G., CôTé, A. M., REY, E., FIROZ, T., KUYPER, L. M., GABOR, J. Y., TOWNSEND, R. R., RABI, D. M. & DASKALOPOULOU, S. S. 2018. Hypertension Canada's 2018 Guidelines for Diagnosis, Risk Assessment, Prevention, and Treatment of Hypertension in Adults and Children. *Can J Cardiol,* 34**,** 506-525.

NOVIELLI, N., COOPER, N. J., ABRAMS, K. R. & SUTTON, A. J. 2010. How is evidence on test performance synthesized for economic decision models of diagnostic tests? A systematic appraisal of Health Technology Assessments in the UK since 1997. *Value Health,* 13**,** 952-7.

NOVIELLI, N., COOPER, N. J. & SUTTON, A. J. 2013a. Evaluating the cost-effectiveness of diagnostic tests in combination: is it important to allow for performance dependency? *Value Health,* 16**,** 536-41.

NOVIELLI, N., SUTTON, A. J. & COOPER, N. J. 2013b. Meta-analysis of the accuracy of two diagnostic tests used in combination: application to the ddimer test and the wells score for the diagnosis of deep vein thrombosis. *Value Health,* 16**,** 619-28.

OKKAOGLU, Y., WELTON, N. J. & JONES, H. E. 2024. Detecting departures from the conditional independence assumption in diagnostic latent class models: a simulation study. *BMC Medical Research Methodology,* 24**,** 299.

PEPE, M. S., ETZIONI, R., FENG, Z., POTTER, J. D., THOMPSON, M. L., THORNQUIST, M., WINGET, M. & YASUI, Y. 2001. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst,* 93**,** 1054-61.

PHILLIPS, L. D. & BANA E COSTA, C. A. 2007. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. *Annals of Operations Research,* 154**,** 51-68.

RANSOHOFF, D. F. & FEINSTEIN, A. R. 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med,* 299**,** 926-30.

REITSMA, J. B., RUTJES, A. W. S., KHAN, K. S., COOMARASAMY, A. & BOSSUYT, P. M. 2009. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology,* 62**,** 797-806.

RELLING, M. V. & EVANS, W. E. 2015. Pharmacogenomics in the clinic. *Nature,* 526**,** 343-50.
RICK, C., MALLETT, S., BROWN, J., OTTRIDGE, R., PALMER, A., PARKER, V., PRIEST, L. & DEEKS, J. J. 2020. Test evaluation trials present different challenges for trial managers compared to intervention trials. *Trials,* 21**,** 987.
RIFKIN, R. D. 1995. Effects of correlated and uncorrelated measurement error on linear regression and correlation in medical method comparison studies. *Statistics in Medicine,* 14**,** 789-798.
RUTJES, A. W., REITSMA, J. B., COOMARASAMY, A., KHAN, K. S. & BOSSUYT, P. M. 2007. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess,* 11**,** iii, ix-51.
SANDERS, G. D., NEUMANN, P. J., BASU, A., BROCK, D. W., FEENY, D., KRAHN, M., KUNTZ, K. M., MELTZER, D. O., OWENS, D. K., PROSSER, L. A., SALOMON, J. A., SCULPHER, M. J., TRIKALINOS, T. A., RUSSELL, L. B., SIEGEL, J. E. & GANIATS, T. G. 2016. Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine. *JAMA,* 316 10**,** 1093-103.
SCHMIDT, R. L. & FACTOR, R. E. 2013. Understanding Sources of Bias in Diagnostic Accuracy Studies. *Archives of Pathology & Laboratory Medicine,* 137**,** 558-565.
SCHUMACHER, S. G., VAN SMEDEN, M., DENDUKURI, N., JOSEPH, L., NICOL, M. P., PAI, M. & ZAR, H. J. 2016. Diagnostic Test Accuracy in Childhood Pulmonary Tuberculosis: A Bayesian Latent Class Analysis. *Am J Epidemiol,* 184**,** 690-700.
SEQUIST, L. V., YANG, J. C., YAMAMOTO, N., O'BYRNE, K., HIRSH, V., MOK, T., GEATER, S. L., ORLOV, S., TSAI, C. M., BOYER, M., SU, W. C., BENNOUNA, J., KATO, T., GORBUNOVA, V., LEE, K. H., SHAH, R., MASSEY, D., ZAZULINA, V., SHAHIDI, M. & SCHULER, M. 2013. Phase III study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with EGFR mutations. *J Clin Oncol,* 31**,** 3327-34.
SEVERENS, J. L. & VAN DER WILT, G.-J. 1999. ECONOMIC EVALUATION OF DIAGNOSTIC TESTS. *International Journal of Technology Assessment in Health Care,* 15**,** 480-496.
SEVILIMEDU, V. & YU, L. 2022. Simulation extrapolation method for measurement error: A review. *Stat Methods Med Res,* 31**,** 1617-1636.
SHAUGHNESSY AF. 2007. Clinical Epidemiology: A Basic Science for Clinical Medicine. *BMJ,* 335**,** 777. doi: 10.1136/bmj.39349.461678.43.
SHERWIN, P. F. 2022. Correcting Diagnostic Test Sensitivity and Specificity for Patient Misclassifications Resulting from Use of an Imperfect Reference Standard. *Diagnostics (Basel),* 13.
SIMEL, D. L., SAMSA, G. P. & MATCHAR, D. B. 1991. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol,* 44**,** 763-70.
SKEDGEL, C., HENDERSON, N., TOWSE, A., MOTT, D. & GREEN, C. 2022. Considering Severity in Health Technology Assessment: Can We Do Better? *Value in Health,* 25**,** 1399-1403.
SNOWSILL, T. 2023. Modelling the Cost-Effectiveness of Diagnostic Tests. *Pharmacoeconomics,* 41**,** 339-351.
SOARES, M. O., WALKER, S., PALMER, S. J. & SCULPHER, M. J. 2018. Establishing the Value of Diagnostic and Prognostic Tests in Health Technology Assessment. *Med Decis Making,* 38**,** 495-508.

STAQUET, M., ROZENCWEIG, M., LEE, Y. J. & MUGGIA, F. M. 1981. Methodology for the assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases,* 34**,** 599-610.

SUTTON, A. J., COOPER, N. J., GOODACRE, S. & STEVENSON, M. 2008. Integration of Meta-analysis and Economic Decision Modeling for Evaluating Diagnostic Tests. *Medical Decision Making,* 28**,** 650-667.

TANG, S., HEMYARI, P., CANCHOLA, J. A. & DUNCAN, J. 2018. Dual composite reference standards (dCRS) in molecular diagnostic research: A new approach to reduce bias in the presence of Imperfect reference. *Journal of Biopharmaceutical Statistics,* 28**,** 951-965.

TORRANCE, G. W. 1986. Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics,* 5**,** 1-30.

TRIKALINOS, T. A. & BALION, C. M. 2012. Chapter 9: options for summarizing medical test performance in the absence of a "gold standard". *J Gen Intern Med,* 27 Suppl 1**,** S67-75.

UMEMNEKU CHIKERE, C. M., WILSON, K., GRAZIADIO, S., VALE, L. & ALLEN, A. J. 2019. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard – An update. *PLOS ONE,* 14**,** e0223832.

UMEMNEKU CHIKERE, C. M., WILSON, K. J., ALLEN, A. J. & VALE, L. 2021. Comparative diagnostic accuracy studies with an imperfect reference standard – a comparison of correction methods. *BMC Medical Research Methodology,* 21**,** 67.

VACEK, P. M. 1985. The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests. *Biometrics,* 41**,** 959-968.

VAN DER POL, S., ROJAS GARCIA, P., ANTOñANZAS VILLAR, F., POSTMA, M. J. & VAN ASSELT, A. D. I. 2021. Health-Economic Analyses of Diagnostics: Guidance on Design and Reporting. *Pharmacoeconomics,* 39**,** 1355-1363.

VARSE, F., JANANI, L., MORADI, Y., SOLAYMANI-DODARAN, M., BARADARAN, H. R. & RIMAZ, S. 2019. Challenges in the design, conduct, analysis, and reporting in randomized clinical trial studies: A systematic review. *Med J Islam Repub Iran,* 33**,** 37.

WHITING, P. F., RUTJES, A. W., WESTWOOD, M. E., MALLETT, S., DEEKS, J. J., REITSMA, J. B., LEEFLANG, M. M., STERNE, J. A. & BOSSUYT, P. M. 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med,* 155**,** 529-36.

WILCKEN, B., WILEY, V., HAMMOND, J. & CARPENTER, K. 2003. Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *N Engl J Med,* 348**,** 2304-12.

WOLFF, A. C., HAMMOND, M. E. H., ALLISON, K. H., HARVEY, B. E., MANGU, P. B., BARTLETT, J. M. S., BILOUS, M., ELLIS, I. O., FITZGIBBONS, P., HANNA, W., JENKINS, R. B., PRESS, M. F., SPEARS, P. A., VANCE, G. H., VIALE, G., MCSHANE, L. M. & DOWSETT, M. 2018. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Arch Pathol Lab Med,* 142**,** 1364-1382.

XIE, X., GAJIC-VELJANOSKI, O., UNGAR, W. J., GAO, C., HUSSAIN, S., TU, H. A. & VOLODIN, A. 2022a. Modeling methods and the degree of parameter uncertainty in probabilistic analyses of economic evaluations. *Network Modeling Analysis in Health Informatics and Bioinformatics,* 12**,** 9.

XIE, X., SCHAINK, A. K., LIU, S., WANG, M. & VOLODIN, A. 2022b. Understanding bias in probabilistic analysis in model-based health economic evaluation. *The European Journal of Health Economics*.

XIE, X., SINCLAIR, A. & DENDUKURI, N. 2017. Evaluating the accuracy and economic value of a new test in the absence of a perfect reference test. *Research Synthesis Methods,* 8**,** 321-332.

XIE, X., TIGGELAAR, S., GUO, J., WANG, M., VANDERSLUIS, S. & UNGAR, W. J. 2022c. Developing Economic Models for Assessing the Cost-Effectiveness of Multiple Diagnostic Tests: Methods and Applications. *Medical Decision Making,* 0**,** 0272989X221089268.

XIE, X., WANG, M., GAJIC-VELJANOSKI, O., YE, C., BLUMBERGER, D. M. & VOLODIN, A. 2022d. Examining the correlation between treatment effects in clinical trials and economic modeling. *Expert Review of Pharmacoeconomics & Outcomes Research,* 22**,** 1071-1078.

YANG, Y., ABEL, L., BUCHANAN, J., FANSHAWE, T. & SHINKINS, B. 2019. Use of Decision Modelling in Economic Evaluations of Diagnostic Tests: An Appraisal and Review of Health Technology Assessments in the UK. *Pharmacoecon Open,* 3**,** 281-291.

ZAKI, R., BULGIBA, A., ISMAIL, R. & ISMAIL, N. A. 2012. Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. *PLOS ONE,* 7**,** e37908.

ZHOU, X. H., MCCLISH, D. & OBUCHOWSKI, N. 2002. Statistical Methods in Diagnostic Medicine John W. *Wiley,* 569.

ZWEIG, M. H. & CAMPBELL, G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem,* 39**,** 561-77.

8 Appendix

1. Staquet et al. (1981)

Description of the method

The Staquet et al. (1981) method (Staquet et al., 1981) is a count-based correction method. It requires 2x2 count data on the classification results of the test against those of the iRS. Table 11 uses a 2x2 table serving as notation.

Table 11. 2x2 table denoting the classification of individuals according to their test results.

	iRS positive	*iRS negative*	*Total*
Test positive	pp	pn	pp+pn=Tp
Test negative	np	nn	np+nn=Tn
Total	pp+np=iRSp	pn+nn=iRSn	pp+pn+np+nn=N

p: positive, n: negative, pp: T^+/iRS^+, pn: T^+/iRS^-, np: T^-/iRS^+, nn: T^-/iRS^-, Tp: number of individuals with a positive test result, Tn: number of individuals with a negative test result, iRSp: number of individuals with a positive iRS result, iRSn: number of individuals with a negative iRS result.

The proposed estimators for adjusted Se and Sp of the test are listed in equation (8) and (9), page 603 in reference (Staquet et al., 1981):

$$Se_T{}' = \frac{Tp\ Sp_{iRS} - pn}{N(Sp_{iRS} - 1) + (pp + np)}$$

$$Sp_T' = \frac{Tn\ Se_{iRS} - np}{NSe_{iRS} - (pp + np)}$$

Where Se_{iRS} and Sp_{iRS} are the sensitivity and specificity for the iRS related to the gold standard.

The paper highlights the influence of disease prevalence on the accuracy of test parameters, noting biases in Sp and predictive values when using iRS. Additionally, the authors stress the importance of appropriate random sampling in trial designs to avoid selection bias, particularly when using iRS for determining true disease status, which may not represent the broader target population. The paper concludes that adjustments for prevalence are crucial in retrospective studies to ensure accurate statistical

assessments, underscoring the complexities in evaluating new diagnostic tests in clinical practice.

2. Brenner (1996)

The Brenner (1996) equations were designed to explore the exposure misclassification in epidemiological studies and validation studies to correct this misclassification. This correction involves the use of an "alloyed gold standard" in the validation studies, though, it is not always perfect. The author proposed two pair of equations, considering CI assumption and positive CD. These equations were designed to estimate the apparent accuracy of an index test, using the true estimations and the alloyed gold standard.

The first pair of estimators was the $Se_{aT}^{ind\prime}$ and $Sp_{aT}^{ind\prime}$, representing the apparent sensitivity and specificity under CI assumption, illustrated below (Brenner (1996), page 407, equations for Se_{ind} and Sp_{ind}).

$$Se_{aT}^{ind\prime} = \frac{p(D^+)Se_{aGS}Se_T + \big(1 - p(D^+)\big)(1 - Sp_{aGS})(1 - Sp_T)}{p(D^+)Se_{iRS} + (1 - p[(D^+))(1 - Sp_{iRS})}$$

$$Sp_{aT}^{ind\prime} = \frac{p(D^+)(1 - Se_{iRS})\ (1 - Se_T) + \big(1 - p(D^+)\big)\ Sp_{iRS}\ Sp_T}{p(D^+)\ (1 - Se_{iRS}) + \big(1 - p(D^+)\big)\ Sp_{iRS}}$$

where $p(D^+)$ is the prevalence, Se_T and Sp_T represent the true sensitivity and specificity, and Se_{aGS} and Sp_{aGS} represent the accuracy for the alloyed gold standard, also known as iRS.

The model considers a positive correlation between tests, which is likely when two tests are based on similar biological processes. The second pair considers a positive correlation between tests, it introduces $Se_{aT}^{dep\prime}$ and $Sp_{aT}^{dep\prime}$ as the apparent sensitivity and specificity, respectively, estimated under positive correlation conditions. The derived estimators which assume positive correlation are presented below (Brenner (1996), page 407, equations for Se_{pos} and Sp_{pos}):

$$Se_{aT}^{dep'} = \frac{p(D^+)Se_T + \big(1 - p(D^+)\big)(1 - Sp_{iRS})}{p(D^+)Se_{iRS} + \big(1 - p(D^+)\big)(1 - Sp_{iRS})}$$

$$Sp_{aT}^{dep'} = \frac{p(D^+)(1 - Se_{iRS}) + \big(1 - p(D^+)\big)\, Sp_T}{p(D^+)(1 - Se_{iRS}) + \big(1 - p(D^+)\big)\, Sp_{iRS}}$$

Unfortunately, as the author does not provide derivations for true accuracy estimates from the apparent accuracy and imperfect RS, we offer a derivation for the equation of Se and Sp under positive CD. Accordingly, we present the derived estimators for Se and Sp as follows:

$$Se_T^{dep'} = \frac{Se_{aT}^{dep'}\big(Se_{iRS}p(D^+) + (p(D^+) - 1)(Sp_{iRS} - 1)\big) - (p(D^+) - 1)(Sp_{iRS} - 1)}{p(D^+)}$$

$$Sp_T^{dep'} = \frac{Sp_{aT}\big(Sp_{iRS}(p(D^+) - 1) + p(D^+)(Sp_{iRS} - 1)\big) - p(D^+)(Sp_{iRS} - 1)}{p(D^+) - 1}$$

3. Equations derivate from Brenner (1996) to estimate corrected accuracy.

4. Derivation for sensitivity parameter

Define the original equation:

$$SE'_{aT} = \frac{p(D^+)Se_{iRS}Se_T + \big(1 - p(D^+)\big)(1 - Sp_{iRS})(1 - Sp_T)}{p(D^+)Se_{iRS} + \big(1 - p(D^+)\big)(1 - Sp_{iRS})}$$

Multiply both sides by the denominator

$$Se'_{aT}\Big(p(D^+)Se_{iRS} + \big(1 - p(D^+)\big)(1 - Sp_{iRS})\Big) = p(D^+)Se_{iRS}Se_T + \big(1 - p(D^+)\big)(1 - Sp_{iRS})(1 - Sp_T)$$

Distribute SE'_{ar} across the term on the left side.

$$Se'_{aT}p(D^+)Se_{iRS} + Se'_{aT}\big(1-p(D^+)\big)(1-Sp_{iRS}) = p(D^+)Se_{iRS}Se_T + \big(1-p(D^+)\big)(1-Sp_{iRS})(1-Sp_T)$$

Isolate terms with Se_T on one side

$$p(D^+)Se_{iRS}Se_T = Se'_{aT}p(D^+)Se_{iRS} + Se'_{aT}\big(1-p(D^+)\big)(1-Sp_{iRS}) - \big(1-p(D^+)\big)(1-Sp_{iRS})(1-Sp_T)$$

Divide both sides by $p(D^+)Se_{iRS}$

$$\mathrm{Se_T} = \frac{Se'_{ar}p(D^+)Se_{iRS} + Se'_{aT}\big(1-p(D^+)\big)(1-Sp_{iRS}) - \big(1-p(D^+)\big)(1-Sp_{iRS})(1-Sp_T)}{\mathrm{p(D^+)Se_{iRS}}}$$

4.1.1. Derivation for specificity parameter

Define the original equation.

$$Sp'_{aT} = \frac{p(D^+)(1-Se_{iRS}) + \big(1-p(D^+)\big)Sp_T}{p(D^+)(1-Se_{iRS}) + \big(1-p(D^+)\big)Sp_{iRS}}$$

Multiply both sides by the denominator

$$Sp'_{aT}\big(p(D^+)(1-Se_{iRS}) + \big(1-p(D^+)\big)Sp_{iRS}\big) = p(D^+)(1-Se_{iRS}) + \big(1-p(D^+)\big)Sp_T$$

Isolate terms with Sp_T on one side

$$\big(1-p(D^+)\big)Sp_T = Sp'_{aT}\cdot p(D^+)(1-Se_{iRS}) + Sp'_{aT}\big(1-p(D^+)\big)Sp_{iRS} - p(D^+)(1-Se_{iRS})$$

Divide both sides by 1- $p(D^+)$ to solve for Sp_T

$$Sp_T = \frac{Sp'_{aT}\cdot p(D^+)(1-Se_{iRS}) + Sp'_{aT}\big(1-p(D^+)\big)Sp_{iRS} - p(D^+)(1-Se_{iRS})}{\big(1-p(D^+)\big)}$$

8.1 Conditions that lead to illogical results using the Staquet approach.

a. Illogical Se estimation

- If $Sp_{iRS} < Sp_{Index}$, then Se greater than 1
- If $Sp_{iRS} < 1 - Sp_{Index}$ and $Sp_{iRS} > 1 - p(D^+)$, then Se less than 0
- If $Sp_{iRS} > 1 - Sp_{Index}$ and $Sp_{iRS} < 1 - p(D^+)$, then Se less than 0

b. Illogical Sp estimation

- If $Se_{iRS} < Se_{Index}$, then Se greater than 1
- If $Se_{iRS} < 1 - Se_{Index}$ and $Se_{iRS} > p(D^+)$, then Se less than 0
- If $Se_{iRS} > 1 - Se_{Index}$ and $Se_{iRS} < p(D^+)$, then Se less than 0

8.2 Conditions that lead to illogical results using the Brenner for conditional dependence approach.

a. Illogical Se estimation greater than 1

- Se_{Index} is high relative to $p(D^+)$
- Se_{iRS} is significatively high to make $Se_{Index}Se_{iRS}p(D^+)$ the dominating term.
- Sp_{iRS} is low enough that the negative impact of $(p(D^+) - 1)(Sp_{iRS} - 1)$ does not significatively reduce the numerator.
- $Se_{Index} > 1$ if $Se_{Index}Se_{iRS} > \frac{1}{p(D^+)} + (1 - p(D^+))(1 - Sp_{iRS})$

b. Illogical Se estimation lower than 0

- Se_{Index} is low, especially in comparison to Se_{iRS} and $p(D^+)$
- Sp_{iRS} is high relative to $p(D^+)$, maximising the negative impact of the subtracted term.
- $Se_{Index} < 0$ if $Se_{Index}Se_{iRS}p(D^+) < (1 - p(D^+))(1 - Sp_{iRS})$

c. Illogical Sp estimation greater than 1

- Given that $Sp_{iRS}(p(D^+) - 1)$ is negative, for the numerator to have a smaller absolute value than the denominator, Sp_{Index} must be sufficiently large to ensure that the positive term $p(D^+)(Sp_{iRS} - 1)$ is dominant.
- Since the numerator includes a subtraction of $p(D^+)(Sp_{iRS} - 1)$, which is negative, Sp_{Index} must no amplify this negative term excessively.
- $Sp_{Index} > 1$ if $Sp_{Index}(Sp_{iRS} - 1) > \frac{1}{1-p(D^+)}$

d. Illogical Sp estimation lower than 0

- The numerator will be positive if the term if $Sp_{Index}(Sp_{iRS} - 1)$ is large enough to overcome the negative value of $Sp_{iRS}(p(D^+) - 1)$, which suggest a high Sp_{iRS} and a low Sp_{Index}
- $Sp_{Index} < 0$ if $\frac{p(D^+)(Sp_{iRS}-1)}{(Sp_{iRS}-1)}$